PAPER

Analysis and Minimization of l_2-Sensitivity for Block-State Realization of IIR Digital Filters

Akimitsu DOI[†a)], ***Member***, **Takao HINAMOTO**[††b)], ***Fellow***, ***and*** **Wu-Sheng LU**[†††c)], ***Nonmember***

SUMMARY Block-state realization of state-space digital filters offers reduced implementation complexity relative to canonical state-space filters while filter's internal structure remains accessible. In this paper, we present a quantitative analysis on l_2 coefficient sensitivity of block-state digital filters. Based on this, we develop two techniques for minimizing average l_2-sensitivity subject to l_2-scaling constraints. One of the techniques is based on a Lagrange function and some matrix-theoretic techniques. The other solution method converts the problem at hand into an unconstrained optimization problem which is solved by using an efficient quasi-Newton algorithm where the key gradient evaluation is done in closed-form formulas for fast and accurate execution of quasi-Newton iterations. A case study is presented to demonstrate the validity and effectiveness of the proposed techniques.

key words: *block-state realization, l_2-scaling constraints, l_2-sensitivity analysis, l_2-sensitivity minimization, state-space digital filters*

1. Introduction

An important issue involved in the implementation of IIR digital filters using fixed-point arithmetic is reducing roundoff noise at the filter's output. As is well known, roundoff noise critically depends on the internal structure of IIR digital filters. In this regard, state-space models for IIR digital filters provide a suitable platform in which the internal structure of a filter can be explored so as to minimize its roundoff noise by choosing an appropriate linear transformation for state-space coordinates without altering the input-output characteristic of the filter. Effective techniques for constructing an optimal filter structure that minimizes the roundoff noise at the filter output subject to l_2-scaling constraints have been explored [1]–[3]. However, such a realization requires $(n+1)^2$ multiplications to compute each output sample for an n-th order filter, an increase of n^2 multiplications over canonical direct-form structures. This is a major disadvantage especially for high-order digital filters as its effect on data throughput rate becomes much greater when n is large. Alternatively, block implementation of digital filters was proposed as a method of increasing data throughput rate through parallelism, and side benefit of using it in terms of decreased roundoff noise due to finite-word-length (FWL) effects was also recognized, see [4]–[6] and the references cited therein. It is true that a block implementation induces additional delay for a given sampling rate. But it is not a concern as long as the system allows a much higher data rate where the sampling period is greatly reduced. So essentially choosing between conventional and block implementations has to do with selecting between sequential (serial) and block (parallel) operations. Parallel processing of block realization schemes makes it well suited for hardware implementations using VLSI circuits as seen in the Texas Instruments TMS320, Motorola 56000, and Analog Devices ADSP-2100 families. Another important and related issue is reducing the effects of quantizing filter coefficients. It is of importance to note that coefficient quantization usually alters filter characteristics. For instance, a stable filter designed under the assumption of infinite precision may become unstable after coefficient quantization. This has motivated the study of coefficient sensitivity and its minimization for digital filters. There are two main classes of techniques for synthesizing an optimal filter structure that minimizes the deviation of a transfer function caused by coefficient quantization, namely l_1/l_2-sensitivity minimization [7]–[11] and l_2-sensitivity minimization [12]-[17]. It has been argued [12]–[15] 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. Alternatively, the problem of minimizing the l_2-sensitivity subject to l_2-scaling constraints for state-space digital filters has been studied to suppress overflow oscillations [16], [17]. The analysis and minimization of the l_1/l_2-sensitivity for the block-state realization of IIR digital filters has also been examined in [18], [19]. A bit earlier, our study on the analysis of l_2-sensitivity and its minimization subject to l_2-scaling constraints for block-state realization of IIR digital filters has been included in [20, Sec. 17.4].

This paper presents a quantitative analysis on l_2-sensitivity for block-state realization of state-space digital filters [20, Sec. 17.4.1]. Following the analysis, we develop two techniques for minimizing a sensitivity measure known as average l_2-sensitivity subject to l_2-scaling constraints for the block-state realization of state-space digital filters [20, Sec. 17.4.2]. One of the techniques developed is based on a Lagrange function, while the other relies on an efficient quasi-Newton algorithm. Unlike the technique based on a Lagrange function in [17], [20, Sec. 17.4.2.1], the present

Manuscript received August 8, 2017.
Manuscript revised September 19, 2017.

[†]The author is with Faculty of Applied Information Science, Hiroshima Institute of Technology, Hiroshima-shi, 731-5193 Japan.
[††]The author is with Hiroshima University, Higashi-hiroshima-shi, 739-8527 Japan.
[†††]The author is with the Department of Electrical and Computer Engineering, University of Victoria, Victoria, B.C, Canada, V8W 3P6.
a) E-mail: doi@cc.it-hiroshima.ac.jp
b) E-mail: hinamoto@ieee.org
c) E-mail: wslu@ece.uvic.ca
DOI: 10.1587/transfun.E101.A.447

paper explores a much easy-to-implement algorithmic steps and proposes a warm-start initial estimate that helps accelerate the optimization process. In [16], the quasi-Newton algorithm was applied where the l_2-sensitivity measure and its gradients are partially expressed by infinite sums, thus these terms can be evaluated only approximately. In the present paper, closed-form formulas for the l_2-sensitivity measure and its gradients are derived for improved accuracy and speed. Experimental results are presented to demonstrate the validity and effectiveness of the proposed methods. This paper can be viewed as a more detailed and partially improved version of the study in [20, Sec. 17.4].

Throughout, $\boldsymbol{I}_n$ denotes the identity matrix of dimension $n\times n$, $\boldsymbol{A}^T$ ($\boldsymbol{A}^H$) denotes transpose (conjugate transpose) of a matrix $\boldsymbol{A}$, and tr[$\boldsymbol{A}$] denotes the trace of square matrix $\boldsymbol{A}$. Bold uppercase, bold lowercase and plain lowercase are used to make the distinction between matrices, vectors and scalar values, respectively.

2. l_2-Sensitivity Analysis and Problem Formulation

A. Block-state realization of a state-space model

Consider a stable, controllable and observable state-space model $(\boldsymbol{A}, \boldsymbol{b}, \boldsymbol{c}, d)_n$ 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 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. Note that in the most pessimistic case where all the elements of coefficient matrices in (1) are nontrivial, $(n+1)^2$ multiplications are required to compute the output $y(k)$ during each sample interval.

Equation (1) can be written as [4]–[6]

$$\begin{aligned} \hat{\boldsymbol{x}}(k+1) &= \hat{\boldsymbol{A}}\hat{\boldsymbol{x}}(k) + \hat{\boldsymbol{B}}\hat{\boldsymbol{u}}(k) \\ \hat{\boldsymbol{y}}(k) &= \hat{\boldsymbol{C}}\hat{\boldsymbol{x}}(k) + \hat{\boldsymbol{D}}\hat{\boldsymbol{u}}(k) \end{aligned} \tag{2a}$$

where $\hat{\boldsymbol{x}}(k) = \boldsymbol{x}(kL)$

$$\hat{\boldsymbol{u}}(k) = \begin{bmatrix} u(kL) \\ u(kL+1) \\ \vdots \\ u(kL+L-1) \end{bmatrix}, \quad \hat{\boldsymbol{y}}(k) = \begin{bmatrix} y(kL) \\ y(kL+1) \\ \vdots \\ y(kL+L-1) \end{bmatrix}$$

$$\hat{\boldsymbol{A}} = \boldsymbol{A}^L, \qquad \hat{\boldsymbol{B}} = \begin{bmatrix} \boldsymbol{A}^{L-1}\boldsymbol{b} & \boldsymbol{A}^{L-2}\boldsymbol{b} & \cdots & \boldsymbol{b} \end{bmatrix}$$

$$\hat{\boldsymbol{C}} = \begin{bmatrix} \boldsymbol{c} \\ \boldsymbol{c}\boldsymbol{A} \\ \vdots \\ \boldsymbol{c}\boldsymbol{A}^{L-1} \end{bmatrix}, \quad \hat{\boldsymbol{D}} = \begin{bmatrix} d & 0 & \cdots & 0 \\ \boldsymbol{c}\boldsymbol{b} & d & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ \boldsymbol{c}\boldsymbol{A}^{L-2}\boldsymbol{b} & \cdots & \boldsymbol{c}\boldsymbol{b} & d \end{bmatrix} \tag{2b}$$

From (2a), it follows that a total of

$$m = n^2 + 2nL + \frac{L(L+1)}{2} \tag{3}$$

multiplications are required in each block of L output samples or, equivalently, an average of

$$\frac{m}{L} = \frac{n^2}{L} + 2n + \frac{L+1}{2} \tag{4}$$

multiplications for each output sample. If the average measure, m/L, in (4) is minimized with respect to L, the optimal block length is obtained as

$$L = \sqrt{2}n \tag{5}$$

which is non-integer and requires rounding to the nearest integer. By substituting (5) into (4), the minimum value of m/L becomes

$$\left(\frac{m}{L}\right)_{min} = (2+\sqrt{2})n + \frac{1}{2} \tag{6}$$

This compares favorably with the processing complexity for the canonical forms of the SISO state-space model in (1), which requires $(2n+1)$ multiplications per output sample, and reveals that the system in (2a) enables us to perform fast processing for high-order digital filters.

In the rest of the paper, the system described by (2a) is referred to as *block-state realization*, $(\hat{\boldsymbol{A}}, \hat{\boldsymbol{B}}, \hat{\boldsymbol{C}}, \hat{\boldsymbol{D}})_n$, that is generated by the state-space model $(\boldsymbol{A}, \boldsymbol{b}, \boldsymbol{c}, d)_n$ in (1). In this way, (2b) defines a mapping $(\boldsymbol{A}, \boldsymbol{b}, \boldsymbol{c}, d)_n \rightarrow (\hat{\boldsymbol{A}}, \hat{\boldsymbol{B}}, \hat{\boldsymbol{C}}, \hat{\boldsymbol{D}})_n$ that transforms the state-space model in (1) to the block-state realization in (2).

The block-state realization was proposed to achieve a high speed algorithm for the state-space model in (1) [4]–[6]. The system in (2a) can be realized by an L-input/L-output digital filter implemented with serial-in/parallel-out and parallel-in/serial-out registers, as shown in Fig. 1. Hence, the number of multiplications per output sample can be drastically reduced by updating the state vector that requires great many multiplications once in L samples. Moreover, because a block-state realization preserves the properties of the state-space model [5], the robust block-state realization can be constructed from the state-space model that minimizes FWL effects. In other words, we can realize the filtering with high speed and accuracy (due to the reduced number of computations per output sample) via the block-state realization. It

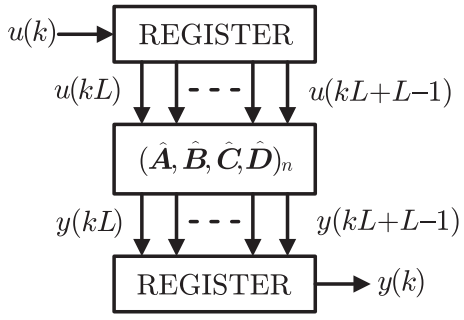


Fig. 1 Block-state realization using serial-in/parallel-out and parallel-in/serial-out registers.

is obvious that the serial input and output sample rates are L times the fundamental clock rate of the system in (2a).

