Minimization of Weighted Pole and Zero Sensitivity for State-Space Digital Filters

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

***Abstract*—This paper presents a systematic study of pole and zero sensitivity minimization for state-space digital filters in several different yet related settings. First, a new weighted measure for pole and zero sensitivity for state-space digital filters is proposed and the problem of minimizing this measure is investigated. To this end, two efficient iterative techniques for minimizing this measure are developed by employing a quasi-Newton algorithm and relying on a recursive matrix equation, respectively. Furthermore, minimization of the proposed sensitivity measure subject to l_2-scaling constraints is examined by extending the two aforementioned solution methods — one converts the contained optimization problem at hand into an unconstrained problem and solves it using a quasi-Newton algorithm, while the other relaxes the constraints into a single constraint on matrix trace and solves the relaxed problem with an effective matrix iteration scheme. In addition, a simple yet novel method for the minimization of a zero sensitivity measure subject to minimal pole sensitivity is explored by pursuing an optimal coordinate transformation matrix. Simulation studies are presented to demonstrate the validity and effectiveness of the proposed techniques.**

***Index Terms*— l_2-scaling, IIR digital filter, Lagrange function, overflow oscillation, pole and zero sensitivity, quasi-Newton algorithm, sensitivity minimization, state-space model**

I. INTRODUCTION

IN the case where a transfer function with infinite accuracy coefficients is designed so as to meet the filter specification requirements and the transfer function is implemented by a state-space model with a finite binary representation, the state-space parameters must be truncated or rounded to fit the finite-word-length (FWL) constraints. This coefficient quantization inevitably changes the characteristics of the digital filter. For instance, it may alter a stable filter to unstable one. This motivates the study of minimizing coefficients sensitivity. As is well known, there are several ways to define sensitivity of a filter with respect to its realization coefficients. Two of them are based on a mixed l_1/l_2 norm and a pure l_2 norm, respectively. One of them measures changes of a certain transfer function, while the other is defined in terms of the poles and zeros of a filter. Several techniques for minimizing the l_1/l_2-sensitivity measure [1]-[6] and the l_2-sensitivity measure [7]-[14] have been proposed. Alternatively, eigenvalue sensitivity of a filter with respect to system matrix has been investigated [15], [16]. It is argued in [15] that the realization of a block diagonal form is preferable to the companion matrix form for the reduction of eigenvalue sensitivity, although the generality of the assertion remains unclear. In [16], it is shown that the eigenvalue sensitivity of a realization whose system matrix is normal is minimal. In [17], [18] pole and zero sensitivity of a filter with respect to state-space parameters is analyzed and its reduction and minimization are addressed. It is demonstrated in [17] that the sensitivities of the poles and zeros of a transfer function with respect to the variations in state-space parameters (due to FWL) are closely related to the sensitivity characteristics of its frequency transfer function. Moreover, minimal pole sensitivity is achieved when the system matrix is normal, and minimal zero sensitivity is achieved when a certain matrix is normal. It should be stressed that minimal pole sensitivity does not guarantee minimal zero sensitivity and vice versa. Also, it is known that it is generally impossible to minimize pole sensitivity and zero sensitivity simultaneously. In [18], a method for minimizing a zero sensitivity measure subject to minimal pole sensitivity for state-space digital filters is explored. However, l_2-scaling for preventing overflow oscillations was not taken into account in [18]. Recently, the minimization problem of weighted pole and zero sensitivity was examined in [19], but l_2-scaling constraints were not imposed.

In this paper, we present a systematic study of pole and zero sensitivity minimization for state-space digital filters in several different yet related settings. First, a new weighted measure for pole and zero sensitivity for state-space digital filters is proposed and the problem of minimizing this measure is investigated. To this end, two efficient iterative techniques for minimizing this measure are developed — one employs a quasi-Newton algorithm, while the other relies on a recursive matrix equation, respectively. Furthermore, minimization of the proposed sensitivity measure subject to l_2-scaling constraints is examined by extending the two aforementioned solution methods — one converts the contained optimization problem at hand into an unconstrained problem and solves it using a quasi-Newton algorithm, while the other relaxes the constraints into a single constraint on matrix trace and solves the relaxed problem with an effective matrix iteration scheme where the Lagrange multiplier is determined via a bisection method. In addition, a simple yet novel method for the minimization of a zero sensitivity measure subject to minimal pole sensitivity is explored by pursuing an optimal coordinate transformation matrix. Simulation studies are presented to demonstrate the validity and effectiveness of the

Manuscript received June 19, 2015

T. Hinamoto is with Hiroshima University Higashi-Hiroshima 739-8527, Japan (e-mail: hinamoto@ieee.org, Phone:+81-86-293-3544, Fax:+81-86-293-3544)

A. Doi is with Hiroshima Institute of Technology Hiroshima 731-5193, Japan (e-mail: 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)

proposed techniques. Unlike [18] and [19], an entire section (Section IV) is devoted to weighted pole and zero sensitivity minimization subject to l_2-scaling constraints. It should be mentioned that the techniques presented in this paper are restricted to proper SISO systems whose transfer functions contain nonzero constant term.

Throughout $\boldsymbol{I}_n$ denotes the identity matrix of dimension $n \times n$, and its ith column is denoted by $\boldsymbol{e}_i$. The transpose (conjugate transpose) of a matrix $\boldsymbol{A}$ and trace of a square matrix $\boldsymbol{A}$ are denoted by $\boldsymbol{A}^T$ ($\boldsymbol{A}^H$) and tr$[\boldsymbol{A}]$, respectively. Boldface uppercase, boldface lowercase and plain lowercase are used to make the distinction among matrices, vectors, and scalar values, respectively.

II. Problem Formulation

Consider a stable, controllable and observable state-space digital filter $(\boldsymbol{A}, \boldsymbol{b}, \boldsymbol{c}, d)_n$ of order n described by

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

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 $\boldsymbol{A}, \boldsymbol{b}, \boldsymbol{c}$ and d are real constant matrices of appropriate dimensions. The transfer function of the digital filter in (1) can be expressed as

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

Assuming that a direct path from the input to the output exists in (1), i.e., $d \neq 0$, it is known [20] that the poles and zeros of $H(z)$ are given by $\{\lambda_l\} = \lambda(\boldsymbol{A})$ and $\{v_l\} = \lambda(\boldsymbol{Z})$, respectively, where an $n \times n$ matrix $\boldsymbol{Z}$ is defined by

$$\boldsymbol{Z} = \boldsymbol{A} - d^{-1}\boldsymbol{b}\boldsymbol{c}. \tag{3}$$

Lemma 1 : Let $\boldsymbol{x}_p(l)$ be a right eigenvector corresponding to λ_l and let $\boldsymbol{y}_p(l)$ be the reciprocal left eigenvector that corresponds to $\boldsymbol{x}_p(l)$. Then, under the assumption that $\boldsymbol{A}$ has a full set of linearly independent eigenvectors, the following holds:

$$\Big(\frac{\partial \lambda_l}{\partial \boldsymbol{A}}\Big)^T = \boldsymbol{x}_p(l)\boldsymbol{y}_p^H(l) \quad \text{for} \quad l = 1, 2, \cdots, n. \tag{4}$$

Proof: See [16], [18].

Lemma 2 : Let $\boldsymbol{x}_z(l)$ be a right eigenvector corresponding to v_l and let $\boldsymbol{y}_z(l)$ be the reciprocal left eigenvector that corresponds to $\boldsymbol{x}_z(l)$. Then, under the assumption that $\boldsymbol{Z}$ has a linearly independent eigenvector set, we have

$$\Big(\frac{\partial v_l}{\partial \boldsymbol{Z}}\Big)^T = \boldsymbol{x}_z(l)\boldsymbol{y}_z^H(l) \quad \text{for} \quad l = 1, 2, \cdots, n. \tag{5}$$

Proof: The proof is identical with that in Lemma 1.

Lemma 3 : If (5) holds, it can be verified that

$$\begin{aligned} \frac{\partial v_l}{\partial \boldsymbol{A}} &= \frac{\partial v_l}{\partial \boldsymbol{Z}}, & \frac{\partial v_l}{\partial \boldsymbol{b}} &= -d^{-1}\frac{\partial v_l}{\partial \boldsymbol{Z}}\boldsymbol{c}^T \\ \frac{\partial v_l}{\partial \boldsymbol{c}} &= -d^{-1}\boldsymbol{b}^T\frac{\partial v_l}{\partial \boldsymbol{Z}}, & \frac{\partial v_l}{\partial d} &= d^{-2}\boldsymbol{b}^T\frac{\partial v_l}{\partial \boldsymbol{Z}}\boldsymbol{c}^T. \end{aligned} \tag{6}$$

Proof: See [9].

We now define the pole sensitivity measure J_p and the zero sensitivity measure J_z for the digital filter in (1) as [18]

$$\begin{aligned} J_p &= \sum_{l=1}^n \Big\|\frac{\partial \lambda_l}{\partial \boldsymbol{A}}\Big\|_F^2 = \sum_{l=1}^n \Big\|\Big(\frac{\partial \lambda_l}{\partial \boldsymbol{A}}\Big)^T\Big\|_F^2 \\ J_z &= \sum_{l=1}^n \Big\{\Big\|\frac{\partial v_l}{\partial \boldsymbol{A}}\Big\|_F^2 + \Big\|\frac{\partial v_l}{\partial \boldsymbol{b}}\Big\|_F^2 + \Big\|\frac{\partial v_l}{\partial \boldsymbol{c}}\Big\|_F^2 + \Big\|\frac{\partial v_l}{\partial d}\Big\|_F^2\Big\} \end{aligned} \tag{7}$$

respectively, where $||\boldsymbol{M}||_F$ denotes the Frobenius norm of an $m \times n$ complex matrix $\boldsymbol{M}$, which is defined by

$$||\boldsymbol{M}||_F = \Big(\sum_{i=1}^m \sum_{j=1}^n |(\boldsymbol{M})_{ij}|^2\Big)^{\frac{1}{2}}.$$

By noting that

$$||\boldsymbol{M}||_F^2 = \text{tr}[\boldsymbol{M}\boldsymbol{M}^H] = \text{tr}[\boldsymbol{M}^H\boldsymbol{M}] = ||\boldsymbol{M}^H||_F^2 \tag{8}$$

and using (4), (5) and (6), we can write (7) as

$$J_p = \sum_{l=1}^n \Big(\boldsymbol{x}_p^H(l)\boldsymbol{x}_p(l)\Big)\Big(\boldsymbol{y}_p^H(l)\boldsymbol{y}_p(l)\Big) \tag{9a}$$

$$J_z = \sum_{l=1}^n \Big(\boldsymbol{x}_z^H(l)\boldsymbol{x}_z(l) + \alpha_l^2\Big)\Big(\boldsymbol{y}_z^H(l)\boldsymbol{y}_z(l) + \beta_l^2\Big) \tag{9b}$$

where

$$\alpha_l = |d^{-1}\boldsymbol{c}\boldsymbol{x}_z(l)|, \qquad \beta_l = |d^{-1}\boldsymbol{y}_z^H(l)\boldsymbol{b}|. \tag{9c}$$

Now if a coordinate transformation defined by

$$\overline{\boldsymbol{x}}(k) = \boldsymbol{T}^{-1}\boldsymbol{x}(k) \tag{10}$$

is applied to the digital filter in (1), the new realization $(\overline{\boldsymbol{A}}, \overline{\boldsymbol{b}}, \overline{\boldsymbol{c}}, d)_n$ can be characterized by

$$\overline{\boldsymbol{A}} = \boldsymbol{T}^{-1}\boldsymbol{A}\boldsymbol{T}, \quad \overline{\boldsymbol{b}} = \boldsymbol{T}^{-1}\boldsymbol{b}, \quad \overline{\boldsymbol{c}} = \boldsymbol{c}\boldsymbol{T}. \tag{11}$$

