Minimization of L_2-Sensitivity for State-Space Digital Filters Subject to L_2-Dynamic-Range-Scaling Constraints

Takao Hinamoto, Hiroaki Ohnishi and Wu-Sheng Lu

Abstract The problem of minimizing an L_2-sensitivity measure subject to L_2-norm dynamic-range scaling constraints for state-space digital filters is formulated. It is shown that the problem can be converted into an unconstrained optimization problem by using linear-algebraic techniques. Next, the unconstrained optimization problem is solved by applying an efficient quasi-Newton algorithm with closed-form formula for gradient evaluation. The coordinate transformation matrix obtained is then used to construct the optimal state-space filter structure that minimizes the L_2-sensitivity measure subject to the scaling constraints. A numerical example is presented to illustrate the utility of the proposed technique.

Keywords L_2-sensitivity, dynamic-range scaling constraints, scaling constrained sensitivity minimization, optimal realization, state-space digital filters

T. Hinamoto and H. Ohnishi are with the Graduate School of Engineering, Hiroshima University, Higashi-Hiroshima, Japan 739-8527.

W.-S. Lu is with the Department of Electrical and Computer Engineering, University of Victoria, B.C., Canada, V8W 3P6.

I. INTRODUCTION

The problem of realizing a fixed-point state-space digital filter with finite word length (FWL) is a significant research topic, since the efficiency and performance of the filter are directly affected by the choice of its state-space filter structure. When a transfer function with infinite accuracy coefficients is designed so as to meet the filter specification requirements and realized by a state-space model, in order to implement the filter in a finite binary representation, the coefficients in the state-space model must be truncated or rounded to fit the FWL constraints. This coefficient quantization usually alters the characteristics of the filter. For instance, it may change a stable filter to an unstable one. This motivates the study of the coefficient sensitivity minimization problem. In the literature, two main classes of techniques have been proposed for constructing state-space digital filters that minimize the coefficient sensitivity in [1]-[10]: L_1/L_2-sensitivity minimization [1]-[5] and L_2-sensitivity minimization [6]-[10]. It has been argued [6]-[10] that the sensitivity measure based on the L_2 norm is more natural and reasonable relative to that based on the L_1/L_2-sensitivity minimization. The L_1/L_2-sensitivity minimization and L_2-sensitivity minimization have also been considered in linear continuous-time systems in [11] and [10], respectively. However, to our best knowledge, there is no study for the minimization of the L_2-sensitivity subject to the L_2-norm dynamic-range scaling constraints for state-space digital filters, although it has been known that the use of scaling constraints can be beneficial for suppressing overflow oscillation [12],[13].

This paper investigates the problem of minimizing the L_2-sensitivity measure subject to L_2-norm dynamic-range scaling constraints for state-space digital filters. To this end, we introduce an expression for evaluating the L_2-sensitivity and formulate the L_2-sensitivity minimization problem subject to the scaling constraints. Next, the constrained optimization problem is converted into an unconstrained optimization problem by using linear-algebraic techniques. The unconstrained optimization problem is then solved using an efficient quasi-Newton algorithm [14]. A numerical example is presented to demonstrate that the proposed algorithms offer reduced L_2-sensitivity compared with that obtained using the conventional methods.

Throughout $\boldsymbol{I}_n$ denotes the identity matrix of dimension $n \times n$. The transpose (conjugate transpose) of a matrix $\boldsymbol{A}$ and trace of a square matrix $\boldsymbol{A}$ are denoted by $\boldsymbol{A}^T$ ($\boldsymbol{A}^*$) and $\mathrm{tr}[\boldsymbol{A}]$, respectively. The direct sum of matrices and ith diagonal element of a square matrix $\boldsymbol{A}$ are denoted by $\oplus$ and $(\boldsymbol{A})_{ii}$, respectively.

II. L_2-SENSITIVITY ANALYSIS

Consider a stable, controllable and observable state-space digital filter

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

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 (1) is given by

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

Definition 1: Let $\boldsymbol{X}$ be an $m \times n$ real matrix and let $f(\boldsymbol{X})$ be a scalar complex function of $\boldsymbol{X}$, differentiable with respect to all the entries of $\boldsymbol{X}$. The sensitivity function of f with respect to $\boldsymbol{X}$ is then defined as

$$\boldsymbol{S}_{\boldsymbol{X}} = \frac{\partial f}{\partial \boldsymbol{X}}, \qquad (\boldsymbol{S}_{\boldsymbol{X}})_{ij} = \frac{\partial f}{\partial x_{ij}} \tag{3}$$

where x_{ij} denotes the (i,j)th entry of matrix $\boldsymbol{X}$.

