MINIMIZED ROUNDOFF NOISE AND POLE SENSITIVITY SUBJECT TO L_2-SCALING CONSTRAINTS FOR IIR FILTERS

Yoichi Hinamoto

Department of Electrical and Computer Engineering
National Institute of Technology, Kagawa College
Takamatsu, Kagawa 761-8058, Japan
Email: hinamoto@t.kagawa-nct.ac.jp

Akimitsu Doi

Department of Computer Science
Hiroshima Institute of Technology
Hiroshima 731-5193, Japan
Emails: doi@cc.it-hiroshima.ac.jp

ABSTRACT

This paper investigates the minimization problem of weighted roundoff noise and pole sensitivity subject to l_2-scaling constraints for state-space digital filters. A new measure for evaluating roundoff noise and pole sensitivity is proposed, and an efficient technique for minimizing this measure is developed. It is shown that the problem can be converted into an unconstrained optimization problem by using linear-algebraic techniques. The unconstrained optimization problem at hand is then solved iteratively by employing an efficient quasi-Newton algorithm with closed-form formulas for key gradient evaluation. Finally a numerical example is presented to demonstrate the validity and effectiveness of the proposed technique.

1. INTRODUCTION

In fixed-point IIR digital filters, roundoff noise occurs due to finite-precision nature of computer arithmetic and in turn degrades the filters' performance. It is well known that the roundoff noise is critically dependent on the internal structure of an IIR digital filter, and several techniques for synthesizing a state-space filter structure that minimizes the roundoff noise at the filter output subject to l_2-scaling constraints have been explored by appropriately choosing a linear transformation to state-space coordinates [1]-[4]. When implementing a state-space model, the coefficients must be truncated or rounded to fit the finite-word-length (FWL) constraints. This coefficient quantization usually changes the characteristics of the filter. For instance, it may alter a stable filter to unstable one. This motivates the study of minimizing coefficients sensitivity. There exist several ways to define sensitivity of a filter with respect to its realization coefficients. Two of them rely 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. Several techniques for minimizing the l_1/l_2-sensitivity measure [5]-[9] and the l_2-sensitivity measure [10]-[13] have been proposed. Alternatively, the pole and zero sensitivity of a filter with respect to state-space parameters has been analyzed, and its reduction or minimization have been considered [14], [15]. It has been stated in [14] that the sensitivities of poles and zeros of a transfer function with respect to the variations in coefficients of a state-space model (due to FWL) are closely related to the sensitivity characteristics of its frequency transfer function, and that minimal pole sensitivity is achieved when matrix $\boldsymbol{A}$ is normal.

In this paper, the problem of minimizing weighted roundoff noise and pole sensitivity subject to l_2-scaling constraints for state-space digital filters is investigated.

2. PROBLEM FORMULATION

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

$$\boldsymbol{x}(k+1) = \boldsymbol{A}\boldsymbol{x}(k) + \boldsymbol{b}u(k), \quad y(k) = \boldsymbol{c}\boldsymbol{x}(k) \tag{1}$$

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 $\boldsymbol{A}$, $\boldsymbol{b}$ and $\boldsymbol{c}$ are real constant matrices of appropriate dimensions.

A. Roundoff Noise Analysis

By taking the quantizations performed before matrix-vector multiplication into account, an FWL implementation of the filter in (1) can be obtained as [14]

$$\begin{aligned} \tilde{\boldsymbol{x}}(k+1) &= \boldsymbol{A}\boldsymbol{Q}[\tilde{\boldsymbol{x}}(k)] + \boldsymbol{b}u(k) \\ \tilde{y}(k) &= \boldsymbol{c}\boldsymbol{Q}[\tilde{\boldsymbol{x}}(k)]. \end{aligned} \tag{2}$$

