Realization of 3-D Separable-Denominator Digital Filters with Low l_2-Sensitivity

Takao Hinamoto, *Life Fellow, IEEE,* Akimitsu Doi and Wu-Sheng Lu, *Fellow, IEEE*

***Abstract*— Three-dimensional (3-D) digital filters find applications in a variety of image and video signal processing problems. This paper presents a coefficient-sensitivity analysis for a wide class of 3-D digital filters with separable denominators in local state space that leads to an analytic formulation for sensitivity minimization, and to present two solution techniques for the sensitivity minimization problem at hand. To this end, a vector-matrix-vector decomposition of a given 3-D transfer function that separates the three variables and leads to a state-space realization in a form convenient for subsequent analysis. An l_2-sensitivity analysis is then performed, the result is a computationally tractable formula of the overall l_2-sensitivity for 3-D digital filters. The l_2-sensitivity is minimized subject to l_2-scaling constraints by using one of the two solution methods proposed – one relaxes the constraints into a single trace constraint and solves the relaxed problem with an effective matrix iteration scheme; while the other converts the contained optimization problem at hand into an unconstrained problem and solves it using a quasi-Newton algorithm. A case study is presented to illustrate the validity and effectiveness of the proposed techniques.**

***Index Terms*— 3-D separable-denominator digital filters, minimal state-space realization, l_2-sensitivity analysis, low l_2-sensitivity, l_2-scaling constraints, no overflow oscillations, Lagrange function, bisection method, quasi-Newton method**

I. INTRODUCTION

It is of practical significance in many applications to construct a filter structure so that the coefficient sensitivity of a digital filter is minimum or nearly minimum in a certain sense. Due to finite-word-length (FWL) effects caused by coefficient truncation or rounding, poor sensitivity may lead to degradation of the transfer characteristics in a FWL implementation of the digital filter. For instance, the characteristics of an originally stable filter might be so altered that the filter may become unstable. This motivates the study of the coefficient sensitivity minimization problem for digital filters. Several techniques have been proposed to analyze l_2-sensitivity and to synthesize the state-space model structures that minimize l_2-sensitivity [1]-[6]. The minimization of l_2-sensitivity for two-dimensional (2-D) state-space digital filters has also been investigated [5],[7]-[9]. More recently, the minimization problem of l_2-sensitivity subject to l_2-scaling constraints has been treated for 1-D and 2-D state-space digital filters [10]-[13]. It is known that the use of scaling constraints can be beneficial for suppressing overflow [14],[15]. In addition, considerable research interest has also been observed in the design of multidimensional (M-D) recursive digital filters [16]-[19]. Our study of 3-D separable-denominator digital filters is motivated as it fits naturally into typical time-space digital filtering settings. An example of this scenario is a video processing task such as compression or de-noising of a video clip, in that the signals of interest assume the form of a time series, with a 2-D spatial-domain signal known as image at each sampling time instant. For a signal compression task, since the signal redundancy in the time domain and spatial domain are inherently different, the filters to be used for time-domain processing and spatial-domain processing have to be distinctly designed, this justifies the use of a 3-D filter of the form $H(z_1, z_2, z_3) = H_1(z_1)H_2(z_2, z_3)$, where z_1 and (z_2, z_3) are associated with the time domain and the 2-D spatial domain, respectively. Assume that for processing efficiency one decides to use IIR filters, then $H_2(z_2, z_3)$ is a 2-D IIR filter. Since for most spatial filtering the desired frequency responses are quadrantally symmetrical, it is well known that $H_2(z_2, z_3)$ possesses separable denominators [20]. As a result, it is quite natural to study 3-D IIR separable-denominator digital filters. Relevant recent studies also include state-space realization of general M-D filters and possible applications of state-space realization in uncertainty modeling [21], and 3-D realization and its applications in distributed grid sensor networks [22]. In addition, a state-space model for general M-D spatially distributed dynamic system is proposed in [23], and new results on stability and stability margin for 2-D systems are reported in [24]. On the other hand, the literature offers only handful results on efficient realization of 3-D state-space digital filters with minimum coefficient sensitivity [25]. This is likely due to the fact that the problems encountered in the 3-D filters are considerably more involved and challenging relative to their 1-D and 2-D counterparts because here one deals with 3-D coefficient arrays instead of coefficient vectors for 1-D and coefficient matrices for 2-D filters.

The objectives of this paper are twofold: to present a coefficient-sensitivity analysis for a wide class of 3-D digital filters with separable denominators in local state space that leads to an analytic formulation for sensitivity minimization, and to present two solution techniques for the sensitivity minimization problem at hand. To this end, we present a vector-matrix-vector decomposition of a given 3-D transfer function that separates the three frequency-dependent variables (z_1, z_2, z_3), and leads in turn to a state-space realization in a form convenient for subsequent analysis. An l_2-sensitivity analysis is then carried out, a central result of the analysis

Manuscript received January 26, 2012; revised mm dd, 2012.

T. Hinamoto and A. Doi are with Hiroshima Institute of Technology Hiroshima 731-5193, Japan (e-mails: hinamoto@ieee.org, doi@cc.it-hiroshima.ac.jp, Phone:+81-82-921-4338, Fax:+81-82-921-8978)

W.-S. Lu is with the Depertment of Electrical and Computer Engineering, University of Victoria, Victoria, B.C, Canada, V8W 3P6. (e-mail: wslu@ece.uvic.ca, Phone:+1-250-721-8692, Fax:+1-250-721-6052)

is a computationally tractable formula of the overall l_2-sensitivity of a 3-D separable-denominator state-space digital filter. We remark that the formula is of considerable difference from that of [25] as it takes into consideration the particular structure of the realization, especially with its fixed 0 and 1 components subject to dynamic-range l_2-scaling constraints. As a result, this paper deals with a non-convex constrained optimization problem to produce a state-space realization with reduced number of nontrivial parameters, guaranteed freedom of internal overflow and reasonably low coefficient sensitivity. We present two solution methods – one relaxes the constraints into a single trace constraint and solves the relaxed problem with an effective matrix iteration scheme where the Lagrange multiplier is determined via a bisection technique; while the other converts the contained optimization problem at hand into an unconstrained problem and solves it using a quasi-Newton algorithm. Closed-form formula for the gradient is derived for efficient evaluation. A numerical example is presented to illustrate the validity and effectiveness of the proposed techniques in Section IV.