Definition 2: Let $\boldsymbol{X}(z)$ be an $m \times n$ complex matrix-valued function of a complex variable z and let $x_{pq}(z)$ be the (p,q)th entry of $\boldsymbol{X}(z)$. The L_2-norm of $\boldsymbol{X}(z)$ is then defined as

$$\begin{aligned} \|\boldsymbol{X}(z)\|_2 &= \left[\frac{1}{2\pi}\int_0^{2\pi}\sum_{p=1}^{m}\sum_{q=1}^{n}\left|x_{pq}(e^{j\omega})\right|^2 d\omega\right]^{\frac{1}{2}} \\ &= \left(\mathrm{tr}\left[\frac{1}{2\pi j}\oint_{|z|=1}\boldsymbol{X}(z)\boldsymbol{X}^*(z)\frac{dz}{z}\right]\right)^{\frac{1}{2}}. \end{aligned} \tag{4}$$

From (2), *Definition 1* and *Definition 2*, the overall L_2-sensitivity measure for the state-space digital filter in (1) is defined as

$$\begin{aligned} S &= \left\|\frac{\partial H(z)}{\partial \boldsymbol{A}}\right\|_2^2 + \left\|\frac{\partial H(z)}{\partial \boldsymbol{b}}\right\|_2^2 + \left\|\frac{\partial H(z)}{\partial \boldsymbol{c}^T}\right\|_2^2 \\ &= \left\|[\boldsymbol{F}(z)\boldsymbol{G}(z)]^T\right\|_2^2 + \left\|\boldsymbol{G}^T(z)\right\|_2^2 + \|\boldsymbol{F}(z)\|_2^2 \end{aligned} \tag{5}$$

where

$$\boldsymbol{F}(z) = (z\boldsymbol{I}_n - \boldsymbol{A})^{-1}\boldsymbol{b}, \qquad \boldsymbol{G}(z) = \boldsymbol{c}(z\boldsymbol{I}_n - \boldsymbol{A})^{-1}.$$

The term d in (2) and the sensitivity with respect to it are coordinate-independent and therefore they are neglected here.

It is easy to show that the L_2-sensitivity measure in (5) can be expressed as

$$S = \mathrm{tr}[\boldsymbol{M}_A] + \mathrm{tr}[\boldsymbol{W}_o] + \mathrm{tr}[\boldsymbol{K}_c] \tag{6}$$

where

$$\begin{aligned}
\boldsymbol{M}_A &= \frac{1}{2\pi j}\oint_{|z|=1} [\boldsymbol{F}(z)\boldsymbol{G}(z)]^T \boldsymbol{F}(z^{-1})\boldsymbol{G}(z^{-1})\frac{dz}{z} \\
\boldsymbol{K}_c &= \frac{1}{2\pi j}\oint_{|z|=1} \boldsymbol{F}(z)\boldsymbol{F}^T(z^{-1})\frac{dz}{z} \\
\boldsymbol{W}_o &= \frac{1}{2\pi j}\oint_{|z|=1} \boldsymbol{G}^T(z)\boldsymbol{G}(z^{-1})\frac{dz}{z}.
\end{aligned}$$

The matrices $\boldsymbol{K}_c$ and $\boldsymbol{W}_o$ are called the controllability and observability Gramians, respectively. The Gramians $\boldsymbol{M}_A$, $\boldsymbol{K}_c$ and $\boldsymbol{W}_o$ can be obtained by solving the Lyapunov equations [15]:

$$\begin{aligned}
\begin{bmatrix} * & * \\ * & \boldsymbol{M}_A \end{bmatrix} &= \begin{bmatrix} \boldsymbol{A} & \boldsymbol{bc} \\ \boldsymbol{0} & \boldsymbol{A} \end{bmatrix}^T \begin{bmatrix} * & * \\ * & \boldsymbol{M}_A \end{bmatrix} \\
&\quad \cdot \begin{bmatrix} \boldsymbol{A} & \boldsymbol{bc} \\ \boldsymbol{0} & \boldsymbol{A} \end{bmatrix} + \begin{bmatrix} \boldsymbol{I}_n & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{0} \end{bmatrix} \\
\boldsymbol{K}_c &= \boldsymbol{A}\boldsymbol{K}_c\boldsymbol{A}^T + \boldsymbol{b}\boldsymbol{b}^T \\
\boldsymbol{W}_o &= \boldsymbol{A}^T\boldsymbol{W}_o\boldsymbol{A} + \boldsymbol{c}^T\boldsymbol{c}
\end{aligned} \tag{7}$$

Utilizing the *Cauchy integral theorem*, matrix $\boldsymbol{M}_A$ in (6) can be written as

$$\boldsymbol{M}_A = \sum_{k=0}^{\infty} \boldsymbol{H}^T(k)\boldsymbol{H}(k) \tag{8}$$

where

$$\boldsymbol{H}(k) = \sum_{p=0}^{k} \boldsymbol{A}^p\boldsymbol{bc}\boldsymbol{A}^{k-p}.$$

If a coordinate transformation defined by

$$\overline{\boldsymbol{x}}(k) = \boldsymbol{T}^{-1}\boldsymbol{x}(k) \tag{9}$$

is applied to the state-space model (1), then the new realization $(\overline{\boldsymbol{A}}, \overline{\boldsymbol{b}}, \overline{\boldsymbol{c}}, d)_n$ can be characterized by

$$\overline{\boldsymbol{A}} = \boldsymbol{T}^{-1}\boldsymbol{A}\boldsymbol{T}, \quad \overline{\boldsymbol{b}} = \boldsymbol{T}^{-1}\boldsymbol{b}, \quad \overline{\boldsymbol{c}} = \boldsymbol{c}\boldsymbol{T}. \tag{10}$$

From (2) and (10), it is clear that the transfer function $H(z)$ is invariant under the coordinate transformation in (9). The coordinate transformation defined by (9) transforms the Gramians $\{\boldsymbol{M}_A, \boldsymbol{K}_c, \boldsymbol{W}_o\}$ to $\{\overline{\boldsymbol{M}}_A, \overline{\boldsymbol{K}}_c, \overline{\boldsymbol{W}}_o\}$, and changes (6) to

$$S(\boldsymbol{T}) = \mathrm{tr}[\overline{\boldsymbol{M}}_A] + \mathrm{tr}[\overline{\boldsymbol{W}}_o] + \mathrm{tr}[\overline{\boldsymbol{K}}_c] \tag{11}$$

where

$$\begin{aligned}
\overline{\boldsymbol{M}}_A &= \sum_{k=0}^{\infty} \boldsymbol{T}^T \boldsymbol{H}^T(k) \boldsymbol{T}^{-T} \boldsymbol{T}^{-1} \boldsymbol{H}(k) \boldsymbol{T} \\
&= \boldsymbol{T}^T \hat{\boldsymbol{M}}_A \boldsymbol{T} \\
\begin{bmatrix} * & * \\ * & \hat{\boldsymbol{M}}_A \end{bmatrix} &= \begin{bmatrix} \boldsymbol{A} & \boldsymbol{b}\boldsymbol{c} \\ \boldsymbol{0} & \boldsymbol{A} \end{bmatrix}^T \begin{bmatrix} * & * \\ * & \hat{\boldsymbol{M}}_A \end{bmatrix} \\
&\quad \cdot \begin{bmatrix} \boldsymbol{A} & \boldsymbol{b}\boldsymbol{c} \\ \boldsymbol{0} & \boldsymbol{A} \end{bmatrix} + \begin{bmatrix} \boldsymbol{T}^{-T}\boldsymbol{T}^{-1} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{0} \end{bmatrix} \\
\overline{\boldsymbol{W}}_o &= \boldsymbol{T}^T \boldsymbol{W}_o \boldsymbol{T}, \quad \overline{\boldsymbol{K}}_c = \boldsymbol{T}^{-1} \boldsymbol{K}_c \boldsymbol{T}^{-T}.
\end{aligned}$$

