Weighted Pole and Zero Sensitivity Minimization Subject to l_2-Scaling Constraints for State-Space Digital Filters

Takao Hinamoto
Hiroshima University
Higashi-Hiroshima 739-8527, Japan
Emails: hinamoto@ieee.org

Akimitsu Doi
Hiroshima Institute of Technology
Hiroshima 731-5193, Japan
Email: doi@cc.it-hiroshima.ac.jp

Wu-Sheng Lu
University of Victoria
Victoria, BC, Canada V8W 3P6
Email: wslu@ece.uvic.ca

***Abstract*—In this paper, the problem of minimizing a weighted pole and zero sensitivity measure subject to l_2-norm dynamic range scaling constraints for state-space digital filters is investigated. First, a new measure for evaluating pole and zero sensitivity is introduced, and it is shown that the constrained optimization problem is converted into an unconstrained optimization problem by using linear algebraic techniques. Then, the unconstrained optimization problem at hand is solved iteratively by employing an efficient quasi-Newton algorithm with closed-form formulas for key gradient evaluation. A numerical example is presented to demonstrate the validity and effectiveness of the proposed iterative optimization technique.**

I. Introduction

When designing a state-space digital filter that meets the specification requirements given by the designer, state-space parameters must be truncated or rounded to implement the filter in a finite binary representation that fits the finite-word-length (FWL) constraints. This coefficient quantization usually changes the characteristics of the filter. For instance, the characteristics of an original stable filter might be so altered that the filter becomes unstable. This motivates the study of the coefficient sensitivity minimization problem for digital filters. As is well known, there are several ways to define sensitivity of a filter with respect to its coefficients. Two of them are based on a mixed l_1/l_2 norm or a pure l_2 norm that measures changes of a certain transfer function, while the other is defined in terms of the poles and zeros of a filter. Over the past several decades, the pole and zero sensitivity of a filter with respect to state-space parameters has been analyzed, and its reduction or minimization have been investigated [1]-[6]. It has been pointed out [3] 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. Although minimal pole sensitivity is achieved when the system matrix $\boldsymbol{A}$ is normal, and minimal zero sensitivity is acheived when a certain matrix is normal, minimal pole sensitivity does not guarantee minimal zero sensitivity and vice versa. In [5], a method for minimizing a zero sensitivity measure subject to minimal ploe sensitivity for state-space digital filters has been explored, since it is generally impossible to minimize pole and zero sensitivity simultaneously. However, l_2-scaling for preventing overflow oscillations was not taken into account in [5]. Recently, the problem of minimizing weighted pole and zero sensitivity was examined and a quasi-Newton algorithm was applied to minimize the weighted pole and zero sensitivity in [6], however l_2-scaling constraints were not imposed.

In this paper, the problem of minimizing weighted pole and zero sensitivity subject to l_2-scaling constraints for state-space digital filters is investigated. An iterative technique for minimizing a weighted pole and zero sensitivity measure subject to l_2-scaling constraints is developed by applying an efficient quasi-Newton algorithm. A numerical example is included to illustrate the validity and effectiveness of the proposed iterative optimization technique.

II. Problem Formulation

A. Weighted Pole and Zero Sensitivity Measure

We 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 $n \times n$, $n \times 1$, $1 \times n$ and 1×1 real constant matrices, respectively. The transfer function of the filter in (1) can be expressed as

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

Let the poles $\{\lambda_l\}$ of $H(z)$ be denoted by $\{\lambda_l\} = \lambda(\boldsymbol{A})$, and the zeros $\{v_l\}$ of $H(z)$ be indecated by $\{v_l\} = \lambda(\boldsymbol{Z})$ with

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

Throughout this paper, we assume that a direct path from the input to the output always exists in (1), i.e., $d \neq 0$ so that $\boldsymbol{Z}$ in (3) is well defined [7].