The matrices $\boldsymbol{A}$, $\boldsymbol{b}$, and $\boldsymbol{c}$ in (2) are assumed to have exact fractional B_c-bit representations. The FWL state-variable vector $\tilde{\boldsymbol{x}}(k)$ and the output $\tilde{y}(k)$ all have B-bit fractional representations, while the input $u(k)$ is a $(B - B_c)$-bit fraction. The quantizer $\boldsymbol{Q}[\cdot]$ in (2) rounds the B-bit fraction $\tilde{\boldsymbol{x}}(k)$ to $(B - B_c)$-bit after the multiplications and additions, where the sign bit is not counted. Subtracting (2) from (1) yields

$$\begin{aligned} \Delta\boldsymbol{x}(k+1) &= \boldsymbol{A}\Delta\boldsymbol{x}(k) + \boldsymbol{A}\boldsymbol{e}(k) \\ \Delta y(k) &= \boldsymbol{c}\Delta\boldsymbol{x}(k) + \boldsymbol{c}\boldsymbol{e}(k) \end{aligned} \tag{3}$$

where $\boldsymbol{e}(k) = \tilde{\boldsymbol{x}}(k) - \boldsymbol{Q}[\tilde{\boldsymbol{x}}(k)]$, $\Delta\boldsymbol{x}(k) = \boldsymbol{x}(k) - \tilde{\boldsymbol{x}}(k)$ and $\Delta y(k) = y(k) - \tilde{y}(k)$. It is assumed that the roundoff error

978-0-9928626-3-3/15/$31.00 ©2015 IEEE

$\boldsymbol{e}(k)$ can be modeled as a zero-mean noise process with covariance $\sigma^2\boldsymbol{I}_n$. By taking the z-transform on both sides of (3) and setting $\Delta\boldsymbol{x}(0)=\boldsymbol{0}$, we have

$$\begin{aligned}\Delta Y(z) &= \boldsymbol{H}_e(z)\boldsymbol{E}(z)\\ \boldsymbol{H}_e(z) &= \boldsymbol{c}(z\boldsymbol{I}_n-\boldsymbol{A})^{-1}\boldsymbol{A}+\boldsymbol{c}=\sum_{k=0}^{\infty}\boldsymbol{c}\boldsymbol{A}^k z^{-k}\end{aligned} \tag{4}$$

where $\Delta Y(z)$ and $\boldsymbol{E}(z)$ stand for the z-transforms of $\Delta y(k)$ and $\boldsymbol{e}(k)$, respectively. Next, the normalized noise gain $J_R=\sigma_{out}^2/\sigma^2$ is defined as

$$J_R=\text{tr}\left[\frac{1}{2\pi j}\oint_{|z|=1}\boldsymbol{H}_e^*(z)\boldsymbol{H}_e(z)\frac{dz}{z}\right]. \tag{5}$$

Substituting (4) into (5) yields

$$J_R=\text{tr}\left[\boldsymbol{W}\right] \tag{6}$$

where $\boldsymbol{W}$ is the observability Gramian of the digital filter in (1) that can be obtained by solving the Lyapunov equation

$$\boldsymbol{W}=\boldsymbol{A}^T\boldsymbol{W}\boldsymbol{A}+\boldsymbol{c}^T\boldsymbol{c}. \tag{7}$$

Applying a coordinate transformation defined by

$$\overline{\boldsymbol{x}}(k)=\boldsymbol{T}^{-1}\boldsymbol{x}(k) \tag{8}$$

to the digital filter in (1), a different yet equivalent state-space realization of (1), $(\overline{\boldsymbol{A}},\overline{\boldsymbol{b}},\overline{\boldsymbol{c}})_n$, can be characterized by

$$\overline{\boldsymbol{A}}=\boldsymbol{T}^{-1}\boldsymbol{A}\boldsymbol{T},\quad \overline{\boldsymbol{b}}=\boldsymbol{T}^{-1}\boldsymbol{b},\quad \overline{\boldsymbol{c}}=\boldsymbol{c}\boldsymbol{T}. \tag{9}$$

With an equivalent state-space realization as specified in (9), the normalized noise gain in (6) is changed to

$$J_R(\boldsymbol{T})=\text{tr}\left[\,\overline{\boldsymbol{W}}\,\right]=\text{tr}\left[\boldsymbol{T}^T\boldsymbol{W}\boldsymbol{T}\right] \tag{10}$$

where $\overline{\boldsymbol{W}}$ denotes the observability Gramian for an equivalent state-space realization as specified in (9).