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

II. REALIZATION AND SENSITIVITY ANALYSIS

A. Minimal Realization

Consider a stable 3-D separable-denominator digital filter

$$H(z_1,z_2,z_3)=\frac{N(z_1,z_2,z_3)}{D_1(z_1)D_2(z_2)D_3(z_3)} \tag{1a}$$

where the denominator and the numerator are assumed to be coprime and

$$\begin{aligned} N(z_1,z_2,z_3)&=\sum_{i=0}^{N_1}\sum_{j=0}^{N_2}\sum_{k=0}^{N_3} a_{ijk}z_1^{-i}z_2^{-j}z_3^{-k}\\ D_l(z_l)&=1+\sum_{\xi=1}^{N_l} b_{l\xi}z_l^{-\xi} \quad \text{for} \quad l=1,2,3. \end{aligned} \tag{1b}$$

Although $H(z_1,z_2,z_3)$ has separable denominator, it is not a separable transfer function because the numerator $N(z_1,z_2,z_3)$ is a non-separable polynomial. A key step towards a minimal state-space realization of $H(z_1,z_2,z_3)$ and subsequent minimization of its l_2-sensitivity is to separate the terms in three variables in $H(z_1,z_2,z_3)$. To this end, algebraic manipulations are performed on $N(z_1,z_2,z_3)$, which lead (1a) to a vector-matrix-vector expression where $H(z_1,z_2,z_3)$ are completely separated in its three variables. Namely,

$$H(z_1,z_2,z_3)=\boldsymbol{f}_1(z_1)\boldsymbol{H}_2(z_2)\,\boldsymbol{g}_3(z_3) \tag{2a}$$

where

$$\boldsymbol{f}_1(z_1)=\frac{[1,z_1^{-1},\cdots,z_1^{-N_1}]}{D_1(z_1)},\quad \boldsymbol{g}_3(z_3)=\frac{[1,z_3^{-1},\cdots,z_3^{-N_3}]^T}{D_3(z_3)}$$

$$\boldsymbol{H}_2(z_2)=\frac{\boldsymbol{\Delta}_0+\boldsymbol{\Delta}_1z_2^{-1}+\cdots+\boldsymbol{\Delta}_{N_2}z_2^{-N_2}}{D_2(z_2)}$$

$$\boldsymbol{\Delta}_l=\begin{bmatrix} a_{0l0} & a_{0l1} & \cdots & a_{0lN_3}\\ a_{1l0} & a_{1l1} & \cdots & a_{1lN_3}\\ \vdots & \vdots & \ddots & \vdots\\ a_{N_1l0} & a_{N_1l1} & \cdots & a_{N_1lN_3}\end{bmatrix},\ l=0,1,\cdots,N_2. \tag{2b}$$

We note that the decomposition in (2a) is somewhat similar to that of a finite 3-D array into the product of two finite 1-D vector arrays and another 1-D finite matrix array in the middle, as done in [26]. Furthermore, if the numerator in (1a) is separable, that is, $N(z_1,z_2,z_3)=N_1(z_1)N_2(z_2)N_3(z_3)$ with $N_l(z_l)=\sum_{\xi=0}^{N_l}a_{l\xi}z_l^{-\xi}$ for $l=1,2,3$, then $a_{ijk}=a_{1i}a_{2j}a_{3k}$ holds for $i=0,1,\cdots,N_1$, $j=0,1,\cdots,N_2$ and $k=0,1,\cdots,N_3$. In such a case, matrices $\boldsymbol{\Delta}_l$ for $l=0,1,\cdots,N_2$ will possess considerably higher sparsity.

The 1-D transfer function $\boldsymbol{H}_2(z_2)$ in (2b) has (N_3+1) inputs and (N_1+1) outputs. It can be realized with a minimal state-space model $(\boldsymbol{A}_2,\boldsymbol{B}_2,\boldsymbol{C}_2,\boldsymbol{\Delta}_0)_p$ as

$$\begin{aligned} \boldsymbol{x}(k+1)&=\boldsymbol{A}_2\boldsymbol{x}(k)+\boldsymbol{B}_2\boldsymbol{u}(k)\\ \boldsymbol{y}(k)&=\boldsymbol{C}_2\boldsymbol{x}(k)+\boldsymbol{\Delta}_0\boldsymbol{u}(k) \end{aligned} \tag{3}$$

where $\boldsymbol{x}(k)$ is a $p\times 1$ state-variable vector, $\boldsymbol{u}(k)$ is an $(N_3+1)\times 1$ input vector, $\boldsymbol{y}(k)$ is an $(N_1+1)\times 1$ output vector, and $\boldsymbol{A}_2,\boldsymbol{B}_2,\boldsymbol{C}_2$ and $\boldsymbol{\Delta}_0$ are real constant matrices of appropriate dimensions. Here, p is the least dimension such that a state-space realization of $\boldsymbol{H}_2(z_2)$ is controllable and observable. Such a realization is called a minimal realization [27]. The reader is referred to Appendix I that explains how this realization can actually be constructed. The transfer function of the linear system in (3) can be expressed as

$$\boldsymbol{H}_2(z_2)=\boldsymbol{C}_2(z_2\boldsymbol{I}_p-\boldsymbol{A}_2)^{-1}\boldsymbol{B}_2+\boldsymbol{\Delta}_0. \tag{4}$$

The 1-D trasfer function $\boldsymbol{f}_1(z_1)$ in (2b) has (N_1+1) inputs and a single output, while the 1-D transfer function $\boldsymbol{g}_3(z_3)$ in (2b) has a single input and (N_3+1) outputs. Consequently they can be realized with minimal state-space models $(\boldsymbol{A}_1,\boldsymbol{B}_1,\overline{\boldsymbol{e}}_{N_1}^T,\overline{\boldsymbol{e}}_1^T)_{N_1}$ and $(\boldsymbol{A}_3,\hat{\boldsymbol{e}}_{N_3},\boldsymbol{C}_3,\hat{\boldsymbol{e}}_1)_{N_3}$ as

$$\begin{aligned} \boldsymbol{f}_1(z_1)&=\frac{(\overline{\boldsymbol{e}}_2^T-b_{11}\overline{\boldsymbol{e}}_1^T)z_1^{-1}+\cdots+(\overline{\boldsymbol{e}}_{N_1}^T-b_{1N_1}\overline{\boldsymbol{e}}_1^T)z_1^{-N_1}}{1+b_{11}z_1^{-1}+\cdots+b_{1N_1}z_1^{-N_1}}+\overline{\boldsymbol{e}}_1^T\\ &=\overline{\boldsymbol{e}}_{N_1}^T(z_1\boldsymbol{I}_{N_1}-\boldsymbol{A}_1)^{-1}\boldsymbol{B}_1+\overline{\boldsymbol{e}}_1^T\\ \boldsymbol{g}_3(z_3)&=\frac{(\hat{\boldsymbol{e}}_2-b_{31}\hat{\boldsymbol{e}}_1)z_3^{-1}+\cdots+(\hat{\boldsymbol{e}}_{N_3}-b_{3N_3}\hat{\boldsymbol{e}}_1)z_3^{-N_3}}{1+b_{31}z_3^{-1}+\cdots+b_{3N_3}z_3^{-N_3}}+\hat{\boldsymbol{e}}_1\\ &=\boldsymbol{C}_3(z_3\boldsymbol{I}_{N_3}-\boldsymbol{A}_3)^{-1}\hat{\boldsymbol{e}}_{N_3}+\hat{\boldsymbol{e}}_1, \end{aligned} \tag{5a}$$

respectively, where [27]

$$\boldsymbol{A}_1=\left[\begin{array}{c:c} \boldsymbol{0} & \vdots\\ \cdots\cdots & \boldsymbol{b}_1\\ \boldsymbol{I}_{N_1-1} & \vdots \end{array}\right],\quad \boldsymbol{B}_1=\begin{bmatrix} & 0 & \cdots & 1\\ \boldsymbol{b}_1 & \vdots & \iddots & \vdots\\ & 1 & \cdots & 0\end{bmatrix}$$

