STATE-SPACE DIGITAL FILTERS WITH MINIMUM L_2-SENSITIVITY SUBJECT TO L_2-SCALING CONSTRAINTS

Takao Hinamoto, Ken-ichi Iwata
Graduate School of Engineering
Hiroshima University
Higashi-Hiroshima 739-8527, Japan
Email: hinamoto@hiroshima-u.ac.jp

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

***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 considered. A novel iterative technique is developed to solve the constraint optimization problem directly. The proposed solution method is largely based on the use of a Lagrange function and some matrix-theoretic techniques. Computer simulation results are also presented to demonstrate the effectiveness of the proposed technique.**

I. INTRODUCTION

In the implementation of fixed-point state-space digital filters with finite word length (FWL), the efficiency and performance of the filter are directly affected by the choice of its state-space filter structure. If a transfer function satisfying specification requirements is designed with infinite accuracy coefficients and realized by a state-space model, the coefficients in the state-space model must be truncated or rounded to fit the FWL constraints. The characteristics of the filter is then altered due to the coefficient quantization, which may turn a stable filter into an unstable one. Therefore, the problem of minimizing the coefficient sensitivity of a digital filter is a significant research topic. Several techniques have been proposed for synthesizing state-space digital filter structures that minimize the coefficient sensitivity. These can be divided into two main classes: the L_1/L_2-sensitivity minimization [1]-[5] and the L_2-sensitivity minimization [6]-[11]. It is noted that the sensitivity measure based on the L_2 norm is more natural and reasonable relative to the L_1/L_2-sensitivity measure. It is well known that applying the L_2-scaling constraints to a state-space digital filter is beneficial for suppressing overflow oscillation [12],[13]. However, not enough research has been done on the minimization of the L_2-sensitivity subject to the L_2-norm dynamic-range scaling constraints [11].

In this paper, the problem of minimizing the L_2-sensitivity measure subject to L_2-norm dynamic-range scaling constraints is investigated for state-space digital filters. To this end, an expression for evaluating the L_2-sensitivity is introduced. An L_2-sensitivity minimization problem subject to the scaling constraints is formulated. An iterative algorithm is then developed to solve the constraint optimization problem directly. Unlike the work reported in [11], the proposed iterative technique relies on neither converting the problem into an unconstrained optimization one nor using a quasi-Newton algorithm. From computer simulation results, it has turned out that the proposed iterative technique requires less than half amount of computations to attain almost the same convergence accuracy as compared to the technique reported in [11].

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 tr$[\boldsymbol{A}]$, respectively. The ith diagonal element of a square matrix $\boldsymbol{A}$ is denoted by $(\boldsymbol{A})_{ii}$.

II. L_2-SENSITIVITY ANALYSIS

Consider a state-space digital filter $(\boldsymbol{A}, \boldsymbol{b}, \boldsymbol{c}, d)_n$ which is stable, controllable and observable

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

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 filter in (1) is given by

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

The L_2-sensitivity of the filter in (1) is defined as follows.

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) and *Definitions 1* and *2*, the overall L_2-sensitivity measure for the 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

$$S = \mathrm{tr}[\boldsymbol{M}(\boldsymbol{I}_n)] + \mathrm{tr}[\boldsymbol{W}_o] + \mathrm{tr}[\boldsymbol{K}_c] \tag{6}$$

where

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

The matrices $\boldsymbol{K}_c$ and $\boldsymbol{W}_o$ are called the controllability and observability Gramians, respectively. The Gramians $\boldsymbol{K}_c$, $\boldsymbol{W}_o$ and $\boldsymbol{M}(\boldsymbol{P})$ with $\boldsymbol{P} = \boldsymbol{I}_n$ can be obtained by solving the Lyapunov equations [14]