B. Pole Sensitivity Analysis

The transfer function of the digital filter in (1) can be expressed as

$$H(z)=\boldsymbol{c}(z\boldsymbol{I}_n-\boldsymbol{A})^{-1}\boldsymbol{b} \tag{11}$$

Suppose that the poles $\{\lambda_l\}$ of $H(z)$ are denoted by $\{\lambda_l\}=\lambda(\boldsymbol{A})$. Moreover, 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, assuming that $\boldsymbol{A}$ has a full set of linearly independent eigenvectors, it follows that [15]

$$\left(\frac{\partial\lambda_l}{\partial\boldsymbol{A}}\right)^T=\boldsymbol{x}_p(l)\boldsymbol{y}_p^H(l)\ \text{ for }\ l=1,2,\cdots,n. \tag{12}$$

Using the Frobenius norm for an $m\times n$ complex matrix $\boldsymbol{M}$ defined by

$$||\boldsymbol{M}||_F=\left(\sum_{i=1}^{m}\sum_{j=1}^{n}|(\boldsymbol{M})_{ij}|^2\right)^{\frac{1}{2}} \tag{13}$$

the pole sensitivity measure for the digital filter in (1) is defined as [15]

$$J_p=\sum_{l=1}^{n}\left\|\frac{\partial\lambda_l}{\partial\boldsymbol{A}}\right\|_F^2=\sum_{l=1}^{n}\left\|\left(\frac{\partial\lambda_l}{\partial\boldsymbol{A}}\right)^T\right\|_F^2. \tag{14}$$

By virtue of

$$||\boldsymbol{M}||_F^2=\text{tr}[\boldsymbol{M}\boldsymbol{M}^H]=\text{tr}[\boldsymbol{M}^H\boldsymbol{M}]=||\boldsymbol{M}^H||_F^2 \tag{15}$$

and using (12), the pole sensitivity measure in (14) is written as

$$J_p=\sum_{l=1}^{n}\Big(\boldsymbol{x}_p^H(l)\boldsymbol{x}_p(l)\Big)\Big(\boldsymbol{y}_p^H(l)\boldsymbol{y}_p(l)\Big). \tag{16}$$

With an equivalent state-space realization as specified in (9), the right eigenvector $\boldsymbol{x}_p(l)$ corresponding to λ_l and the reciprocal left eigenvector $\boldsymbol{y}_p(l)$ corresponding to $\boldsymbol{x}_p(l)$ are changed to

$$\overline{\boldsymbol{x}}_p(l)=\boldsymbol{T}^{-1}\boldsymbol{x}_p(l),\qquad \overline{\boldsymbol{y}}_p(l)=\boldsymbol{T}^T\boldsymbol{y}_p(l) \tag{17}$$

respectively, for $l=1,2,\cdots,n$. Thus, referring to (16), the pole sensitivity measure $J_p(\boldsymbol{T})$ for an equivalent state-space realization can be expressed as

$$J_p(\boldsymbol{T})=\sum_{l=1}^{n}\Big(\boldsymbol{x}_p^H(l)\boldsymbol{T}^{-T}\boldsymbol{T}^{-1}\boldsymbol{x}_p(l)\Big)\Big(\boldsymbol{y}_p^H(l)\boldsymbol{T}\boldsymbol{T}^T\boldsymbol{y}_p(l)\Big). \tag{18}$$

It is known that an inequality $J_p(\boldsymbol{T})\geq n$ are always satisfied for any $n\times n$ nonsingular matrix $\boldsymbol{T}$ [15].

Remark 1 [14], [15]: It is noted that zero sensitivity for the digital filter in (1) can be examined in the case where a direct path from the input to the output in (1) exists.

C. Problem Statement

It should be noted that the l_2-scaling constraints on the state-variable vector $\boldsymbol{x}(k)$ involve the controllability Gramian $\boldsymbol{K}$ of the digital filter in (1), which can be computed by solving the Lyapunov equation

$$\boldsymbol{K}=\boldsymbol{A}\boldsymbol{K}\boldsymbol{A}^T+\boldsymbol{b}\boldsymbol{b}^T. \tag{19}$$

In order to suppress overflow oscillations, l_2-scaling constraints are imposed on the state-variable vector $\overline{\boldsymbol{x}}(k)$ so that

$$(\overline{\boldsymbol{K}})_{ii}=(\boldsymbol{T}^{-1}\boldsymbol{K}\boldsymbol{T}^{-T})_{ii}=1,\quad i=1,2,\cdots,n \tag{20}$$

where $\overline{\boldsymbol{K}}$ indicates the controllability Gramian for an equivalent state-space realization as specified in (9).