$$\boldsymbol{A}_3=\left[\begin{array}{c:c:c} \boldsymbol{0} & \vdots & \boldsymbol{I}_{N_3-1}\\ \cdots\cdots & \cdots & \\ \multicolumn{3}{c}{\boldsymbol{c}_3}\end{array}\right],\quad \boldsymbol{C}_3=\begin{bmatrix} \multicolumn{4}{c}{\boldsymbol{c}_3}\\ \multicolumn{4}{c}{\cdots\cdots\cdots\cdots}\\ 0 & \cdots & 0 & 1\\ \vdots & & \iddots & \vdots\\ 1 & 0 & \cdots & 0\end{bmatrix}$$

$$[\overline{\boldsymbol{e}}_1, \overline{\boldsymbol{e}}_2, \cdots, \overline{\boldsymbol{e}}_{N_1}] = \boldsymbol{I}_{N_1}, \qquad [\hat{\boldsymbol{e}}_1, \hat{\boldsymbol{e}}_2, \cdots, \hat{\boldsymbol{e}}_{N_3}] = \boldsymbol{I}_{N_3}$$
$$\boldsymbol{b}_1 = -\left[b_{1N_1}, \cdots, b_{12}, b_{11}\right]^T, \quad \boldsymbol{c}_3 = -\left[b_{3N_3}, \cdots, b_{32}, b_{31}\right]. \tag{5b}$$

The state-space model in (3) contains $(p + N_1 + 1)(p + N_3 + 1)$ nontrivial and independent parameters, while the numbers of independent parameters in (5a) are N_1 and N_3, respectively. We stress that although the parameter vectors in each of $(\boldsymbol{A}_1, \boldsymbol{B}_1)$ and $(\boldsymbol{A}_3, \boldsymbol{C}_3)$ appear twice, as independent parameters we count each of them only once because, when it varies, the two copies vary in exactly the same way.

As a result, a local state-space model for the 3-D digital filter with separable denominator in (1a) can be realized by

$$\begin{aligned} \boldsymbol{x}'(i,j,k) &= \boldsymbol{A}\boldsymbol{x}(i,j,k) + \boldsymbol{b}u(i,j,k) \\ y(i,j,k) &= \boldsymbol{c}\boldsymbol{x}(i,j,k) + du(i,j,k) \end{aligned} \tag{6a}$$

where $i, j, k \geq 0$ and

$$\boldsymbol{x}'(i,j,k) = \begin{bmatrix} \boldsymbol{x}^h(i+1,j,k) \\ \boldsymbol{x}^v(i,j+1,k) \\ \boldsymbol{x}^a(i,j,k+1) \end{bmatrix}, \quad \boldsymbol{x}(i,j,k) = \begin{bmatrix} \boldsymbol{x}^h(i,j,k) \\ \boldsymbol{x}^v(i,j,k) \\ \boldsymbol{x}^a(i,j,k) \end{bmatrix}$$
$$\boldsymbol{A} = \begin{bmatrix} \boldsymbol{A}_1 & \boldsymbol{A}_4 & \boldsymbol{A}_6 \\ \boldsymbol{0} & \boldsymbol{A}_2 & \boldsymbol{A}_5 \\ \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{A}_3 \end{bmatrix}, \qquad \boldsymbol{b} = \begin{bmatrix} \boldsymbol{B}_1\boldsymbol{\Delta}_0\hat{\boldsymbol{e}}_1 \\ \boldsymbol{B}_2\hat{\boldsymbol{e}}_1 \\ \hat{\boldsymbol{e}}_{N_3} \end{bmatrix}$$
$$\boldsymbol{c} = \begin{bmatrix} \overline{\boldsymbol{e}}_{N_1}^T & \overline{\boldsymbol{e}}_1^T\boldsymbol{C}_2 & \overline{\boldsymbol{e}}_1^T\boldsymbol{\Delta}_0\boldsymbol{C}_3 \end{bmatrix}, \qquad d = a_{000}$$
$$\boldsymbol{A}_4 = \boldsymbol{B}_1\boldsymbol{C}_2, \quad \boldsymbol{A}_5 = \boldsymbol{B}_2\boldsymbol{C}_3, \quad \boldsymbol{A}_6 = \boldsymbol{B}_1\boldsymbol{\Delta}_0\boldsymbol{C}_3. \tag{6b}$$

Here, $\boldsymbol{x}^h(i,j,k)$ is an $N_1 \times 1$ horizontal state vector, $\boldsymbol{x}^v(i,j,k)$ is a $p \times 1$ vertical state vector, $\boldsymbol{x}^a(i,j,k)$ is an $N_3 \times 1$ additional state vector, $u(i,j,k)$ is a scalar input, and $y(i,j,k)$ is a scalar output.

In summary, the 3-D digital filter under consideration admits an implementation scheme as illustrated in Fig. 1 showing vividly a system structure that allows one to focus on optimizing its dominating 1-D MIMO subsystem $(\boldsymbol{A}_2, \boldsymbol{B}_2, \boldsymbol{C}_2, \boldsymbol{\Delta}_0)_p$ in order to reduce its coefficient sensitivity.

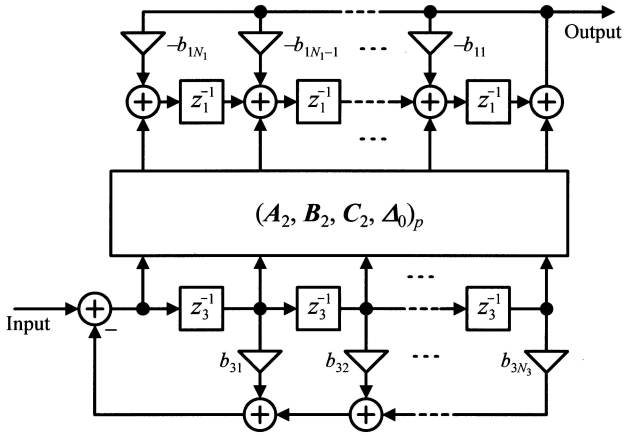

Fig. 1. Block diagram of a 3-D separable-denominator digital filter.

B. l_2-Sensitivity Analysis

The l_2-sensitivities of the 3-D transfer function in (2a) with respect to coefficient matrices $\boldsymbol{A}_2$, $\boldsymbol{B}_2$, $\boldsymbol{C}_2$, $\boldsymbol{\Delta}_0$, $\boldsymbol{b}_1$, and $\boldsymbol{c}_3^T$ are computed as follows.