$$\begin{aligned} \boldsymbol{K}_c &= \boldsymbol{A}\boldsymbol{K}_c\boldsymbol{A}^T + \boldsymbol{b}\boldsymbol{b}^T \\ \boldsymbol{W}_o &= \boldsymbol{A}^T\boldsymbol{W}_o\boldsymbol{A} + \boldsymbol{c}^T\boldsymbol{c} \\ \boldsymbol{Y} &= \begin{bmatrix} \boldsymbol{A} & \boldsymbol{bc} \\ \boldsymbol{0} & \boldsymbol{A} \end{bmatrix}^T \boldsymbol{Y} \begin{bmatrix} \boldsymbol{A} & \boldsymbol{bc} \\ \boldsymbol{0} & \boldsymbol{A} \end{bmatrix} + \begin{bmatrix} \boldsymbol{P}^{-1} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{0} \end{bmatrix} \end{aligned} \tag{7}$$

and taking the lower-right $n \times n$ block of $\boldsymbol{Y}$ as $\boldsymbol{M}(\boldsymbol{P})$, i.e.,

$$\boldsymbol{M}(\boldsymbol{P}) = \begin{bmatrix} \boldsymbol{0} & \boldsymbol{I}_n \end{bmatrix} \boldsymbol{Y} \begin{bmatrix} \boldsymbol{0} \\ \boldsymbol{I}_n \end{bmatrix}. \tag{8}$$

If a coordinate transformation defined by

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

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

$$\begin{aligned} \overline{\boldsymbol{A}} &= \boldsymbol{T}^{-1}\boldsymbol{A}\boldsymbol{T}, \qquad \overline{\boldsymbol{b}} = \boldsymbol{T}^{-1}\boldsymbol{b}, \qquad \overline{\boldsymbol{c}} = \boldsymbol{c}\boldsymbol{T} \\ \overline{\boldsymbol{K}}_c &= \boldsymbol{T}^{-1}\boldsymbol{K}_c\boldsymbol{T}^{-T}, \qquad \overline{\boldsymbol{W}}_o = \boldsymbol{T}^T\boldsymbol{W}_o\boldsymbol{T}. \end{aligned} \tag{10}$$

From (2) and (10), it is clear that the transfer function $H(z)$ is invariant under the coordinate transformation in (9). Noting that the coordinate transformation in (9) transforms the Gramian $\boldsymbol{M}(\boldsymbol{I}_n)$ into $\boldsymbol{T}^T\boldsymbol{M}(\boldsymbol{P})\boldsymbol{T}$, the L_2-sensitivity measure in (6) is changed to

$$S(\boldsymbol{P}) = \mathrm{tr}[\boldsymbol{M}(\boldsymbol{P})\boldsymbol{P}] + \mathrm{tr}[\boldsymbol{W}_o\boldsymbol{P}] + \mathrm{tr}[\boldsymbol{K}_c\boldsymbol{P}^{-1}] \tag{11}$$

where $\boldsymbol{P} = \boldsymbol{T}\boldsymbol{T}^T$.

Moreover, if the L_2-norm dynamic-range scaling constraints are imposed on the new state-variable vector $\overline{\boldsymbol{x}}(k)$, 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 $\boldsymbol{c}$, obtain an $n \times n$ nonsingular matrix $\boldsymbol{T}$ which minimizes (11) subject to the scaling constraints in (12).*

III. L_2-SENSITIVITY MINIMIZATION

The problem of minimizing $S(\boldsymbol{P})$ in (11) subject to the constraints in (12) is a constrained nonlinear optimization problem where the variable matrix is $\boldsymbol{P}$. If we sum the n constraints in (12) up, then we have

$$\mathrm{tr}[\boldsymbol{T}^{-1}\boldsymbol{K}_c\boldsymbol{T}^{-T}] = \mathrm{tr}[\boldsymbol{K}_c\boldsymbol{P}^{-1}] = n. \tag{13}$$