The pole sensitivity matrix for the lth eigenvalue λ_l of matrix $\boldsymbol{A}$ is defined as [2]

$$\frac{\partial \lambda_l}{\partial \boldsymbol{A}} = \begin{bmatrix} \dfrac{\partial \lambda_l}{\partial a_{11}} & \cdots & \dfrac{\partial \lambda_l}{\partial a_{1n}} \\ \vdots & \ddots & \vdots \\ \dfrac{\partial \lambda_l}{\partial a_{n1}} & \cdots & \dfrac{\partial \lambda_l}{\partial a_{nn}} \end{bmatrix} \tag{4}$$

where a_{ij} for $i, j = 1, 2, \cdots, n$ denote the (i, j)th element of matrix $\boldsymbol{A}$. It is known that [2], [5]

$$\Big(\frac{\partial \lambda_l}{\partial \boldsymbol{A}}\Big)^T = \boldsymbol{x}_p(l)\boldsymbol{y}_p^H(l) \quad \text{for} \quad l = 1, 2, \cdots, n \tag{5}$$

holds if $\boldsymbol{A}$ has a full set of linearly independent eigenvectors, where $\boldsymbol{x}_p(l)$ is a right eigenvector corresponding to λ_l and $\boldsymbol{y}_p(l)$ is the reciprocal left eigenvector corresponding to $\boldsymbol{x}_p(l)$. Similarly, if $\boldsymbol{Z}$ has a linearly independent eigenvector set, then we have

$$\Big(\frac{\partial v_l}{\partial \boldsymbol{Z}}\Big)^T = \boldsymbol{x}_z(l)\boldsymbol{y}_z^H(l) \quad \text{for} \quad l = 1, 2, \cdots, n \tag{6}$$

where $\boldsymbol{x}_z(l)$ is a right eigenvector corresponding to v_l and $\boldsymbol{y}_z(l)$ is the reciprocal left eigenvector corresponding to $\boldsymbol{x}_z(l)$. Moreover, it is known that [4]

$$\begin{aligned} \frac{\partial v_l}{\partial \boldsymbol{A}} &= \frac{\partial v_l}{\partial \boldsymbol{Z}}, & \frac{\partial v_l}{\partial \boldsymbol{b}} &= -d^{-1}\frac{\partial v_l}{\partial \boldsymbol{Z}}\boldsymbol{c}^T \\ \frac{\partial v_l}{\partial \boldsymbol{c}} &= -d^{-1}\boldsymbol{b}^T\frac{\partial v_l}{\partial \boldsymbol{Z}}, & \frac{\partial v_l}{\partial d} &= d^{-2}\boldsymbol{b}^T\frac{\partial v_l}{\partial \boldsymbol{Z}}\boldsymbol{c}^T. \end{aligned} \tag{7}$$

Using the Frobenius norm, the pole sensitivity measure J_p and the zero sensitivity measure J_z for the filter in (1) can be defined as [5]

$$\begin{aligned} J_p &= \sum_{l=1}^{n}\Big\|\frac{\partial \lambda_l}{\partial \boldsymbol{A}}\Big\|_F^2 = \sum_{l=1}^{n}\Big\|\Big(\frac{\partial \lambda_l}{\partial \boldsymbol{A}}\Big)^T\Big\|_F^2 \\ J_z &= \sum_{l=1}^{n}\Big\{\Big\|\frac{\partial v_l}{\partial \boldsymbol{A}}\Big\|_F^2 + \Big\|\frac{\partial v_l}{\partial \boldsymbol{b}}\Big\|_F^2 + \Big\|\frac{\partial v_l}{\partial \boldsymbol{c}}\Big\|_F^2 + \Big\|\frac{\partial v_l}{\partial d}\Big\|_F^2\Big\} \end{aligned} \tag{8}$$

respectively. By making use of (5), (6) and (7), we can write (8) 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

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

Applying a coordinate transformation defined by

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

to the digital filter in (1), the new realization associated with state variable $\overline{\boldsymbol{x}}(k)$ can be characterized by $(\overline{\boldsymbol{A}}, \overline{\boldsymbol{b}}, \overline{\boldsymbol{c}}, d)_n$ with