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

$$\boldsymbol{S}_{\boldsymbol{X}} = \frac{\partial f(\boldsymbol{X})}{\partial \boldsymbol{X}}, \qquad (\boldsymbol{S}_{\boldsymbol{X}})_{ij} = \frac{\partial f(\boldsymbol{X})}{\partial x_{ij}} \tag{7}$$

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

By means of (2a), (4), (5a), Definition 1, and the formula

$$\frac{\partial \boldsymbol{A}^{-1}}{\partial t} = -\boldsymbol{A}^{-1}\frac{\partial \boldsymbol{A}}{\partial t}\boldsymbol{A}^{-1}, \tag{8}$$

the sensitivities of $H(z_1, z_2, z_3)$ with respect to matrices $\boldsymbol{A}_2$, $\boldsymbol{B}_2$, $\boldsymbol{C}_2$, $\boldsymbol{\Delta}_0$, $\boldsymbol{b}_1$, and $\boldsymbol{c}_3^T$ are evaluated as

$$\begin{aligned} \frac{\partial H(z_1,z_2,z_3)}{\partial \boldsymbol{A}_2} &= [\boldsymbol{f}(z_2,z_3)\boldsymbol{g}(z_1,z_2)]^T \\ \frac{\partial H(z_1,z_2,z_3)}{\partial \boldsymbol{B}_2} &= [\boldsymbol{g}_3(z_3)\boldsymbol{g}(z_1,z_2)]^T \\ \frac{\partial H(z_1,z_2,z_3)}{\partial \boldsymbol{C}_2} &= [\boldsymbol{f}(z_2,z_3)\boldsymbol{f}_1(z_1)]^T \\ \frac{\partial H(z_1,z_2,z_3)}{\partial \boldsymbol{\Delta}_0} &= [\boldsymbol{g}_3(z_3)\boldsymbol{f}_1(z_1)]^T \\ \frac{\partial H(z_1,z_2,z_3)}{\partial \boldsymbol{b}_1} &= [\boldsymbol{f}_1(z_1)\boldsymbol{H}_2(z_2)\boldsymbol{g}_3(z_3)\boldsymbol{g}_1(z_1)]^T \\ \frac{\partial H(z_1,z_2,z_3)}{\partial \boldsymbol{c}_3^T} &= \boldsymbol{f}_3(z_3)\boldsymbol{f}_1(z_1)\boldsymbol{H}_2(z_2)\boldsymbol{g}_3(z_3) \end{aligned} \tag{9a}$$

where

$$\begin{aligned} \boldsymbol{f}(z_2,z_3) &= (z_2\boldsymbol{I}_p - \boldsymbol{A}_2)^{-1}\boldsymbol{B}_2\,\boldsymbol{g}_3(z_3) \\ \boldsymbol{g}(z_1,z_2) &= \boldsymbol{f}_1(z_1)\boldsymbol{C}_2(z_2\boldsymbol{I}_p - \mathbf{A}_2)^{-1} \\ \boldsymbol{g}_1(z_1) &= \overline{\boldsymbol{e}}_{N_1}^T(z_1\boldsymbol{I}_{N_1} - \boldsymbol{A}_1)^{-1} \\ \boldsymbol{f}_3(z_3) &= (z_3\boldsymbol{I}_{N_3} - \boldsymbol{A}_3)^{-1}\hat{\boldsymbol{e}}_{N_3}. \end{aligned} \tag{9b}$$

Definition 2: Let $\boldsymbol{X}(z_1, z_2, z_3)$ be an $m \times n$ complex-valued matrix function of complex variables z_1, z_2 and z_3, and $x_{pq}(z_1, z_2, z_3)$ be the (p,q)th entry of $\boldsymbol{X}(z_1, z_2, z_3)$. The l_2-norm of $\boldsymbol{X}(z_1, z_2, z_3)$ is defined as

$$\begin{aligned} \|\boldsymbol{X}(z_1,z_2,z_3)\|_2 &= \left[\frac{1}{(2\pi)^3}\int_0^{2\pi}\int_0^{2\pi}\int_0^{2\pi} \right. \\ &\left. \sum_{p=1}^{m}\sum_{q=1}^{n}|x_{pq}(e^{j\omega_1}, e^{j\omega_2}, e^{j\omega_3})|^2 d\omega_1 d\omega_2 d\omega_3\right]^{\frac{1}{2}} \\ &= \left(\mathrm{tr}\left[\frac{1}{(2\pi j)^3}\oint_{|z_1|=1}\oint_{|z_2|=1}\oint_{|z_3|=1} \right.\right. \\ &\left.\left. \boldsymbol{X}(z_1,z_2,z_3)\boldsymbol{X}^*(z_1,z_2,z_3)\frac{dz_1}{z_1}\frac{dz_2}{z_2}\frac{dz_3}{z_3}\right]\right)^{\frac{1}{2}}. \end{aligned} \tag{10}$$

With (10), the overall l_2-sensitivity measure for the 3-D transfer function in (2a) is defined using (4) and (5a) by

$$S = \left\| \frac{\partial H(z_1,z_2,z_3)}{\partial \boldsymbol{A}_2} \right\|_2^2 + \left\| \frac{\partial H(z_1,z_2,z_3)}{\partial \boldsymbol{B}_2} \right\|_2^2 + \left\| \frac{\partial H(z_1,z_2,z_3)}{\partial \boldsymbol{C}_2} \right\|_2^2 + \left\| \frac{\partial H(z_1,z_2,z_3)}{\partial \boldsymbol{\Delta}_0} \right\|_2^2 + \left\| \frac{\partial H(z_1,z_2,z_3)}{\partial \boldsymbol{b}_1} \right\|_2^2 + \left\| \frac{\partial H(z_1,z_2,z_3)}{\partial \boldsymbol{c}_3^T} \right\|_2^2. \tag{11}$$

We remark that measure S in (11) is a natural 3-D extension of the classical l_2-sensitivity measure for 1-D state-space digital filters [1]. We also remark that coefficient sensitivity is closely related to FWL effects, the reader is referred to Appendix II for further details.

To derive a computationally tractable formula for sensitivity S, we need to evaluate the terms in (9a) explicitly. To this end, we write the impulse responses of the following three 2-D transfer functions as

$$\begin{aligned} \boldsymbol{g}_3(z_3)\boldsymbol{f}_1(z_1) &= \sum_{i=0}^{\infty}\sum_{j=0}^{\infty} \boldsymbol{R}_{ij} z_1^{-i} z_3^{-j} \\ \boldsymbol{H}_2(z_2)\boldsymbol{g}_3(z_3) &= \sum_{i=0}^{\infty}\sum_{j=0}^{\infty} \boldsymbol{r}_{ij} z_2^{-i} z_3^{-j} \\ \boldsymbol{f}_1(z_1)\boldsymbol{H}_2(z_2) &= \sum_{i=0}^{\infty}\sum_{j=0}^{\infty} \hat{\boldsymbol{r}}_{ij} z_1^{-i} z_2^{-j} \end{aligned} \tag{12a}$$