Consequently, the problem of minimizing (11) subject to the constraints in (12) can be *relaxed* into the following problem:

$$\begin{aligned} &\text{minimize } S(\boldsymbol{P}) \text{ in (11)} \\ &\text{subject to } tr[\boldsymbol{K}_c\boldsymbol{P}^{-1}] = n. \end{aligned} \tag{14}$$

Although clearly a solution of problem (14) is not necessarily a solution of the problem of minimizing (11) subject to the constraints in (12), it is important to stress that the ultimate solution we seek for is not matrix $\boldsymbol{P}$ but a nonsingular matrix $\boldsymbol{T}$ that is related to the solution of the problem of minimizing (11) subject to the constraints in (12) as $\boldsymbol{P} = \boldsymbol{T}\boldsymbol{T}^T$. If matrix $\boldsymbol{P}$ is a solution of problem (14) and $\boldsymbol{P}^{1/2}$ denotes a matrix square root of $\boldsymbol{P}$, i.e., $\boldsymbol{P} = \boldsymbol{P}^{1/2}\boldsymbol{P}^{1/2}$, then it is easy to see that any matrix $\boldsymbol{T}$ of the form $\boldsymbol{T} = \boldsymbol{P}^{1/2}\boldsymbol{U}$ where $\boldsymbol{U}$ is an arbitrary orthogonal matrix still holds the relation $\boldsymbol{P} = \boldsymbol{T}\boldsymbol{T}^T$. As will be shown shortly, under the constraints in (12) there exists an orthogonal matrix $\boldsymbol{U}$ such that matrix $\boldsymbol{T} = \boldsymbol{P}^{1/2}\boldsymbol{U}$ satisfies the constraints in (12), where $\boldsymbol{P}^{1/2}$ is a square root of the solution matrix $\boldsymbol{P}$ for problem (14).

It is for these reasons we now address problem (14) as the first step of our solution strategy. To solve (14), we define the Lagrange function of the problem as

$$\begin{aligned} J(\boldsymbol{P}, \lambda) = {}& \mathrm{tr}[\boldsymbol{M}(\boldsymbol{P})\boldsymbol{P}] + \mathrm{tr}[\boldsymbol{W}_o\boldsymbol{P}] \\ &+ \mathrm{tr}[\boldsymbol{K}_c\boldsymbol{P}^{-1}] + \lambda(\mathrm{tr}[\boldsymbol{K}_c\boldsymbol{P}^{-1}] - n) \end{aligned} \tag{15}$$

where λ is a Lagrange multiplier. It is well known that the solution of problem (14) must satisfy the Karush-Kuhn-Tucker (KKT) conditions $\partial J(\boldsymbol{P}, \lambda)/\partial \boldsymbol{P} = \boldsymbol{0}$ and $\partial J(\boldsymbol{P}, \lambda)/\partial \lambda = 0$ where the gradients are found to be

$$\begin{aligned} \frac{\partial J(\boldsymbol{P}, \lambda)}{\partial \boldsymbol{P}} &= \boldsymbol{M}(\boldsymbol{P}) - \boldsymbol{P}^{-1}\boldsymbol{N}(\boldsymbol{P})\boldsymbol{P}^{-1} + \boldsymbol{W}_o \\ &\quad -(\lambda + 1)\boldsymbol{P}^{-1}\boldsymbol{K}_c\boldsymbol{P}^{-1} \\ \frac{\partial J(\boldsymbol{P}, \lambda)}{\partial \lambda} &= \mathrm{tr}[\boldsymbol{K}_c\boldsymbol{P}^{-1}] - n \end{aligned} \tag{16}$$

where $\boldsymbol{N}(\boldsymbol{P})$ is obtained by solving the Lyapunov equation

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

and then taking the upper-left $n \times n$ block of $\boldsymbol{Z}$, i.e.,

$$\boldsymbol{N}(\boldsymbol{P}) = \begin{bmatrix} \boldsymbol{I}_n & \boldsymbol{0} \end{bmatrix} \boldsymbol{Y} \begin{bmatrix} \boldsymbol{I}_n \\ \boldsymbol{0} \end{bmatrix}.$$