In this paper, we consider a weighted roundoff noise and pole sensitivity measure $J_\gamma(\boldsymbol{T})$ defined by

$$J_\gamma(\boldsymbol{T})=(1-\gamma)J_R(\boldsymbol{T})+\gamma J_p(\boldsymbol{T}) \tag{21}$$

where $0\leq\gamma\leq 1$ is a weighting factor. It is evident 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 roundoff noise. In addition, by setting γ to unity or zero $J_\gamma(\boldsymbol{T})$ becomes $J_p(\boldsymbol{T})$ or $J_R(\boldsymbol{T})$, respectively.

The problem being considered here is to design the optimal coordinate transformation matrix $\boldsymbol{T}$ that minimize (21) subject to l_2-scaling constraints in (20) where $0 \le \gamma \le 1$ is a weighting factor specified by the designer.

Remark 2 [10]: It is noted that an l_2-sensitivity measure for the equivalent state-space realization $(\overline{\boldsymbol{A}}, \overline{\boldsymbol{b}}, \overline{\boldsymbol{c}})_n$ in (9) can be evaluated by

$$\begin{aligned} S_2(\overline{\boldsymbol{A}}, \overline{\boldsymbol{b}}, \overline{\boldsymbol{c}}) &= \left\| \frac{\partial H(z)}{\partial \overline{\boldsymbol{A}}} \right\|_2^2 + \left\| \frac{\partial H(z)}{\partial \overline{\boldsymbol{b}}} \right\|_2^2 + \left\| \frac{\partial H(z)}{\partial \overline{\boldsymbol{c}}^T} \right\|_2^2 \\ &= \mathrm{tr}\Big\{ [\boldsymbol{T}^{-1} \;\; \boldsymbol{0}]\, \boldsymbol{R} \begin{bmatrix} \boldsymbol{T}^{-T} \\ \boldsymbol{0} \end{bmatrix} \Big\} + \mathrm{tr}[\overline{\boldsymbol{W}}] + \mathrm{tr}[\overline{\boldsymbol{K}}] \end{aligned} \tag{22}$$

where matrix $\boldsymbol{R}$ is obtained by solving the Lyapunov equation

$$\boldsymbol{R} = \begin{bmatrix} \boldsymbol{A} & \boldsymbol{bc} \\ \boldsymbol{0} & \boldsymbol{A} \end{bmatrix} \boldsymbol{R} \begin{bmatrix} \boldsymbol{A} & \boldsymbol{bc} \\ \boldsymbol{0} & \boldsymbol{A} \end{bmatrix}^T + \begin{bmatrix} \boldsymbol{0} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{T}\boldsymbol{T}^T \end{bmatrix}.$$

3. WEIGHTED ROUNDOFF NOISE AND POLE SENSITIVITY MINIMIZATION

In this section, we define

$$\hat{\boldsymbol{T}} = \boldsymbol{T}^T \boldsymbol{K}^{-\frac{1}{2}} \tag{23}$$

which leads (20) to

$$(\hat{\boldsymbol{T}}^{-T} \hat{\boldsymbol{T}}^{-1})_{ii} = 1 \;\text{ for }\; i = 1, 2, \cdots, n. \tag{24}$$

It is obvious that the conditions in (24) are always satisfied by choosing matrix $\hat{\boldsymbol{T}}^{-1}$ as

$$\hat{\boldsymbol{T}}^{-1} = \left[\frac{\boldsymbol{t}_1}{||\boldsymbol{t}_1||}, \frac{\boldsymbol{t}_2}{||\boldsymbol{t}_2||}, \cdots, \frac{\boldsymbol{t}_n}{||\boldsymbol{t}_n||} \right]. \tag{25}$$