where for $i \geq 1$ and $i \geq 1$,

$$\begin{aligned} \boldsymbol{R}_{00} &= \hat{\boldsymbol{e}}_1 \overline{\boldsymbol{e}}_1^T, \quad \boldsymbol{R}_{ij} = \boldsymbol{C}_3 \boldsymbol{A}_3^{j-1} \hat{\boldsymbol{e}}_{N_3} \overline{\boldsymbol{e}}_{N_1}^T \boldsymbol{A}_1^{i-1} \boldsymbol{B}_1 \\ \boldsymbol{R}_{i0} &= \hat{\boldsymbol{e}}_1 \overline{\boldsymbol{e}}_{N_1}^T \boldsymbol{A}_1^{i-1} \boldsymbol{B}_1, \quad \boldsymbol{R}_{0j} = \boldsymbol{C}_3 \boldsymbol{A}_3^{j-1} \hat{\boldsymbol{e}}_{N_3} \overline{\boldsymbol{e}}_1^T \\ \boldsymbol{r}_{00} &= \boldsymbol{\Delta}_0 \hat{\boldsymbol{e}}_1, \quad \boldsymbol{r}_{ij} = \boldsymbol{C}_2 \boldsymbol{A}_2^{i-1} \boldsymbol{B}_2 \boldsymbol{C}_3 \boldsymbol{A}_3^{j-1} \hat{\boldsymbol{e}}_{N_3} \\ \boldsymbol{r}_{i0} &= \boldsymbol{C}_2 \boldsymbol{A}_2^{i-1} \boldsymbol{B}_2 \hat{\boldsymbol{e}}_1, \quad \boldsymbol{r}_{0j} = \boldsymbol{\Delta}_0 \boldsymbol{C}_3 \boldsymbol{A}_3^{j-1} \hat{\boldsymbol{e}}_{N_3} \\ \hat{\boldsymbol{r}}_{00} &= \overline{\boldsymbol{e}}_1^T \boldsymbol{\Delta}_0, \quad \hat{\boldsymbol{r}}_{ij} = \overline{\boldsymbol{e}}_{N_1}^T \boldsymbol{A}_1^{i-1} \boldsymbol{B}_1 \boldsymbol{C}_2 \boldsymbol{A}_2^{j-1} \boldsymbol{B}_2 \\ \hat{\boldsymbol{r}}_{i0} &= \overline{\boldsymbol{e}}_{N_1}^T \boldsymbol{A}_1^{i-1} \boldsymbol{B}_1 \boldsymbol{\Delta}_0, \quad \hat{\boldsymbol{r}}_{0j} = \overline{\boldsymbol{e}}_1^T \boldsymbol{C}_2 \boldsymbol{A}_2^{j-1} \boldsymbol{B}_2. \end{aligned} \tag{12b}$$

Since the 3-D filter in (1a) is assumed to be stable, each series in (12a) is convergent, hence the infinite sum can be approximated by a finite sum with $(0,0) \leq (i,j) \leq (I,J)$ provided that positive integers I and J are sufficiently large. From (12b), it follows that the adequate numerical values of such I and J depend on the spectral radiuses of matrices $\boldsymbol{A}_1$, $\boldsymbol{A}_2$, and $\boldsymbol{A}_3$. A practical approach to identify the right values of I and J is by trial-and-error in that one computes a truncated series with certain (I,J), then repeating the computation with both I and J increased by a certain amount and compare the two results. The process continues until the difference in norm between the two resulting matrices is less than a prescribed tolerance.

From (9a), (9b) and (12a) it follows that

$$\begin{aligned} \boldsymbol{f}(z_2,z_3)\boldsymbol{g}(z_1,z_2) &= \sum_{i=0}^{\infty}\sum_{j=0}^{\infty} \begin{bmatrix} \boldsymbol{I}_p & \boldsymbol{0} \end{bmatrix} \cdot \left(z_2 \boldsymbol{I}_{2p} - \begin{bmatrix} \boldsymbol{A}_2 & \boldsymbol{B}_2 \boldsymbol{R}_{ij} \boldsymbol{C}_2 \\ \boldsymbol{0} & \boldsymbol{A}_2 \end{bmatrix} \right)^{-1} \begin{bmatrix} \boldsymbol{0} \\ \boldsymbol{I}_p \end{bmatrix} z_1^{-i} z_3^{-j} \\ \boldsymbol{g}_3(z_3)\boldsymbol{g}(z_1,z_2) &= \sum_{i=0}^{\infty}\sum_{j=0}^{\infty} \boldsymbol{R}_{ij} \boldsymbol{C}_2 (z_2 \boldsymbol{I}_p - \boldsymbol{A}_2)^{-1} z_1^{-i} z_3^{-j} \\ \boldsymbol{f}(z_2,z_3)\boldsymbol{f}_1(z_1) &= \sum_{i=0}^{\infty}\sum_{j=0}^{\infty} (z_2 \boldsymbol{I}_p - \boldsymbol{A}_2)^{-1} \boldsymbol{B}_2 \boldsymbol{R}_{ij} z_1^{-i} z_3^{-j} \\ \boldsymbol{f}_1(z_1)\boldsymbol{H}_2(z_2)\boldsymbol{g}_3(z_3)\boldsymbol{g}_1(z_1) &= \sum_{i=0}^{\infty}\sum_{j=0}^{\infty} \begin{bmatrix} \overline{\boldsymbol{e}}_{N_1}^T & \overline{\boldsymbol{e}}_1^T \boldsymbol{r}_{ij} \overline{\boldsymbol{e}}_{N_1}^T \end{bmatrix} \cdot \left(z_1 \boldsymbol{I}_{2N_1} - \begin{bmatrix} \boldsymbol{A}_1 & \boldsymbol{B}_1 \boldsymbol{r}_{ij} \overline{\boldsymbol{e}}_{N_1}^T \\ \boldsymbol{0} & \boldsymbol{A}_1 \end{bmatrix} \right)^{-1} \begin{bmatrix} \boldsymbol{0} \\ \boldsymbol{I}_{N_1} \end{bmatrix} z_2^{-i} z_3^{-j} \\ \boldsymbol{f}_3(z_3)\boldsymbol{f}_1(z_1)\boldsymbol{H}_2(z_2)\boldsymbol{g}_3(z_3) &= \sum_{i=0}^{\infty}\sum_{j=0}^{\infty} \begin{bmatrix} \boldsymbol{I}_{N_3} & \boldsymbol{0} \end{bmatrix} \cdot \left(z_3 \boldsymbol{I}_{2N_3} - \begin{bmatrix} \boldsymbol{A}_3 & \hat{\boldsymbol{e}}_{N_3} \hat{\boldsymbol{r}}_{ij} \boldsymbol{C}_3 \\ \boldsymbol{0} & \boldsymbol{A}_3 \end{bmatrix} \right)^{-1} \begin{bmatrix} \hat{\boldsymbol{e}}_{N_3} \hat{\boldsymbol{r}}_{ij} \hat{\boldsymbol{e}}_1 \\ \hat{\boldsymbol{e}}_{N_3} \end{bmatrix} z_1^{-i} z_2^{-j}. \end{aligned} \tag{13}$$