Moreover, if the L_2-norm dynamic-range scaling constraints are imposed on the new state-variable vector $\overline{\boldsymbol{x}}(k)$, then it is required that for $i = 1, 2, \cdots, n$

$$(\overline{\boldsymbol{K}}_c)_{ii} = (\boldsymbol{T}^{-1} \boldsymbol{K}_c \boldsymbol{T}^{-T})_{ii} = 1. \tag{12}$$

The problem of L_2-sensitivity minimization subject to L_2-norm dynamic-range scaling constraints is now formulated as follows: For given $\boldsymbol{A}$, $\boldsymbol{b}$ and $\boldsymbol{c}$ (and therefore, $\boldsymbol{M}_A$, $\boldsymbol{K}_c$ and $\boldsymbol{W}_o$), obtain an $n \times n$ nonsingular matrix $\boldsymbol{T}$ which minimizes (11) subject to the constraints in (12).

Remark 1: By definition, the sensitivity measure described in this paper has no upper bound but does have a lower bound (say zero) because the sensitivity measure is always nonnegative. This observation, in conjunction with the fact that both the objective function (defined by (11)) and constraint functions (defined by (12)) are continuously differentiable, leads us to conclude that the problem of minimizing $S(\boldsymbol{T})$ subject to (12) admits local solutions. On the other hand, since $S(\boldsymbol{T})$ and

the constraint functions are nonconvex, the best that an optimization algorithm can claim is that it identifies a local minimizer of the problem at hand.

Remark 2: It is well known that constrained minimization problems, especially the minimization of nonconvex function subject to nonconvex constraints, are in general considerably more involved than their unconstrained counterpart, but the problem is that it is not always possible to convert a nonconvex constrained problem to an unconstrained one. What is done in the next section is to show that in the present case we do have an equivalent unconstrained problem to work with.

III. L_2-SENSITIVITY MINIMIZATION UNDER SCALING CONSTRAINTS

When the state-space model (1) is assumed to be stable and controllable, the controllability Gramian $\boldsymbol{K}_c$ is symmetric and positive-definite [15]. This implies that $\boldsymbol{K}_c^{1/2}$ satisfying $\boldsymbol{K}_c = \boldsymbol{K}_c^{1/2}\boldsymbol{K}_c^{1/2}$ is also symmetric and positive-definite. Defining

$$\hat{\boldsymbol{T}} = \boldsymbol{T}^T \boldsymbol{K}_c^{-\frac{1}{2}}, \tag{13}$$

the constraints in (12) can be expressed as

$$(\hat{\boldsymbol{T}}^{-T}\hat{\boldsymbol{T}}^{-1})_{ii} = 1, \qquad i = 1, 2, \cdots, n. \tag{14}$$

The constraints in (14) simply state that each column in $\hat{\boldsymbol{T}}^{-1}$ must be a unity vector. If matrix $\hat{\boldsymbol{T}}^{-1}$ is assumed to have the form

$$\hat{\boldsymbol{T}}^{-1} = \left[\frac{\boldsymbol{t}_1}{||\boldsymbol{t}_1||}, \frac{\boldsymbol{t}_2}{||\boldsymbol{t}_2||}, \cdots, \frac{\boldsymbol{t}_n}{||\boldsymbol{t}_n||}\right], \tag{15}$$

then (14) is always satisfied. From (13), it follows that (11) is changed to

$$\begin{aligned} J_o(\hat{\boldsymbol{T}}) = \mathrm{tr}\left[\sum_{k=0}^{\infty} \hat{\boldsymbol{T}}\hat{\boldsymbol{H}}^T(k)\hat{\boldsymbol{T}}^{-1}\hat{\boldsymbol{T}}^{-T}\hat{\boldsymbol{H}}(k)\hat{\boldsymbol{T}}^T\right] \\ +\mathrm{tr}[\hat{\boldsymbol{T}}\hat{\boldsymbol{W}}_o\hat{\boldsymbol{T}}^T] + \mathrm{tr}[\hat{\boldsymbol{T}}^{-T}\hat{\boldsymbol{T}}^{-1}] \end{aligned} \tag{16}$$

where

$$\hat{\boldsymbol{H}}(k) = \boldsymbol{K}_c^{-\frac{1}{2}}\boldsymbol{H}(k)\boldsymbol{K}_c^{\frac{1}{2}}, \qquad \hat{\boldsymbol{W}}_o = \boldsymbol{K}_c^{\frac{1}{2}}\boldsymbol{W}_o\boldsymbol{K}_c^{\frac{1}{2}}.$$

From the foregoing arguments, the problem of obtaining an $n \times n$ nonsingular matrix $\boldsymbol{T}$ which minimizes (11) subject to the constraints in (12) can be converted into an unconstrained optimization problem of obtaining an $n \times n$ nonsingular matrix $\hat{\boldsymbol{T}}$ which minimizes (16).