B. Analysis of l_2-sensitivity

B.1 Definition of sensitivity function

Consider an MIMO transfer function $\boldsymbol{H}(z)$ that contains N parameters $\{p_1, p_2, \cdots, p_N\}$. Let $\{\tilde{p}_i\}$ be the FWL version of $\{p_i\}$ where $\tilde{p}_i = p_i + \Delta p_i$ with Δp_i the parameter perturbation, and $\boldsymbol{H}_{act}(z)$ be the transfer function associated with perturbed parameters $\{\tilde{p}_i\}$. The first-order approximation of $\boldsymbol{H}_{act}(z)$ then gives

$$\boldsymbol{H}_{act}(z) = \boldsymbol{H}(z) + \Delta\boldsymbol{H}(z) \tag{7a}$$

where $\Delta\boldsymbol{H}(z)$ will be

$$\Delta\boldsymbol{H}(z) = \sum_{i=1}^{N} \frac{\partial \boldsymbol{H}(z)}{\partial p_i}\Delta p_i \tag{7b}$$

For a fixed-point implementation of B bits, the parameter perturbations are considered to be independent uniformly distributed random variables within a range $[-2^{-B-1}, 2^{-B-1}]$. Then, a measure of the deviation in the transfer function can statistically be defined as

$$\begin{aligned}\sigma^2_{\Delta H} &= \frac{1}{2\pi}\int_0^{2\pi} E\left(\mathrm{tr}[\Delta\boldsymbol{H}^H(e^{j\omega})\Delta\boldsymbol{H}(e^{j\omega})]\right) d\omega \\ &= \frac{1}{2\pi j}\oint_{|z|=1} E\left(\mathrm{tr}[\Delta\boldsymbol{H}^H(z)\Delta\boldsymbol{H}(z)]\right)\frac{dz}{z}\end{aligned} \tag{8}$$

where $E(\cdot)$ denotes the ensemble average operation. Since $\{\Delta p_i\}$ are independent uniformly distributed random variables, it follows that

$$E\left(\mathrm{tr}[\Delta\boldsymbol{H}^H(z)\Delta\boldsymbol{H}(z)]\right) = \sum_{i=1}^{N}\mathrm{tr}\Big[\frac{\partial \boldsymbol{H}^H(z)}{\partial p_i}\cdot\frac{\partial \boldsymbol{H}(z)}{\partial p_i}\Big]\sigma^2 \tag{9}$$

where $\sigma^2 = E((\Delta p_i)^2) = 2^{-2B}/12$. Equation (9) establishes an analytic relationship between deviation in the transfer function induced by an FWL block-state realization and parameter sensitivity.

We are now in a position to define the l_2-sensitivity of the block-state realization in (2a).

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 entries of $\boldsymbol{X}$. The sensitivity function of $f(\boldsymbol{X})$ with respect to $\boldsymbol{X}$ is then defined as

$$\frac{\partial f(\boldsymbol{X})}{\partial \boldsymbol{X}} = \begin{bmatrix} \dfrac{\partial f(\boldsymbol{X})}{\partial x_{11}} & \cdots & \dfrac{\partial f(\boldsymbol{X})}{\partial x_{1n}} \\ \vdots & \ddots & \vdots \\ \dfrac{\partial f(\boldsymbol{X})}{\partial x_{m1}} & \cdots & \dfrac{\partial f(\boldsymbol{X})}{\partial x_{mn}} \end{bmatrix} \tag{10}$$

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

B.2 l_2-sensitivity of system (2)

We now apply the above definition to evaluate the sensitivity of the transfer function of the block-state realization in (2a) which is given by

$$\boldsymbol{H}(z) = \hat{\boldsymbol{C}}(z\boldsymbol{I}_n - \hat{\boldsymbol{A}})^{-1}\hat{\boldsymbol{B}} + \hat{\boldsymbol{D}} \tag{11}$$

whose (i,l)th element is described by

$$H_{il}(z) = \hat{\boldsymbol{c}}_i(z\boldsymbol{I}_n - \hat{\boldsymbol{A}})^{-1}\hat{\boldsymbol{b}}_l + \hat{d}_{il} \tag{12a}$$

where

$$\hat{\boldsymbol{B}} = \begin{bmatrix} \hat{\boldsymbol{b}}_1 & \hat{\boldsymbol{b}}_2 & \cdots & \hat{\boldsymbol{b}}_L \end{bmatrix}$$

$$\hat{\boldsymbol{C}} = \begin{bmatrix} \hat{\boldsymbol{c}}_1 \\ \hat{\boldsymbol{c}}_2 \\ \vdots \\ \hat{\boldsymbol{c}}_L \end{bmatrix}, \quad \hat{d}_{il} = \begin{cases} 0 & \text{for} \quad i < l \\ d & \text{for} \quad i = l \\ \boldsymbol{c}\boldsymbol{A}^{i-l-1}\boldsymbol{b} & \text{for} \quad i > l \end{cases} \tag{12b}$$

By virtue of (12a), Definition 1, and the formula

$$\frac{\partial \boldsymbol{A}^{-1}}{\partial a} = -\boldsymbol{A}^{-1}\frac{\partial \boldsymbol{A}}{\partial a}\boldsymbol{A}^{-1} \tag{13}$$

with a an entry of any invertible matrix $\boldsymbol{A}$, the sensitivities of $H_{il}(z)$ with respect to matrices $\hat{\boldsymbol{A}}$, $\hat{\boldsymbol{b}}_l$, $\hat{\boldsymbol{c}}_i^T$ and $\hat{d}_{il}$ are found to be

$$\begin{aligned} \frac{\partial H_{il}(z)}{\partial \hat{\boldsymbol{A}}} &= [\boldsymbol{f}_l(z)\boldsymbol{g}_i(z)]^T, & \frac{\partial H_{il}(z)}{\partial \hat{\boldsymbol{b}}_l} &= \boldsymbol{g}_i^T(z) \\ \frac{\partial H_{il}(z)}{\partial \hat{\boldsymbol{c}}_i^T} &= \boldsymbol{f}_l(z), & \frac{\partial H_{il}(z)}{\partial \hat{d}_{il}} &= u_o(i-l) \end{aligned} \tag{14a}$$

respectively, where

$$\begin{aligned} \boldsymbol{f}_l(z) &= (z\boldsymbol{I}_n - \hat{\boldsymbol{A}})^{-1}\hat{\boldsymbol{b}}_l = (z\boldsymbol{I}_n - \boldsymbol{A}^L)^{-1}\boldsymbol{A}^{L-l}\boldsymbol{b} \\ &= \sum_{k=0}^{\infty}\boldsymbol{A}^{(k+1)L-l}\boldsymbol{b}\, z^{-(k+1)} \\ \boldsymbol{g}_i(z) &= \hat{\boldsymbol{c}}_i(z\boldsymbol{I}_n - \hat{\boldsymbol{A}})^{-1} = \boldsymbol{c}\boldsymbol{A}^{i-1}(z\boldsymbol{I}_n - \boldsymbol{A}^L)^{-1} \\ &= \sum_{k=0}^{\infty}\boldsymbol{c}\boldsymbol{A}^{kL+i-1}z^{-(k+1)} \\ u_o(i) &= \begin{cases} 1 & \text{for} \quad i \geq 0 \\ 0 & \text{for} \quad i < 0 \end{cases} \end{aligned} \tag{14b}$$

In the rest of this section, $\boldsymbol{f}_l(z)$ and $\boldsymbol{g}_i(z)$ are referred to as *intermediate functions*.

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

$$\begin{aligned}\|\boldsymbol{X}(z)\|_2 &= \left[\frac{1}{2\pi}\int_0^{2\pi}\sum_{p=1}^{m}\sum_{q=1}^{n}\left|x_{pq}(e^{j\omega})\right|^2 d\omega\right]^{\frac{1}{2}} \\ &= \left(\mathrm{tr}\left[\frac{1}{2\pi j}\oint_{|z|=1}\boldsymbol{X}(z)\boldsymbol{X}^H(z)\frac{dz}{z}\right]\right)^{\frac{1}{2}}\end{aligned} \tag{15}$$

From (14) and Definition 2, the l_2-sensitivity of the subsystem (12a) can be expressed as

$$\begin{aligned} S_{il} &= \left\| \frac{\partial H_{il}(z)}{\partial \hat{\boldsymbol{A}}} \right\|_2^2 + \left\| \frac{\partial H_{il}(z)}{\partial \hat{\boldsymbol{b}}_l} \right\|_2^2 + \left\| \frac{\partial H_{il}(z)}{\partial \hat{\boldsymbol{c}}_i^T} \right\|_2^2 + \left\| \frac{\partial H_{il}(z)}{\partial \hat{d}_{il}} \right\|_2^2 \\ &= \left\| [\boldsymbol{f}_l(z)\boldsymbol{g}_i(z)]^T \right\|_2^2 + \left\| \boldsymbol{g}_i^T(z) \right\|_2^2 + \| \boldsymbol{f}_l(z) \|_2^2 + u_o(i-l) \end{aligned} \tag{16}$$

In order to evaluate the integrals involved in (16), we utilize *Cauchy's integral theorem*

$$\frac{1}{2\pi j} \oint_C z^i \frac{\partial z}{z} = \begin{cases} 1, & i = 0 \\ 0, & i \neq 0 \end{cases} \tag{17}$$

with C a counterclockwise contour that encircles the origin. Then by virtue of (15), (16) becomes

$$\begin{aligned} S_{il} = \mathrm{tr}[\boldsymbol{N}_{il}] &+ \sum_{k=0}^{\infty} \mathrm{tr}\left[(\boldsymbol{c}\boldsymbol{A}^{kL+i-1})^T \boldsymbol{c}\boldsymbol{A}^{kL+i-1} \right] \\ &+ \sum_{k=0}^{\infty} \mathrm{tr}\left[\boldsymbol{A}^{(k+1)L-l} \boldsymbol{b} \left(\boldsymbol{A}^{(k+1)L-l} \boldsymbol{b} \right)^T \right] + u_o(i-l) \end{aligned} \tag{18a}$$

where $\mathrm{tr}[\boldsymbol{N}_{il}] = \mathrm{tr}[\boldsymbol{M}_{il}]$ and

$$\begin{aligned} \boldsymbol{N}_{il} &= \frac{1}{2\pi j} \oint_{|z|=1} [\boldsymbol{f}_l(z)\boldsymbol{g}_i(z)]^T \boldsymbol{f}_l(z^{-1}) \boldsymbol{g}_i(z^{-1}) \frac{dz}{z} \\ \boldsymbol{M}_{il} &= \frac{1}{2\pi j} \oint_{|z|=1} \boldsymbol{f}_l(z^{-1}) \boldsymbol{g}_i(z^{-1}) [\boldsymbol{f}_l(z)\boldsymbol{g}_i(z)]^T \frac{dz}{z} \end{aligned} \tag{18b}$$

A closed-form solution for evaluating matrices $\boldsymbol{N}_{il}$ and $\boldsymbol{M}_{il}$ will be deduced shortly.

We note that each output $y(kL+i-1)$ for $i = 1, 2, \cdots, L$ in the output vector $\boldsymbol{y}(k)$ is generated by the subsystem

$$\begin{aligned} \boldsymbol{H}_i(z) &= [H_{i1}(z), H_{i2}(z), \cdots, H_{iL}(z)] \\ &= \hat{\boldsymbol{c}}_i (z\boldsymbol{I} - \hat{\boldsymbol{A}})^{-1} \hat{\boldsymbol{B}} + \hat{\boldsymbol{d}}_i \end{aligned} \tag{19}$$

where $\hat{\boldsymbol{d}}_i = [\hat{d}_{i1}, \hat{d}_{i2}, \cdots, \hat{d}_{iL}]$. From this in conjunction with (18a), the l_2-sensitivity measure for the subsystem in (19) is found to be

$$\begin{aligned} S_i &= \sum_{l=1}^{L} S_{il} \\ &= \sum_{l=1}^{L} \mathrm{tr}[\boldsymbol{N}_{il}] + \sum_{l=1}^{L} \sum_{k=0}^{\infty} \mathrm{tr}\left[(\boldsymbol{c}\boldsymbol{A}^{kL+i-1})^T \boldsymbol{c}\boldsymbol{A}^{kL+i-1} \right] \\ &\quad + \sum_{k=0}^{\infty} \mathrm{tr}\left[\boldsymbol{A}^k \boldsymbol{b} (\boldsymbol{A}^k \boldsymbol{b})^T \right] + i \quad \text{for } i = 1, 2, \cdots, L \end{aligned} \tag{20}$$

Hence, the overall l_2-sensitivity for the block-state realization in (11) can be written as

$$\begin{aligned} S = \sum_{i=1}^{L} S_i = \sum_{i=1}^{L} \sum_{l=1}^{L} \mathrm{tr}[\boldsymbol{N}_{il}] &+ \sum_{l=1}^{L} \sum_{k=0}^{\infty} \mathrm{tr}\left[(\boldsymbol{c}\boldsymbol{A}^k)^T \boldsymbol{c}\boldsymbol{A}^k \right] \\ &+ \sum_{i=1}^{L} \sum_{k=0}^{\infty} \mathrm{tr}\left[\boldsymbol{A}^k \boldsymbol{b} (\boldsymbol{A}^k \boldsymbol{b})^T \right] + \frac{L(L+1)}{2} \end{aligned} \tag{21}$$

which is equivalent to

$$\begin{aligned} S = \sum_{i=1}^{L} \sum_{l=1}^{L} \mathrm{tr}[\boldsymbol{N}_{il}] &+ L\,\mathrm{tr}[\boldsymbol{W}_o] + L\,\mathrm{tr}[\boldsymbol{K}_c] \\ &+ \frac{L(L+1)}{2} \end{aligned} \tag{22a}$$

where $\boldsymbol{K}_c$ and $\boldsymbol{W}_o$ are the controllability and observability Grammians of the state-space model in (1), defined by

$$\boldsymbol{K}_c = \sum_{k=0}^{\infty} \boldsymbol{A}^k \boldsymbol{b} \, (\boldsymbol{A}^k \boldsymbol{b})^T, \quad \boldsymbol{W}_o = \sum_{k=0}^{\infty} (\boldsymbol{c}\boldsymbol{A}^k)^T \boldsymbol{c}\boldsymbol{A}^k \tag{22b}$$

that can be obtained by solving the Lyapunov equations

$$\boldsymbol{K}_c = \boldsymbol{A}\boldsymbol{K}_c\boldsymbol{A}^T + \boldsymbol{b}\boldsymbol{b}^T, \quad \boldsymbol{W}_o = \boldsymbol{A}^T\boldsymbol{W}_o\boldsymbol{A} + \boldsymbol{c}^T\boldsymbol{c} \tag{22c}$$