Hence the KKT conditions become

$$\boldsymbol{P}\,\boldsymbol{F}(\boldsymbol{P})\boldsymbol{P} = \boldsymbol{G}(\boldsymbol{P}, \lambda), \qquad \text{tr}[\boldsymbol{K}_c \boldsymbol{P}^{-1}] = n \tag{17}$$

where

$$\begin{aligned} \boldsymbol{F}(\boldsymbol{P}) &= \boldsymbol{M}(\boldsymbol{P}) + \boldsymbol{W}_o \\ \boldsymbol{G}(\boldsymbol{P}, \lambda) &= \boldsymbol{N}(\boldsymbol{P}) + (\lambda + 1)\boldsymbol{K}_c. \end{aligned}$$

The first equation in (17) is highly nonlinear with respect to $\boldsymbol{P}$. An effective approach to solving the first equation in (17) is to *relax* it into the following recursive second-order matrix equation:

$$\boldsymbol{P}_{i+1}\boldsymbol{F}(\boldsymbol{P}_i)\boldsymbol{P}_{i+1} = \boldsymbol{G}(\boldsymbol{P}_i, \lambda_i) \tag{18}$$

where $\boldsymbol{P}_i$ is assumed to be known from the previous recursion and the solution $\boldsymbol{P}_{i+1}$ is given by [10]

$$\boldsymbol{P}_{i+1} = \boldsymbol{F}(\boldsymbol{P}_i)^{-\frac{1}{2}} [\boldsymbol{F}(\boldsymbol{P}_i)^{\frac{1}{2}} \boldsymbol{G}(\boldsymbol{P}_i, \lambda_i) \boldsymbol{F}(\boldsymbol{P}_i)^{\frac{1}{2}}]^{\frac{1}{2}} \boldsymbol{F}(\boldsymbol{P}_i)^{-\frac{1}{2}}. \tag{19}$$

To derive a recursive formula for the Lagrange multiplier λ, we use (17) to write

$$\text{tr}[\boldsymbol{P}\boldsymbol{F}(\boldsymbol{P})] = \text{tr}[\boldsymbol{N}(\boldsymbol{P})\boldsymbol{P}^{-1}] + n(\lambda + 1) \tag{20}$$

which naturally suggests the following recursion for λ:

$$\lambda_{i+1} = \frac{\text{tr}[\boldsymbol{P}_i \boldsymbol{F}(\boldsymbol{P}_i)] - \text{tr}[\boldsymbol{N}(\boldsymbol{P}_i)\boldsymbol{P}_i^{-1}]}{n} - 1. \tag{21}$$

In the above algorithm, λ_i is the solution of the previous iteration. The initial estimates are given by $\boldsymbol{P}_0 = \boldsymbol{I}_n$ and any value of $\lambda_0 > 0$. This iteration process continues until (17) is satisfied within a prescribed numerical tolerance.

As the second step of the solution strategy, we now turn our attention to the construction of the optimal coordinate transformation matrix $\boldsymbol{T}$ that solves the problem of minimizing (11) subject to the constraints in (12). As analyzed earlier, the optimal $\boldsymbol{T}$ assumes the form

$$\boldsymbol{T} = \boldsymbol{P}^{\frac{1}{2}}\boldsymbol{U} \tag{22}$$

where $\boldsymbol{P}^{1/2}$ is the square root of the matrix $\boldsymbol{P}$ obtained above, and $\boldsymbol{U}$ is an $n \times n$ orthogonal matrix to be determined as follows. From (10) and (22) it follows that

$$\begin{aligned} \overline{\boldsymbol{K}}_c &= \boldsymbol{T}^{-1}\boldsymbol{K}_c\boldsymbol{T}^{-T} \\ &= \boldsymbol{U}^T \boldsymbol{P}^{-\frac{1}{2}} \boldsymbol{K}_c \boldsymbol{P}^{-\frac{1}{2}} \boldsymbol{U}. \end{aligned} \tag{23}$$