From (2) and (11), it is observed that the transfer function $H(z)$ is invariant under the coordinate transformation in (10). For the new realization $(\overline{\boldsymbol{A}}, \overline{\boldsymbol{b}}, \overline{\boldsymbol{c}}, d)_n$, the right eigenvector $\boldsymbol{x}_p(l)$ corresponding to λ_l and the reciprocal left eigenvector $\boldsymbol{y}_p(l)$ corresponding to $\boldsymbol{x}_p(l)$ become [17], [18]

$$\overline{\boldsymbol{x}}_p(l) = \boldsymbol{T}^{-1}\boldsymbol{x}_p(l), \qquad \overline{\boldsymbol{y}}_p(l) = \boldsymbol{T}^T\boldsymbol{y}_p(l) \tag{12}$$

for $l = 1, 2, \cdots, n$, respectively. Similarly, the right eigenvector $\boldsymbol{x}_z(l)$ corresponding to v_l and the reciprocal left eigenvector $\boldsymbol{y}_z(l)$ corresponding to $\boldsymbol{x}_z(l)$ become

$$\overline{\boldsymbol{x}}_z(l) = \boldsymbol{T}^{-1}\boldsymbol{x}_z(l), \qquad \overline{\boldsymbol{y}}_z(l) = \boldsymbol{T}^T\boldsymbol{y}_z(l) \tag{13}$$

for $l = 1, 2, \cdots, n$, respectively. Therefore, for the new realization $(\overline{\boldsymbol{A}}, \overline{\boldsymbol{b}}, \overline{\boldsymbol{c}}, d)_n$ specified in (11), the pole and zero sensitivity measures in (9) can be expressed as

$$\begin{aligned} J_p(\boldsymbol{T}) &= \sum_{l=1}^n \Big(\boldsymbol{x}_p^H(l)\boldsymbol{T}^{-T}\boldsymbol{T}^{-1}\boldsymbol{x}_p(l)\Big)\Big(\boldsymbol{y}_p^H(l)\boldsymbol{T}\boldsymbol{T}^T\boldsymbol{y}_p(l)\Big) \\ J_z(\boldsymbol{T}) &= \sum_{l=1}^n \Big(\boldsymbol{x}_z^H(l)\boldsymbol{T}^{-T}\boldsymbol{T}^{-1}\boldsymbol{x}_z(l) + \alpha_l^2\Big) \\ &\qquad \cdot \Big(\boldsymbol{y}_z^H(l)\boldsymbol{T}\boldsymbol{T}^T\boldsymbol{y}_z(l) + \beta_l^2\Big) \end{aligned} \tag{14}$$

respectively. It is noted that for any $n \times n$ nonsingular matrix $\boldsymbol{T}$, the following inequalities hold true [16]-[18]:

$$J_p(\boldsymbol{T}) \geq n, \qquad J_z(\boldsymbol{T}) \geq \sum_{l=1}^{n} (1+|\alpha_l \beta_l|)^2. \tag{15}$$

There are several ways joint optimization of pole and zero sensitivity can be carried out. One approach is to minimize one sensitivity measure, say, $J_p(\boldsymbol{T})$, subject to the other sensitivity, $J_z(\boldsymbol{T})$, being bounded from above. This leads to a constrained problem where the upper bound of $J_z(\boldsymbol{T})$ needs to be determined very carefully—because of (15), an upper bound too small can make the problem infeasible, but an upper bound too large is obviously undesirable as it would defeat the purpose of keeping $J_z(\boldsymbol{T})$ as small as possible. Nevertheless, in Section III.C we consider a special circumstance and show that the minimization of zero sensitivity subject to a constraint on pole sensitivity admits a closed-form solution (see (33) and (34)). Another approach is to deal with the problem at hand as a multiobjective optimization problem with a vectorized objective function $\boldsymbol{J}(\boldsymbol{T}) = [J_p(\boldsymbol{T}), J_z(\boldsymbol{T})]^T$, whose goal is to minimize $\boldsymbol{J}(\boldsymbol{T})$ in a certain sense, for example via the notion of Pareto boundary [21]. One of realistic and effective techniques to deal with multiobjective optimization involving complicated and nonconvex component objectives is to scalarize the problem by forming a weighted sum of the components of the vector objective.

In the present case, we consider a weighted pole and zero sensitivity measure $J_\gamma(\boldsymbol{T})$ defined by

$$J_\gamma(\boldsymbol{T}) = \gamma J_p(\boldsymbol{T}) + (1-\gamma) J_z(\boldsymbol{T}) \tag{16}$$

where $0 \leq \gamma \leq 1$ is a weighting factor to control the trade-off between the two component sensitivities in the sense that a $J_\gamma(\boldsymbol{T})$ with a greater γ represents a measure that places more emphasis on pole sensitivity, while a $J_\gamma(\boldsymbol{T})$ with a smaller γ serves as a measure that weights more heavily on zero sensitivity. In addition, by setting γ to unity or zero $J_\gamma(\boldsymbol{T})$ becomes $J_p(\boldsymbol{T})$ or $J_z(\boldsymbol{T})$, respectively.

The problem of minimizing the weighted pole and zero sensitivity measure is now formulated as follows: *For given $\boldsymbol{A}$, $\boldsymbol{b}$, $\boldsymbol{c}$ and d, obtain an $n \times n$ transformation matrix $\boldsymbol{T}$ which minimizes $J_\gamma(\boldsymbol{T})$ in (16) with γ specified by the designer*.

III. POLE AND ZERO SENSITIVITY MINIMIZATION

A. Minimization of (16) Using A Quasi-Newton (QN) Algorithm

To minimize (16) with respect to an $n \times n$ transformation matrix $\boldsymbol{T}$, we define

$$\boldsymbol{T}^{-1} = [\boldsymbol{t}_1, \boldsymbol{t}_2, \cdots, \boldsymbol{t}_n], \quad \boldsymbol{x} = [\boldsymbol{t}_1^T, \boldsymbol{t}_2^T, \cdots, \boldsymbol{t}_n^T]^T \tag{17}$$

which establishes a mapping that uniquely connects a nonsingular $n \times n$ matrix $\boldsymbol{T}$ with a column vector $\boldsymbol{x}$ of length n^2 and denote $J_\gamma(\boldsymbol{T})$ in (16) as function $J_\gamma(\boldsymbol{x})$. In the kth iteration, a quasi-Newton algorithm, known as the Broyden-Fletcher-Goldfarb-Shanno (BFGS) algorithm [22], updates the current point $\boldsymbol{x}_k$ to point $\boldsymbol{x}_{k+1}$ as [23]

$$\boldsymbol{x}_{k+1} = \boldsymbol{x}_k + \alpha_k \boldsymbol{d}_k \tag{18}$$

where

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

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

$$|J_\gamma(\boldsymbol{x}_{k+1}) - J_\gamma(\boldsymbol{x}_k)| < \varepsilon \tag{19}$$

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. 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.

The gradient of $J_\gamma(\boldsymbol{x})$ can be evaluated using closed-form expressions as

$$\begin{aligned} \frac{\partial J_\gamma(\boldsymbol{x})}{\partial t_{ij}} &= \lim_{\Delta \to 0} \frac{J_\gamma(\boldsymbol{T}_{ij}) - J_\gamma(\boldsymbol{T})}{\Delta} \\ &= \gamma \frac{\partial J_p(\boldsymbol{x})}{\partial t_{ij}} + (1-\gamma) \frac{\partial J_z(\boldsymbol{x})}{\partial t_{ij}} \end{aligned} \tag{20}$$

where $\boldsymbol{T}_{ij}$ is the matrix obtained from $\boldsymbol{T}$ with a perturbed (i,j)th component. It follows that [24, p. 655]

$$\begin{aligned} \boldsymbol{T}_{ij} &= \boldsymbol{T} - \frac{\Delta \boldsymbol{T} \boldsymbol{e}_i \boldsymbol{e}_j^T \boldsymbol{T}}{1 + \Delta \boldsymbol{e}_j^T \boldsymbol{T} \boldsymbol{e}_i} \simeq \boldsymbol{T} - \Delta \boldsymbol{T} \boldsymbol{e}_i \boldsymbol{e}_j^T \boldsymbol{T} \\ \boldsymbol{T}_{ij}^{-1} &= \boldsymbol{T}^{-1} + \Delta \frac{\partial \boldsymbol{t}_j}{\partial t_{ij}} \boldsymbol{e}_j^T = \boldsymbol{T}^{-1} + \Delta \boldsymbol{e}_i \boldsymbol{e}_j^T \end{aligned} \tag{21}$$

and

$$\begin{aligned} \frac{\partial J_p(\boldsymbol{x})}{\partial t_{ij}} &= \sum_{l=1}^{n} \Big\{ \Big(\boldsymbol{x}_p^H(l) \tilde{\boldsymbol{M}}(\boldsymbol{T}) \boldsymbol{x}_p(l)\Big)\Big(\boldsymbol{y}_p^H(l) \boldsymbol{T}\boldsymbol{T}^T \boldsymbol{y}_p(l)\Big) \\ &\qquad - \Big(\boldsymbol{x}_p^H(l) \boldsymbol{T}^{-T}\boldsymbol{T}^{-1} \boldsymbol{x}_p(l)\Big)\Big(\boldsymbol{y}_p^H(l) \tilde{\boldsymbol{N}}(\boldsymbol{T}) \boldsymbol{y}_p(l)\Big) \Big\} \\ \frac{\partial J_z(\boldsymbol{x})}{\partial t_{ij}} &= \sum_{l=1}^{n} \Big\{ \boldsymbol{x}_z^H(l) \tilde{\boldsymbol{M}}(\boldsymbol{T}) \boldsymbol{x}_z(l) \Big(\boldsymbol{y}_z^H(l) \boldsymbol{T}\boldsymbol{T}^T \boldsymbol{y}_z(l) + \beta_l^2\Big) \\ &\qquad - \Big(\boldsymbol{x}_z^H(l) \boldsymbol{T}^{-T}\boldsymbol{T}^{-1} \boldsymbol{x}_z(l) + \alpha_l^2\Big) \boldsymbol{y}_z^H(l) \tilde{\boldsymbol{N}}(\boldsymbol{T}) \boldsymbol{y}_z(l) \Big\} \end{aligned} \tag{22}$$

where t_{ij} denotes the jth (for $j = 1, 2, \cdots, n$) element of vector $\boldsymbol{t}_i$ with $\boldsymbol{t}_i$ defined by (17) for $i = 1, 2, \cdots, n$ and

$$\begin{aligned} \tilde{\boldsymbol{M}}(\boldsymbol{T}) &= \boldsymbol{e}_j \boldsymbol{e}_i^T \boldsymbol{T}^{-1} + \boldsymbol{T}^{-T} \boldsymbol{e}_i \boldsymbol{e}_j^T \\ \tilde{\boldsymbol{N}}(\boldsymbol{T}) &= \boldsymbol{T}\big[\boldsymbol{e}_i \boldsymbol{e}_j^T \boldsymbol{T} + \boldsymbol{T}^T \boldsymbol{e}_j \boldsymbol{e}_i^T\big] \boldsymbol{T}^T. \end{aligned}$$