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

For this realization, the right eigenvector $\boldsymbol{x}_p(l)$ corresponding to λ_l and the reciprocal left eigenvector $\boldsymbol{y}_p(l)$ corresponding to $\boldsymbol{x}_p(l)$ can be written as [3], [5]

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

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

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

respectively, for $l = 1, 2, \cdots, n$. Hence, for the realization $(\overline{\boldsymbol{A}}, \overline{\boldsymbol{b}}, \overline{\boldsymbol{c}}, d)_n$ specified in (11), the pole and zero sensitivity measures in (9a) and (9b) 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. Note that for any $n \times n$ nonsingular matrix $\boldsymbol{T}$

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

always hold true [2]-[5]. In this paper, we consider a weighted pole and zero sensitivity measure $J_\gamma(\boldsymbol{P})$ 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. Our consideration of this weighted measure is motivated by the fact that it allows to study various combined pole and zero sensitivity to its parameters by assigning weight γ to various values, with the extreme circumstances of "minimizing pole sensitivity only" and "minimizing zero sensitivity only" as two special cases of $\gamma = 1$ and $\gamma = 0$, respectively.

B. l_2-Scaling Constraints and Problem Statement

The controllability Gramian $\boldsymbol{K}_c$ of the filter in (1) can be obtained by solving the Lyapunov equation

$$\boldsymbol{K}_c = \boldsymbol{A}\boldsymbol{K}_c\boldsymbol{A}^T + \boldsymbol{b}\boldsymbol{b}^T. \tag{17}$$

This matrix plays an important role in the dynamic-range scaling of the state-variable vector $\overline{\boldsymbol{x}}(k)$ in (10) [8].

With a new realization as specified in (11), the controllability Gramian assumes the form

$$\overline{\boldsymbol{K}}_c = \boldsymbol{T}^{-1}\boldsymbol{K}_c\boldsymbol{T}^{-T}. \tag{18}$$

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

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

Summarizing, our purpose in this paper is to obtain an $n \times n$ coordinate transformation matrix $\boldsymbol{T}$ that minimizes (16) subject to the l_2-scaling constraints in (19).

III. Pole and Zero Sensitivity Minimization Subject to l_2-Scaling Constraints

Let us define

$$\hat{\boldsymbol{T}} = \boldsymbol{T}^T\boldsymbol{K}_c^{-\frac{1}{2}} \tag{20}$$

so that the l_2-scaling constraints in (19) become

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

It is obvious that the constraints in (21) are always satisfied if matrix $\hat{\boldsymbol{T}}^{-1}$ assumes the form

$$\hat{\boldsymbol{T}}^{-1} = \left[\frac{\boldsymbol{t}_1}{||\boldsymbol{t}_1||}, \frac{\boldsymbol{t}_2}{||\boldsymbol{t}_2||}, \cdots, \frac{\boldsymbol{t}_n}{||\boldsymbol{t}_n||} \right]. \tag{22}$$

Defining an $n^2 \times 1$ vector $\boldsymbol{x} = (\boldsymbol{t}_1^T, \boldsymbol{t}_2^T, \cdots, \boldsymbol{t}_n^T)^T$, the pole and zero sensitivity measures $J_p(\boldsymbol{T})$ and $J_z(\boldsymbol{T})$ in (14) are functions of $\boldsymbol{x}$ and can be written as

$$\begin{aligned} J_p(\boldsymbol{x}) &= \sum_{l=1}^{n} \Big(\hat{\boldsymbol{x}}_p^H(l) \hat{\boldsymbol{T}}^{-1} \hat{\boldsymbol{T}}^{-T} \hat{\boldsymbol{x}}_p(l) \Big) \Big(\hat{\boldsymbol{y}}_p^H(l) \hat{\boldsymbol{T}}^T \hat{\boldsymbol{T}} \hat{\boldsymbol{y}}_p(l) \Big) \\ J_z(\boldsymbol{x}) &= \sum_{l=1}^{n} \Big(\hat{\boldsymbol{x}}_z^H(l) \hat{\boldsymbol{T}}^{-1} \hat{\boldsymbol{T}}^{-T} \hat{\boldsymbol{x}}_z(l) + \alpha_l^2 \Big) \\ &\qquad \cdot \Big(\hat{\boldsymbol{y}}_z^H(l) \hat{\boldsymbol{T}}^T \hat{\boldsymbol{T}} \hat{\boldsymbol{y}}_z(l) + \beta_l^2 \Big) \end{aligned} \tag{23}$$