Remark 3: The use of a structured $\boldsymbol{T}$ matrix as specified in (15) is merely to eliminate the constraints in (12). Needless to say, this increases the degree of nonlinearity for the objective function in (16). However, due to the equivalence of the two optimization problem, it can be concluded that the unconstrained minimization of the objective function in (16) admits at least local solutions.

Now we apply a quasi-Newton algorithm to minimize (16) with respect to matrix $\hat{\boldsymbol{T}}$ given by (15). Let $\boldsymbol{x}$ be the column vector that collects the variables in matrix $\hat{\boldsymbol{T}}$. Then $J_o(\hat{\boldsymbol{T}})$ is a function of $\boldsymbol{x}$, which we denote by $J(\boldsymbol{x})$. The algorithm starts with a trivial initial point $\boldsymbol{x}_0$ obtained from an initial assignment $\hat{\boldsymbol{T}} = \boldsymbol{I}_n$. Then, in the kth iteration a quasi-Newton algorithm updates the most recent point $\boldsymbol{x}_k$ to point $\boldsymbol{x}_{k+1}$ as

$$\boldsymbol{x}_{k+1} = \boldsymbol{x}_k + \alpha_k \boldsymbol{d}_k \tag{17}$$

where [14]

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

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

$$|J(\boldsymbol{x}_{k+1}) - J(\boldsymbol{x}_k)| < \varepsilon \tag{18}$$

where $\varepsilon > 0$ is a prescribed tolerance. If the iteration is terminated at step k, then $\boldsymbol{x}_k$ is viewed as a solution point.

The implementation of (17) requires the gradient of $J(\boldsymbol{x})$. Closed-form expressions for $\nabla J(\boldsymbol{x})$ are given below.

$$\begin{aligned}\frac{\partial J_o(\hat{\boldsymbol{T}})}{\partial t_{ij}} &= \lim_{\Delta\to\infty}\frac{J_o(\hat{\boldsymbol{T}}_{ij})-J_o(\hat{\boldsymbol{T}})}{\Delta} \\ &= 2\beta_1-\beta_2+2\beta_3\end{aligned} \tag{19}$$

where $\hat{\boldsymbol{T}}_{ij}$ is the matrix obtained from $\hat{\boldsymbol{T}}$ with a perturbed (i,j)th component, which is given by

$$\hat{\boldsymbol{T}}_{ij} = \hat{\boldsymbol{T}} + \frac{\Delta\hat{\boldsymbol{T}}\boldsymbol{g}_{ij}\boldsymbol{e}_j^T\hat{\boldsymbol{T}}}{1-\Delta\boldsymbol{e}_j^T\hat{\boldsymbol{T}}\boldsymbol{g}_{ij}}$$

$$\beta_1 = \boldsymbol{e}_j^T\sum_{k=0}^{\infty}\hat{\boldsymbol{T}}\hat{\boldsymbol{H}}^T(k)\hat{\boldsymbol{T}}^{-1}\hat{\boldsymbol{T}}^{-T}\hat{\boldsymbol{H}}(k)\hat{\boldsymbol{T}}^T\hat{\boldsymbol{T}}\boldsymbol{g}_{ij}$$

$$\beta_2 = \boldsymbol{e}_j^T\sum_{k=0}^{\infty}\hat{\boldsymbol{T}}^{-T}\hat{\boldsymbol{H}}(k)\hat{\boldsymbol{T}}^T\hat{\boldsymbol{T}}\hat{\boldsymbol{H}}^T(k)\boldsymbol{g}_{ij}$$

$$\beta_3 = \boldsymbol{e}_j^T\hat{\boldsymbol{T}}\,\hat{\boldsymbol{W}}_o\hat{\boldsymbol{T}}^T\hat{\boldsymbol{T}}\boldsymbol{g}_{ij}$$

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

where $\boldsymbol{e}_i$ denotes an $n\times 1$ unit vector whose ith element equals unity.