B. Minimization of (16) Using A Recursive Matrix Equation (RME)

The pole and zero sensitivity measures in (14) can also be expressed in terms of matrix $\boldsymbol{P} = \boldsymbol{T}\boldsymbol{T}^T$ as

$$J_p(\boldsymbol{T}) = \sum_{l=1}^{n} \Big(\mathrm{tr}\big[\boldsymbol{x}_p(l)\boldsymbol{x}_p^H(l)\boldsymbol{P}^{-1}\big]\Big)\Big(\mathrm{tr}\big[\boldsymbol{y}_p(l)\boldsymbol{y}_p^H(l)\boldsymbol{P}\big]\Big)$$
$$J_z(\boldsymbol{T}) = \sum_{l=1}^{n} \Big(\mathrm{tr}\big[\boldsymbol{x}_z(l)\boldsymbol{x}_z^H(l)\boldsymbol{P}^{-1}\big] + \alpha_l^2\Big) \cdot\Big(\mathrm{tr}\big[\boldsymbol{y}_z(l)\boldsymbol{y}_z^H(l)\boldsymbol{P}\big] + \beta_l^2\Big) \tag{23}$$

respectively. From (23) we see that $J_\gamma(\boldsymbol{T})$ is differentiable with respect to $\boldsymbol{P}$ for nonsingular $\boldsymbol{P}$. Therfore, the optimal $\boldsymbol{P}$ must satisfy $\partial J_\gamma(\boldsymbol{T})/\partial \boldsymbol{P} = \mathbf{0}$, and $\partial J_\gamma(\boldsymbol{T})/\partial \boldsymbol{P}$ is found to be

$$\frac{\partial J_\gamma(\boldsymbol{T})}{\partial \boldsymbol{P}} = \boldsymbol{M}_\gamma(\boldsymbol{P}) - \boldsymbol{P}^{-1}\boldsymbol{N}_\gamma(\boldsymbol{P})\boldsymbol{P}^{-1} \tag{24}$$

where

$$\boldsymbol{M}_\gamma(\boldsymbol{P}) = \gamma \sum_{l=1}^{n} \mathrm{tr}\big[\boldsymbol{x}_p(l)\boldsymbol{x}_p^H(l)\boldsymbol{P}^{-1}\big]\boldsymbol{y}_p(l)\boldsymbol{y}_p^H(l) + (1-\gamma)\sum_{l=1}^{n}\Big(\mathrm{tr}\big[\boldsymbol{x}_z(l)\boldsymbol{x}_z^H(l)\boldsymbol{P}^{-1}\big] + \alpha_l^2\Big)\boldsymbol{y}_z(l)\boldsymbol{y}_z^H(l)$$
$$\boldsymbol{N}_\gamma(\boldsymbol{P}) = \gamma \sum_{l=1}^{n} \mathrm{tr}\big[\boldsymbol{y}_p(l)\boldsymbol{y}_p^H(l)\boldsymbol{P}\big]\boldsymbol{x}_p(l)\boldsymbol{x}_p^H(l) + (1-\gamma)\sum_{l=1}^{n}\Big(\mathrm{tr}\big[\boldsymbol{y}_z(l)\boldsymbol{y}_z^H(l)\boldsymbol{P}\big] + \beta_l^2\Big)\boldsymbol{x}_z(l)\boldsymbol{x}_z^H(l).$$

Hence, $\partial J_\gamma(\boldsymbol{T})/\partial \boldsymbol{P} = \mathbf{0}$ is equivalent to

$$\boldsymbol{P}\,\boldsymbol{M}_\gamma(\boldsymbol{P})\boldsymbol{P} = \boldsymbol{N}_\gamma(\boldsymbol{P}). \tag{25}$$

The equation in (25) is highly nonlinear with respect to $\boldsymbol{P}$ and, to the authors' best knowledge, a closed-form solution does not exist. Described below is a recursive approach that is found effective in solving (25). A recursive equation is generated by temporarily fixing matrix $\boldsymbol{P} = \boldsymbol{P}_k$ in both $\boldsymbol{M}_\gamma(\boldsymbol{P})$ and $\boldsymbol{N}_\gamma(\boldsymbol{P})$, which relaxes (25) to

$$\boldsymbol{P}\,\boldsymbol{M}_\gamma(\boldsymbol{P}_k)\boldsymbol{P} = \boldsymbol{N}_\gamma(\boldsymbol{P}_k) \tag{26}$$

It can be readily verified that (26) admits a closed-form solution, denoted by $\boldsymbol{P}_{k+1}$, as

$$\boldsymbol{P}_{k+1} = \boldsymbol{M}_\gamma(\boldsymbol{P}_k)^{-\frac{1}{2}}[\boldsymbol{M}_\gamma(\boldsymbol{P}_k)^{\frac{1}{2}}\boldsymbol{N}_\gamma(\boldsymbol{P}_k)\boldsymbol{M}_\gamma(\boldsymbol{P}_k)^{\frac{1}{2}}]^{\frac{1}{2}} \cdot \boldsymbol{M}_\gamma(\boldsymbol{P}_k)^{-\frac{1}{2}}. \tag{27}$$

In other words, matrix $\boldsymbol{P}_{k+1}$ given by (27) satisfies

$$\boldsymbol{P}_{k+1}\boldsymbol{M}_\gamma(\boldsymbol{P}_k)\boldsymbol{P}_{k+1} = \boldsymbol{N}_\gamma(\boldsymbol{P}_k). \tag{28}$$

With an initial $\boldsymbol{P}_0$, say $\boldsymbol{P}_0 = \boldsymbol{I}_n$, (27) recursively generates a sequence of solutions $\{\boldsymbol{P}_k | \, k = 1, 2, \cdots\}$ for the relaxed equation (26). If $\boldsymbol{P}_k$ converges to a matrix $\boldsymbol{P}^*$ as k grows, then taking limit $k \to \infty$ in (28) yields

$$\boldsymbol{P}^*\,\boldsymbol{M}_\gamma(\boldsymbol{P}_k^*)\boldsymbol{P}^* = \boldsymbol{N}_\gamma(\boldsymbol{P}_k^*) \tag{29}$$

Therefore, $\boldsymbol{P}^*$ solves the original nonlinear matrix equation (25). Although a theoretical argument for the convergence of $\{\boldsymbol{P}_k | \, k = 1, 2, \cdots\}$ is not yet available, simulations on a large number of instances have indicated that the recursive approach described here is reliable to secure an excellent approximate solution of (25) with a small number of iterations. A sample of these examples will be illustrated in Section V. In practice, the recursive process continues until

$$\Big| J_\gamma(\boldsymbol{P}_{k+1}^{\frac{1}{2}}) - J_\gamma(\boldsymbol{P}_k^{\frac{1}{2}}) \Big| < \varepsilon \tag{30}$$

is satisfied for a prescribed tolerance $\varepsilon > 0$ where $\boldsymbol{P}^{\frac{1}{2}}$ stands for the square root of matrix $\boldsymbol{P}$. If the iteration is terminated at step k, $\boldsymbol{P}_k$ is claimed to be a solution point.

We now turn our attention to the construction of a coordinate transformation matrix $\boldsymbol{T}$ that solves the problem of minimizing the weighted pole and zero sensitivity measure in (16). Since $\boldsymbol{P} = \boldsymbol{T}\boldsymbol{T}^T$, an optimizing $\boldsymbol{T}$ assumes the form

$$\boldsymbol{T} = \boldsymbol{P}^{\frac{1}{2}}\boldsymbol{U} \tag{31}$$

where $\boldsymbol{U}$ is an arbitrary $n \times n$ orthogonal matrix.

C. Minimization of Zero Sensitivity Subject to a Constraint on Pole Sensitivity

Supposing the coordinate transformation matrix $\boldsymbol{T}_o$ yields $J_p(\boldsymbol{T}_o) = n$, it can be verified that $J_p(\pm\sqrt{\zeta}\,\boldsymbol{T}_o) = n$ holds for any scalar $\pm\sqrt{\zeta}$. Therefore, one can reduce the zero sensitivity as much as possible while keep the pole sensitivity unaltered by adequately tuning the value of ζ. To this end, we compute

$$\frac{\partial J_z(\pm\sqrt{\zeta}\,\boldsymbol{T}_o)}{\partial \zeta} = \zeta^{-2}\sum_{l=1}^{n}\Big\{\zeta^2\alpha_l^2\boldsymbol{y}_z^H(l)\boldsymbol{T}_o\boldsymbol{T}_o^T\boldsymbol{y}_z(l) - \beta_l^2\boldsymbol{x}_z^H(l)\boldsymbol{T}_o^{-T}\boldsymbol{T}_o^{-1}\boldsymbol{x}_z(l)\Big\}. \tag{32}$$

By solving $\partial J_z(\pm\sqrt{\zeta}\,\boldsymbol{T}_o)/\partial\zeta = 0$, the optimal value of ζ is found to be

$$\zeta = \sqrt{\frac{\sum_{l=1}^{n}\beta_l^2\boldsymbol{x}_z^H(l)\boldsymbol{T}_o^{-T}\boldsymbol{T}_o^{-1}\boldsymbol{x}_z(l)}{\sum_{l=1}^{n}\alpha_l^2\boldsymbol{y}_z^H(l)\boldsymbol{T}_o\boldsymbol{T}_o^T\boldsymbol{y}_z(l)}}. \tag{33}$$

This implies that the optimal coordinate transformation matrix $\boldsymbol{T}^{opt}$ that minimizes $J_z(\pm\sqrt{\zeta}\,\boldsymbol{T}_o)$ with respect to ζ subject to $J_p(\boldsymbol{T}_o) = n$ can be obtained by

$$\boldsymbol{T}^{opt} = \pm\sqrt{\zeta}\,\boldsymbol{T}_o \tag{34}$$

where ζ is given by (33). Notice that by applying the arithmetic-geometric mean inequality, the following relation holds:

$$\begin{aligned} &J_z(\boldsymbol{T}_o) - J_z(\boldsymbol{T}^{opt}) \\ &= \sum_{l=1}^{n}\alpha_l^2\boldsymbol{y}_z^H(l)\boldsymbol{T}_o\boldsymbol{T}_o^T\boldsymbol{y}_z(l) + \sum_{l=1}^{n}\beta_l^2\boldsymbol{x}_z^H(l)\boldsymbol{T}_o^{-T}\boldsymbol{T}_o^{-1}\boldsymbol{x}_z(l) \\ &\quad -2\sqrt{\sum_{l=1}^{n}\alpha_l^2\boldsymbol{y}_z^H(l)\boldsymbol{T}_o\boldsymbol{T}_o^T\boldsymbol{y}_z(l)\sum_{l=1}^{n}\beta_l^2\boldsymbol{x}_z^H(l)\boldsymbol{T}_o^{-T}\boldsymbol{T}_o^{-1}\boldsymbol{x}_z(l)} \\ &\quad \geq 0. \end{aligned} \tag{35}$$

Therefore, as expected, by using the optimal $\boldsymbol{T}^{opt}$ instead of matrix $\boldsymbol{T}_o$ the zero sensitivity always gets reduced.

IV. POLE AND ZERO SENSITIVITY MINIMIZATION SUBJECT TO l_2-SCALING CONSTRAINTS

A. l_2-Scaling Constraints and Problem Statement

The controllability Gramian $\boldsymbol{K}_c$ of the digital filter in (1) plays an important role in the dynamic-range scaling of the state-variable vector $\overline{\boldsymbol{x}}(k)$, and matrix $\boldsymbol{K}_c$ can be obtained by solving the Lyapunov equation

$$\boldsymbol{K}_c = \boldsymbol{A}\boldsymbol{K}_c\boldsymbol{A}^T + \boldsymbol{b}\boldsymbol{b}^T. \tag{36}$$

With an equivalent state-space realization as specified in (11), the controllability Gramian assumes the form

$$\overline{\boldsymbol{K}}_c = \boldsymbol{T}^{-1}\boldsymbol{K}_c\boldsymbol{T}^{-T}. \tag{37}$$

If l_2-scaling constraints are imposed on the new state-variable vector $\overline{\boldsymbol{x}}(k)$ in (10), it is required that

$$\boldsymbol{e}_i^T\overline{\boldsymbol{K}}_c\boldsymbol{e}_i = \boldsymbol{e}_i^T\boldsymbol{T}^{-1}\boldsymbol{K}_c\boldsymbol{T}^{-T}\boldsymbol{e}_i = 1 \tag{38}$$

for $i = 1, 2, \cdots, n$.

The problem being considered here is to obtain an $n \times n$ transformation matrix $\boldsymbol{T}$ that minimizes (16) subject to the l_2-scaling constraints in (38).