The l_2-sensitivity measure for the state-space model in (1) was derived in [17], namely,

$$S_o = \mathrm{tr}[\boldsymbol{N}_{11}] + \mathrm{tr}[\boldsymbol{W}_o] + \mathrm{tr}[\boldsymbol{K}_c] + 1 \tag{23}$$

The l_2-sensitivity of a conventional state-space model, S_o, can be regarded as a special case of the l_2-sensitivity of a block realization with $L = 1$. However, for the general case with $L > 1$, the two sensitivity measures S and S_o are not directly comparable because the realization scheme has changed from serial to parallel. The average sensitivity S_{ave} comes as a reasonable measure as comparing it with S_o becomes meaningful and it has a straightforward connection to sensitivity S by a scaling factor of $1/L$. That is

$$S_{ave} = \frac{1}{L} \sum_{i=1}^{L} S_i = \frac{S}{L} \tag{24}$$

which in conjunction with (22a) gives

$$\begin{aligned} S_{ave} = \frac{1}{L} \sum_{i=1}^{L} \sum_{l=1}^{L} \mathrm{tr}[\boldsymbol{N}_{il}] &+ \mathrm{tr}[\boldsymbol{W}_o] + \mathrm{tr}[\boldsymbol{K}_c] \\ &+ \frac{L+1}{2} \end{aligned} \tag{25}$$

Notice that the same idea was applied to evaluate the roundoff noise [4] and the l_1/l_2-sensitivity [18], [19] in the block-state realization.

C. Problem formulation

To identify an optimal internal structure of a given IIR filter that achieves the minimum average l_2-sensitivity, consider a coordinate transformation defined by

$$\overline{\boldsymbol{x}}(k) = \boldsymbol{T}^{-1}\boldsymbol{x}(k) \tag{26}$$

that is applied to the state-space model in (1). The new realization associated with state vector $\overline{\boldsymbol{x}}(k)$, denoted by $(\overline{\boldsymbol{A}}, \overline{\boldsymbol{b}}, \overline{\boldsymbol{c}}, d)_n$, is related to the original realization as

$$\overline{\boldsymbol{A}} = \boldsymbol{T}^{-1}\boldsymbol{A}\boldsymbol{T}, \qquad \overline{\boldsymbol{b}} = \boldsymbol{T}^{-1}\boldsymbol{b}, \qquad \overline{\boldsymbol{c}} = \boldsymbol{c}\boldsymbol{T} \tag{27}$$

For this realization, the new controllability Grammian $\overline{\boldsymbol{K}}_c$ and new observability Grammian $\overline{\boldsymbol{W}}_o$ are related to the original Grammians as

$$\overline{\boldsymbol{K}}_c = \boldsymbol{T}^{-1}\boldsymbol{K}_c\boldsymbol{T}^{-T}, \qquad \overline{\boldsymbol{W}}_o = \boldsymbol{T}^{T}\boldsymbol{W}_o\boldsymbol{T} \tag{28}$$

respectively, the new intermediate functions $\overline{\boldsymbol{f}}_j(z)$ and $\overline{\boldsymbol{g}}_i(z)$ are found to be [14], [17]

$$\overline{\boldsymbol{f}}_l(z) = \boldsymbol{T}^{-1}\boldsymbol{f}_l(z), \qquad \overline{\boldsymbol{g}}_i(z) = \boldsymbol{g}_i(z)\,\boldsymbol{T} \tag{29}$$

respectively, and the canonical-state to block-state mapping is given by

$$(\overline{\boldsymbol{A}}, \overline{\boldsymbol{b}}, \overline{\boldsymbol{c}}, d)_n \rightarrow (\boldsymbol{T}^{-1}\hat{\boldsymbol{A}}\boldsymbol{T}, \boldsymbol{T}^{-1}\hat{\boldsymbol{B}}, \hat{\boldsymbol{C}}\boldsymbol{T}, \hat{\boldsymbol{D}})_n \tag{30}$$

Moreover, matrices $\boldsymbol{N}_{il}$ and $\boldsymbol{M}_{il}$ introduced in (18b) are transformed to $\overline{\boldsymbol{N}}_{il}$ and $\overline{\boldsymbol{M}}_{il}$ as

$$\begin{aligned}
\overline{\boldsymbol{N}}_{il} &= \frac{1}{2\pi j}\oint_{|z|=1} [\,\overline{\boldsymbol{f}}_l(z)\overline{\boldsymbol{g}}_i(z)\,]^T\,\overline{\boldsymbol{f}}_l(z^{-1})\overline{\boldsymbol{g}}_i(z^{-1})\frac{dz}{z} \\
&= \boldsymbol{T}^T\boldsymbol{N}_{il}(\boldsymbol{P})\,\boldsymbol{T} \\
\overline{\boldsymbol{M}}_{il} &= \frac{1}{2\pi j}\oint_{|z|=1} \overline{\boldsymbol{f}}_l(z^{-1})\overline{\boldsymbol{g}}_i(z^{-1})\,[\,\overline{\boldsymbol{f}}_l(z)\overline{\boldsymbol{g}}_i(z)\,]^T\,\frac{dz}{z} \\
&= \boldsymbol{T}^{-1}\boldsymbol{M}_{il}(\boldsymbol{P})\,\boldsymbol{T}^{-T}
\end{aligned} \tag{31a}$$

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

$$\begin{aligned}
\boldsymbol{N}_{il}(\boldsymbol{P}) &= \frac{1}{2\pi j}\oint_{|z|=1} [\boldsymbol{f}_l(z)\boldsymbol{g}_i(z)]^T\boldsymbol{P}^{-1}\boldsymbol{f}_l(z^{-1})\boldsymbol{g}_i(z^{-1})\frac{dz}{z} \\
\boldsymbol{M}_{il}(\boldsymbol{P}) &= \frac{1}{2\pi j}\oint_{|z|=1} \boldsymbol{f}_l(z^{-1})\boldsymbol{g}_i(z^{-1})\boldsymbol{P}\,[\boldsymbol{f}_l(z)\boldsymbol{g}_i(z)]^T\frac{dz}{z}
\end{aligned} \tag{31b}$$

Hence $\boldsymbol{N}_{il} = \boldsymbol{N}_{il}(\boldsymbol{I}_n)$ and $\boldsymbol{M}_{il} = \boldsymbol{M}_{il}(\boldsymbol{I}_n)$ follow from (18b). It is noted that

$$\begin{aligned}
\overline{\boldsymbol{f}}_l(z)\overline{\boldsymbol{g}}_i(z) &= \boldsymbol{T}^{-1}\boldsymbol{f}_l(z)\boldsymbol{g}_i(z)\,\boldsymbol{T} \\
&= \begin{bmatrix} \boldsymbol{T}^{-1} & \boldsymbol{0} \end{bmatrix} \begin{bmatrix} z\boldsymbol{I}_n - \boldsymbol{A}^L & -\boldsymbol{A}^{L-l}\boldsymbol{b}\boldsymbol{c}\boldsymbol{A}^{i-1} \\ \boldsymbol{0} & z\boldsymbol{I}_n - \boldsymbol{A}^L \end{bmatrix}^{-1} \begin{bmatrix} \boldsymbol{0} \\ \boldsymbol{T} \end{bmatrix}
\end{aligned} \tag{32}$$

If we denote the observability Grammian of a composite system $\overline{\boldsymbol{f}}_l(z)\overline{\boldsymbol{g}}_i(z)$ in (32) by $\boldsymbol{Y}_{il}$, matrix $\boldsymbol{N}_{il}(\boldsymbol{P})$ can be obtained by solving the Lyapunov equation

$$\begin{aligned}
\boldsymbol{Y}_{il} = &\begin{bmatrix} \boldsymbol{A}^L & \boldsymbol{A}^{L-l}\boldsymbol{b}\boldsymbol{c}\boldsymbol{A}^{i-1} \\ \boldsymbol{0} & \boldsymbol{A}^L \end{bmatrix}^T \boldsymbol{Y}_{il} \begin{bmatrix} \boldsymbol{A}^L & \boldsymbol{A}^{L-l}\boldsymbol{b}\boldsymbol{c}\boldsymbol{A}^{i-1} \\ \boldsymbol{0} & \boldsymbol{A}^L \end{bmatrix} \\
&+ \begin{bmatrix} \boldsymbol{P}^{-1} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{0} \end{bmatrix}
\end{aligned} \tag{33}$$

and then taking the lower-right $n \times n$ block of $\boldsymbol{Y}_{il}$ as $\boldsymbol{N}_{il}(\boldsymbol{P})$:

$$\boldsymbol{N}_{il}(\boldsymbol{P}) = \begin{bmatrix} \boldsymbol{0} & \boldsymbol{I}_n \end{bmatrix} \boldsymbol{Y}_{il} \begin{bmatrix} \boldsymbol{0} \\ \boldsymbol{I}_n \end{bmatrix} \tag{34}$$

Therefore, under coordinate transformation $\overline{\boldsymbol{x}}(k) = \boldsymbol{T}^{-1}\boldsymbol{x}(k)$ in (26), the average l_2-sensitivity measure in (25) becomes

$$\begin{aligned}
S(\boldsymbol{T})_{ave} = &\frac{1}{L}\sum_{i=1}^{L}\sum_{l=1}^{L} \mathrm{tr}[\boldsymbol{T}^T\boldsymbol{N}_{il}(\boldsymbol{T}\boldsymbol{T}^T)\boldsymbol{T}] + \mathrm{tr}[\boldsymbol{T}^T\boldsymbol{W}_o\boldsymbol{T}] \\
&+ \mathrm{tr}[\boldsymbol{T}^{-1}\boldsymbol{K}_c\boldsymbol{T}^{-T}] + \frac{L+1}{2}
\end{aligned} \tag{35}$$

From (2) and (27), we can show that both transfer functions $\boldsymbol{H}(z)$ in (11) and $H_{il}(z)$ in (12a) are invariant under the coordinate transformation in (26).

In order to prevent overflow on the states, we impose a scaling condition that controls l_2-norm of the impulse response from input to state, hence the energy flow in the signal path. Analytically this is performed by examining the input-to-state transfer function of the system as

$$(z\boldsymbol{I}_n - \hat{\boldsymbol{A}})^{-1}\hat{\boldsymbol{B}} = \sum_{k=0}^{\infty}\hat{\boldsymbol{A}}^k\hat{\boldsymbol{B}}z^{-(k+1)} \tag{36}$$

hence the input-to-state impulse response is given by $\{\hat{\boldsymbol{B}}, \hat{\boldsymbol{A}}\hat{\boldsymbol{B}}, \hat{\boldsymbol{A}}^2\hat{\boldsymbol{B}}, \cdots\}$. The impulse response has n sequences (rows) of the form $\boldsymbol{e}_i^T \cdot \{\hat{\boldsymbol{B}}, \hat{\boldsymbol{A}}\hat{\boldsymbol{B}}, \hat{\boldsymbol{A}}^2\hat{\boldsymbol{B}}, \cdots\}$ where $\boldsymbol{e}_i$ denotes the ith column of the identity matrix $\boldsymbol{I}_n$, and the l_2-scaling requires that the l_2-norm of each sequence be unity, namely

$$\boldsymbol{e}_i^T\left[\sum_{k=0}^{\infty}\hat{\boldsymbol{A}}^k\hat{\boldsymbol{B}}(\hat{\boldsymbol{A}}^k\hat{\boldsymbol{B}})^T\right]\boldsymbol{e}_i = 1 \quad \text{for} \quad i = 1, 2, \cdots, n \tag{37a}$$

From system model (2), it can readily be verified that

$$\sum_{k=0}^{\infty}\hat{\boldsymbol{A}}^k\hat{\boldsymbol{B}}(\hat{\boldsymbol{A}}^k\hat{\boldsymbol{B}})^T = \sum_{k=0}^{\infty}\boldsymbol{A}^k\boldsymbol{b}(\boldsymbol{A}^k\boldsymbol{b})^T = \boldsymbol{K}_c \tag{37b}$$

Therefore, the scaling conditions in (37a) can be reduced to

$$\boldsymbol{e}_i^T\boldsymbol{K}_c\boldsymbol{e}_i = 1 \quad \text{for} \quad i = 1, 2, \cdots, n \tag{37c}$$

Hence the l_2-scaling on the vector impulse response $\overline{\boldsymbol{A}}^k\overline{\boldsymbol{b}}$ can be achieved by choosing a coordinate transformation in (26) so that

$$(\overline{\boldsymbol{K}}_c)_{ii} = (\boldsymbol{T}^{-1}\boldsymbol{K}_c\boldsymbol{T}^{-T})_{ii} = 1 \quad \text{for} \quad i = 1, 2, \cdots, n \tag{38}$$

are satisfied.