Remark 4: Although a global solution cannot be claimed, it may be worthwhile to report that we have applied the proposed algorithm to a significant number of state-space filters of various sizes, and without exception the algorithm converges to a solution that outperforms their counterparts obtained by the methods in [10] and [13]. In this regard, it is interesting to note that the unconstrained solution offered by the method in [13] is known to be globally optimal.

IV. NUMERICAL EXAMPLE

Consider a state-space digital filter, (1), specified by

$$\boldsymbol{A} = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0.453770 & -1.556160 & 1.974860 \end{bmatrix}$$

$$\boldsymbol{b} = \begin{bmatrix} 0 & 0 & 0.242096 \end{bmatrix}^T$$

$$\boldsymbol{c} = \begin{bmatrix} 0.095706 & 0.095086 & 0.327556 \end{bmatrix}$$

$$d = 0.015940$$

whose poles are at $z = 0.6578817$ and $z = 0.6584892 \pm j0.5060989$.

Using (7), the Gramians $\boldsymbol{M}_A$, $\boldsymbol{K}_c$ and $\boldsymbol{W}_o$ are calculated as

$$\boldsymbol{M}_A = \begin{bmatrix} 8.921380 & -22.046457 & 17.916285 \\ -22.046457 & 55.671710 & -46.052011 \\ 17.916285 & -46.052011 & 42.522082 \end{bmatrix}$$

$$\boldsymbol{K}_c = \begin{bmatrix} 1.000000 & 0.872501 & 0.562821 \\ 0.872501 & 1.000000 & 0.872501 \\ 0.562821 & 0.872501 & 1.000000 \end{bmatrix}$$

$$\boldsymbol{W}_o = \begin{bmatrix} 0.820741 & -2.035328 & 1.628161 \\ -2.035328 & 5.307273 & -4.264903 \\ 1.628161 & -4.264903 & 3.941491 \end{bmatrix}$$

and the L_2-sensitivity measure S in (6) is found to be

$$S = 120.184677.$$

Choosing $\hat{\boldsymbol{T}} = \boldsymbol{I}_n$ (therefore $\boldsymbol{T} = \boldsymbol{K}_c^{1/2}$ in (13)) as the initial estimate and $\varepsilon = 10^{-7}$, it took the proposed quasi-Newton algorithm 15 iterations to converge to

$$\hat{\boldsymbol{T}}^{opt} = \begin{bmatrix} 0.376709 & -0.319162 & 0.256412 \\ 0.910212 & 0.154833 & -0.218993 \\ 0.172058 & 0.934967 & 0.941433 \end{bmatrix}$$

or equivalently,

$$\boldsymbol{T}^{opt} = \begin{bmatrix} 0.913655 & -0.857313 & 0.877296 \\ 0.905773 & -0.121938 & 0.493844 \\ 0.576905 & 0.415235 & 0.377361 \end{bmatrix}$$

where (16) is used by truncating the infinite sum with $k = 100$. In this case, the L_2-sensitivity measure in (11) is minimized subject to the scaling constraints in (12) to

$$S(\boldsymbol{T}) = 8.683279.$$

The optimal state-space filter structure that minimizes the L_2-sensitivity measure, (11), subject to the scaling constraints in (12) is then constructed by substituting $\boldsymbol{T} = \boldsymbol{T}^{opt}$ into (10) as

$$\overline{\boldsymbol{A}} = \begin{bmatrix} 0.586086 & 0.567626 & 0.080361 \\ -0.450289 & 0.725542 & 0.188298 \\ -0.017947 & -0.021129 & 0.663233 \end{bmatrix}$$

$$\overline{\boldsymbol{b}} = \begin{bmatrix} -0.362927 \\ 0.393927 \\ 0.762923 \end{bmatrix}$$

$$\overline{\boldsymbol{c}} = \begin{bmatrix} 0.362538 & 0.042368 & 0.254527 \end{bmatrix}$$

$$d = 0.015940.$$

The L_2-sensitivity profile of first 18 iterations is given in Table I, where $J_o(\hat{\boldsymbol{T}})$ in (16) is used to evaluate the L_2-sensitivity under the same truncation of the infinite sum. From Table I, it is seen that with a tolerance $\varepsilon = 10^{-7}$ the algorithm converges with 15 iterations.