B. Minimization of (16) Subject to l_2-Scaling Constraints — Method 1: Using A Quasi-Newton (QN) Algorithm

We start by defining

$$\hat{\boldsymbol{T}} = \boldsymbol{T}^T\boldsymbol{K}_c^{-\frac{1}{2}} \tag{39}$$

which leads (38) to

$$\boldsymbol{e}_i^T\hat{\boldsymbol{T}}^{-T}\hat{\boldsymbol{T}}^{-1}\boldsymbol{e}_i = 1 \quad \text{for} \quad i = 1, 2, \cdots, n. \tag{40}$$

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

$$\hat{\boldsymbol{T}}^{-1} = \left[\frac{\boldsymbol{t}_1}{||\boldsymbol{t}_1||}, \frac{\boldsymbol{t}_2}{||\boldsymbol{t}_2||}, \cdots, \frac{\boldsymbol{t}_n}{||\boldsymbol{t}_n||}\right]. \tag{41}$$

By defing an $n^2 \times 1$ vector $\boldsymbol{x} = (\boldsymbol{t}_1^T, \boldsymbol{t}_2^T, \cdots, \boldsymbol{t}_n^T)^T$ consisting of independent variables in $\hat{\boldsymbol{T}}^{-1}$, the pole and zero sensitivity measures $J_p(\boldsymbol{T})$ and $J_z(\boldsymbol{T})$ in (14) can be written as

$$\begin{aligned} J_p(\boldsymbol{x}) &= \sum_{l=1}^{n}\Big(\hat{\boldsymbol{x}}_p^H(l)\hat{\boldsymbol{T}}^{-1}\hat{\boldsymbol{T}}^{-T}\hat{\boldsymbol{x}}_p(l)\Big)\Big(\hat{\boldsymbol{y}}_p^H(l)\hat{\boldsymbol{T}}^T\hat{\boldsymbol{T}}\hat{\boldsymbol{y}}_p(l)\Big) \\ J_z(\boldsymbol{x}) &= \sum_{l=1}^{n}\Big(\hat{\boldsymbol{x}}_z^H(l)\hat{\boldsymbol{T}}^{-1}\hat{\boldsymbol{T}}^{-T}\hat{\boldsymbol{x}}_z(l) + \alpha_l^2\Big) \\ &\qquad \cdot\Big(\hat{\boldsymbol{y}}_z^H(l)\hat{\boldsymbol{T}}^T\hat{\boldsymbol{T}}\hat{\boldsymbol{y}}_z(l) + \beta_l^2\Big) \end{aligned} \tag{42}$$

respectively, where

$$\begin{aligned} \hat{\boldsymbol{x}}_p(l) &= \boldsymbol{K}_c^{-\frac{1}{2}}\boldsymbol{x}_p(l), \qquad \hat{\boldsymbol{y}}_p(l) = \boldsymbol{K}_c^{\frac{1}{2}}\boldsymbol{y}_p(l) \\ \hat{\boldsymbol{x}}_z(l) &= \boldsymbol{K}_c^{-\frac{1}{2}}\boldsymbol{x}_z(l), \qquad \hat{\boldsymbol{y}}_z(l) = \boldsymbol{K}_c^{\frac{1}{2}}\boldsymbol{y}_z(l). \end{aligned}$$

The original constrained optimization problem can now be converted into an unconstrained optimization problem of obtaining an $n^2 \times 1$ vector $\boldsymbol{x}$ which minimizes

$$J_\gamma(\boldsymbol{x}) = \gamma J_p(\boldsymbol{x}) + (1-\gamma)J_z(\boldsymbol{x}), \quad 0 \le \gamma \le 1. \tag{43}$$

We propose to solve the unconstrained problem by an iterative algorithm that starts with an initial point $\boldsymbol{x}_0$ obtained from the assignment $\hat{\boldsymbol{T}}^{-1} = \boldsymbol{I}_n$, updates vector $\boldsymbol{x}_k$ in the kth iteration by using (18), and terminates with a criterion stated in (19).

The gradient of $J_\gamma(\boldsymbol{x})$ can be evaluated using closed-form expressions as

$$\begin{aligned} \frac{\partial J_\gamma(\boldsymbol{x})}{\partial t_{ij}} &= \lim_{\Delta\to 0}\frac{J_\gamma(\hat{\boldsymbol{T}}_{ij}) - J_\gamma(\hat{\boldsymbol{T}})}{\Delta} \\ &= \gamma\frac{\partial J_p(\boldsymbol{x})}{\partial t_{ij}} + (1-\gamma)\frac{\partial J_z(\boldsymbol{x})}{\partial t_{ij}} \end{aligned} \tag{44}$$

where $\hat{\boldsymbol{T}}_{ij}$ is the matrix obtained from $\hat{\boldsymbol{T}}$ with a perturbed (i,j)th component. It follows that [24, p. 655]

$$\begin{aligned} \hat{\boldsymbol{T}}_{ij} &= \hat{\boldsymbol{T}} + \frac{\Delta\hat{\boldsymbol{T}}\boldsymbol{g}_{ij}\boldsymbol{e}_j^T\hat{\boldsymbol{T}}}{1-\Delta\boldsymbol{e}_j^T\hat{\boldsymbol{T}}\boldsymbol{g}_{ij}}, \quad \hat{\boldsymbol{T}}_{ij}^{-1} = \hat{\boldsymbol{T}}^{-1} - \Delta\boldsymbol{g}_{ij}\boldsymbol{e}_j^T \\ \boldsymbol{g}_{ij} &= -\partial\left\{\frac{\boldsymbol{t}_j}{||\boldsymbol{t}_j||}\right\}/\partial t_{ij} = \tfrac{1}{||\boldsymbol{t}_j||^3}(t_{ij}\boldsymbol{t}_j - ||\boldsymbol{t}_j||^2\boldsymbol{e}_i) \end{aligned} \tag{45}$$

and

$$\begin{aligned} \frac{\partial J_p(\boldsymbol{x})}{\partial t_{ij}} &= \sum_{l=1}^{n}\Big\{\Big(\hat{\boldsymbol{x}}_p^H(l)\hat{\boldsymbol{M}}(\hat{\boldsymbol{T}})\hat{\boldsymbol{x}}_p(l)\Big)\Big(\hat{\boldsymbol{y}}_p^H(l)\hat{\boldsymbol{T}}^T\hat{\boldsymbol{T}}\hat{\boldsymbol{y}}_p(l)\Big) \\ &\quad + \Big(\hat{\boldsymbol{x}}_p^H(l)\hat{\boldsymbol{T}}^{-1}\hat{\boldsymbol{T}}^{-T}\hat{\boldsymbol{x}}_p(l)\Big)\Big(\hat{\boldsymbol{y}}_p^H(l)\hat{\boldsymbol{N}}(\hat{\boldsymbol{T}})\hat{\boldsymbol{y}}_p(l)\Big)\Big\} \\ \frac{\partial J_z(\boldsymbol{x})}{\partial t_{ij}} &= \sum_{l=1}^{n}\Big\{\hat{\boldsymbol{x}}_z^H(l)\hat{\boldsymbol{M}}(\hat{\boldsymbol{T}})\hat{\boldsymbol{x}}_z(l)\Big(\hat{\boldsymbol{y}}_z^H(l)\hat{\boldsymbol{T}}^T\hat{\boldsymbol{T}}\hat{\boldsymbol{y}}_z(l) + \beta_l^2\Big) \\ &\quad + \Big(\hat{\boldsymbol{x}}_z^H(l)\hat{\boldsymbol{T}}^{-1}\hat{\boldsymbol{T}}^{-T}\hat{\boldsymbol{x}}_z(l) + \alpha_l^2\Big)\hat{\boldsymbol{y}}_z^H(l)\hat{\boldsymbol{N}}(\hat{\boldsymbol{T}})\hat{\boldsymbol{y}}_z(l)\Big\} \end{aligned} \tag{46}$$

where

$$\begin{aligned} \hat{\boldsymbol{M}}(\hat{\boldsymbol{T}}) &= -[\,\boldsymbol{g}_{ij}\boldsymbol{e}_j^T\hat{\boldsymbol{T}}^{-T} + \hat{\boldsymbol{T}}^{-1}\boldsymbol{e}_j\boldsymbol{g}_{ij}^T\,] \\ \hat{\boldsymbol{N}}(\hat{\boldsymbol{T}}) &= \hat{\boldsymbol{T}}^T\big[\boldsymbol{e}_j\boldsymbol{g}_{ij}^T\hat{\boldsymbol{T}}^T + \hat{\boldsymbol{T}}\boldsymbol{g}_{ij}\boldsymbol{e}_j^T\big]\hat{\boldsymbol{T}}. \end{aligned}$$

C. Minimization of (16) Subject to l_2-Scaling Constraints — Method 2: Using A Lagrange Function (LF)

In order to minimize (16) with $J_p(\boldsymbol{T})$ and $J_z(\boldsymbol{T})$ given by (23) with respect to an $n \times n$ symmetric positive-definite matrix $\boldsymbol{P}$ subject to l_2-scaling constraints in (38), we define the Lagrange function

$$I_\gamma(\boldsymbol{P},\xi) = J_\gamma(\boldsymbol{T}) + \xi\big(\mathrm{tr}[\boldsymbol{K}_c\boldsymbol{P}^{-1}] - n\big) \tag{47}$$

where ξ is a Lagrange multiplier, and compute the gradients

$$\begin{aligned} \frac{\partial I_\gamma(\boldsymbol{P},\xi)}{\partial\boldsymbol{P}} &= \boldsymbol{M}_\gamma(\boldsymbol{P}) - \boldsymbol{P}^{-1}\boldsymbol{N}_\gamma(\boldsymbol{P})\boldsymbol{P}^{-1} \\ &\quad - \xi\boldsymbol{P}^{-1}\boldsymbol{K}_c\boldsymbol{P}^{-1} \\ \frac{\partial I_\gamma(\boldsymbol{P},\xi)}{\partial\xi} &= \mathrm{tr}[\boldsymbol{K}_c\boldsymbol{P}^{-1}] - n \end{aligned} \tag{48}$$

where $\boldsymbol{M}_\gamma(\boldsymbol{P})$ and $\boldsymbol{N}_\gamma(\boldsymbol{P})$ are defined in (24). By setting $\partial I_\gamma(\boldsymbol{P},\xi)/\partial\boldsymbol{P} = \boldsymbol{0}$ and $\partial I_\gamma(\boldsymbol{P},\xi)/\partial\xi = 0$, it follows that

$$\boldsymbol{P}\,\boldsymbol{M}_\gamma(\boldsymbol{P})\boldsymbol{P} = \boldsymbol{G}_\gamma(\boldsymbol{P},\xi), \qquad \mathrm{tr}[\boldsymbol{K}_c\boldsymbol{P}^{-1}] = n \tag{49}$$

where

$$\boldsymbol{G}_\gamma(\boldsymbol{P},\xi) = \boldsymbol{N}_\gamma(\boldsymbol{P}) + \xi\boldsymbol{K}_c.$$

In order to solve the highly nonlinear equations in (49), we propose an iterative technique which starts with an initial iterate $\boldsymbol{P}_0$ and relaxes the first equation in (49) into a recursive second-order matrix equation

$$\boldsymbol{P}_{k+1}\boldsymbol{M}_\gamma(\boldsymbol{P}_k)\boldsymbol{P}_{k+1} = \boldsymbol{G}_\gamma(\boldsymbol{P}_k, \xi_{k+1}) \tag{50}$$

where $\boldsymbol{P}_k$ is known from the previous recursion. Solving (50) for a symmetric and positive-definite $\boldsymbol{P}_{k+1}$, we obtain

$$\begin{aligned}\boldsymbol{P}_{k+1} = {} & \boldsymbol{M}_\gamma(\boldsymbol{P}_k)^{-\frac{1}{2}}[\boldsymbol{M}_\gamma(\boldsymbol{P}_k)^{\frac{1}{2}}\boldsymbol{G}_\gamma(\boldsymbol{P}_k,\xi_{k+1})\boldsymbol{M}_\gamma(\boldsymbol{P}_k)^{\frac{1}{2}}]^{\frac{1}{2}} \\ & \cdot \boldsymbol{M}_\gamma(\boldsymbol{P}_k)^{-\frac{1}{2}}.\end{aligned} \tag{51}$$

The Lagrange multiplier ξ_{k+1} can be efficiently obtained using a bisection method [25] so that

$$f(\gamma,\xi_{k+1}) = \Big| n - \mathrm{tr}\big[\tilde{\boldsymbol{K}}_k \tilde{\boldsymbol{G}}_k(\gamma,\xi_{k+1})\big]\Big| < \varepsilon \tag{52}$$

where

$$\begin{aligned}\tilde{\boldsymbol{G}}_k(\gamma,\xi_{k+1}) &= [\boldsymbol{M}_\gamma(\boldsymbol{P}_k)^{\frac{1}{2}}\boldsymbol{G}_\gamma(\boldsymbol{P}_k,\xi_{k+1})\boldsymbol{M}_\gamma(\boldsymbol{P}_k)^{\frac{1}{2}}]^{-\frac{1}{2}} \\ \tilde{\boldsymbol{K}}_k &= \boldsymbol{M}_\gamma(\boldsymbol{P}_k)^{\frac{1}{2}}\boldsymbol{K}_c\boldsymbol{M}_\gamma(\boldsymbol{P}_k)^{\frac{1}{2}}.\end{aligned}$$

This iteration process continues until

$$\Big| I_\gamma\big(\boldsymbol{P}_{k+1},\xi_{k+1}\big) - I_\gamma\big(\boldsymbol{P}_k,\xi_k\big)\Big| < \varepsilon \tag{53}$$

is satisfied for a prescribed tolerance $\varepsilon > 0$. If the iteration is terminated at step k, we set $\boldsymbol{P} = \boldsymbol{P}_k$ and claim it to be a solution.