Under these circumstances, the problem being considered here is to obtain an $n \times n$ coordinate transformation matrix $\boldsymbol{T}$ that minimizes the average l_2-sensitivity measure $S(\boldsymbol{T})_{ave}$ in (35) subject to the l_2-scaling constraints in (38).

Remark 1: The l_2-sensitivity measure for the new realization $(\overline{\boldsymbol{A}}, \overline{\boldsymbol{b}}, \overline{\boldsymbol{c}}, d)_n$ characterized by (27) derived in [17], namely,

$$\begin{aligned} S_o(\boldsymbol{T}) = {} & \mathrm{tr}[\boldsymbol{T}^T\boldsymbol{N}_{11}(\boldsymbol{T}\boldsymbol{T}^T)\boldsymbol{T}] + \mathrm{tr}[\boldsymbol{T}^T\boldsymbol{W}_o\boldsymbol{T}] \\ & + \mathrm{tr}[\boldsymbol{T}^{-1}\boldsymbol{K}_c\boldsymbol{T}^{-T}] + 1 \end{aligned} \tag{39}$$

can be regarded as a special case of formula (35) with $L = 1$.

Remark 2: An average l_1/l_2-sensitivity measure for the block-state realization in (2a) was proposed in [18],[19] as

$$M_{ave} = \mathrm{tr}[\hat{\boldsymbol{W}}_o]\,\mathrm{tr}[\hat{\boldsymbol{K}}_c] + \mathrm{tr}[\hat{\boldsymbol{W}}_o] + \mathrm{tr}[\hat{\boldsymbol{K}}_c] + \frac{L+1}{2} \tag{40a}$$

where

$$\begin{aligned} \hat{\boldsymbol{K}}_c &= \sum_{k=0}^{\infty} \hat{\boldsymbol{A}}^k\hat{\boldsymbol{B}}(\hat{\boldsymbol{A}}^k\hat{\boldsymbol{B}})^T = \sum_{l=1}^{L} \boldsymbol{K}_l \\ \hat{\boldsymbol{W}}_o &= \sum_{k=0}^{\infty} (\hat{\boldsymbol{C}}\hat{\boldsymbol{A}}^k)^T\hat{\boldsymbol{C}}\hat{\boldsymbol{A}}^k = \sum_{i=1}^{L} \boldsymbol{W}_i \end{aligned} \tag{40b}$$

with $\boldsymbol{K}_l$ and $\boldsymbol{W}_i$ obtained by solving the Lyapunov equations

$$\boldsymbol{K}_l = \hat{\boldsymbol{A}}\boldsymbol{K}_l\hat{\boldsymbol{A}}^T + \hat{\boldsymbol{b}}_l\hat{\boldsymbol{b}}_l^T, \quad \boldsymbol{W}_i = \hat{\boldsymbol{A}}^T\boldsymbol{W}_i\hat{\boldsymbol{A}} + \hat{\boldsymbol{c}}_i^T\hat{\boldsymbol{c}}_i \tag{40c}$$

for $i, l = 1, 2, \cdots, L$.

3. Minimization of l_2-Sensitivity Subject to l_2-Scaling Constraints

A. Method 1: Using A Lagrange Function

We now consider the problem of minimizing the average l_2-sensitivity measure in (35) subject to l_2-scaling constraints in (38). Since the measure in (35) can be expressed in terms of matrix $\boldsymbol{P} = \boldsymbol{T}\boldsymbol{T}^T$ as

$$\begin{aligned} S(\boldsymbol{P})_{ave} = {} & \frac{1}{L}\sum_{i=1}^{L}\sum_{l=1}^{L} \mathrm{tr}[\boldsymbol{N}_{il}(\boldsymbol{P})\boldsymbol{P}] + \mathrm{tr}[\boldsymbol{W}_o\boldsymbol{P}] \\ & + \mathrm{tr}[\boldsymbol{K}_c\boldsymbol{P}^{-1}] + \frac{L+1}{2} \end{aligned} \tag{41}$$

the problem we deal with is a constrained nonlinear optimization problem where the variable is matrix $\boldsymbol{P}$.

It turns out the constraints in (38) is quite challenging. Here we handle this technical problem by a two-step procedure. First, by summing up the constraints in (38), we are led to a relaxed constraint as

$$\mathrm{tr}[\boldsymbol{T}^{-1}\boldsymbol{K}_c\boldsymbol{T}^{-T}] = \mathrm{tr}[\boldsymbol{K}_c\boldsymbol{P}^{-1}] = n \tag{42}$$

Consequently, the problem of minimizing the average l_2-sensitivity measure in (35) subject to the constraints in (38) can be *relaxed* into the following problem:

$$\begin{aligned} & \text{minimize } S(\boldsymbol{P})_{ave} \text{ in (41) with respect to } \boldsymbol{P} \\ & \text{subject to } \mathrm{tr}[\boldsymbol{K}_c\boldsymbol{P}^{-1}] = n \end{aligned} \tag{43}$$

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

Note that although the solution $\boldsymbol{P}_{opt}$ to problem (43) satisfies constraint (42) which is equivalent to $\mathrm{tr}[\boldsymbol{T}_{opt}^{-1}\boldsymbol{K}_c\boldsymbol{T}_{opt}^{-T}] = n$ where $\boldsymbol{T}_{opt} = \boldsymbol{P}_{opt}^{1/2}$. However, $\boldsymbol{T}_{opt}$ does not necessarily satisfy the l_2-scaling constraints in (38). In the second step of our solution procedure detailed below, we will present a technique for obtaining an $n \times n$ orthogonal matrix $\boldsymbol{U}$ such that (38) is satisfied by $\boldsymbol{T} = \boldsymbol{P}_{opt}^{1/2}\boldsymbol{U}$ while the relation $\boldsymbol{T}\boldsymbol{T}^T = \boldsymbol{P}_{opt}$ is obviously maintained.

We now address problem (43) as the first step of our solution procedure. To this end, we define the Lagrange function of the problem as

$$\begin{aligned} J(\boldsymbol{P}, \lambda) = {} & \frac{1}{L}\sum_{i=1}^{L}\sum_{l=1}^{L} \mathrm{tr}[\boldsymbol{N}_{il}(\boldsymbol{P})\boldsymbol{P}] + \mathrm{tr}[\boldsymbol{W}_o\boldsymbol{P}] \\ & + n + \frac{L+1}{2} + \lambda\,(\mathrm{tr}[\boldsymbol{K}_c\boldsymbol{P}^{-1}] - n) \end{aligned} \tag{44}$$

where λ is a Lagrange multiplier. It is well known that the solution of problem (43) 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

$$\begin{aligned} \frac{\partial J(\boldsymbol{P}, \lambda)}{\partial \boldsymbol{P}} = {} & \frac{1}{L}\sum_{i=1}^{L}\sum_{l=1}^{L} \boldsymbol{N}_{il}(\boldsymbol{P}) + \boldsymbol{W}_o \\ & - \boldsymbol{P}^{-1}\left[\frac{1}{L}\sum_{i=1}^{L}\sum_{l=1}^{L} \boldsymbol{M}_{il}(\boldsymbol{P}) + \lambda\boldsymbol{K}_c\right]\boldsymbol{P}^{-1} \\ \frac{\partial J(\boldsymbol{P}, \lambda)}{\partial \lambda} = {} & \mathrm{tr}[\boldsymbol{K}_c\boldsymbol{P}^{-1}] - n \end{aligned} \tag{45}$$

The matrices $\boldsymbol{M}_{il}(\boldsymbol{P})$ for $i, l = 1, 2, \cdots, L$ in (45) or (31b) can be obtained by solving the Lyapunov equations

$$\begin{aligned} \boldsymbol{X}_{il} = {} & \begin{bmatrix} \boldsymbol{A}^L & \boldsymbol{A}^{L-l}\boldsymbol{b}\boldsymbol{c}\boldsymbol{A}^{i-1} \\ \boldsymbol{0} & \boldsymbol{A}^L \end{bmatrix} \boldsymbol{X}_{il} \begin{bmatrix} \boldsymbol{A}^L & \boldsymbol{A}^{L-l}\boldsymbol{b}\boldsymbol{c}\boldsymbol{A}^{i-1} \\ \boldsymbol{0} & \boldsymbol{A}^L \end{bmatrix}^T \\ & + \begin{bmatrix} \boldsymbol{0} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{P} \end{bmatrix} \end{aligned} \tag{46}$$

and then taking the upper-left $n \times n$ block of $\boldsymbol{X}_{il}$ as $\boldsymbol{M}_{il}(\boldsymbol{P})$:

$$\boldsymbol{M}_{il}(\boldsymbol{P}) = \begin{bmatrix} \boldsymbol{I}_n & \boldsymbol{0} \end{bmatrix} \boldsymbol{X}_{il} \begin{bmatrix} \boldsymbol{I}_n \\ \boldsymbol{0} \end{bmatrix} \tag{47}$$

From (45), the KKT conditions can be expressed compactly as

$$\boldsymbol{P}\,\boldsymbol{F}(\boldsymbol{P})\boldsymbol{P} = \boldsymbol{G}(\boldsymbol{P}, \lambda), \qquad \mathrm{tr}[\boldsymbol{K}_c \boldsymbol{P}^{-1}] = n \tag{48a}$$

where

$$\begin{aligned} \boldsymbol{F}(\boldsymbol{P}) &= \frac{1}{L}\sum_{i=1}^{L}\sum_{l=1}^{L} \boldsymbol{N}_{il}(\boldsymbol{P}) + \boldsymbol{W}_o \\ \boldsymbol{G}(\boldsymbol{P}, \lambda) &= \frac{1}{L}\sum_{i=1}^{L}\sum_{l=1}^{L} \boldsymbol{M}_{il}(\boldsymbol{P}) + \lambda \boldsymbol{K}_c \end{aligned} \tag{48b}$$

Note that the first equation in (48a) is highly nonlinear with respect to $\boldsymbol{P}$. An effective approach to solving the first equation in (48a) is to *relax* it into the recursive second-order matrix equation

$$\boldsymbol{P}_{k+1}\boldsymbol{F}(\boldsymbol{P}_k)\boldsymbol{P}_{k+1} = \boldsymbol{G}(\boldsymbol{P}_k, \lambda_{k+1}) \tag{49}$$

where $\boldsymbol{P}_k$ is assumed to be known from the previous recursion. Note that if the matrix sequence $\{\boldsymbol{P}_k\}$ converges to its limit matrix, say $\boldsymbol{P}$, then (49) converges the first equation in (48a) as k goes to infinity. The solution $\boldsymbol{P}_{k+1}$ of (49) is then given by

$$\begin{aligned} &\boldsymbol{P}_{k+1} \\ &\quad = \boldsymbol{F}(\boldsymbol{P}_k)^{-\frac{1}{2}}\left[\boldsymbol{F}(\boldsymbol{P}_k)^{\frac{1}{2}}\boldsymbol{G}(\boldsymbol{P}_k, \lambda_{k+1})\boldsymbol{F}(\boldsymbol{P}_k)^{\frac{1}{2}}\right]^{\frac{1}{2}}\boldsymbol{F}(\boldsymbol{P}_k)^{-\frac{1}{2}} \end{aligned} \tag{50}$$

To derive a recursive formula for the Lagrange multiplier λ, we use (48a) to write

$$\mathrm{tr}[\boldsymbol{P}\boldsymbol{F}(\boldsymbol{P})] = \frac{1}{L}\sum_{i=1}^{L}\sum_{l=1}^{L}\mathrm{tr}[\boldsymbol{M}_{il}(\boldsymbol{P})\boldsymbol{P}^{-1}] + n\lambda \tag{51}$$

which naturally suggests the recursion for λ

$$\lambda_{k+1} = \frac{\mathrm{tr}[\boldsymbol{P}_k\boldsymbol{F}(\boldsymbol{P}_k)] - \dfrac{1}{L}\displaystyle\sum_{i=1}^{L}\sum_{l=1}^{L}\mathrm{tr}[\boldsymbol{M}_{il}(\boldsymbol{P}_k)\boldsymbol{P}_k^{-1}]}{n} \tag{52}$$

where $\boldsymbol{P}_0$ is an initial estimate such that $\boldsymbol{P}_0 = \boldsymbol{P}_0^T$.