By defining an $n^2 \times 1$ vector $\boldsymbol{x} = (\boldsymbol{t}_1^T, \boldsymbol{t}_2^T, \cdots, \boldsymbol{t}_n^T)^T$, the normalized noise gain $J_R(\boldsymbol{T})$ in (10) can be expressed as

$$J_R(\boldsymbol{x}) = \mathrm{tr}\left[\hat{\boldsymbol{T}} \hat{\boldsymbol{W}} \hat{\boldsymbol{T}}^T \right] \tag{26}$$

where

$$\hat{\boldsymbol{W}} = \boldsymbol{K}^{\frac{1}{2}} \boldsymbol{W} \boldsymbol{K}^{\frac{1}{2}}.$$

Moreover, the pole sensitivity measure $J_p(\boldsymbol{T})$ in (18) can be written as

$$J_p(\boldsymbol{x}) = \sum_{l=1}^{n} \Big(\hat{\boldsymbol{x}}_p^H(l) \hat{\boldsymbol{T}}^{-1} \hat{\boldsymbol{T}}^{-T} \hat{\boldsymbol{x}}_p(l) \Big) \Big(\hat{\boldsymbol{y}}_p^H(l) \hat{\boldsymbol{T}}^T \hat{\boldsymbol{T}} \hat{\boldsymbol{y}}_p(l) \Big) \tag{27}$$

where

$$\hat{\boldsymbol{x}}_p(l) = \boldsymbol{K}^{-\frac{1}{2}} \boldsymbol{x}_p(l), \qquad \hat{\boldsymbol{y}}_p(l) = \boldsymbol{K}^{\frac{1}{2}} \boldsymbol{y}_p(l).$$

In this way, the original constrained optimization problem can be converted into an unconstrained optimization problem of obtaining an $n^2 \times 1$ vector $\boldsymbol{x}$ which minimizes

$$J_\gamma(\boldsymbol{x}) = (1-\gamma) J_R(\boldsymbol{x}) + \gamma J_p(\boldsymbol{x}). \tag{28}$$

Applying a quasi-Newton algorithm to minimize $J_\gamma(\boldsymbol{x})$ in (28), in the kth iteration the most recent point $\boldsymbol{x}_k$ is updated to point $\boldsymbol{x}_{k+1}$ as [16]

$$\boldsymbol{x}_{k+1} = \boldsymbol{x}_k + \alpha_k \boldsymbol{d}_k \tag{29}$$

where

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

In the above, $\nabla J_\gamma(\boldsymbol{x})$ is 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 $J_\gamma(\boldsymbol{x})$. The above algorithm starts with a trivial initial point $\boldsymbol{x}_0$ obtained from an initial assignment $\hat{\boldsymbol{T}}^{-1} = \boldsymbol{I}_n$, and this iteration process continues until

$$|J_\gamma(\boldsymbol{x}_{k+1}) - J_\gamma(\boldsymbol{x}_k)| < \varepsilon \tag{30}$$

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

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

$$\begin{aligned} \frac{\partial J_\gamma(\boldsymbol{x})}{\partial t_{ij}} &= \lim_{\Delta \to 0} \frac{J_\gamma(\hat{\boldsymbol{T}}_{ij}) - J_\gamma(\hat{\boldsymbol{T}})}{\Delta} \\ &= (1-\gamma) \frac{\partial J_R(\boldsymbol{x})}{\partial t_{ij}} + \gamma \frac{\partial J_p(\boldsymbol{x})}{\partial t_{ij}} \end{aligned} \tag{31}$$

where $\hat{\boldsymbol{T}}_{ij}$ is the matrix obtained from $\hat{\boldsymbol{T}}$ with a perturbed (i,j)th component, it follows that [17, p. 655]