In order to find an $n \times n$ orthogonal matrix $\boldsymbol{U}$ such that the matrix $\overline{\boldsymbol{K}}_c$ in (23) satisfies the scaling constraints in (12), we perform the eigenvalue-eigenvector decomposition for the positive definite matrix $\boldsymbol{P}^{-1/2}\boldsymbol{K}_c\boldsymbol{P}^{-1/2}$ as

$$\boldsymbol{P}^{-\frac{1}{2}}\boldsymbol{K}_c\boldsymbol{P}^{-\frac{1}{2}} = \boldsymbol{R}\boldsymbol{\Theta}\boldsymbol{R}^T \tag{24}$$

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

$$\boldsymbol{S}\boldsymbol{\Theta}\boldsymbol{S}^T = \begin{bmatrix} 1 & * & \cdots & * \\ * & 1 & \ddots & \vdots \\ \vdots & \ddots & \ddots & * \\ * & \cdots & * & 1 \end{bmatrix} \tag{25}$$

can be obtained by numerical manipulations [13, p.278]. Using (23), (24) and (25), it can be readily verified that the orthogonal matrix $\boldsymbol{U} = \boldsymbol{R}\boldsymbol{S}^T$ leads to a $\overline{\boldsymbol{K}}_c$ in (23) whose diagonal elements are equal to unity, hence the constraints in (12) are now satisfied. This matrix $\boldsymbol{T}$ together with (22) gives the solution of the problem of minimizing (11) subject to the constraints in (12) as

$$\boldsymbol{T} = \boldsymbol{P}^{\frac{1}{2}}\boldsymbol{R}\boldsymbol{S}^T. \tag{26}$$

IV. NUMERICAL EXAMPLE

Let a state-space digital filter in (1) be specified by

$$\begin{aligned} \boldsymbol{A} &= \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0.453770 & -1.556160 & 1.974860 \end{bmatrix} \\ \boldsymbol{b} &= \begin{bmatrix} 0 & 0 & 0.242096 \end{bmatrix}^T \\ \boldsymbol{c} &= \begin{bmatrix} 0.095706 & 0.095086 & 0.327556 \end{bmatrix} \\ d &= 0.015940. \end{aligned}$$

Performing the computation of (7) and (8), the Gramians $\boldsymbol{K}_c$, $\boldsymbol{W}_o$ and $\boldsymbol{M}(\boldsymbol{I}_3)$ are calculated as

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

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

$$\boldsymbol{M}(\boldsymbol{I}_3) = \begin{bmatrix} 8.921380 & -22.046457 & 17.916285 \\ -22.046457 & 55.671710 & -46.052011 \\ 17.916285 & -46.052011 & 42.522082 \end{bmatrix}.$$

The L_2-sensitivity measure in (6) is computed as

$$S = 120.184677.$$

Choosing $\boldsymbol{P}_0 = \boldsymbol{I}_3$ and $\lambda_0 = 100$ as the initial estimates, it took the proposed iterative algorithm 500 iterations to converge to

$$\boldsymbol{P}^{opt} = \begin{bmatrix} 2.307529 & 1.375667 & 0.514400 \\ 1.375667 & 1.103115 & 0.678193 \\ 0.514400 & 0.678193 & 0.666912 \end{bmatrix}$$

which yields

$$\boldsymbol{T}^{opt} = \begin{bmatrix} 0.906372 & 0.756223 & 0.956110 \\ 0.196978 & 0.857123 & 0.574155 \\ -0.369823 & 0.597630 & 0.415910 \end{bmatrix}.$$