Remark 3: If the initial estimate is chosen as $\boldsymbol{P}_0 = \boldsymbol{K}_c$, the related matrix $\boldsymbol{T}_0 = \boldsymbol{K}_c^{1/2}\boldsymbol{V}$ produces an input-normal realization $(\boldsymbol{T}_0^{-1}\boldsymbol{A}\boldsymbol{T}_0, \boldsymbol{T}_0^{-1}\boldsymbol{b}, \boldsymbol{c}\boldsymbol{T}_0, d)_n$ satisfying $\boldsymbol{T}_0^{-1}\boldsymbol{K}_c\boldsymbol{T}_0^{-T} = \boldsymbol{I}_n$ and $\boldsymbol{T}_0^T\boldsymbol{W}_o\boldsymbol{T}_0 = \boldsymbol{\Sigma}^2$ [21] where $\boldsymbol{V}$ and $\boldsymbol{\Sigma}$ are obtained by eigenvalue-eigenvector decomposition of $\boldsymbol{K}_c^{1/2}\boldsymbol{W}_o\boldsymbol{K}_c^{1/2}$ as

$$\boldsymbol{K}_c^{1/2}\boldsymbol{W}_o\boldsymbol{K}_c^{1/2} = \boldsymbol{V}\boldsymbol{\Sigma}^2\boldsymbol{V}^T \tag{53}$$

with $n \times n$ diagonal $\boldsymbol{\Sigma}^2$ and orthogonal $\boldsymbol{V}$ composed of the eigenvalues and eigenvectors of $\boldsymbol{K}_c^{1/2}\boldsymbol{W}_o\boldsymbol{K}_c^{1/2}$, respectively.

The above iteration process continues until

$$|S(\boldsymbol{P}_{k+1})_{ave} - S(\boldsymbol{P}_k)_{ave}| + |n - \mathrm{tr}[\boldsymbol{K}_c\boldsymbol{P}_{k+1}^{-1}]| < \varepsilon \tag{54}$$

is satisfied where $\varepsilon > 0$ is a prescribed tolerance. If the iteration is terminated at step k, then we take $\boldsymbol{P} = \boldsymbol{P}_k$ as the solution. Since $\boldsymbol{P} = \boldsymbol{T}\boldsymbol{T}^T$, the optimal $\boldsymbol{T}$ assumes the form

$$\boldsymbol{T} = \boldsymbol{P}^{\frac{1}{2}}\boldsymbol{U} \tag{55}$$

where $\boldsymbol{P}^{1/2}$ is the square root of matrix $\boldsymbol{P}$ obtained above, and $\boldsymbol{U}$ is an $n \times n$ orthogonal matrix.

As the second step of the solution procedure, we now turn our attention to the construction of the optimal coordinate transformation matrix $\boldsymbol{T}$ that solves the problem of minimizing the measure in (35) subject to the constraints in (38). To this end, we examine a procedure for determining the $n \times n$ orthogonal matrix $\boldsymbol{U}$ in (55) in order for the nonsingular matrix $\boldsymbol{T}$ to satisfy the l_2-scaling constraints in (38) without altering the optimal $\boldsymbol{P}$.

From (28) and (55), it follows that

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

In order to find an $n \times n$ orthogonal matrix $\boldsymbol{U}$ such that the matrix $\overline{\boldsymbol{K}}_c$ in (56) satisfies the scaling constraints in (38), we perform the eigenvalue-eigenvector decomposition for the positive-definite matrix $\boldsymbol{P}^{-1/2}\boldsymbol{K}_c\boldsymbol{P}^{-1/2}$ as

$$\boldsymbol{P}^{-\frac{1}{2}}\boldsymbol{K}_c\boldsymbol{P}^{-\frac{1}{2}} = \boldsymbol{R}\boldsymbol{\Theta}\boldsymbol{R}^T \tag{57}$$

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

$$(\boldsymbol{S}\boldsymbol{\Theta}\boldsymbol{S}^T)_{ii} = 1 \ \text{ for } \ i = 1, 2, \cdots, n \tag{58}$$

can be obtained by numerical manipulations [3] (p.278). By using (56)-(58), it can be readily verified that the orthogonal matrix $\boldsymbol{U} = \boldsymbol{R}\boldsymbol{S}^T$ leads to a $\overline{\boldsymbol{K}}_c$ in (56) whose diagonal elements are equal to unity, hence the constraints in (38) are now satisfied. This matrix $\boldsymbol{T}$ together with (55) gives the solution to the problem of minimizing (35) subject to the constraints in (38) as

$$\boldsymbol{T} = \boldsymbol{P}^{\frac{1}{2}}\boldsymbol{R}\boldsymbol{S}^T \tag{59}$$

B. Method 2: Using A Quasi-Newton Algorithm

Since the state-space model in (1) is stable and controllable, the controllability Grammian $\boldsymbol{K}_c$ is positive-definite. Hence $\boldsymbol{K}_c^{1/2}$ satisfying $\boldsymbol{K}_c = \boldsymbol{K}_c^{1/2}\boldsymbol{K}_c^{1/2}$ is also positive-definite.

By defining

$$\hat{\boldsymbol{T}} = \boldsymbol{T}^T\boldsymbol{K}_c^{-\frac{1}{2}} \tag{60}$$

the l_2-scaling constraints in (38) can be written as

$$(\hat{\boldsymbol{T}}^{-T}\hat{\boldsymbol{T}}^{-1})_{ii} = 1 \ \text{ for } \ i = 1, 2, \cdots, n \tag{61}$$

The constraints in (61) simply state that each column in

matrix $\hat{\boldsymbol{T}}^{-1}$ must be a unit vector. If matrix $\hat{\boldsymbol{T}}^{-1}$ is assumed to have the form

$$\hat{\boldsymbol{T}}^{-1} = \left[\frac{\boldsymbol{t}_1}{||\boldsymbol{t}_1||}, \frac{\boldsymbol{t}_2}{||\boldsymbol{t}_2||}, \cdots, \frac{\boldsymbol{t}_n}{||\boldsymbol{t}_n||}\right] \tag{62}$$

then (61) is always satisfied. Using (60), it follows from (35) that

$$S(\hat{\boldsymbol{T}})_{ave} = \mathrm{tr}[\hat{\boldsymbol{T}}\hat{\boldsymbol{N}}(\hat{\boldsymbol{T}})\hat{\boldsymbol{T}}^T] + \mathrm{tr}[\hat{\boldsymbol{T}}\tilde{\boldsymbol{W}}_o\hat{\boldsymbol{T}}^T] + n + \frac{L+1}{2} \tag{63a}$$

where

$$\begin{aligned}
&\mathrm{tr}[\hat{\boldsymbol{T}}\hat{\boldsymbol{N}}(\hat{\boldsymbol{T}})\hat{\boldsymbol{T}}^T] = \mathrm{tr}[\hat{\boldsymbol{T}}^{-T}\hat{\boldsymbol{M}}(\hat{\boldsymbol{T}})\hat{\boldsymbol{T}}^{-1}] \\
&[\hat{\boldsymbol{N}}(\hat{\boldsymbol{T}}),\ \hat{\boldsymbol{M}}(\hat{\boldsymbol{T}})] = \frac{1}{L}\sum_{i=1}^{L}\sum_{l=1}^{L}[\hat{\boldsymbol{N}}_{il}(\hat{\boldsymbol{T}}),\ \hat{\boldsymbol{M}}_{il}(\hat{\boldsymbol{T}})] \\
&\hat{\boldsymbol{N}}_{il}(\hat{\boldsymbol{T}}) = \boldsymbol{K}_c^{\frac{1}{2}}\boldsymbol{N}_{il}(\boldsymbol{K}_c^{\frac{1}{2}}\hat{\boldsymbol{T}}^T\hat{\boldsymbol{T}}\boldsymbol{K}_c^{\frac{1}{2}})\boldsymbol{K}_c^{\frac{1}{2}} \\
&\hat{\boldsymbol{M}}_{il}(\hat{\boldsymbol{T}}) = \boldsymbol{K}_c^{-\frac{1}{2}}\boldsymbol{M}_{il}(\boldsymbol{K}_c^{\frac{1}{2}}\hat{\boldsymbol{T}}^T\hat{\boldsymbol{T}}\boldsymbol{K}_c^{\frac{1}{2}})\boldsymbol{K}_c^{-\frac{1}{2}} \\
&\tilde{\boldsymbol{W}}_o = \boldsymbol{K}_c^{\frac{1}{2}}\boldsymbol{W}_o\boldsymbol{K}_c^{\frac{1}{2}}
\end{aligned} \tag{63b}$$

From the foregoing arguments, the problem of obtaining an $n \times n$ nonsingular matrix $\boldsymbol{T}$ which minimizes $S(\boldsymbol{T})_{ave}$ in (35) subject to the l_2-scaling constraints in (38) can be converted into an unconstrained optimization problem of obtaining an $n \times n$ nonsingular matrix $\hat{\boldsymbol{T}}$ which minimizes $S(\hat{\boldsymbol{T}})_{ave}$ in (63a).

We now apply a quasi-Newton (known as the Broyden-Fletcher-Goldfarb-Shanno (BFGS)) algorithm [22] to minimize (63a) with respect to $\hat{\boldsymbol{T}}$ in (62). Let $\boldsymbol{x}$ be the column vector that collects the independent variables in matrix $\hat{\boldsymbol{T}}$, i.e.,

$$\boldsymbol{x} = (\boldsymbol{t}_1^T, \boldsymbol{t}_2^T, \cdots, \boldsymbol{t}_n^T)^T \tag{64}$$

Then, $S(\hat{\boldsymbol{T}})_{ave}$ is a function of $\boldsymbol{x}$ and is denoted by $J_o(\boldsymbol{x})$. The algorithm starts with a trivial initial point $\boldsymbol{x}_0$ obtained from an initial assignment $\hat{\boldsymbol{T}} = \boldsymbol{I}_n$ (therefore $\boldsymbol{T} = \boldsymbol{K}_c^{1/2}$ in (60)). Then, in the kth iteration, a quasi-Newton algorithm updates the most recent point $\boldsymbol{x}_k$ to point $\boldsymbol{x}_{k+1}$ as [22]

$$\boldsymbol{x}_{k+1} = \boldsymbol{x}_k + \alpha_k \boldsymbol{d}_k \tag{65a}$$

where

$$\begin{aligned}
&\boldsymbol{d}_k = -\boldsymbol{S}_k\nabla J_o(\boldsymbol{x}_k), \quad \alpha_k = arg[\min_{\alpha}\ J_o(\boldsymbol{x}_k + \alpha\boldsymbol{d}_k)] \\
&\boldsymbol{S}_{k+1} = \boldsymbol{S}_k + \left(1 + \frac{\boldsymbol{\gamma}_k^T\boldsymbol{S}_k\boldsymbol{\gamma}_k}{\boldsymbol{\gamma}_k^T\boldsymbol{\delta}_k}\right)\frac{\boldsymbol{\delta}_k\boldsymbol{\delta}_k^T}{\boldsymbol{\gamma}_k^T\boldsymbol{\delta}_k} - \frac{\boldsymbol{\delta}_k\boldsymbol{\gamma}_k^T\boldsymbol{S}_k + \boldsymbol{S}_k\boldsymbol{\gamma}_k\boldsymbol{\delta}_k^T}{\boldsymbol{\gamma}_k^T\boldsymbol{\delta}_k} \\
&\boldsymbol{S}_0 = \boldsymbol{I}, \quad \boldsymbol{\delta}_k = \boldsymbol{x}_{k+1} - \boldsymbol{x}_k, \quad \boldsymbol{\gamma}_k = \nabla J_o(\boldsymbol{x}_{k+1}) - \nabla J_o(\boldsymbol{x}_k)
\end{aligned} \tag{65b}$$

In the above, $\nabla J_o(\boldsymbol{x})$ is the gradient of $J_o(\boldsymbol{x})$ with respect to $\boldsymbol{x}$, and $\boldsymbol{S}_k$ is a positive-definite approximation of the inverse Hessian matrix of $J_o(\boldsymbol{x})$. Outlined below is an algorithm to compute an approximate local minimizer α_k for $J_o(\boldsymbol{x}_k + \alpha\boldsymbol{d}_k)$. In Section 4, $\alpha_{init} = 2^{-25}$ and $N = 30$ were chosen.