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

$$J_\gamma(\boldsymbol{x}) = \gamma J_p(\boldsymbol{x}) + (1-\gamma) J_z(\boldsymbol{x}), \quad 0 \le \gamma \le 1. \tag{24}$$

We propose to solve the unconstrained optimization problem by applying an iterative algorithm that starts with an initial point $\boldsymbol{x}_0$ obtained from the assignment $\hat{\boldsymbol{T}}^{-1} = \boldsymbol{I}_n$. In the kth iteration, a quasi-Newton algorithm updates the current point $\boldsymbol{x}_k$ to point $\boldsymbol{x}_{k+1}$ as [9]

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

where

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

$\nabla J_\gamma(\boldsymbol{x})$ 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

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

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

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

$$\begin{aligned} \frac{\partial J_\gamma(\boldsymbol{x})}{\partial t_{ij}} &= \lim_{\Delta \to 0} \frac{J_\gamma(\hat{\boldsymbol{T}}_{ij}) - J_\gamma(\hat{\boldsymbol{T}})}{\Delta} \\ &= \gamma \frac{\partial J_p(\boldsymbol{x})}{\partial t_{ij}} + (1-\gamma) \frac{\partial J_z(\boldsymbol{x})}{\partial t_{ij}} \end{aligned} \tag{27}$$

where $\hat{\boldsymbol{T}}_{ij}$ is the matrix obtained from $\hat{\boldsymbol{T}}$ with a perturbed (i,j)th component. It follows that [10, p. 655]

$$\begin{aligned} \hat{\boldsymbol{T}}_{ij} &= \hat{\boldsymbol{T}} + \frac{\Delta \hat{\boldsymbol{T}} \boldsymbol{g}_{ij} \boldsymbol{e}_j^T \hat{\boldsymbol{T}}}{1 - \Delta \boldsymbol{e}_j^T \hat{\boldsymbol{T}} \boldsymbol{g}_{ij}}, \quad \hat{\boldsymbol{T}}_{ij}^{-1} = \hat{\boldsymbol{T}}^{-1} - \Delta \boldsymbol{g}_{ij} \boldsymbol{e}_j^T \\ \boldsymbol{g}_{ij} &= -\partial \left\{ \frac{\boldsymbol{t}_j}{||\boldsymbol{t}_j||} \right\} / \partial t_{ij} = \tfrac{1}{||\boldsymbol{t}_j||^3} (t_{ij} \boldsymbol{t}_j - ||\boldsymbol{t}_j||^2 \boldsymbol{e}_i) \\ \frac{\partial J_p(\boldsymbol{x})}{\partial t_{ij}} &= \sum_{l=1}^{n} \Big\{ \Big(\hat{\boldsymbol{x}}_p^H(l) \hat{\boldsymbol{M}}(\hat{\boldsymbol{T}}) \hat{\boldsymbol{x}}_p(l) \Big) \Big(\hat{\boldsymbol{y}}_p^H(l) \hat{\boldsymbol{T}}^T \hat{\boldsymbol{T}} \hat{\boldsymbol{y}}_p(l) \Big) \\ &\quad + \Big(\hat{\boldsymbol{x}}_p^H(l) \hat{\boldsymbol{T}}^{-1} \hat{\boldsymbol{T}}^{-T} \hat{\boldsymbol{x}}_p(l) \Big) \Big(\hat{\boldsymbol{y}}_p^H(l) \hat{\boldsymbol{N}}(\hat{\boldsymbol{T}}) \hat{\boldsymbol{y}}_p(l) \Big) \Big\} \\ \frac{\partial J_z(\boldsymbol{x})}{\partial t_{ij}} &= \sum_{l=1}^{n} \Big\{ \hat{\boldsymbol{x}}_z^H(l) \hat{\boldsymbol{M}}(\hat{\boldsymbol{T}}) \hat{\boldsymbol{x}}_z(l) \Big(\hat{\boldsymbol{y}}_z^H(l) \hat{\boldsymbol{T}}^T \hat{\boldsymbol{T}} \hat{\boldsymbol{y}}_z(l) + \beta_l^2 \Big) \\ &\quad + \Big(\hat{\boldsymbol{x}}_z^H(l) \hat{\boldsymbol{T}}^{-1} \hat{\boldsymbol{T}}^{-T} \hat{\boldsymbol{x}}_z(l) + \alpha_l^2 \Big) \hat{\boldsymbol{y}}_z^H(l) \hat{\boldsymbol{N}}(\hat{\boldsymbol{T}}) \hat{\boldsymbol{y}}_z(l) \Big\} \end{aligned} \tag{28}$$