Having obtained an optimal $\boldsymbol{P}$, we now turn our attention to the construction of the optimal coordinate transformation matrix $\boldsymbol{T}$ that solves the problem of minimizing (16) subject to l_2-scaling constraints in (38). As is analyzed earlier, the optimal $\boldsymbol{T}$ assumes the form

$$\boldsymbol{T} = \boldsymbol{P}^{\frac{1}{2}}\boldsymbol{U} \tag{54}$$

where $\boldsymbol{U}$ is an $n\times n$ orthogonal matrix that can be determined to satisfy the l_2-scaling constraints as follows. From (37) and (54) it follows that

$$\overline{\boldsymbol{K}}_c = \boldsymbol{U}^T\boldsymbol{P}^{-\frac{1}{2}}\boldsymbol{K}_c\boldsymbol{P}^{-\frac{1}{2}}\boldsymbol{U}. \tag{55}$$

In order to find an $n\times n$ orthogonal matrix $\boldsymbol{U}$ such that the matrix $\overline{\boldsymbol{K}}_c$ satisfies l_2-scaling constraints in (38), we perform the eigenvalue-eigenvector decomposition for the symmetric positive-definite matrix $\boldsymbol{P}^{-\frac{1}{2}}\boldsymbol{K}_c\boldsymbol{P}^{-\frac{1}{2}}$ as

$$\boldsymbol{P}^{-\frac{1}{2}}\boldsymbol{K}_c\boldsymbol{P}^{-\frac{1}{2}} = \boldsymbol{R}\boldsymbol{\Theta}\boldsymbol{R}^T \tag{56}$$

where $\boldsymbol{\Theta} = \mathrm{diag}\{\theta_1,\theta_2,\cdots,\theta_n\}$ with $\theta_i > 0$ for all i and $\boldsymbol{R}$ is an $n\times n$ orthogonal matrix. Next, an $n\times n$ orthogonal matrix $\boldsymbol{S}$ such that

$$\boldsymbol{e}_i^T\boldsymbol{S}\boldsymbol{\Theta}\boldsymbol{S}^T\boldsymbol{e}_i = 1 \quad \text{for} \quad i = 1,2,\cdots,n \tag{57}$$

can be obtained by a numerical procedure [27, p.278]. Using (55)-(57), it can be readily verified that the orthogonal matrix $\boldsymbol{U} = \boldsymbol{R}\boldsymbol{S}^T$ leads to a $\overline{\boldsymbol{K}}_c$ in (55) whose diagonal elements are equal to unity, hence the l_2-scaling constraints in (38) are satisfied. This matrix $\boldsymbol{U}$ together with (54) yields the solution of the problem of minimizing (16) subject to the l_2-scaling constraints in (38) as

$$\boldsymbol{T} = \boldsymbol{P}^{\frac{1}{2}}\boldsymbol{R}\boldsymbol{S}^T. \tag{58}$$

V. SIMULATION STUDIES

Example 1:

A. Filter Description and Pole and Zero Sensitivity

Consider a digital filter $(\boldsymbol{A},\boldsymbol{b},\boldsymbol{c},d)_4$ in (1) studied in [18]

$$\boldsymbol{A} = \begin{bmatrix} 3.7183 & 1 & 0 & 0 \\ -5.2153 & 0 & 1 & 0 \\ 3.2689 & 0 & 0 & 1 \\ -0.7724 & 0 & 0 & 0 \end{bmatrix}, \quad \boldsymbol{b} = 10^{-3}\begin{bmatrix} 0.1755 \\ 0.0178 \\ 0.1652 \\ 0.0052 \end{bmatrix}$$

$$\boldsymbol{c} = \begin{bmatrix} 1 & 0 & 0 & 0 \end{bmatrix}, \qquad d = 0.0227$$

for which the eigenvalues of $\boldsymbol{A}$ and $\boldsymbol{Z}$ were found to be

$$\begin{aligned}\lambda &= 0.963556 \pm j0.156437, \quad 0.895594 \pm j0.092081 \\ v &= 1.147681 \pm j0.329541, \quad 0.707604 \pm j0.202980\end{aligned}$$

respectively. By using (9c), we obtained

$$\begin{bmatrix} \alpha_1 \\ \alpha_2 \\ \alpha_3 \\ \alpha_4 \end{bmatrix} = \begin{bmatrix} 12.215615 \\ 12.215615 \\ 9.687213 \\ 9.687213 \end{bmatrix}, \quad \begin{bmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \\ \beta_4 \end{bmatrix} = \begin{bmatrix} 0.386463 \\ 0.386463 \\ 0.310390 \\ 0.310390 \end{bmatrix}.$$

The original pole and zero sensitivity measures in (9) were computed as $J_p = 7.209829\times 10^6$ and $J_z = 1.135788\times 10^6$, respectively. The controllability Gramian was computed using (36) as

$$\boldsymbol{K}_c = \begin{bmatrix} 0.071608 & -0.195274 & 0.178646 & -0.054830 \\ -0.195274 & 0.533746 & -0.489317 & 0.150472 \\ 0.178646 & -0.489317 & 0.449437 & -0.138452 \\ -0.054830 & 0.150472 & -0.138452 & 0.042721 \end{bmatrix}.$$

B. Sensitivity Minimization Free from l_2-Scaling Constraints

B-1. The Use of A Quasi-Newton (QN) Algorithm

By choosing $\gamma = 1$ in (16), $\varepsilon = 10^{-8}$ in (19), and $\boldsymbol{T}^{-1} = [\boldsymbol{t}_1,\boldsymbol{t}_2,\boldsymbol{t}_3,\boldsymbol{t}_4] = \boldsymbol{I}_4$ in (17) as an initial estimate in the iteration vector $\boldsymbol{x}_k$ of (18), it took the proposed algorithm 29 iterations to converge to

$$\boldsymbol{T}^{-1} = \begin{bmatrix} 0.231986 & 0.185815 & 0.142997 & 0.105745 \\ 0.693092 & 0.716739 & 0.725929 & 0.720384 \\ 0.690995 & 0.761630 & 0.824477 & 0.876111 \\ 0.072256 & 0.206669 & 0.360409 & 0.533445 \end{bmatrix}$$

with $J_p(\boldsymbol{T}) = 4.000000$, $J_z(\boldsymbol{T}) = 1.858629\times 10^8$. In this case, from (11) it was observed that $\overline{\boldsymbol{A}}\,\overline{\boldsymbol{A}}^T = \overline{\boldsymbol{A}}^T\overline{\boldsymbol{A}}$ holds, hence matrix $\overline{\boldsymbol{A}}$ is normal and the lower bound $n = 4$ of $J_p(\boldsymbol{T})$ is achieved. The pole sensitivity performance of 29 iterations is shown in Fig. 1, from which it is seen that the iterative algorithm converges with 29 iterations.

Since $J_p(\boldsymbol{T}) = 4$ is satisfied, we set $\boldsymbol{T} = \boldsymbol{T}_o$. Then, by applying (33) and (34) it was found that

$$\boldsymbol{T}^{opt} = \begin{bmatrix} 0.073901 & -0.330545 & 0.317280 & -0.089358 \\ -0.150176 & 0.919582 & -0.899730 & 0.265615 \\ 0.090774 & -0.856028 & 0.856192 & -0.268161 \\ -0.013158 & 0.266860 & -0.272865 & 0.091927 \end{bmatrix}$$

where $J_p(\boldsymbol{T}^{opt}) = 4.000000$, $J_z(\boldsymbol{T}^{opt}) = 593.244132$ and $\zeta = 6.856655\times 10^{-7}$. It is noted that $\boldsymbol{T}^{opt}$ obtained above

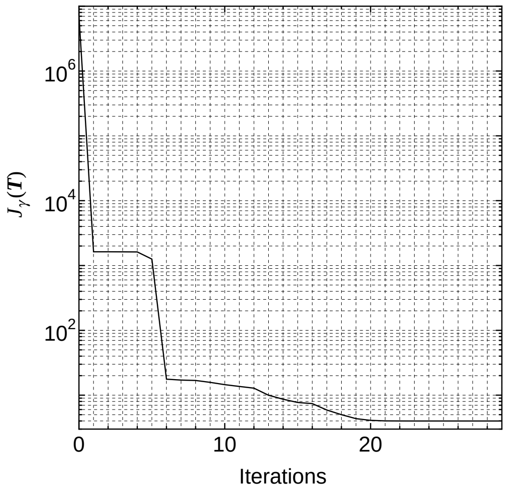

Fig. 1. Profile of $J_\gamma(\boldsymbol{T})$ during the first 29 iterations with $\gamma = 1$

corresponds to the optimal coordinate transformation matrix that minimizes a zero sensitivity measure $J_z(\pm\sqrt{\zeta}\,\boldsymbol{T}_o)$ with respect to ζ subject to minimal pole sensitivity satisfying $J_p(\boldsymbol{T}_o) = 4$.

Detailed numerical results of the optimized pole and zero sensitivity measures corresponding to various weight values of γ are summarized in Table I where $J_z(\boldsymbol{T}) \geq \sum_{l=1}^{4}(1 + |\alpha_l \beta_l|)^2 = 97.566165$ always holds. From Table I, it is observed that the lower bound of $J_z(\boldsymbol{T})$ was achieved with $\gamma = 0$. It is also observed that the sum of pole sensitivity and zero sensitivity reaches minimum when the weight value $\gamma = 0.5$ is chosen. Moreover, with γ as an adjustable design parameter, the proposed weighted optimization provides a variety of options to suit the need of a particular application where the designer may prefer smaller pole sensitivity or smaller zero sensitivity.

TABLE I
PERFORMANCE COMPARISON USING *QN* METHOD

γ	$J_\gamma(\boldsymbol{T})$	$J_p(\boldsymbol{T})$	$J_z(\boldsymbol{T})$	$J_p(\boldsymbol{T})+J_z(\boldsymbol{T})$
1.0	4.000000	4.000000	1.858629×10^8	1.858629×10^8
0.8	40.878190	13.926837	148.683601	162.610439
0.6	64.780537	22.715994	127.877352	150.593345
0.5	74.608206	27.598917	121.617495	149.216412
0.4	83.435795	34.001531	116.391971	150.393503
0.2	97.096840	56.211283	107.318230	163.529512
0.0	97.566165	283.699595	97.566165	381.265759

B-2. The Use of A Recursive Matrix Equation (RME)

By choosing $\gamma = 1$ in (16), $\varepsilon = 10^{-8}$ in (30), and $\boldsymbol{P}_0 = \boldsymbol{I}_4$ in (27) as an initial estimate, it took the proposed iterative algorithm iterations to converge to

$$\boldsymbol{P} = 10^3 \begin{bmatrix} 0.318569 & -0.890465 & 0.835038 & -0.262540 \\ -0.890465 & 2.494146 & -2.343591 & 0.738320 \\ 0.835038 & -2.343591 & 2.206461 & -0.696498 \\ -0.262540 & 0.738320 & -0.696498 & 0.220301 \end{bmatrix}$$

which yields

$$\boldsymbol{T} = 10 \begin{bmatrix} 0.504660 & -1.279387 & 1.093553 & -0.313564 \\ -1.279387 & 3.491572 & -3.188277 & 0.973880 \\ 1.093553 & -3.188277 & 3.110342 & -1.014605 \\ -0.313564 & 0.973880 & -1.014605 & 0.356118 \end{bmatrix}$$

where $\boldsymbol{U} = \boldsymbol{I}_4$ was used in (31). From (14) it follows that $J_p(\boldsymbol{T}) = 4.000000$, $J_z(\boldsymbol{T}) = 1.819305 \times 10^5$. The weighted pole and zero sensitivity performance $J_\gamma(\boldsymbol{P}_k^{\frac{1}{2}})$ were obtained as $J_\gamma(\boldsymbol{P}_0^{\frac{1}{2}}) = 7.209829 \times 10^6$ and $J_\gamma(\boldsymbol{P}_k^{\frac{1}{2}}) = 4.000000$ for $k = 1, 2$ from which it is seen that the iterative algorithm converges to the optimal solution with two iterations.