Referring to (9a), (10) and (13), the l_2-sensitivity measure in (11) can be expressed as

$$\begin{aligned} S = {} & \mathrm{tr}[\boldsymbol{M}_A(\boldsymbol{I}_p)] + \mathrm{tr}[\boldsymbol{W}_B] + \mathrm{tr}[\boldsymbol{K}_C] \\ & + \mathrm{tr}[\boldsymbol{N}_{\Delta_0}] + \mathrm{tr}[\boldsymbol{W}_1] + \mathrm{tr}[\boldsymbol{K}_3] \end{aligned} \tag{14a}$$

where Gramians $\boldsymbol{M}_A(\boldsymbol{P})$, $\boldsymbol{W}_B$, $\boldsymbol{K}_C$, $\boldsymbol{N}_{\Delta_0}$, $\boldsymbol{W}_1$, and $\boldsymbol{K}_3$ can be computed using

$$\begin{aligned} \boldsymbol{M}_A(\boldsymbol{P}) &= \sum_{i=0}^{\infty}\sum_{j=0}^{\infty} \begin{bmatrix} \boldsymbol{0} & \boldsymbol{I}_p \end{bmatrix} \boldsymbol{M}_{ij}^A \begin{bmatrix} \boldsymbol{0} \\ \boldsymbol{I}_p \end{bmatrix} \\ \boldsymbol{M}_{ij}^A &= \begin{bmatrix} \boldsymbol{A}_2 & \boldsymbol{B}_2 \boldsymbol{R}_{ij} \boldsymbol{C}_2 \\ \boldsymbol{0} & \boldsymbol{A}_2 \end{bmatrix}^T \boldsymbol{M}_{ij}^A \begin{bmatrix} \boldsymbol{A}_2 & \boldsymbol{B}_2 \boldsymbol{R}_{ij} \boldsymbol{C}_2 \\ \boldsymbol{0} & \boldsymbol{A}_2 \end{bmatrix} + \begin{bmatrix} \boldsymbol{P}^{-1} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{0} \end{bmatrix} \\ \boldsymbol{W}_B &= \sum_{i=0}^{\infty}\sum_{j=0}^{\infty} \boldsymbol{W}_{ij}^B, \qquad \boldsymbol{K}_C = \sum_{i=0}^{\infty}\sum_{j=0}^{\infty} \boldsymbol{K}_{ij}^C \\ \boldsymbol{W}_{ij}^B &= \boldsymbol{A}_2^T \boldsymbol{W}_{ij}^B \boldsymbol{A}_2 + \boldsymbol{C}_2^T \boldsymbol{R}_{ij}^T \boldsymbol{R}_{ij} \boldsymbol{C}_2 \\ \boldsymbol{K}_{ij}^C &= \boldsymbol{A}_2 \boldsymbol{K}_{ij}^C \boldsymbol{A}_2^T + \boldsymbol{B}_2 \boldsymbol{R}_{ij} \boldsymbol{R}_{ij}^T \boldsymbol{B}_2^T \\ \boldsymbol{N}_{\Delta_0} &= \sum_{i=0}^{\infty}\sum_{j=0}^{\infty} \boldsymbol{R}_{ij}^T \boldsymbol{R}_{ij} \\ \boldsymbol{W}_1 &= \sum_{i=0}^{\infty}\sum_{j=0}^{\infty} \begin{bmatrix} \boldsymbol{0} & \boldsymbol{I}_{N_1} \end{bmatrix} \boldsymbol{W}_{ij}^1 \begin{bmatrix} \boldsymbol{0} \\ \boldsymbol{I}_{N_1} \end{bmatrix} \\ \boldsymbol{W}_{ij}^1 &= \begin{bmatrix} \boldsymbol{A}_1 & \boldsymbol{B}_1 \boldsymbol{r}_{ij} \overline{\boldsymbol{e}}_{N_1}^T \\ \boldsymbol{0} & \boldsymbol{A}_1 \end{bmatrix}^T \boldsymbol{W}_{ij}^1 \begin{bmatrix} \boldsymbol{A}_1 & \boldsymbol{B}_1 \boldsymbol{r}_{ij} \overline{\boldsymbol{e}}_{N_1}^T \\ \boldsymbol{0} & \boldsymbol{A}_1 \end{bmatrix} + \begin{bmatrix} \overline{\boldsymbol{e}}_{N_1}^T & \overline{\boldsymbol{e}}_1^T \boldsymbol{r}_{ij} \overline{\boldsymbol{e}}_{N_1}^T \end{bmatrix}^T \begin{bmatrix} \overline{\boldsymbol{e}}_{N_1}^T & \overline{\boldsymbol{e}}_1^T \boldsymbol{r}_{ij} \overline{\boldsymbol{e}}_{N_1}^T \end{bmatrix} \\ \boldsymbol{K}_3 &= \sum_{i=0}^{\infty}\sum_{j=0}^{\infty} \begin{bmatrix} \boldsymbol{I}_{N_3} & \boldsymbol{0} \end{bmatrix} \boldsymbol{K}_{ij}^3 \begin{bmatrix} \boldsymbol{I}_{N_3} \\ \boldsymbol{0} \end{bmatrix} \\ \boldsymbol{K}_{ij}^3 &= \begin{bmatrix} \boldsymbol{A}_3 & \hat{\boldsymbol{e}}_{N_3} \hat{\boldsymbol{r}}_{ij} \boldsymbol{C}_3 \\ \boldsymbol{0} & \boldsymbol{A}_3 \end{bmatrix} \boldsymbol{K}_{ij}^3 \begin{bmatrix} \boldsymbol{A}_3 & \hat{\boldsymbol{e}}_{N_3} \hat{\boldsymbol{r}}_{ij} \boldsymbol{C}_3 \\ \boldsymbol{0} & \boldsymbol{A}_3 \end{bmatrix}^T + \begin{bmatrix} \hat{\boldsymbol{e}}_{N_3} \hat{\boldsymbol{r}}_{ij} \hat{\boldsymbol{e}}_1 \\ \hat{\boldsymbol{e}}_{N_3} \end{bmatrix} \begin{bmatrix} \hat{\boldsymbol{e}}_{N_3} \hat{\boldsymbol{r}}_{ij} \hat{\boldsymbol{e}}_1 \\ \hat{\boldsymbol{e}}_{N_3} \end{bmatrix}^T. \end{aligned} \tag{14b}$$

III. REALIZATION WITH LOW SENSITIVITY

Transfer function $H(z_1,z_2,z_3)$ in (2a) consists of three 1-D factors. As shown in (5a) and (5b), realizations of the first and last factors, i.e. $\boldsymbol{f}_1(z_1)$ and $\boldsymbol{g}_3(z_3)$, possess rather

simple structures as they involve only N_1 and N_3 independent parameters, respectively, in addition to the fixed ones and zeros. We note, however, that if coordinate transformations would be applied to each of these two sub-transfer functions so as to reduce coefficient sensitivity, the resulting state-space realizations would in general involve as many as $2[N_1(N_1+1)+N_3(N_3+1)]$ parameters, which implies a drastic increase in realization complexity. On the contrary, the sub-transfer function $\boldsymbol{H}_2(z_2)$ contains most of the filter's nontrivial coefficients and its state-space realization $(\boldsymbol{A}_2, \boldsymbol{B}_2, \boldsymbol{C}_2, \boldsymbol{\Delta}_0)_p$ hardly contains many fixed trivial components like ones and zeros (see (A.4) in Appendix I), therefore, application of a coordinate transformation to $(\boldsymbol{A}_2, \boldsymbol{B}_2, \boldsymbol{C}_2, \boldsymbol{\Delta}_0)_p$ will not increase realization complexity in a significant manner. For these reasons, we in the rest of the paper shall seek to minimize the sensitivity associated with matrix transfer function $\boldsymbol{H}_2(z_2)$.