where

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

If the iteration is terminated at step k, $\boldsymbol{x}_k$ is claimed to be a solution point.

IV. A Numerical Example

Consider a 4th-order state-space digital filter $(\boldsymbol{A}, \boldsymbol{b}, \boldsymbol{c}, d)_4$ in (1) studied in [4] (p. 151) with

$$\boldsymbol{A}^T = \begin{bmatrix} 3.5906 & 1 & 0 & 0 \\ -4.8535 & 0 & 1 & 0 \\ 2.9260 & 0 & 0 & 1 \\ -0.6636 & 0 & 0 & 0 \end{bmatrix} \qquad \boldsymbol{b} = \begin{bmatrix} 1 \\ 0 \\ 0 \\ 0 \end{bmatrix}$$

$$\begin{aligned} \boldsymbol{c} &= 10^{-3} \left[\; 0.2353 \quad 0.0355 \quad 0.2147 \quad 0.0104 \; \right] \\ d &= 0.1425 \end{aligned}$$

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

$$\begin{aligned} \lambda &= 0.932750 \pm j0.134752, \quad 0.862550 \pm j0.056149 \\ v &= 1.044533 \pm j0.212868, \quad 0.749941 \pm j0.147041 \end{aligned}$$

respectively. By using (9c), we obtained

$$\begin{bmatrix} \alpha_1 \\ \alpha_2 \\ \alpha_3 \\ \alpha_4 \end{bmatrix} = 10^{-3} \begin{bmatrix} 1.755627 \\ 1.755627 \\ 1.419538 \\ 1.419538 \end{bmatrix}, \qquad \begin{bmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \\ \beta_4 \end{bmatrix} = 10^{2} \begin{bmatrix} 2.597616 \\ 2.597616 \\ 2.477293 \\ 2.477293 \end{bmatrix}$$

which give

$$J_p(\boldsymbol{T}) \ge 4, \qquad J_z(\boldsymbol{T}) \ge \sum_{l=1}^{4} (1 + |\alpha_l \beta_l|)^2 = 7.894107.$$

The original pole and zero sensitivity measures in (9a) and (9b) were computed as

$$J_p = 1.723110 \times 10^7, \qquad J_z = 3.352758 \times 10^5.$$

The controllability Gramian was then computed using (17) as

$$\boldsymbol{K}_c = 10^5 \begin{bmatrix} 2.114795 & 2.103883 & 2.071415 & 2.018180 \\ 2.103883 & 2.114795 & 2.103883 & 2.071415 \\ 2.071415 & 2.103883 & 2.114795 & 2.103883 \\ 2.018180 & 2.071415 & 2.103883 & 2.114795 \end{bmatrix}.$$