Algorithm 1 Search $\alpha_k = arg[\min_{\alpha}\ J_o(\boldsymbol{x}_k + \alpha\boldsymbol{d}_k)]$

```
α_k ← 0
α ← α_init                          ▷ α_init is a small positive value
J_o,min ← J_o(x_k)
for n = 1 ... N do
    if J_o,min > J_o(x_k + α d_k) then
        J_o,min ← J_o(x_k + α d_k)
        α_k ← α
    end if
    α ← 2α
end for
Δ_α ← α_k/2
for n = 1 ... N do
    if J_o,min > J_o(x_k + (α_k + Δ_α) d_k) then
        J_o,min ← J_o(x_k + (α_k + Δ_α) d_k)
        α_k ← α_k + Δ_α
    else if J_o,min > J_o(x_k + (α_k − Δ_α) d_k) then
        J_o,min ← J_o(x_k + (α_k − Δ_α) d_k)
        α_k ← α_k − Δ_α
    end if
    Δ_α ← Δ_α/2
end for
```

This iteration process continues until

$$|J_o(\boldsymbol{x}_{k+1}) - J_o(\boldsymbol{x}_k)| < \varepsilon \tag{66}$$

is satisfied where $\varepsilon > 0$ is a prescribed tolerance. If the iteration is terminated at step k, then $\boldsymbol{x}_k$ is taken to be the solution to the minimization problem.

The implementation of (65a) requires the computation of $\nabla J_o(\boldsymbol{x})$ in each iteration, and hence the availability of a closed-form formulation to compute $\nabla J_o(\boldsymbol{x})$ will make the algorithm considerably faster relative to the evaluation of $\nabla J_o(\boldsymbol{x})$ based on numerical differentiation.

Each term of the objective function in (63a) has the form $J(\boldsymbol{x}) = \mathrm{tr}[\hat{\boldsymbol{T}}\boldsymbol{N}\hat{\boldsymbol{T}}^T]$ (or $J(\boldsymbol{x}) = \mathrm{tr}[\hat{\boldsymbol{T}}^{-T}\boldsymbol{M}\hat{\boldsymbol{T}}^{-1}]$) which, in the light of (62), can be expressed as

$$\begin{aligned}
J(\boldsymbol{x}) = \mathrm{tr}\Big[&\Big[\frac{\boldsymbol{t}_1}{||\boldsymbol{t}_1||}, \frac{\boldsymbol{t}_2}{||\boldsymbol{t}_2||}, \cdots, \frac{\boldsymbol{t}_n}{||\boldsymbol{t}_n||}\Big]^{-1} \\
&\cdot \boldsymbol{N}\Big[\frac{\boldsymbol{t}_1}{||\boldsymbol{t}_1||}, \frac{\boldsymbol{t}_2}{||\boldsymbol{t}_2||}, \cdots, \frac{\boldsymbol{t}_n}{||\boldsymbol{t}_n||}\Big]^{-T}\Big]
\end{aligned} \tag{67}$$

To compute $\partial J(\boldsymbol{x})/\partial t_{pq}$, we perturb the pth component of vector $\boldsymbol{t}_q$ by a small amount, say Δ, and keep the rest of $\hat{\boldsymbol{T}}$ unchanged. If we denote the perturbed qth column of $\hat{\boldsymbol{T}}^{-1}$ by $\tilde{\boldsymbol{t}}_q/||\tilde{\boldsymbol{t}}_q||$, then a linear approximation of $\tilde{\boldsymbol{t}}_q/||\tilde{\boldsymbol{t}}_q||$ can be obtained as

$$\frac{\tilde{\boldsymbol{t}}_q}{||\tilde{\boldsymbol{t}}_q||} \simeq \frac{\boldsymbol{t}_q}{||\boldsymbol{t}_q||} + \Delta\,\partial\Big\{\frac{\boldsymbol{t}_q}{||\boldsymbol{t}_q||}\Big\}/\partial t_{pq} = \frac{\boldsymbol{t}_q}{||\boldsymbol{t}_q||} - \Delta\,\boldsymbol{g}_{pq} \tag{68a}$$

where

$$\boldsymbol{g}_{pq} = -\partial\Big\{\frac{\boldsymbol{t}_q}{||\boldsymbol{t}_q||}\Big\}/\partial t_{pq} = \frac{1}{||\boldsymbol{t}_q||^3}(t_{pq}\boldsymbol{t}_q - ||\boldsymbol{t}_q||^2\boldsymbol{e}_p) \tag{68b}$$

Now let $\hat{\boldsymbol{T}}_{pq}$ be the matrix obtained from $\hat{\boldsymbol{T}}$ with a perturbed (p, q)th component, then we obtain

$$\hat{\boldsymbol{T}}_{pq}^{-1} = \hat{\boldsymbol{T}}^{-1} - \Delta \boldsymbol{g}_{pq} \boldsymbol{e}_q^T \tag{69}$$

and the matrix inversion formula [23] (p. 655) gives

$$\hat{\boldsymbol{T}}_{pq} = \hat{\boldsymbol{T}} + \frac{\Delta \hat{\boldsymbol{T}} \boldsymbol{g}_{pq} \boldsymbol{e}_q^T \hat{\boldsymbol{T}}}{1 - \Delta \boldsymbol{e}_q^T \hat{\boldsymbol{T}} \boldsymbol{g}_{pq}} \simeq \hat{\boldsymbol{T}} + \Delta \hat{\boldsymbol{T}} \boldsymbol{g}_{pq} \boldsymbol{e}_q^T \hat{\boldsymbol{T}} \tag{70}$$

For convenience, we define $\hat{\boldsymbol{T}}_{pq} = \hat{\boldsymbol{T}} + \Delta \boldsymbol{S}$ and write

$$\hat{\boldsymbol{T}}_{pq} \boldsymbol{N} \hat{\boldsymbol{T}}_{pq}^T = \hat{\boldsymbol{T}} \boldsymbol{N} \hat{\boldsymbol{T}}^T + \Delta \boldsymbol{S} \boldsymbol{N} \hat{\boldsymbol{T}}^T + \Delta \hat{\boldsymbol{T}} \boldsymbol{N} \boldsymbol{S}^T + \Delta^2 \boldsymbol{S} \boldsymbol{N} \boldsymbol{S}^T \tag{71}$$

which implies that

$$\mathrm{tr}[\hat{\boldsymbol{T}}_{pq} \boldsymbol{N} \hat{\boldsymbol{T}}_{pq}^T] - \mathrm{tr}[\hat{\boldsymbol{T}} \boldsymbol{N} \hat{\boldsymbol{T}}^T] \simeq \Delta \, \mathrm{tr}[\boldsymbol{S}(\boldsymbol{N} + \boldsymbol{N}^T) \hat{\boldsymbol{T}}^T] \tag{72}$$

provided that Δ is sufficiently small. Hence

$$\begin{aligned} \frac{\partial \mathrm{tr}[\hat{\boldsymbol{T}} \boldsymbol{N} \hat{\boldsymbol{T}}^T]}{\partial t_{pq}} &= \lim_{\Delta \to 0} \frac{\mathrm{tr}[\hat{\boldsymbol{T}}_{pq} \boldsymbol{N} \hat{\boldsymbol{T}}_{pq}^T] - \mathrm{tr}[\hat{\boldsymbol{T}} \boldsymbol{N} \hat{\boldsymbol{T}}^T]}{\Delta} \\ &= \mathrm{tr}[\boldsymbol{S}(\boldsymbol{N} + \boldsymbol{N}^T) \hat{\boldsymbol{T}}^T] \end{aligned} \tag{73}$$

Similarly, if we define $\hat{\boldsymbol{T}}_{pq}^{-1} = \hat{\boldsymbol{T}}^{-1} - \Delta \boldsymbol{S}$, then we arrive at

$$\begin{aligned} \frac{\partial \mathrm{tr}[\hat{\boldsymbol{T}}^{-T} \boldsymbol{M} \hat{\boldsymbol{T}}^{-1}]}{\partial t_{pq}} &= \lim_{\Delta \to 0} \frac{\mathrm{tr}[\hat{\boldsymbol{T}}_{pq}^{-T} \boldsymbol{M} \hat{\boldsymbol{T}}_{pq}^{-1}] - \mathrm{tr}[\hat{\boldsymbol{T}}^{-T} \boldsymbol{M} \hat{\boldsymbol{T}}^{-1}]}{\Delta} \\ &= -\mathrm{tr}[\boldsymbol{S}^T (\boldsymbol{M} + \boldsymbol{M}^T) \hat{\boldsymbol{T}}^{-1}] \end{aligned} \tag{74}$$

Referring to (73), it follows from (63a) and (70) that

$$\begin{aligned} \frac{\partial \mathrm{tr}[\hat{\boldsymbol{T}} \hat{\boldsymbol{N}}(\hat{\boldsymbol{T}}) \hat{\boldsymbol{T}}^T]}{\partial t_{pq}} &= 2 \, \mathrm{tr}[\hat{\boldsymbol{T}} \boldsymbol{g}_{pq} \boldsymbol{e}_q^T \hat{\boldsymbol{T}} \hat{\boldsymbol{N}}(\hat{\boldsymbol{T}}) \hat{\boldsymbol{T}}^T] \\ &= 2 \, \boldsymbol{e}_q^T \hat{\boldsymbol{T}} \hat{\boldsymbol{N}}(\hat{\boldsymbol{T}}) \hat{\boldsymbol{T}}^T \hat{\boldsymbol{T}} \boldsymbol{g}_{pq} \\ \frac{\partial \mathrm{tr}[\hat{\boldsymbol{T}} \tilde{\boldsymbol{W}}_o \hat{\boldsymbol{T}}^T]}{\partial t_{pq}} &= 2 \, \mathrm{tr}[\hat{\boldsymbol{T}} \boldsymbol{g}_{pq} \boldsymbol{e}_q^T \hat{\boldsymbol{T}} \tilde{\boldsymbol{W}}_o \hat{\boldsymbol{T}}^T] \\ &= 2 \, \boldsymbol{e}_q^T \hat{\boldsymbol{T}} \tilde{\boldsymbol{W}}_o \hat{\boldsymbol{T}}^T \hat{\boldsymbol{T}} \boldsymbol{g}_{pq} \end{aligned} \tag{75}$$

Referring to (74), it follows from (63a) and (69) that

$$\begin{aligned} &\frac{\partial \mathrm{tr}[\hat{\boldsymbol{T}}^{-T} \hat{\boldsymbol{M}}(\hat{\boldsymbol{T}}) \hat{\boldsymbol{T}}^{-1}]}{\partial t_{pq}} = -2 \, \mathrm{tr}[\boldsymbol{e}_q \boldsymbol{g}_{pq}^T \hat{\boldsymbol{M}}(\hat{\boldsymbol{T}}) \hat{\boldsymbol{T}}^{-1}] \\ &= -2 \, \boldsymbol{g}_{pq}^T \hat{\boldsymbol{M}}(\hat{\boldsymbol{T}}) \hat{\boldsymbol{T}}^{-1} \boldsymbol{e}_q = -2 \, \boldsymbol{e}_q^T \hat{\boldsymbol{T}}^{-T} \hat{\boldsymbol{M}}(\hat{\boldsymbol{T}}) \boldsymbol{g}_{pq} \end{aligned} \tag{76}$$