TABLE I

L_2-SENSITIVITY PROFILE OF FIRST 18 ITERATIONS

k	L_2-Sensitivity	k	L_2-Sensitivity
	120.18467700	9	8.80368098
0	10.71346288	10	8.70867352
1	10.70375499	11	8.70588841
2	10.70375483	12	8.69200561
3	10.33912612	13	8.69014918
4	10.27174422	14	8.68509811
5	9.59476520	15	8.68327357
6	8.99317571	16	8.68327359
7	8.93852568	17	8.68327362
8	8.88777828	18	8.68327366

For comparison purposes, the method reported in [10] is applied to minimize the L_2-sensivivity measure in (11) (without considering the scaling constraints in (12)) and the resulting optimal coordinate transformation matrix is scaled by an appropriate nonsingular diagonal matrix, so that (12) is satisfied. The result is

$$S(\boldsymbol{T}) = 9.817579$$

where

$$\boldsymbol{T} = \begin{bmatrix} 1.0 & 0.0 & 0.0 \\ 0.594723 & 0.562052 & 0.0 \\ 0.221714 & 0.736136 & 0.306792 \end{bmatrix}$$

Moreover, by applying the method reported in [13], the optimal coordinate transformation matrix $\boldsymbol{T}$ is constructed as

$$\boldsymbol{T} = \begin{bmatrix} -0.605406 & -0.119653 & 1.219423 \\ 0.107851 & 0.097317 & 0.941720 \\ 0.540830 & -0.071898 & 0.569047 \end{bmatrix}$$

which minimizes the roundoff noise at the filter output subject to the scaling constraints in (12). The L_2-sensitivity of the resulting filter with minimum roundoff noise is computed as

$$S(\boldsymbol{T}) = 8.797931.$$

It is noted that these values of the L_2-sensitivity are larger than $S(\boldsymbol{T}) = 8.683279$ obtained by the proposed method.

Finally, the results obtained in the simulation are summarized in Table II. From this table, it is observed that the proposed technique offers the smallest L_2-sensitivity subject to the L_2-norm dynamic-range scaling constraints relative to the existing methods presented in [10] and [13] for state-space digital filters.

TABLE II

L_2-SENSITIVITY COMPARISON

Realization	L_2-Sensitivity
Original	120.184661
Proposed	8.683279
Method in [10]	9.817579
Method in [13]	8.797931

V. CONCLUSION

The problem of minimizing the L_2-sensitivity of a state-space digital filter subject to L_2-norm dynamic-range scaling constraints has been investigated. It has been shown that the L_2-sensitivity minimization problem subject to the scaling constraints can be converted into an unconstrained optimization problem by using linear algebraic techniques. An efficient quasi-Newton algorithm has been applied to solve the unconstrained optimization problem. The coordinate transformation matrix obtained has allowed us to construct the optimal state-space filter structure. Our computer simulation results have demonstrated the effectiveness of the proposed technique compared with the existing methods [10],[13].

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

[3] 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, pp.72-80, Jan. 1989.

[4] 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.

[5] 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.

[6] 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.

[7] G. Li and M. Gevers, "Optimal synthetic FWL design of state-space digital filters", in *Proc. 1992 IEEE Int. Conf. Acoust., Speech, Signal Processing*, vol.4, pp.429-432.

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

[9] U. Helmke and J. B. Moore, *Optimization and Dynamical Systems*, Springer-Verlag, London, 1994.

[10] T. Hinamoto, S. Yokoyama, T. Inoue, W. Zeng and W.-S. Lu, "Analysis and minimization of L_2-sensitivity for linear systems and two-dimensional state-space filters using general controllability and observability Gramians," *IEEE Trans. Circuits Syst. I*, vol.49, pp.1279-1289, Sept. 2002.

[11] W. J. Lutz and S. L. Hakimi, "Design of multi-input multi-output systems with minimum sensitivity", *IEEE Trans. Circuits Syst.*, vol.35, pp.1114-1122, Sept. 1988.

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

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

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

[15] T. Kailath, *Linear System*, Englewood Cliffs, N.J.: Prentice Hall, 1980.