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

and

$$\begin{aligned} \frac{\partial J_R(\boldsymbol{x})}{\partial t_{ij}} &= 2\, \boldsymbol{e}_j^T \hat{\boldsymbol{T}} \hat{\boldsymbol{W}} \hat{\boldsymbol{T}}^T \hat{\boldsymbol{T}} \boldsymbol{g}_{ij} \\ \frac{\partial J_p(\boldsymbol{x})}{\partial t_{ij}} &= \sum_{l=1}^{n} \Big\{ \Big(\hat{\boldsymbol{x}}_p^H(l) \hat{\boldsymbol{M}}(\hat{\boldsymbol{T}}) \hat{\boldsymbol{x}}_p(l) \Big) \Big(\hat{\boldsymbol{y}}_p^H(l) \hat{\boldsymbol{T}}^T \hat{\boldsymbol{T}} \hat{\boldsymbol{y}}_p(l) \Big) \\ &\quad + \Big(\hat{\boldsymbol{x}}_p^H(l) \hat{\boldsymbol{T}}^{-1} \hat{\boldsymbol{T}}^{-T} \hat{\boldsymbol{x}}_p(l) \Big) \Big(\hat{\boldsymbol{y}}_p^H(l) \hat{\boldsymbol{N}}(\hat{\boldsymbol{T}}) \hat{\boldsymbol{y}}_p(l) \Big) \Big\} \end{aligned}$$

with

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

4. AN ILLUSTRATIVE EXAMPLE

Consider the 4th-order Butterworth lowpass filter $(\boldsymbol{A}, \boldsymbol{b}, \boldsymbol{c})_4$ with a narrow normalized bassband of 0.05, described by

$$\boldsymbol{A} = \begin{bmatrix} 3.589734 & 1 & 0 & 0 \\ -4.851276 & 0 & 1 & 0 \\ 2.924053 & 0 & 0 & 1 \\ -0.663010 & 0 & 0 & 0 \end{bmatrix}, \quad \boldsymbol{b} = 10^{-3} \begin{bmatrix} 0.237096 \\ 0.035885 \\ 0.216300 \\ 0.010527 \end{bmatrix}$$
$$\boldsymbol{c} = \begin{bmatrix} 1 & 0 & 0 & 0 \end{bmatrix}$$

whose numerator b and denominator a was found using MATLAB as $[b, a] = \text{butter}(4, 0.05)$.

When a coordinate transformation defined by

$$\boldsymbol{T}_o = \text{diag}\{\, 0.226458,\ 0.588059,\ 0.513017,\ 0.150144 \,\}$$

was applied to the above filter $(\boldsymbol{A}, \boldsymbol{b}, \boldsymbol{c})_4$, the resulting controllability and observability Gramians $\boldsymbol{K}_o = \boldsymbol{T}_o^{-1} \boldsymbol{K} \boldsymbol{T}_o^{-T}$ and $\boldsymbol{W}_o = \boldsymbol{T}_o^T \boldsymbol{W} \boldsymbol{T}_o$ were computed using (19) and (7) as

$$\boldsymbol{K}_o = \begin{bmatrix} 1.000000 & -0.999248 & 0.997433 & -0.994918 \\ -0.999248 & 1.000000 & -0.999452 & 0.998047 \\ 0.997433 & -0.999452 & 1.000000 & -0.999566 \\ -0.994918 & 0.998047 & -0.999566 & 1.000000 \end{bmatrix}$$

$$\boldsymbol{W}_o = 10^4 \begin{bmatrix} 1.063597 & 2.747677 & 2.360090 & 0.672994 \\ 2.747677 & 7.172055 & 6.224575 & 1.793652 \\ 2.360090 & 6.224575 & 5.458399 & 1.589267 \\ 0.672994 & 1.793652 & 1.589267 & 0.467539 \end{bmatrix}.$$

Then, the normalized noise gain in (10) subject to l_2-scaling constraints: $(\boldsymbol{K}_o)_{ii} = 1$ for $i = 1, 2, \cdots, n$ was found to be

$$J_R(\boldsymbol{T}_o) = \text{tr}[\boldsymbol{W}_o] = 1.416159 \times 10^5.$$

Moreover, the l_2-sensitivity measure in (22) was computed as

$$S_2(\boldsymbol{T}_o^{-1} \boldsymbol{A} \boldsymbol{T}_o, \boldsymbol{T}_o^{-1} \boldsymbol{b}, \boldsymbol{c} \boldsymbol{T}_o) = 9.779175 \times 10^6.$$

Next, the eigenvalues of matrix $\boldsymbol{A}$ were found to be

$$\lambda = 0.931900 \pm j0.136363, \quad 0.862967 \pm j0.052305.$$

The original pole sensitivity measures in (16) and (18) were computed as

$$J_p = 1.863101 \times 10^7 \ \text{ and } \ J_p(\boldsymbol{T}_o) = 1.774671 \times 10^7$$

respectively.