Applying a coordinate transformation $\overline{\boldsymbol{x}}(k) = \boldsymbol{T}^{-1}\boldsymbol{x}(k)$ to the linear system $(\boldsymbol{A}_2, \boldsymbol{B}_2, \boldsymbol{C}_2, \boldsymbol{\Delta}_0)_p$ in (3), we obtain a new realization $(\overline{\boldsymbol{A}}_2, \overline{\boldsymbol{B}}_2, \overline{\boldsymbol{C}}_2, \boldsymbol{\Delta}_0)_p$ characterized by

$$\overline{\boldsymbol{A}}_2 = \boldsymbol{T}^{-1}\boldsymbol{A}_2\boldsymbol{T}, \quad \overline{\boldsymbol{B}}_2 = \boldsymbol{T}^{-1}\boldsymbol{B}_2, \quad \overline{\boldsymbol{C}}_2 = \boldsymbol{C}_2\boldsymbol{T} \tag{15}$$

where $\boldsymbol{T}$ is a $p\times p$ nonsingular matrix. For the new realization, the l_2-sensitivity measure in (14a) is written as

$$S(\boldsymbol{P}) = J(\boldsymbol{P}) + \text{tr}[\boldsymbol{N}_{\Delta_0}] + \text{tr}[\boldsymbol{W}_1] + \text{tr}[\boldsymbol{K}_3] \tag{16a}$$

where

$$J(\boldsymbol{P}) = \text{tr}[\boldsymbol{M}_A(\boldsymbol{P})\boldsymbol{P}] + \text{tr}[\boldsymbol{W}_B\boldsymbol{P}] + \text{tr}[\boldsymbol{K}_C\boldsymbol{P}^{-1}] \tag{16b}$$

with $\boldsymbol{P} = \boldsymbol{T}\boldsymbol{T}^T$. For the reasons stated above, our attention shall now be focused on the minimization of sensitivity measure $J(\boldsymbol{P})$.

Since $\boldsymbol{f}(z_2, z_3)$ is the transfer function from the filter input to the state-variable vector $\boldsymbol{x}(k)$, a controllability Gramian $\boldsymbol{K}$ is defined by

$$\boldsymbol{K} = \frac{1}{(2\pi j)^2}\oint_{|z_2|=1}\oint_{|z_3|=1}\boldsymbol{f}(z_2, z_3)\boldsymbol{f}^*(z_2, z_3)\frac{dz_2}{z_2}\frac{dz_3}{z_3} \tag{17}$$

that can be obtained by solving the following Lyapunov equations:

$$\begin{aligned}\boldsymbol{K}_0 &= \boldsymbol{A}_3\boldsymbol{K}_0\boldsymbol{A}_3^T + \hat{\boldsymbol{e}}_{N_3}\hat{\boldsymbol{e}}_{N_3}^T \\ \boldsymbol{K} &= \boldsymbol{A}_2\boldsymbol{K}\boldsymbol{A}_2^T + \boldsymbol{B}_2(\boldsymbol{C}_3\boldsymbol{K}_0\boldsymbol{C}_3^T + \hat{\boldsymbol{e}}_1\hat{\boldsymbol{e}}_1^T)\boldsymbol{B}_2^T.\end{aligned} \tag{18}$$

For the new realization $(\overline{\boldsymbol{A}}_2, \overline{\boldsymbol{B}}_2, \overline{\boldsymbol{C}}_2, \boldsymbol{\Delta}_0)_p$, l_2-scaling constraints on the state-variable vector $\overline{\boldsymbol{x}}(k)$ are given by

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

The reader is referred to Appendix III that explains (19) in details. Summarizing, our sensitivity minimization problem is to find a coordinate transformation matrix $\boldsymbol{T}$ that minimizes $J(\boldsymbol{P})$ in (16b) subject to the l_2-scaling constraints in (19). Two solution methods for this problem are presented below.

We stress that apart from the difference in technical details in implementing these methods, they are analogous to their 1-D and 2-D counterparts investigated in [10]-[11].

A. A Constrained Optimization Method

Directly dealing with the l_2-scaling constraints in (19) was found technically unfeasible. Instead, we consider a relaxation of (19) to a single constraint as

$$\text{tr}[\boldsymbol{T}^{-1}\boldsymbol{K}\boldsymbol{T}^{-T}] = \text{tr}[\boldsymbol{K}\boldsymbol{P}^{-1}] = p. \tag{20}$$

The relaxation proposed in (20) for the constraints in (19) has two advantages: First, rather than dealing with a total of p constraints in (19), with (20) we have only one single equality constraint that, as will be demonstrated below, can be handled by a Lagrange function with one additional parameter; second, from (20) it is seen that the constraint is now expressed in terms of matrix $\boldsymbol{P}$, therefore, once an optimal $\boldsymbol{P}$ is identified, an optimal coordinate transformation matrix $\boldsymbol{T}$ can be set to

$$\boldsymbol{T} = \boldsymbol{P}^{\frac{1}{2}}\boldsymbol{U} \tag{21}$$

with $\boldsymbol{U}$ a $p\times p$ orthogonal matrix. Note that changing $\boldsymbol{U}$ in (21) to a different orthogonal matrix does not affect the optimality of $\boldsymbol{T}$ in (21) in the sense that, for any orthogonal $\boldsymbol{U}$, $\boldsymbol{P} = \boldsymbol{T}\boldsymbol{T}^T$ remains valid, hence it does not alter the optimal value of the sensitivity measure in (16b). Further notice that with (21), (19) becomes

$$(\boldsymbol{U}^T\boldsymbol{P}^{-\frac{1}{2}}\boldsymbol{K}\boldsymbol{P}^{-\frac{T}{2}}\boldsymbol{U})_{ii} = 1 \quad \text{for} \quad i = 1, 2, \cdots, p. \tag{22}$$

With $\boldsymbol{K}$ fixed and $\boldsymbol{P}$ determined, it is straightforward to find an orthogonal matrix $\boldsymbol{U}$ so that the constraints in (22) (hence (19)) are satisfied. In words, via (21) a solution of the relaxed problem can be readily concerted to a solution that accurately satisfies the l_2-scaling constraints in (19). The reader is referred to Section IV.A for a numerical example that illustrates the technique described above. This justifies the relaxation made in (20) and in this way, we now focus on the problem

$$\begin{aligned}&\text{minimize } J(\boldsymbol{P}) \text{ in (16b)} \\ &\text{subject to } \text{tr}[\boldsymbol{K}\boldsymbol{P}^{-1}] = p.\end{aligned} \tag{23}$$