Since matrix $\boldsymbol{T}$ shown above yields $J_p(\boldsymbol{T}) = 4$, we set $\boldsymbol{T} = \boldsymbol{T}_o$. Then, by applying (33) and (34) it was found that

$$\boldsymbol{T}^{opt} = \begin{bmatrix} 0.133686 & -0.338914 & 0.289686 & -0.083064 \\ -0.338914 & 0.924930 & -0.844586 & 0.257984 \\ 0.289686 & -0.844586 & 0.823940 & -0.268772 \\ -0.083064 & 0.257984 & -0.268772 & 0.094337 \end{bmatrix}$$

and $J_z(\boldsymbol{T}^{opt}) = 593.186873$ where $\zeta = 7.017385 \times 10^{-4}$.

Detailed numerical results for the optimized pole and zero sensitivity measures corresponding to various weight values γ are summarized in Table II.

TABLE II
PERFORMANCE COMPARISON USING *RME* METHOD

γ	$J_\gamma(\boldsymbol{T})$	$J_p(\boldsymbol{T})$	$J_z(\boldsymbol{T})$	$J_p(\boldsymbol{T})+J_z(\boldsymbol{T})$
1.0	4.000000	4.000000	1.819305×10^5	1.819345×10^5
0.8	40.878190	13.926822	148.683663	162.610485
0.6	64.638191	22.453675	127.914966	150.368641
0.5	74.608206	27.598918	121.617494	149.216412
0.4	83.435795	34.001550	116.391959	150.393509
0.2	97.096840	56.211294	107.318227	163.529521
0.0	97.566165	283.698424	97.566165	381.264589

C. Sensitivity Minimization Subject to l_2-Scaling Constraints

C-1. The Use of A Quasi-Newton (QN) Algorithm

By choosing $\gamma = 1$ in (43), $\varepsilon = 10^{-8}$ in (19), and starting with an initial point $\boldsymbol{x}_0$ obtained from the assignment $\hat{\boldsymbol{T}}^{-1} = \boldsymbol{I}_4$ in (41), it took the proposed algorithm 19 iterations to converge to

$$\hat{\boldsymbol{T}}^{-1} = \begin{bmatrix} 0.864243 & 0.900469 & 0.786701 & 0.542671 \\ -0.490448 & 0.391427 & 0.345446 & 0.722078 \\ -0.094086 & -0.168073 & 0.507431 & -0.076744 \\ 0.060769 & -0.087704 & 0.065444 & 0.422163 \end{bmatrix}$$

or equivalently,

$$\boldsymbol{T} = \begin{bmatrix} 0.149803 & -0.161958 & 0.261061 & -0.203040 \\ -0.366793 & 0.433947 & -0.742778 & 0.584686 \\ 0.299031 & -0.385952 & 0.706504 & -0.563004 \\ -0.080834 & 0.113909 & -0.224536 & 0.181708 \end{bmatrix}$$

where $J_p(\boldsymbol{T}) = 4.000000$ and $J_z(\boldsymbol{T}) = 936.412522$. With an equivalent state-space realization as specified in (11), the controllability Gramian was computed from (37) as

$$\overline{\boldsymbol{K}}_c = \begin{bmatrix} 1.000000 & 0.596733 & 0.466712 & 0.147733 \\ 0.596733 & 1.000000 & 0.752591 & 0.747172 \\ 0.466712 & 0.752591 & 1.000000 & 0.665045 \\ 0.147733 & 0.747172 & 0.665045 & 1.000000 \end{bmatrix}$$

where the l_2-scaling constraints are satisfied. In addition, from (11) it was observed that $\overline{\boldsymbol{A}}\,\overline{\boldsymbol{A}}^T = \overline{\boldsymbol{A}}^T\overline{\boldsymbol{A}}$ holds. This means that matrix $\overline{\boldsymbol{A}}$ is normal and the lower bound $n = 4$ of $J_p(\boldsymbol{T})$ is achieved. The pole sensitivity performance of 19 iterations is shown in Fig. 2, from which it is seen that the iterative algorithm converges with 19 iterations.

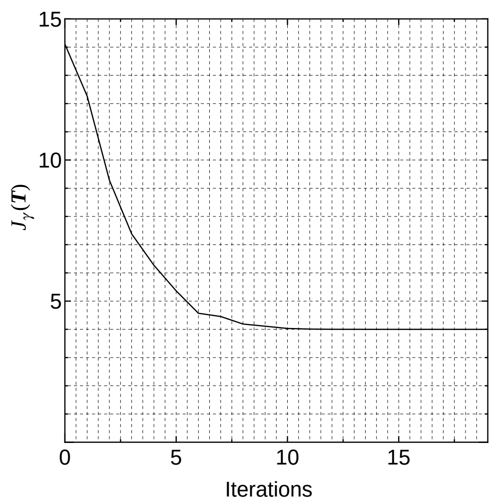

Fig. 2. Profile of $J_\gamma(\boldsymbol{T})$ during the first 19 iterations with $\gamma = 1$

Detailed numerical results of the optimized pole and zero sensitivity measures subject to the l_2-scaling constraints corresponding to various values of γ are summarized in Table III.

TABLE III
QN METHOD SUBJECT TO SCALING CONSTRAINTS

γ	$J_\gamma(\boldsymbol{T})$	$J_p(\boldsymbol{T})$	$J_z(\boldsymbol{T})$	$J_p(\boldsymbol{T})+J_z(\boldsymbol{T})$
1.0	4.000000	4.000000	936.412522	940.412522
0.8	41.489644	14.827462	148.138371	162.965833
0.6	64.863774	23.442437	126.995780	150.438217
0.4	83.435914	33.963942	116.417229	150.381170
0.2	97.629962	51.388153	109.190414	160.578567
0.0	104.016409	125.930606	104.016409	229.947015

C-2. The Use of A Lagrange Function (LF)

By setting $\gamma = 1$ in (16), $\varepsilon = 10^{-8}$ in (53), and choosing $\boldsymbol{P}_0 = \boldsymbol{I}_4$ in (51) as an initial estimate, it took the proposed iterative algorithm 8 iterations to converge to

$$\boldsymbol{P} = 10\begin{bmatrix} 0.336146 & -0.948542 & 0.899849 & -0.286556 \\ -0.948542 & 2.679553 & -2.544864 & 0.811398 \\ 0.899849 & -2.544864 & 2.419739 & -0.772465 \\ -0.286556 & 0.811398 & -0.772465 & 0.246930 \end{bmatrix}$$

which from (58) yields

$$\boldsymbol{T} = \begin{bmatrix} 1.163353 & -0.501460 & 1.162615 & -0.636344 \\ -3.325136 & 1.323961 & -3.211571 & 1.916231 \\ 3.196696 & -1.163301 & 2.981285 & -1.933181 \\ -1.032291 & 0.337201 & -0.927826 & 0.655060 \end{bmatrix}$$

where $J_p(\boldsymbol{T}) = 4.000000$ and $J_z(\boldsymbol{T}) = 5452.531195$. For the new realization, the controllability Gramian was found from (37) to be

$$\overline{\boldsymbol{K}}_c = \begin{bmatrix} 1.000000 & 0.296865 & -0.585417 & 0.804018 \\ 0.296865 & 1.000000 & 0.585417 & 0.804018 \\ -0.585417 & 0.585417 & 1.000000 & 0.000000 \\ 0.804018 & 0.804018 & 0.000000 & 1.000000 \end{bmatrix}$$

where the l_2-scaling constraints are satisfied. $I_\gamma(\boldsymbol{P}, \xi)$ performance of 8 iterations is shown in Fig. 3, from which it is seen that the iterative algorithm converges with 8 iterations where

$$\begin{bmatrix} \xi_1 & \xi_2 \\ \xi_3 & \xi_4 \\ \xi_5 & \xi_6 \\ \xi_7 & \xi_8 \end{bmatrix} = \begin{bmatrix} -1.069263 \times 10^6 & -4.213525 \times 10^2 \\ -7.117870 \times 10^0 & -4.496284 \times 10^{-1} \\ -1.817527 \times 10^{-2} & -5.577404 \times 10^{-4} \\ -1.676641 \times 10^{-5} & -5.025417 \times 10^{-7} \end{bmatrix}.$$

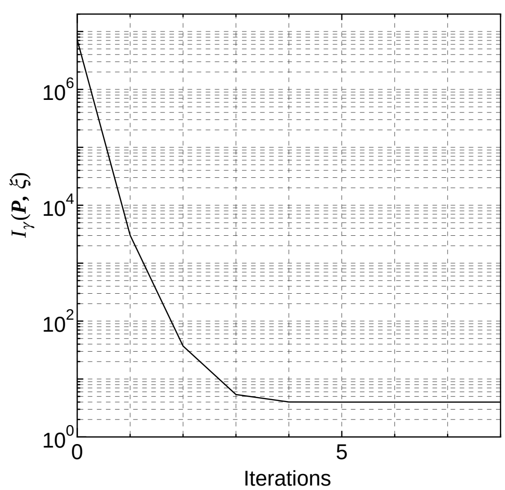

Fig. 3. Profile of $I_\gamma(\boldsymbol{P}, \xi)$ during the first 8 iterations with $\gamma = 1$

Detailed numerical results of the optimized pole and zero sensitivity measures subject to l_2-scaling constraints corresponding to various values of γ are summarized in Table IV.

TABLE IV
LF METHOD SUBJECT TO SCALING CONSTRAINTS

γ	$J_\gamma(\boldsymbol{T})$	$J_p(\boldsymbol{T})$	$J_z(\boldsymbol{T})$	$J_p(\boldsymbol{T})+J_z(\boldsymbol{T})$
1.0	4.000000	4.000000	5452.531195	5456.531195
0.8	41.489644	14.827471	148.138333	162.965804
0.6	64.863774	23.442505	126.995679	150.438183
0.4	83.435914	33.963990	116.417196	150.381186
0.2	97.629962	51.388289	109.190380	160.578669
0.0	104.016409	125.930718	104.016409	229.947127

We now present some remarks on the numerical results summarized in Tables III and IV. First, for a γ in the entire range $0 \leq \gamma \leq 1$, both the QN and LF methods yield identical minimized weighted sensitivity $J_\gamma(\boldsymbol{T})$. Second, excluding the case of $\gamma = 1$, for each $\gamma \in [0, 1)$ the individual zero and pole sensitivities (hence their sum) associated the corresponding optimized transformation matrix $\boldsymbol{T}$ obtained by both methods are practically identical as well. Finally, concerning the case of $\gamma = 1$, it follows from (16) that the use of $\gamma = 1$ simply excludes $J_z(\boldsymbol{T})$ in the optimization procedure and, as a result, the zero sensitivity went wildly large as shown in both Tables IV and V. For practical system implementations, therefore, the use of γ in the range $0 \leq \gamma < 1$ is recommended.

In [17], the pole and zero sensitivity was examined for the realizations with minimum roundoff nosie subject to l_2-scaling constraints where the MRH realization [26],[27] does not equip identity state residue correction, while the optimal realization [17] does. In [10] and [11], the problem of minimizing the l_2-sensitivity of a transfer function subject to l_2-scaling constraints was solved by different techniques. In Table V the performances of the proposed methods are compared with those acheieved by the methods in [26], [27], [17], [10] and [11] for six implementation settings where S_p and S_z are defined by $S_p = ||\boldsymbol{A}\boldsymbol{A}^T - \boldsymbol{A}^T\boldsymbol{A}||_F$ and $S_z = ||\boldsymbol{Z}\boldsymbol{Z}^T - \boldsymbol{Z}^T\boldsymbol{Z}||_F$ which approximately measure pole and zero sensitivities, respectively [17]. It is noted that S_p and S_z express the degree of normality of matrices $\boldsymbol{A}$ and $\boldsymbol{Z}$, respectively.