By choosing $\gamma = 1$ in (16), $\varepsilon = 10^{-8}$ in (26), and starting with an initial point $\boldsymbol{x}_0$ obtained from the assignment $\hat{\boldsymbol{T}}^{-1} = \boldsymbol{I}_4$ in (22), it took the proposed algorithm 46 iterations to converge to an optimal solution

$$\hat{\boldsymbol{T}}^{-1} = \begin{bmatrix} 0.938360 & 0.017908 & -0.755062 & -0.937579 \\ -0.312622 & 0.997747 & 0.510786 & 0.326915 \\ -0.145701 & 0.043182 & 0.321273 & -0.118611 \\ -0.022776 & -0.048121 & 0.256443 & -0.002029 \end{bmatrix}$$

which leads to

$$\boldsymbol{T} = 10^3 \begin{bmatrix} 0.582203 & 0.136146 & 0.780107 & -0.335331 \\ 0.691413 & 0.088922 & 0.900359 & -0.290230 \\ 0.822348 & 0.034922 & 1.024101 & -0.226571 \\ 0.978001 & -0.024284 & 1.150141 & -0.140859 \end{bmatrix}$$

where

$$J_p(\boldsymbol{T}) = 4.000000, \quad J_z(\boldsymbol{T}) = 94.475457.$$

With a new realization as specified in (11), the controllability Gramian was computed from (18) as

$$\overline{\boldsymbol{K}}_c = \begin{bmatrix} 1.000000 & -0.300309 & -0.920853 & -0.964660 \\ -0.300309 & 1.000000 & 0.497647 & 0.304364 \\ -0.920853 & 0.497647 & 1.000000 & 0.836286 \\ -0.964660 & 0.304364 & 0.836286 & 1.000000 \end{bmatrix}$$

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 46 iterations is shown in Fig. 1 where $\gamma = 1$, i.e., $J_\gamma(\boldsymbol{T}) = J_p(\boldsymbol{T})$, from which it is seen that the iterative algorithm converges with 46 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 I. From this table, it is observed that the sum of pole sensitivity and zero sensitivity reaches minimum when the weight value $\gamma = 0.5$ is chosen.

TABLE I. 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	94.475456	98.475456
0.9	8.146143	5.418404	32.695793	38.114197
0.8	10.371509	6.706023	25.033451	31.739475
0.7	11.925343	8.009567	21.062153	29.071720
0.6	13.019405	9.444195	18.382219	27.826415
0.5	13.725046	11.127511	16.322580	27.450091
0.4	14.055390	13.239362	14.599408	27.838770
0.3	13.977582	16.127920	13.056008	29.183928
0.2	13.390428	20.649254	11.575721	32.224976
0.1	12.012883	29.946357	10.020275	39.966632
0.0	8.224923	101.266506	8.224923	109.491429

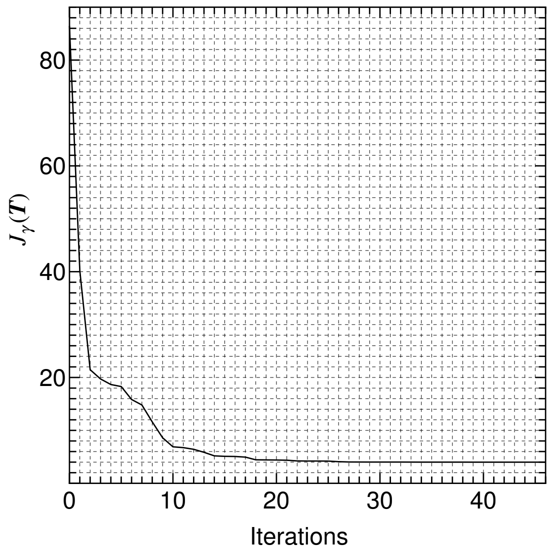

Fig. 1. Profile of $J_\gamma(\boldsymbol{T})$ during the first 46 iterations with $\gamma = 1$

V. CONCLUSION

This paper has investigated the problem of minimizing weighted pole and zero sensitivity subject to l_2-scaling constraints for state-space digital filters. The problem at hand has been converted into an unconstrained optimization problem by using linear-algebraic techniques. The unconstrained optimization problem has then been solved iteratively by applying an efficient quasi-Newton algorithm. The results of a numerical example have demonstrated the validity and effectiveness of the proposed iterative optimization technique.

REFERENCES

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

[2] 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.

[3] 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.

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

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

[6] 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.

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

[8] 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.

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

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