The quasi-Newton algorithm was applied to minimize (28) with $\gamma = 0.7$ by choosing $\hat{\boldsymbol{T}}^{-1} = \boldsymbol{I}_4$ as an initial assignment, and setting tolerance to $\varepsilon = 10^{-8}$ in (30). It took the algorithm 67 iterations to converge to the solution

$$\hat{\boldsymbol{T}} = \begin{bmatrix} -0.005208 & 1.163747 & -0.111659 & -0.787204 \\ -0.037351 & -0.404560 & 1.799691 & 2.040456 \\ -0.164510 & -0.518875 & -1.552980 & 4.982677 \\ 1.193078 & 0.558202 & -0.064418 & -6.380411 \end{bmatrix}$$

which is equivalent to

$$\boldsymbol{T} = \begin{bmatrix} -0.178135 & 0.230044 & -0.317367 & 0.217107 \\ 0.449984 & -0.602281 & 0.885719 & -0.599864 \\ -0.380296 & 0.526722 & -0.824583 & 0.559049 \\ 0.107642 & -0.153385 & 0.256254 & -0.175300 \end{bmatrix}$$

and the objective function in (21) was found to be

$$J_\gamma(\boldsymbol{T}) = 3.246633$$

with $J_R(\boldsymbol{T}) = 1.327903$ and $J_p(\boldsymbol{T}) = 4.068946$.

The profile of $J_\gamma(\boldsymbol{x})$ during the first 67 iterations of the algorithm is depicted in Fig. 1. From this figure it is observed that the weighted roundoff noise and pole sensitivity for a state-space digital filter were drastically reduced subject to l_2-scaling constraints. In this case, the coefficient matrices

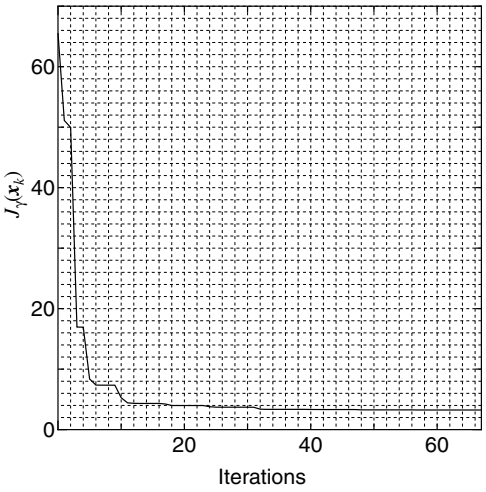

Fig. 1. Profile of $J_\gamma(\boldsymbol{x})$ during the first 67 iterations.

$\overline{\boldsymbol{A}}$, $\overline{\boldsymbol{b}}$ and $\overline{\boldsymbol{c}}$ for the equivalent state-space realization specified by (9) were obtained as

$$\overline{\boldsymbol{A}} = \begin{bmatrix} 0.925785 & 0.005903 & -0.123828 & -0.066614 \\ -0.042551 & 0.870369 & 0.037577 & -0.044076 \\ 0.097415 & -0.053644 & 0.909232 & 0.044846 \\ 0.074370 & 0.033706 & 0.019872 & 0.884347 \end{bmatrix}$$

$$\overline{\boldsymbol{b}} = \begin{bmatrix} 0.472882 & 0.535104 & 0.318385 & 0.287515 \end{bmatrix}^T$$

$$\overline{\boldsymbol{c}} = \begin{bmatrix} -0.178135 & 0.230044 & -0.317367 & 0.217107 \end{bmatrix}$$

and the controllability and observability Gramians $\overline{\boldsymbol{K}}$ and $\overline{\boldsymbol{W}}$ were computed from (20) and (10) as

$$\overline{\boldsymbol{K}} = \begin{bmatrix} 1.000000 & 0.430444 & 0.368731 & 0.281793 \\ 0.430444 & 1.000000 & 0.825645 & 0.884131 \\ 0.368731 & 0.825645 & 1.000000 & 0.969484 \\ 0.281793 & 0.884131 & 0.969484 & 1.000000 \end{bmatrix}$$