TABLE V
PERFORMANCE COMPARISON AMONG SIX METHODS

Realization	J_p	J_z	S_p	S_z
QN ($\gamma = 1$)	4	936.41	5.677×10^{-9}	241.64
QN ($\gamma = 0$)	125.93	104.02	0.1389	0.3459
LF ($\gamma = 1$)	4	5452.53	1.067×10^{-8}	1776.11
LF ($\gamma = 0$)	125.93	104.02	0.1389	0.3459
MRH [26],[27]	5.25	312.53	0.0225	37.89
Optimal [17]	10.06	196.95	0.0633	11.18
Optimal [10],[11]	4.63	398.79	0.0257	60.12

From this table it is obseved that in case $\gamma = 1$, the proposed methods provide lower pole sensitivity than those in [26], [27], [17], [10] and [11], and in case $\gamma = 0$, the proposed techniques produce lower zero sensitivity than those in [26], [27], [17], [10] and [11]. It is noted that the results obtained using the technique in [10] practically coincided with those derived from the technique in [11].

Example 2:

A. Filter Description and Pole and Zero Sensitivity

Consider a digital filter $(\boldsymbol{A}, \boldsymbol{b}, \boldsymbol{c}, d)_3$ in (1) studied in [28]

$$\boldsymbol{A}^T = \begin{bmatrix} 0 & 0 & 0.339377 \\ 1 & 0 & -1.152652 \\ 0 & 1 & 1.520167 \end{bmatrix}, \qquad \boldsymbol{b} = \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}$$

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

$$d = 0.065959$$

for which the eigenvalues of $\boldsymbol{A}$ and $\boldsymbol{Z}$ were found to be

$$\lambda = 0.486510 \pm j0.619334, \qquad 0.547146$$

$$v = -0.658423 \pm 0.349187, \qquad -1.934315$$

respectively. By using (9c), we obtained

$$\begin{bmatrix} \alpha_1 \\ \alpha_2 \\ \alpha_3 \end{bmatrix} = \begin{bmatrix} 1.621793 \\ 1.621793 \\ 3.579089 \end{bmatrix}, \qquad \begin{bmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \end{bmatrix} = 10 \begin{bmatrix} 2.240574 \\ 2.240574 \\ 3.750803 \end{bmatrix}.$$

The original pole and zero sensitivity measures in (9) were computed as $J_p = 3.939670 \times 10$ and $J_z = 2.348783 \times 10^4$, respectively. The controllability Gramian was computed using (36) as

$$\boldsymbol{K}_c = \begin{bmatrix} 5.215397 & 3.869762 & 1.184455 \\ 3.869762 & 5.215397 & 3.869762 \\ 1.184455 & 3.869762 & 5.215397 \end{bmatrix}.$$

B. Sensitivity Minimization Free from l_2-Scaling Constraints

B-1. The Use of A Quasi-Newton (QN) Algorithm

By choosing $\gamma = 1$ in (16), $\varepsilon = 10^{-8}$ in (19), and $\boldsymbol{T}^{-1} = [\boldsymbol{t}_1, \boldsymbol{t}_2, \boldsymbol{t}_3] = \boldsymbol{I}_3$ in (17) as an initial estimate in the iteration vector $\boldsymbol{x}_k$ of (18), it took the proposed algorithm 9 iterations to converge to

$$\boldsymbol{T}^{-1} = \begin{bmatrix} 0.231714 & -1.014743 & 0.324011 \\ -0.988289 & 1.639910 & -1.166028 \\ 0.313418 & -1.185126 & 1.639743 \end{bmatrix}$$

with $J_p(\boldsymbol{T}) = 3.000000$ and $J_z(\boldsymbol{T}) = 4.449920 \times 10^4$. In this case, from (11) it was observed that $\overline{\boldsymbol{A}}\,\overline{\boldsymbol{A}}^T = \overline{\boldsymbol{A}}^T\overline{\boldsymbol{A}}$ holds, hence matrix $\overline{\boldsymbol{A}}$ is normal and the lower bound $n = 3$ of $J_p(\boldsymbol{T})$ is achieved. The sensitivity performance $J_\gamma(\boldsymbol{T})$ of 9 iterations is shown in Fig. 4, from which it is seen that the iterative algorithm converges with 9 iterations.

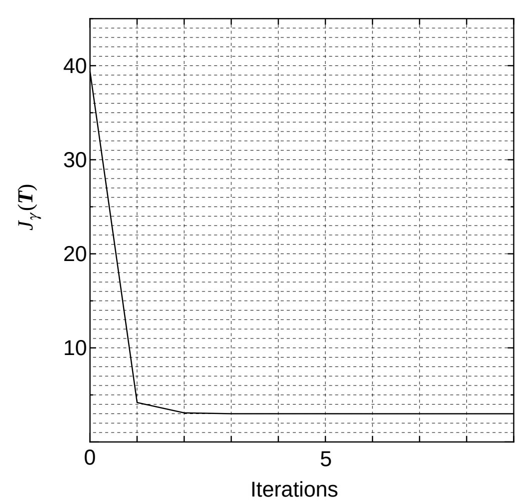

Fig. 4. Profile of $J_\gamma(\boldsymbol{T})$ during the first 9 iterations with $\gamma = 1$

Since $J_p(\boldsymbol{T}) = 3$ is satisfied, we set $\boldsymbol{T} = \boldsymbol{T}_o$. Then, by applying (33) and (34) it was found that

$$\boldsymbol{T}^{opt} = \begin{bmatrix} -2.943887 & -2.882586 & -1.468110 \\ -2.826649 & -0.627000 & 0.112680 \\ -1.480273 & 0.097808 & 1.402799 \end{bmatrix}$$

where $J_p(\boldsymbol{T}^{opt}) = 3.000000$, $J_z(\boldsymbol{T}^{opt}) = 3.643451 \times 10^4$ and $\zeta = 2.9123379$. It is noted that $\boldsymbol{T}^{opt}$ obtained above corresponds to the optimal coordinate transformation matrix that minimizes a zero sensitivity measure $J_z(\pm\sqrt{\zeta}\,\boldsymbol{T}_o)$ with respect to ζ subject to minimal pole sensitivity satisfying $J_p(\boldsymbol{T}_o) = 3$.

Detailed numerical results of the optimized pole and zero sensitivity measures corresponding to various weight values of γ are summarized in Table VI where $J_z(\boldsymbol{T}) \geq \sum_{l=1}^{3}(1+|\alpha_l\beta_l|)^2 = 2.107926 \times 10^4$ always holds. From Table VI, it is observed that the lower bound of $J_z(\boldsymbol{T})$ was achieved with $\gamma = 0$. Moreover, with γ as an adjustable design parameter, the proposed weighted optimization provides a variety of options to suit the need of a particular application where the designer may prefer smaller pole sensitivity or smaller zero sensitivity.

TABLE VI
PERFORMANCE COMPARISON USING *QN* METHOD

γ	$J_\gamma(\boldsymbol{T})$	$J_p(\boldsymbol{T})$	$J_z(\boldsymbol{T})$
1.0	3.000000	3.000000	4.449920×10^4
0.8	4.405814×10^3	1.036408×10^2	2.161451×10^4
0.6	8.676070×10^3	1.895820×10^2	2.140580×10^4
0.4	1.289555×10^4	3.099084×10^2	2.128598×10^4
0.2	1.706066×10^4	5.488701×10^2	2.118861×10^4
0.0	2.107926×10^4	3.208545×10^3	2.107926×10^4

B-2. The Use of A Recursive Matrix Equation (RME)

By choosing $\gamma = 1$ in (16), $\varepsilon = 10^{-8}$ in (30), and $\boldsymbol{P}_0 = \boldsymbol{I}_3$ in (27) as an initial estimate, it took the proposed iterative algorithm two iterations to converge to

$$\boldsymbol{P} = \begin{bmatrix} 6.321617 & 3.278396 & 0.564446 \\ 3.278396 & 2.847494 & 1.446066 \\ 0.564446 & 1.446066 & 1.444805 \end{bmatrix}$$

which yields

$$\boldsymbol{T} = \begin{bmatrix} 2.345821 & 0.904843 & -0.001243 \\ 0.904843 & 1.278210 & 0.628437 \\ -0.001243 & 0.628437 & 1.024632 \end{bmatrix}$$

where $\boldsymbol{U} = \boldsymbol{I}_3$ was used in (31). From (14) it follows that $J_p(\boldsymbol{T}) = 3.000000$, $J_z(\boldsymbol{T}) = 4.490190 \times 10^4$. The weighted pole and zero sensitivity performance $J_\gamma(\boldsymbol{P}_k^{1/2})$ were obtained as $J_\gamma(\boldsymbol{P}_0^{1/2}) = 3.939670 \times 10$ and $J_\gamma(\boldsymbol{P}_k^{1/2}) = 3.000000$ for $k = 1, 2$ from which it is seen that the iterative algorithm converges to the optimal solution with two iterations.

Since matrix $\boldsymbol{T}$ shown above yields $J_p(\boldsymbol{T}) = 3$, we set $\boldsymbol{T} = \boldsymbol{T}_o$. Then, by applying (33) and (34) it was found that

$$\boldsymbol{T}^{opt} = \begin{bmatrix} 4.060448 & 1.566218 & -0.002152 \\ 1.566218 & 2.212490 & 1.087780 \\ -0.002152 & 1.087780 & 1.773564 \end{bmatrix}$$

and $J_z(\boldsymbol{T}^{opt}) = 3.638392 \times 10^4$ where $\zeta = 2.996114$.

Detailed numerical results for the optimized pole and zero sensitivity measures corresponding to various weight values γ are summarized in Table VII.

TABLE VII
PERFORMANCE COMPARISON USING *RME* METHOD

γ	$J_\gamma(\boldsymbol{T})$	$J_p(\boldsymbol{T})$	$J_z(\boldsymbol{T})$
1.0	3.000000	3.000000	4.490190×10^4
0.8	4.405814×10^3	1.036407×10^2	2.161451×10^4
0.6	8.676070×10^3	1.895825×10^2	2.140580×10^4
0.4	1.289555×10^4	3.099089×10^2	2.128598×10^4
0.2	1.706066×10^4	5.488713×10^2	2.118861×10^4
0.0	2.107926×10^4	3.208545×10^3	2.107926×10^4

C. Sensitivity Minimization Subject to l_2-Scaling Constraints

C-1. The Use of A Quasi-Newton (QN) Algorithm

By choosing $\gamma = 1$ in (43), $\varepsilon = 10^{-8}$ in (19), and starting with an initial point $\boldsymbol{x}_0$ obtained from the assignment $\hat{\boldsymbol{T}}^{-1} = \boldsymbol{I}_3$ in (41), it took the proposed algorithm 17 iterations to converge to

$$\hat{\boldsymbol{T}}^{-1} = \begin{bmatrix} 0.197596 & -0.185863 & -0.393290 \\ 0.933859 & 0.041955 & 0.085987 \\ -0.298101 & -0.981680 & 0.915385 \end{bmatrix}$$

or equivalently,

$$\boldsymbol{T} = \begin{bmatrix} 1.476859 & -3.253423 & -2.953198 \\ 2.072059 & -2.089248 & -0.485641 \\ 1.025119 & -1.420567 & 1.058596 \end{bmatrix}$$

where $J_p(\boldsymbol{T}) = 3.000000$ and $J_z(\boldsymbol{T}) = 3.694619 \times 10^4$. With an equivalent state-space realization as specified in (11), the controllability Gramian was computed from (37) as

$$\overline{\boldsymbol{K}}_c = \begin{bmatrix} 1.000000 & 0.295094 & -0.270289 \\ 0.295094 & 1.000000 & -0.821909 \\ -0.270289 & -0.821909 & 1.000000 \end{bmatrix}$$

where the l_2-scaling constraints are satisfied. In addition, from (11) it was observed that $\overline{\boldsymbol{A}}\,\overline{\boldsymbol{A}}^T = \overline{\boldsymbol{A}}^T\overline{\boldsymbol{A}}$ holds. This means that matrix $\overline{\boldsymbol{A}}$ is normal and the lower bound $n = 3$ of $J_p(\boldsymbol{T})$ is achieved. The pole sensitivity performance of 17 iterations is shown in Fig. 5, from which it is seen that the iterative algorithm converges with 17 iterations.