The gradient of $J_o(\boldsymbol{x})$ with respect to t_{pq} can be evaluated using closed-form expressions as

$$\begin{aligned} \frac{\partial J_o(\boldsymbol{x})}{\partial t_{pq}} &= \lim_{\Delta \to 0} \frac{S(\hat{\boldsymbol{T}}_{pq})_{ave} - S(\hat{\boldsymbol{T}})_{ave}}{\Delta} \\ &= 2(\beta_1 - \beta_2 + \beta_3) \end{aligned} \tag{77a}$$

where

$$\begin{aligned} \beta_1 &= \boldsymbol{e}_q^T \hat{\boldsymbol{T}} \hat{\boldsymbol{N}}(\hat{\boldsymbol{T}}) \hat{\boldsymbol{T}}^T \hat{\boldsymbol{T}} \boldsymbol{g}_{pq} \\ \beta_2 &= \boldsymbol{e}_q^T \hat{\boldsymbol{T}}^{-T} \hat{\boldsymbol{M}}(\hat{\boldsymbol{T}}) \boldsymbol{g}_{pq}, \quad \beta_3 = \boldsymbol{e}_q^T \hat{\boldsymbol{T}} \tilde{\boldsymbol{W}}_o \hat{\boldsymbol{T}}^T \hat{\boldsymbol{T}} \boldsymbol{g}_{pq} \end{aligned} \tag{77b}$$

4. A Case Study

Consider the 4th-order Butterworth low-pass digital filter $(\boldsymbol{A}_o, \boldsymbol{b}_o, \boldsymbol{c}_o, d)_4$ with a narrow normalized passband of 0.05, described by

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

$$\boldsymbol{c}_o = \begin{bmatrix} 1 & 0 & 0 & 0 \end{bmatrix}, \quad d = 3.123898 \times 10^{-5}$$

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

Applying a coordinate transformation defined by

$$\boldsymbol{T}_s = \mathrm{diag}\{\, 0.226458,\ 0.588059,\ 0.513017,\ 0.150144 \,\}$$

to the above state-space model, we obtained an equivalent state-space model $(\boldsymbol{A}, \boldsymbol{b}, \boldsymbol{c}, d)_4$ that satisfies l_2-scaling constraints as

$$\boldsymbol{A} = \begin{bmatrix} 3.589734 & 2.596768 & 0 & 0 \\ -1.868197 & 0 & 0.872390 & 0 \\ 1.290747 & 0 & 0 & 0.292669 \\ -1.000000 & 0 & 0 & 0 \end{bmatrix}$$

$$\boldsymbol{b} = 10^{-3} \begin{bmatrix} 1.046973 & 0.061023 & 0.421624 & 0.070114 \end{bmatrix}^T$$

$$\boldsymbol{c} = \begin{bmatrix} 0.226458 & 0 & 0 & 0 \end{bmatrix}, \quad d = 3.123898 \times 10^{-5}$$

where $\boldsymbol{A} = \boldsymbol{T}_s^{-1} \boldsymbol{A}_o \boldsymbol{T}_s$, $\boldsymbol{b} = \boldsymbol{T}_s^{-1} \boldsymbol{b}_o$, $\boldsymbol{c} = \boldsymbol{c}_o \boldsymbol{T}_s$ and the eigenvalues of $\boldsymbol{A}$ have the magnitudes 0.941824 and 0.864550. By solving the Lyapunov equations in (22c), the controllability and observability Grammians $\boldsymbol{K}_c$ and $\boldsymbol{W}_o$ were computed as

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

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

Since $\sqrt{2}n = 5.656854$ for the filter order $n = 4$, the blocklength is set to $L = 6$ in this section. The block-state realization was then constructed from (2b) as

$$\hat{\boldsymbol{A}} = \begin{bmatrix} 42.27560 & 82.60016 & 50.68938 & 9.51583 \\ -37.89405 & -71.90962 & -42.32252 & -7.44160 \\ 34.50854 & 64.08980 & 36.64013 & 6.17955 \\ -31.80883 & -58.10401 & -32.51401 & -5.32723 \end{bmatrix}$$

$$\hat{\boldsymbol{B}} = 10^{-3} \begin{bmatrix} 50.9309 & 33.5423 & 19,7751 & \\ -42.9331 & -26.7551 & -14.4198 & \\ 37.5070 & 22.6166 & 11.6789 & \\ -33.5423 & -19.7751 & -9.93632 & \end{bmatrix}$$

$$\begin{bmatrix} 9.93632 & 3.91682 & 1.04697 \\ -6.12056 & -1.58813 & 0.06102 \\ 4.74920 & 1.37190 & 0.42162 \\ -3.91682 & -1.04697 & 0.07011 \end{bmatrix}$$

$$\hat{\boldsymbol{C}} = \begin{bmatrix} 0.226458 & 0 & 0 & 0 \\ 0.812924 & 0.588059 & 0 & 0 \\ 1.819571 & 2.110976 & 0.513017 & 0 \\ 3.250232 & 4.725005 & 1.841595 & 0.150144 \\ 5.067114 & 8.440099 & 4.122049 & 0.538977 \\ 7.203366 & 13.158123 & 7.363061 & 1.206395 \end{bmatrix}$$

$$\begin{bmatrix} \hat{d}_{11} & \hat{d}_{21} & \hat{d}_{31} \\ \hat{d}_{41} & \hat{d}_{51} & \hat{d}_{61} \end{bmatrix} = 10^{-3} \begin{bmatrix} 0.031239 & 0.237096 & 0.886995 \\ 2.250160 & 4.478225 & 7.595914 \end{bmatrix}$$

where the eigenvalues of $\hat{\boldsymbol{A}}$ have the magnitudes 0.697942 and 0.417582.

The average l_2-sensitivity measure for the block-state realization described by (11) was then computed from (25) as

$$S_{ave} = 40.933372 \times 10^4 \quad (27.330613 \times 10^4)$$

which was compared with the l_2-sensitivity of the state-space model, shown in (23) as

$$S_o = 977.917589 \times 10^4 \quad (488.229633 \times 10^4)$$

where the values in the round brackets are the sensitivities that have taken the trivial elements with 0 and ±1 values in the coefficient matrices into account [24].

A) The Use of A Lagrange Function

The recursive matrix equation in (50) together with (52) was applied to minimize (44) with tolerance $\varepsilon = 10^{-8}$ in (54) and initial estimate $\boldsymbol{P}_0 = \boldsymbol{K}_c$. It took the algorithm 76 iterations to converge to the solution

$$\boldsymbol{P} = \begin{bmatrix} 0.875338 & -0.904805 & 0.932402 & -0.958480 \\ -0.904805 & 0.938156 & -0.969673 & 0.999738 \\ 0.932402 & -0.969673 & 1.005245 & -1.039528 \\ -0.958480 & 0.999738 & -1.039528 & 1.078311 \end{bmatrix}$$

and from (59)

$$\boldsymbol{T} = \begin{bmatrix} 0.327836 & 0.559715 & 0.663273 & 0.121038 \\ -0.313941 & -0.586064 & -0.699777 & -0.080238 \\ 0.301577 & 0.604055 & 0.740336 & 0.036271 \\ -0.294198 & -0.614717 & -0.783389 & 0.013515 \end{bmatrix}$$

was obtained. A new block-state realization was obtained from (27) and (30) as

$$\hat{\boldsymbol{A}}_{new} = \begin{bmatrix} 0.397376 & -0.473485 & 0.232855 & 0.220533 \\ 0.537510 & 0.388153 & -0.130919 & 0.201674 \\ -0.028026 & 0.015221 & 0.446037 & 0.327314 \\ 0.021347 & -0.406390 & 0.101068 & 0.447310 \end{bmatrix}$$

$$\hat{\boldsymbol{B}}_{new} = \begin{bmatrix} 0.040743 & -0.020980 & -0.090083 \\ -0.268273 & -0.280345 & -0.281986 \\ 0.241915 & 0.256942 & 0.271631 \\ 0.225341 & 0.222337 & 0.222858 \end{bmatrix}$$

$$\begin{bmatrix} -0.166178 & -0.248790 & -0.337393 \\ -0.271737 & -0.248370 & -0.210986 \\ 0.284592 & 0.293878 & 0.296858 \\ 0.229256 & 0.244342 & 0.271408 \end{bmatrix}$$

$$\hat{\boldsymbol{C}}_{new} = \begin{bmatrix} 0.074241 & 0.126752 & 0.150203 & 0.027410 \\ 0.081890 & 0.110365 & 0.127680 & 0.051210 \\ 0.088513 & 0.091165 & 0.109465 & 0.069463 \\ 0.093382 & 0.070177 & 0.095119 & 0.083100 \\ 0.096040 & 0.048327 & 0.084164 & 0.092889 \\ 0.096263 & 0.026433 & 0.076108 & 0.099465 \end{bmatrix}$$

and the new controllability Grammian was found to be

$$\overline{\boldsymbol{K}}_c = \begin{bmatrix} 1.000000 & 0.300486 & 0.342142 & 0.474457 \\ 0.300486 & 1.000000 & -0.342142 & -0.474457 \\ 0.342142 & -0.342142 & 1.000000 & 0.910932 \\ 0.474457 & -0.474457 & 0.910932 & 1.000000 \end{bmatrix}$$

The average l_2-sensitivity for the block-state realization described by (11) was then computed from (35) as

$$S(\boldsymbol{T})_{ave} = 8.759262$$

which was compared with the l_2-sensitivity of a new state-space model $(\overline{\boldsymbol{A}}, \overline{\boldsymbol{b}}, \overline{\boldsymbol{c}}, d)_n$, shown in (39) as

$$S_o(\boldsymbol{T}) = 29.500327$$

The profiles of $S(\boldsymbol{P})_{ave}$ in (41) and $\mathrm{tr}[\boldsymbol{K}_c\boldsymbol{P}^{-1}]$ during the first 76 iterations of the algorithm are depicted in Fig. 2 and Fig. 3, respectively.

B) The Use of A Quasi-Newton Algorithm

The quasi-Newton algorithm was applied to minimize (63a) with initial assignment $\hat{\boldsymbol{T}} = \boldsymbol{I}_4$, and tolerance $\varepsilon = 10^{-8}$ in (66). It took the algorithm 22 iterations to converge to the solution

$$\hat{\boldsymbol{T}} = \begin{bmatrix} 0.571134 & -0.415538 & -1.326551 & 0.389172 \\ 0.407792 & 0.933965 & -1.087187 & -2.120849 \\ 0.043983 & -0.117207 & 3.040058 & 2.480486 \\ -0.373909 & 0.428710 & 1.758732 & 2.561697 \end{bmatrix}$$

which is equivalent to

$$\boldsymbol{T} = \begin{bmatrix} -0.307934 & 0.205102 & 0.403183 & -0.758913 \\ 0.350071 & -0.233366 & -0.379012 & 0.785847 \\ -0.385730 & 0.267285 & 0.350018 & -0.813980 \\ 0.414649 & -0.303751 & -0.313832 & 0.845979 \end{bmatrix}$$

where $\boldsymbol{P} = \boldsymbol{T}\boldsymbol{T}^T$ was computed as

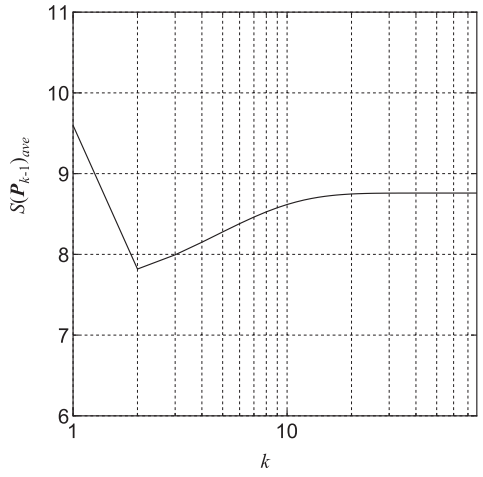


Fig. 2 Profile of $S(\boldsymbol{P})_{ave}$ during the first 76 iterations.

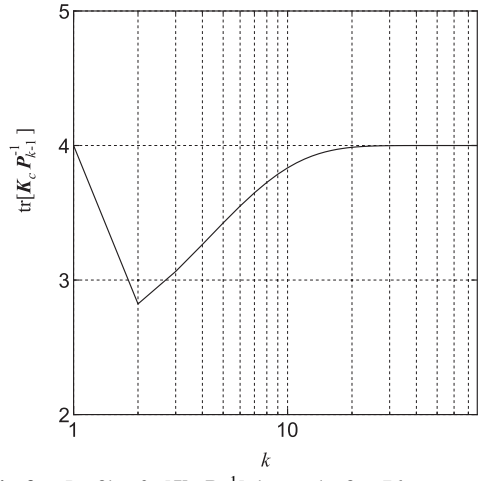


Fig. 3 Profile of $\mathrm{tr}[\boldsymbol{K}_c \boldsymbol{P}^{-1}]$ during the first 76 iterations.

$$\boldsymbol{P} = \begin{bmatrix} 0.875395 & -0.904863 & 0.932461 & -0.958540 \\ -0.904863 & 0.938216 & -0.969734 & 0.999798 \\ 0.932461 & -0.969734 & 1.005306 & -1.039588 \\ -0.958540 & 0.999798 & -1.039588 & 1.078370 \end{bmatrix}$$

which is practically identical to matrix $\boldsymbol{P}$ in Section IV-A.