$$\overline{\boldsymbol{W}} = \begin{bmatrix} 0.402694 & -0.345058 & 0.155350 & -0.241403 \\ -0.345058 & 0.339744 & -0.177349 & 0.255916 \\ 0.155350 & -0.177349 & 0.357751 & -0.087177 \\ -0.241403 & 0.255916 & -0.087177 & 0.227714 \end{bmatrix}$$

respectively. For the above matrix $\overline{\boldsymbol{A}}$, it follows from (13) that

$$||\overline{\boldsymbol{A}}\,\overline{\boldsymbol{A}}^T - \overline{\boldsymbol{A}}^T\overline{\boldsymbol{A}}||_F = 7.731601 \times 10^{-3}$$

which implies that matrix $\overline{\boldsymbol{A}}$ is practically normal. Moreover, the controllability Gramian $\overline{\boldsymbol{K}}$ satisfies l_2-scaling constraints. For reference, the l_2-sensitivity measure in (22) was computed as

$$S_2(\overline{\boldsymbol{A}}, \overline{\boldsymbol{b}}, \overline{\boldsymbol{c}}) = 45.179954.$$

This reveals that minimization of weighted roundoff noise and pole sensitivity subject to l_2-scaling constraints also reduce the l_2- sensitivity for a state-space digital filter considerably.

The whole numerical results of the roundoff noise measure $J_R(\boldsymbol{T})$ in (10), the pole sensitivity measure $J_p(\boldsymbol{T})$ in (18), $J_R(\boldsymbol{T}) + J_p(\boldsymbol{T})$, and the objective function $J_\gamma(\boldsymbol{T})$ in (21) are summarized in Table 1.

Table 1. Performance comparison

γ	$J_\gamma(\boldsymbol{T})$	$J_R(\boldsymbol{T})$	$J_p(\boldsymbol{T})$	$J_R(\boldsymbol{T})+J_p(\boldsymbol{T})$
1.0	4.000000	6.678752	4.000000	10.678752
0.9	3.765801	1.565948	4.010229	5.576178
0.8	3.513441	1.435279	4.032982	5.468261
0.7	3.246633	1.327903	4.068946	5.396849
0.6	2.965042	1.227214	4.123594	5.350808
0.5	2.666454	1.126471	4.206436	5.332908
0.4	2.347839	1.027726	4.328008	5.355734
0.3	2.004220	0.928102	4.515163	5.443265
0.2	1.625958	0.823537	4.835642	5.659179
0.1	1.189538	0.706245	5.539181	6.245426
0.0	0.555541	0.555541	13.604578	14.160119

5. CONCLUSION

The problem of minimizing a weighted roundoff noise and pole sensitivity measure subject to l_2-scaling constraints for state-space filters has been investigated. An efficient iterative technique for minimizing a weighted roundoff noise and pole sensitivity measure subject to l_2-scaling constraints has been developed by employing a quasi-Newton algorithm. Computer simulation results have demonstrated the validity and effectiveness of the proposed technique.

REFERENCES

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

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

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

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

[5] 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.

[6] 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, pp.884-888, Oct. 1984.

[7] L. Thiele, "On the sensitivity of linear state-space systems," *IEEE Trans. Circuits Syst*., vol.CAS-33, pp.502-510, May 1986.

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

[9] 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, pp.365-377, May 1992.

[10] W.-Y. Yan and J. B. Moore, "On L^2-sensitivity minimization of linear state-space systems," *IEEE Trans. Circuits Syst. I*, vol.39, pp.641-648, Aug. 1992.

[11] G. Li and M. Gevers, "Optimal synthetic FWL design of state-space digital filters," in *Proc. ICASSP 1992*, San Francisco, CA, August 24-28, 1992, vol.4, pp.429-432.

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

[13] 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. II*, vol.52, pp.641-645, Oct. 2005.

[14] 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, pp.1210-1220, Oct. 1986.

[15] G. Li, "On pole and zero sensitivity of linear systems," *IEEE Trans. Circuits Syst. I*, vol. 44, pp.583-590, Jul. 1997.

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

[17] T. Kailath, *Linear Systems*, Prentice Hall, 1980.