Detailed numerical results of the optimized pole and zero sensitivity measures subject to the l_2-scaling constraints corresponding to various values of γ are summarized in Table VIII.

TABLE VIII
QN METHOD SUBJECT TO SCALING CONSTRAINTS

γ	$J_\gamma(\boldsymbol{T})$	$J_p(\boldsymbol{T})$	$J_z(\boldsymbol{T})$
1.0	3.000000	3.000000	3.694619×10^4
0.8	4.536731×10^3	6.107676×10	2.243935×10^4
0.6	9.000127×10^3	9.222873×10	2.236197×10^4
0.4	1.344729×10^4	1.243733×10^2	2.232923×10^4
0.2	1.788259×10^4	1.641194×10^2	2.231220×10^4
0.0	2.230594×10^4	2.237873×10^2	2.230594×10^4

C-2. The Use of A Lagrange Function (LF)

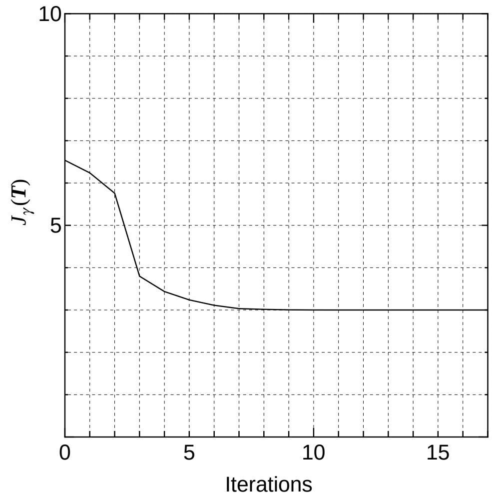

Fig. 5. Profile of $J_\gamma(\boldsymbol{T})$ during the first 17 iterations with $\gamma = 1$

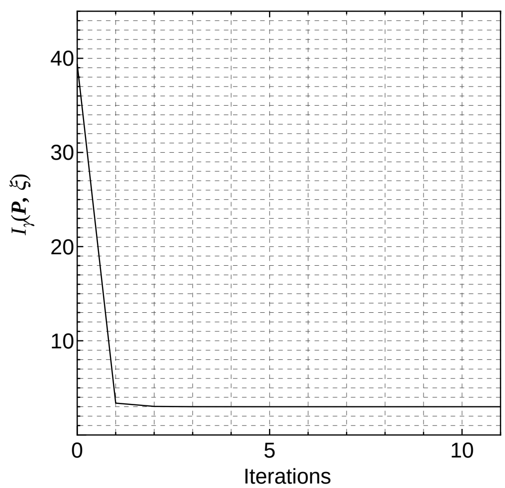

Fig. 6. Profile of $I_\gamma(\boldsymbol{P}, \xi)$ during the first 11 iterations with $\gamma = 1$

By setting $\gamma = 1$ in (16), $\varepsilon = 10^{-8}$ in (53), and choosing $\boldsymbol{P}_0 = \boldsymbol{I}_3$ in (51) as an initial estimate, it took the proposed iterative algorithm 11 iterations to converge to

$$\boldsymbol{P} = \begin{bmatrix} 17.924303 & 9.235070 & 1.155092 \\ 9.235070 & 8.394091 & 4.237920 \\ 1.155092 & 4.237920 & 4.391196 \end{bmatrix}$$

which from (58) yields

$$\boldsymbol{T} = \begin{bmatrix} 0.567932 & -3.070934 & 2.858518 \\ -0.972170 & -0.742317 & 2.626393 \\ -0.629230 & 1.075664 & 1.684700 \end{bmatrix}$$

where $J_p(\boldsymbol{T}) = 3.000000$ and $J_z(\boldsymbol{T}) = 3.651390 \times 10^4$. For the new realization, the controllability Gramian was found from (37) to be

$$\overline{\boldsymbol{K}}_c = \begin{bmatrix} 1.000000 & 0.651587 & 0.309100 \\ 0.651587 & 1.000000 & 0.651587 \\ 0.309100 & 0.651587 & 1.000000 \end{bmatrix}$$

where the l_2-scaling constraints are satisfied. $I_\gamma(\boldsymbol{P}, \xi)$ performance of 11 iterations is shown in Fig. 6, from which it is seen that the iterative algorithm converges with 11 iterations where

$$\begin{bmatrix} \xi_1 & \xi_2 & \xi_3 \\ \xi_4 & \xi_5 & \xi_6 \\ \xi_7 & \xi_8 & \xi_9 \\ \xi_{10} & \xi_{11} & \end{bmatrix} = \begin{bmatrix} 7.000912 & 0.318485 & 0.098598 \\ 0.035715 & 0.013400 & 0.005085 \\ 0.001937 & 0.000739 & 0.000282 \\ 0.000108 & 0.000041 & \end{bmatrix}.$$

Detailed numerical results of the optimized pole and zero sensitivity measures subject to l_2-scaling constraints corresponding to various values of γ are summarized in Table IX.

In Table X the performances of the proposed methods are compared with those acheieved by the methods in [26], [27], [17], [10] and [11] for six implementation settings. The results shown in this table reveal exactly the same tendency as in Table V.

TABLE IX
LF Method Subject to Scaling Constraints

γ	$J_\gamma(\boldsymbol{T})$	$J_p(\boldsymbol{T})$	$J_z(\boldsymbol{T})$
1.0	3.000000	3.000000	3.651390×10^4
0.8	4.536731×10^3	6.107673×10	2.243935×10^4
0.6	9.000127×10^3	9.222873×10	2.236197×10^4
0.4	1.344729×10^4	1.243733×10^2	2.232923×10^4
0.2	1.788259×10^4	1.641194×10^2	2.231220×10^4
0.0	2.230594×10^4	2.237873×10^2	2.230594×10^4

VI. Conclusion

The problem of minimizing pole and zero sensitivity for state-space filters has been investigated. Two efficient iterative techniques for minimizing a weighted pole and zero sensitivity measure have been developed by employing a quasi-Newton algorithm or relying on a recursive matrix equation. A novel and simple method has also been explored to obtain the optimal coordinate transformation matrix which minimizes a zero sensitivity measure subject to minimal pole sensitivity. Moreover, two iterative methods for minimizing the weighted pole and zero sensitivity measure subject to l_2-scaling constraints have been proposed. Simulation results have demonstrated the validity and effectiveness of the proposed techniques.

TABLE X
Performance Comparison among Six Methods

Realization	J_p	J_z	S_p	S_z
QN ($\gamma = 1$)	3	36946	8.957×10^{-6}	693.5
QN ($\gamma = 0$)	223.79	22306	24.367	5.033
LF ($\gamma = 1$)	3	36514	1.613×10^{-5}	664.3
LF ($\gamma = 0$)	223.79	22306	24.367	5.033
MRH [26],[27]	3.586	29411	0.4325	243.4
Optimal [17]	4.802	26993	0.6951	150.2
Optimal [10],[11]	3.297	31147	0.4686	307.8

REFERENCES

[1] L. Thiele, “Design of sensitivity and round-off noise optimal state-space discrete systems,” *Int. J. Circuit Theory Appl.*, vol. 12, pp. 39-46, Jan. 1984.

[2] V. Tavsanoglu and L. Thiele, “Optimal design of state-space digital filters by simultaneous minimization of sensitivity and roundoff noise,” *IEEE Trans. Circuits Syst.*, vol. CAS-31, no. 10, pp. 884-888, Oct. 1984.

[3] L. Thiele, “On the sensitivity of linear state-space systems,” *IEEE Trans. Circuits Syst.*, vol. CAS-33, no. 5, pp. 502-510, May 1986.

[4] M. Iwatsuki, M. Kawamata and T. Higuchi, “Statistical sensitivity and minimum sensitivity structures with fewer coefficients in discrete time linear systems,” *IEEE Trans. Circuits Syst.*, vol.37, no. 1, pp.72-80, Jan. 1989.

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

[6] G. Li, B. D. O. Anderson, M. Gevers and J. E. Perkins, “Optimal FWL design of state-space digital systems with weighted sensitivity minimization and sparseness consideration,” *IEEE Trans. Circuits Syst. I*, vol. 39, no. 5, pp. 365-377, May 1992.

[7] W.-Y. Yan and J. B. Moore, “On L^2-sensitivity minimization of linear state-space systems,” *IEEE Trans. Circuits Syst. I*, vol. 39, no. 8, pp. 641-648, Aug. 1992.

[8] G. Li and M. Gevers, “Optimal synthetic FWL design of state-space digital filters,” in *Proc. ICASSP 1992*, vol.4, pp. 429-432.

[9] M. Gevers and G. Li, *Parameterizations in Control, Estimation and Filtering Problems: Accuracy Aspects*, New York: Springer-Verlag, 1993.

[10] T. Hinamoto, H. Ohnishi and W.-S. Lu, “Minimization of L_2-sensitivity for state-space digital filters subject to L_2-dynamic-range scaling constraints,” *IEEE Trans. Circuits Syst.-II*, vol. 52, no. 10, pp. 641-645, Oct. 2005.

[11] T. Hinamoto, K. Iwata and W.-S. Lu, “L_2-sensitivity minimization of one- and two-dimensional state-space digital filters subject to L_2-scaling constraints,” *IEEE Trans. Signal Process.*, vol. 54, no. 5, pp. 1804-1812, May 2006.

[12] S. Yamaki, M. Abe and M. Kawamata, “A novel approach to L_2-sensitivity minimization of digital filters subject to L_2-scaling constraints,” in *Proc. ISCAS 2006*, pp. 5219-5222.

[13] S. Yamaki, M. Abe and M. Kawamata, “A closed form solution to L_2-sensitivity minimization of second-order state-space digital filters,” in *Proc. ISCAS 2006*, pp. 5223-5226.

[14] T. Hinamoto, T. Kawagoe, “Optimal synthesis of state-estimate feedback controllers with minimum l_2-sensitivity,” *IEEE Trans. Circuits Syst.-I.*, vol. 55, no. 8, pp. 2402-2410, Sep. 2008.

[15] P. E. Mantey, “Eigenvalue sensitivity and state-variable selection,” *IEEE Trans. Automatic Contr.*, vol. AC-13, no. 3, pp. 263-269, Jun. 1968.

[16] R. E. Skelton and D. A. Wagie, “Minimal root sensitivity in linear systems,” *J. Guidance Contr.*, vol. 7, no. 5, pp. 570-574, Sep.-Oct. 1984.

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

[18] G. Li, “On pole and zero sensitivity of linear systems,” *IEEE Trans. Circuits Syst. I*, vol. 44, no. 7, pp. 583-590, Jul. 1997.

[19] T. Hinamoto, A. Doi and W.-S. Lu, “Weighted pole and zero sensitivity minimization for state-space digital filters,” in *Proc. ISCAS 2015*, pp. 2193-2196.

[20] R. A. Roberts and C. T. Mullis, *Digital Signal Processing*, Addison-Wesley Publishing Company, Inc., 1987.

[21] S. Boyd and L. Vandenberghe, *Convex Optimization*, Cambridge University Press, 2004.

[22] J. E. Dennis and J. J. More, “Quasi-Newton methods, motivation and theory,” *SIAM Rev.*, vol. 19, no. 1, pp. 46-89, 1977.

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

[24] T. Kailath, *Linear System*. Englewood Cliffs, NJ: Prentice-Hall, 1980.

[25] H. Togawa, *Handbook of Numerical Methods*, Tokyo, Japan, Saiencesha, 1992.

[26] 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, Sep. 1976.

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

[28] 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. 1, 2013.

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 an Emeritus Professor 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 430 papers in these areas. He is co-editor and co-author of Two-Dimensional Signal and Image Processing (Tokyo, Japan: SICE, 1996). Dr. Hinamoto is a Fellow of the Institue of Electronics, Information and Communication Engineers (IEICE) and a Fellow of the Society of Instrument and Control Engineers (SICE).

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. From 2005 to 2015, he was an Associate Professor and since October 2015, he has been a 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).