A new block-state realization was then obtained from (27) and (30) as

$$\hat{\boldsymbol{A}}_{new} = \begin{bmatrix} 0.426110 & 0.516400 & -0.130405 & 0.290374 \\ -0.269864 & 0.389027 & 0.160428 & 0.110379 \\ 0.450255 & 0.105445 & 0.449322 & -0.002678 \\ -0.390308 & -0.117906 & -0.316171 & 0.414417 \end{bmatrix}$$

$$\hat{\boldsymbol{B}}_{new} = \begin{bmatrix} 0.285271 & 0.262568 & 0.228614 \\ 0.268791 & 0.312840 & 0.356168 \\ 0.170038 & 0.164974 & 0.166233 \\ -0.019883 & 0.021455 & 0.065752 \end{bmatrix}$$

$$\begin{bmatrix} 0.183139 & 0.126272 & 0.058614 \\ 0.396886 & 0.432672 & 0.460722 \\ 0.176240 & 0.197682 & 0.233484 \\ 0.113488 & 0.165557 & 0.223393 \end{bmatrix}$$

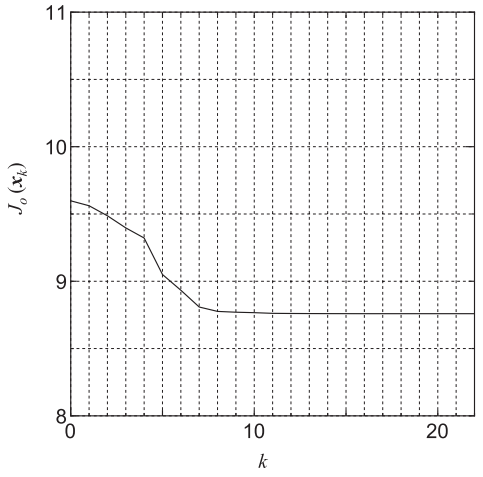


Fig. 4 Profile of $J_o(\boldsymbol{x})$ during the first 22 iterations.

$$\hat{\boldsymbol{C}}_{new} = \begin{bmatrix} -0.069734 & 0.046447 & 0.091304 & -0.171862 \\ -0.044464 & 0.029499 & 0.104875 & -0.154814 \\ -0.019201 & 0.017688 & 0.113099 & -0.139578 \\ 0.005132 & 0.010595 & 0.117075 & -0.125515 \\ 0.027790 & 0.007686 & 0.117717 & -0.112175 \\ 0.048199 & 0.008352 & 0.115785 & -0.099257 \end{bmatrix}$$

Moreover, the new controllability Grammian was found to be

$$\overline{\boldsymbol{K}}_c = \begin{bmatrix} 1.000000 & 0.609669 & 0.656806 & -0.159006 \\ 0.609669 & 1.000000 & 0.530787 & 0.197938 \\ 0.656806 & 0.530787 & 1.000000 & -0.674646 \\ -0.159006 & 0.197938 & -0.674646 & 1.000000 \end{bmatrix}$$

and the average l_2-sensitivity measure for the block-state realization described by (11) was computed from (35) as

$$S(\boldsymbol{T})_{ave} = 8.759262$$

which was compared with the l_2-sensitivity of a new state-space model $(\overline{\boldsymbol{A}}, \overline{\boldsymbol{b}}, \overline{\boldsymbol{c}}, d)_n$, shown in (39) as

$$S_o(\boldsymbol{T}) = 29.500303$$

The profile of the average l_2-sensitivity measure $J_o(\boldsymbol{x})$ during the first 22 iterations of the algorithm is depicted in Fig. 4.

We see that as far as the digital filter examined in this example is concerned, the two algorithms offer practically the same performance in terms of l_2-sensitivity minimization. Concerning the computational complexity, the 76 iterations required by the constrained algorithm consumed 4.09375 seconds of CPU time versus 41.12500 seconds for the unconstrained algorithm to finish its 22 iterations on a PC with a 3.3GHz i5-3550 CPU and 8GB memory. Obviously, the constrained method is considerably more efficient than its unconstrained counterpart. However, for digital filter of large scales, it is quite likely that the algorithms' complexity will change due to the fact that the constrained algorithm heavily involves expensive matrix manipulations such as computing square root as well as inverse of matrices. For this reason,

we believe that together the two algorithms can serve for l_2-sensitivity minimization for a wide range of filter scales.

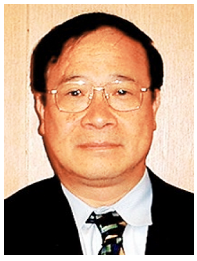
5. Conclusion

The l_2-sensitivity in the block-state realization derived from a given SISO state-space model have been analyzed. The problem of minimizing the average l_2-sensitivity measure subject to the l_2-norm dynamic-range scaling constraints has then been solved by two techniques. One has been based on a Lagrange function, while the other has relied on an efficient quasi-Newton algorithm. Finally, a numerical example has been presented to demonstrate the validity and effectiveness of the proposed techniques.

References

[1] S.Y. Hwang, "Roundoff noise in state-space digital filtering: A general analysis," IEEE Trans. Acoust., Speech, Signal Process., vol.ASSP-24, no.3, 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.CAS-23, no.9, 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.ASSP-25, no.4, pp.273–281, Aug. 1977.
[4] C.W. Barnes and S. Shinnaka, "Finite word effects in block-state realization of fixed-point digital filters," IEEE Trans. Circuits Syst., vol.CAS-27, no.5, pp.345–349, May 1980.
[5] C.W. Barnes and S. Shinnaka, "Block-shift invariance and block implementation of discrete-time filters," IEEE Trans. Circuits Syst., vol.CAS-27, no.8, pp.667–672, Aug. 1980.
[6] J. Zeman and A.G. Lindgren, "Fast digital filters with low roundoff noise," IEEE Trans. Circuits Syst., vol.CAS-28, no.7, pp.716–723, July 1981.
[7] L. Thiele, "Design of sensitivity and round-off noise optimal state-space discrete systems," Int. J. Circuit Theory Appl., vol.12, no.1, pp.39–46, Jan. 1984.
[8] V. Tavsanoglu and L. Thiele, "Optimal design of state-space digital filters by simultaneous minimization of sensitivity and roundoff noise," IEEE Trans. Circuits Syst., vol.CAS-31, no.10, pp.884–888, Oct. 1984.
[9] L. Thiele, "On the sensitivity of linear state-space systems," IEEE Trans. Circuits Syst., vol.CAS-33, no.5, pp.502–510, May 1986.
[10] 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.CAS-37, no.1, pp.72–80, Jan. 1989.
[11] 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, Fundam. Theory Appl., vol.39, no.5, pp.365–377, May 1992.
[12] W.-Y. Yan and J.B. Moore, "On L^2-sensitivity minimization of linear state-space systems," IEEE Trans. Circuits Syst. I, Fundam. Theory Appl., vol.39, no.8, pp.641–648, Aug. 1992.
[13] G. Li and M. Gevers, "Optimal synthetic FWL design of state-space digital filters," Proc. ICASSP 1992, vol.4, pp.429–432, 1992.
[14] M. Gevers and G. Li, Parameterizations in Control, Estimation and Filtering Problems: Accuracy Aspects, Springer-Verlag, New York, 1993.
[15] 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, Fundam. Theory Appl., vol.49, no.9, pp.1279–1289, Sept. 2002.
[16] T. Hinamoto, H. Ohnishi, and W.-S. Lu, "Minimization of L_2-sensitivity for state-space digital filters subject to L_2-dynamic-range scaling constraints," IEEE Trans. Circuits Syst.-II, Exp. Briefs, vol.52, no.10, pp.641–645, Oct. 2005.
[17] T. Hinamoto, K. Iwata, and W.-S. Lu, "L_2-sensitivity minimization of one- and two-dimensional state-space digital filters subject to L_2-scaling constraints," IEEE Trans. Signal Process., vol.54, no.5, pp.1804–1812, May 2006.
[18] Y. Tsunekawa, J. Shida, M. Kawamata, and T. Higuchi, "Coefficient sensitivity analysis for block-state realizations of state-space digital filters," IEICE Trans. Fundamentals (Japanese edition), vol.J74-A, no.7, pp.996–1005, July 1991, also Electronics and Communications in Japan, part 3, vol.75, no.3, pp.13–25, 1992.
[19] Y. Tsunekawa, J. Shida, M. Kawamata, and T. Higuchi, "Minimization of coefficient sensitivity for block-state realizations of state-space digital filters," IEICE Trans. Fundamentals (Japanese edition), vol.J75-A, no.3, pp.506–515, March 1992, also Electronics and Communications in Japan, part 3, vol.75, no.11, pp.88–101, 1992.
[20] T. Hinamoto and W.-S. Lu, Digital Filter Design and Realization, River Publishers, Denmark/The Netherlands, 2017.
[21] L.M. Silverman, "Optimal approximation of linear systems," Proc. Joint Automat. Contr. Conf., S.F., FA8-A, 1980.
[22] R. Fletcher, Practical Methods of Optimization, 2nd ed., Wiley, New York, 1987.
[23] T. Kailath, Linear System, Prentice-Hall, Englewood Cliffs, N.J., 1980.
[24] C. Xiao, "Improved L_2-sensitivity for state-space digital systems," IEEE Trans. Signal Process., vol.45, no.4, pp.837–840, April 1997.

Akimitsu Doi received the B.E. degree in Electronic Engineering from Tottori University in 1992, and M.E. and Ph.D. degrees in System Engineering from Hiroshima University in 1994 and 1997, respectively. From 1997 to 2002, he was a Research Associate with the Faculty of Engineering, Hiroshima University, Hiroshima, Japan. From 2002 to 2005, he was a Lecturer in the Faculty of Engineering, Hiroshima Institute of Technology, Hiroshima, Japan. From 2005 to 2015, he was an Associate Professor and since October 2015, he has been a Professor at the same Institute. His research interests are in the area of digital signal and image processing.



Takao Hinamoto received the B.E. degree from Okayama University, Okayama, Japan, in 1969, the M.E. degree from Kobe University, Kobe, Japan, in 1971, and the Dr. Eng. degree from Osaka University, Osaka, Japan, in 1977, all in electrical engineering. From 1972 to 1988, he was with the Faculty of Engineering, Kobe University, Kobe, Japan. From 1979 to 1981, he was a Visiting Member of Staff in the Department of Electrical Engineering, Queen's University, Kingston, ON, Canada, on leave from Kobe University. During 1988–1991, he was Professor of Electronic Circuits in the Faculty of Engineering, Tottori University, Tottori, Japan. During 1992–2009, he was Professor of Electronic Control in the Department of Electrical Engineering, Hiroshima University, Hiroshima, Japan. Since 2009, he has been a Professor Emeritus of Hiroshima University. His research interests include digital signal processing, system theory, and control engineering. He has published over 400 papers in these areas. He is co-editor and co-author of Two-Dimensional Signal and Image Processing (Tokyo, Japan: SICE, 1996). Dr. Hinamoto is a Life Fellow of the IEEE and a Fellow of the Society of Instrument and Control Engineers (SICE).

Wu-Sheng Lu received his undergraduate degree in mathematics from Fudan University, Shanghai, China, in 1964, the M.S. degree in electrical engineering, and the Ph.D. degree in control science from University of Minnesota, Minneapolis, USA, in 1983 and 1984, respectively. He was a post-doctoral fellow at University of Victoria, Victoria, B.C., Canada, in 1985, and a visiting assistant professor at University of Minnesota from January 1986 to April 1987. Since May 1987, he has been with the Electrical and Computer Engineering Department, University of Victoria where he is a professor. His current research interests include analysis and design of digital filters, signal and image processing, and methods of convex optimization. He is the co-author with A. Antoniou of Two-Dimensional Digital Filters (Marcel Dekker, 1992) and Practical Optimization: Algorithms and Engineering Applications (Springer, 2007). Dr. Lu served as editor for the Canadian Journal of Electrical and Computer Engineering and associate editor for several journals including IEEE Transactions on Circuits and Systems I and II, International Journal of Multidimensional Systems and Signal Processing, and Journal of Circuits, Systems, and Signal Processing. Dr. Lu is a Life Fellow of the IEEE, a Fellow of the Engineering Institute of Canada, and a registered professional engineer in British Columbia, Canada.