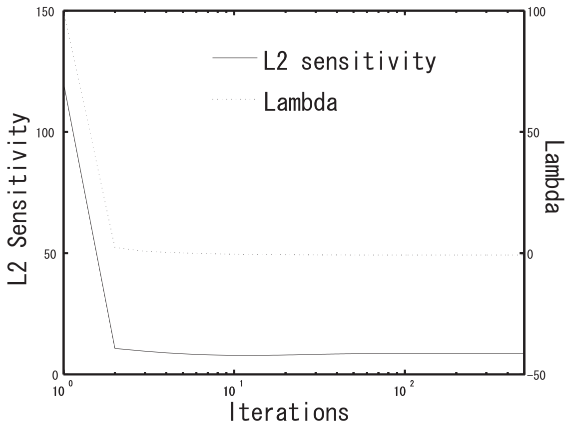

Fig. 1. L_2-Sensitivity and λ Performances

In this case, $\boldsymbol{M}(\boldsymbol{P})$ is computed from (7) and (8) as

$$\boldsymbol{M}(\boldsymbol{P}) = \begin{bmatrix} 1.908677 & -0.301984 & -1.313686 \\ -0.301984 & 1.701052 & 0.430349 \\ -1.313686 & 0.430349 & 1.395025 \end{bmatrix}$$

and the L_2-sensitivity measure in (11) is minimized subject to the scaling constraints in (12) to

$$S(\boldsymbol{P}^{opt}) = 8.672129.$$

The L_2-sensitivity and λ performances of 500 iterations are shown in Fig.1, from which it is seen that the proposed iterative algorithm sufficiently converges with 500 iterations.

For comparison purposes, only the iterative algorithm in (19) is applied by letting $\lambda_i = 0$ for any i and setting $\boldsymbol{P}_0 = \boldsymbol{I}_3$ to minimize the L_2-sensivivity measure in (11) (without considering the scaling constraints in (12)) and after 500 iterations it converges to

$$\boldsymbol{P} = \begin{bmatrix} 4.774934 & 2.835816 & 1.053819 \\ 2.835816 & 2.287705 & 1.415049 \\ 1.053819 & 1.415049 & 1.403809 \end{bmatrix}$$

which yields

$$\boldsymbol{T} = \begin{bmatrix} 2.185162 & 0.0 & 0.0 \\ 1.297760 & 0.776868 & 0.0 \\ 0.482261 & 1.015861 & 0.373174 \end{bmatrix}$$

and $S(\boldsymbol{P}) = 7.832680$. The above coordinate transformation matrix $\boldsymbol{T}$ is then scaled by an appropriate nonsingular diagonal matrix, so that the scaling constraints in (12) are satisfied. Then the result is

$$S(\boldsymbol{P}) = 9.822372$$

where $\boldsymbol{P} = \boldsymbol{T}\boldsymbol{T}^T$ and

$$\boldsymbol{T} = \begin{bmatrix} 1.0 & 0.0 & 0.0 \\ 0.593896 & 0.562461 & 0.0 \\ 0.220698 & 0.735495 & 0.307225 \end{bmatrix}.$$

Applying the technique reported in [11] yields

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

Moreover, by applying the method in [13], the coordinate transformation matrix $\boldsymbol{T}$ which yields the optimal filter structure 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 optimal filter structure is computed as

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

VI. CONCLUSION

This paper has considered the problem of minimizing an L_2-sensitivity measure subject to L_2-norm dynamic range scaling constraints for state-space digital filters. An efficient iterative technique has been developed by using a Lagrange function and some matrix-theoretic techniques in order to solve the constraint optimization problem directly. Our computer simulation results have demonstrated the effectiveness of the proposed technique compared with several existing methods.

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] T. Hinamoto, H. Ohnishi and W.-S. Lu, "Minimization of L_2-sensitivity for state-space digital filters subject to L_2-scaling constraints," *in Proc. 2004 IEEE Int. Symp. Circuits Syst.*, vol.III, pp.137-140.

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