To solve problem (23), we define the following Lagrange function of the problem:

$$\begin{aligned}J_o(\boldsymbol{P}, \lambda) = &\text{tr}[\boldsymbol{M}_A(\boldsymbol{P})\boldsymbol{P}] + \text{tr}[\boldsymbol{W}_B\boldsymbol{P}] \\ &+\text{tr}[\boldsymbol{K}_C\boldsymbol{P}^{-1}] + \lambda(\text{tr}[\boldsymbol{K}\boldsymbol{P}^{-1}] - p)\end{aligned} \tag{24}$$

where λ is the Lagrange multiplier. Computing $J_o(\boldsymbol{P}, \lambda)/\partial\boldsymbol{P}$ by using [28, p.275]

$$\frac{d[\text{tr}(\boldsymbol{M}\boldsymbol{X})]}{d\boldsymbol{X}} = \boldsymbol{M}^T, \quad \frac{d[\text{tr}(\boldsymbol{M}\boldsymbol{X}^{-1})]}{d\boldsymbol{X}} = -(\boldsymbol{X}^{-1}\boldsymbol{M}\boldsymbol{X}^{-1})^T \tag{25}$$

and setting $\partial J_o(\boldsymbol{P}, \lambda)/\partial\boldsymbol{P} = \boldsymbol{0}$, it follows that

$$\boldsymbol{P}\boldsymbol{F}(\boldsymbol{P})\boldsymbol{P} = \boldsymbol{G}(\boldsymbol{P}, \lambda) \tag{26a}$$

where

$$\begin{aligned}\boldsymbol{F}(\boldsymbol{P}) &= \boldsymbol{M}_A(\boldsymbol{P}) + \boldsymbol{W}_B \\ \boldsymbol{G}(\boldsymbol{P}, \lambda) &= \boldsymbol{N}_A(\boldsymbol{P}) + \boldsymbol{K}_C + \lambda\boldsymbol{K}\end{aligned} \tag{26b}$$

with

$$\begin{aligned}\boldsymbol{N}_A(\boldsymbol{P}) &= \sum_{i=0}^{\infty}\sum_{j=0}^{\infty}\begin{bmatrix}\boldsymbol{I}_p & \boldsymbol{0}\end{bmatrix}\boldsymbol{N}_{ij}^A\begin{bmatrix}\boldsymbol{I}_p \\ \boldsymbol{0}\end{bmatrix} \\ \boldsymbol{N}_{ij}^A &= \begin{bmatrix}\boldsymbol{A}_2 & \boldsymbol{B}_2\boldsymbol{R}_{ij}\boldsymbol{C}_2 \\ \boldsymbol{0} & \boldsymbol{A}_2\end{bmatrix}\boldsymbol{N}_{ij}^A\begin{bmatrix}\boldsymbol{A}_2 & \boldsymbol{B}_2\boldsymbol{R}_{ij}\boldsymbol{C}_2 \\ \boldsymbol{0} & \boldsymbol{A}_2\end{bmatrix}^T \\ &\quad + \begin{bmatrix}\boldsymbol{0} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{P}\end{bmatrix}.\end{aligned} \tag{26c}$$

The equation in (26a) is highly nonlinear with respect to $\boldsymbol{P}$. An effective approach to solving the equation in (26a) is to relax it into the following recursive second-order matrix equation:

$$\boldsymbol{P}^{(k+1)}\boldsymbol{F}(\boldsymbol{P}^{(k)})\boldsymbol{P}^{(k+1)} = \boldsymbol{G}(\boldsymbol{P}^{(k)}, \lambda^{(k+1)}) \tag{27}$$

where $\boldsymbol{P}^{(k)}$ is assumed to be known from the previous recursion and the solution $\boldsymbol{P}^{(k+1)}$ is given by

$$\begin{aligned}\boldsymbol{P}^{(k+1)} = \boldsymbol{F}(\boldsymbol{P}^{(k)})^{-\frac{1}{2}}[\boldsymbol{F}(\boldsymbol{P}^{(k)})^{\frac{1}{2}} \\ \cdot\boldsymbol{G}(\boldsymbol{P}^{(k)}, \lambda^{(k+1)})\boldsymbol{F}(\boldsymbol{P}^{(k)})^{\frac{1}{2}}]^{\frac{1}{2}}\boldsymbol{F}(\boldsymbol{P}^{(k)})^{-\frac{1}{2}}.\end{aligned} \tag{28}$$

Here, the Lagrange multiplier $\lambda^{(k+1)}$ can be efficiently obtained using a bisection method [29] so that

$$f(\lambda^{(k+1)}) = |\, p - \mathrm{tr}[\tilde{\boldsymbol{K}}^{(k)}\tilde{\boldsymbol{G}}^{(k)}(\lambda^{(k+1)})]\,| < \varepsilon \tag{29a}$$

where

$$\begin{aligned}\tilde{\boldsymbol{G}}^{(k)}(\lambda^{(k+1)}) &= [\boldsymbol{F}(\boldsymbol{P}^{(k)})^{\frac{1}{2}}\mathbf{G}(\boldsymbol{P}^{(k)}, \lambda^{(k+1)})\boldsymbol{F}(\boldsymbol{P}^{(k)})^{\frac{1}{2}}]^{-\frac{1}{2}} \\ \tilde{\boldsymbol{K}}^{(k)} &= \boldsymbol{F}(\boldsymbol{P}^{(k)})^{\frac{1}{2}}\boldsymbol{K}\boldsymbol{F}(\boldsymbol{P}^{(k)})^{\frac{1}{2}}.\end{aligned} \tag{29b}$$

This iteration process continues until

$$|J(\boldsymbol{P}^{(k)}, \lambda^{(k+1)}) - J(\boldsymbol{P}^{(k-1)}, \lambda^{(k)})| < \varepsilon \tag{30}$$

is satisfied for a prescribed tolerance $\varepsilon > 0$. If the iteration is terminated at step k, $\boldsymbol{P}^{(k)}$ is claimed to be a solution point.

B. An Unconstrained Optimization Method

Defining

$$\hat{\boldsymbol{T}} = \boldsymbol{T}^T\boldsymbol{K}^{-\frac{1}{2}} \tag{31}$$

the l_2-scaling constraints in (19) can be written as

$$(\hat{\boldsymbol{T}}^{-T}\hat{\boldsymbol{T}}^{-1})_{ii} = 1 \quad \text{for} \quad i = 1, 2, \cdots, p. \tag{32}$$

It is obvious that the conditions in (32) are always satisfied by choosing $\hat{\boldsymbol{T}}^{-1}$ as

$$\hat{\boldsymbol{T}}^{-1} = \left[\frac{\boldsymbol{t}_1}{||\boldsymbol{t}_1||}, \frac{\boldsymbol{t}_2}{||\boldsymbol{t}_2||}, \cdots, \frac{\boldsymbol{t}_p}{||\boldsymbol{t}_p||}\right]. \tag{33}$$

By writing $J(\boldsymbol{P})$ in (16b) as

$$J(\boldsymbol{T}) = \mathrm{tr}[\boldsymbol{T}^T\boldsymbol{M}_A(\boldsymbol{T}\boldsymbol{T}^T)\boldsymbol{T}] + \mathrm{tr}[\boldsymbol{T}^T\boldsymbol{W}_B\boldsymbol{T}] + \mathrm{tr}[\boldsymbol{T}^{-1}\boldsymbol{K}_C\boldsymbol{T}^{-T}] \tag{34}$$

and then using (31), the l_2-sensitivity measure can be expressed as

$$J(\boldsymbol{x}) = \mathrm{tr}[\hat{\boldsymbol{T}}\hat{\boldsymbol{M}}_A(\hat{\boldsymbol{T}})\hat{\boldsymbol{T}}^T] + \mathrm{tr}[\hat{\boldsymbol{T}}\hat{\boldsymbol{W}}_B\hat{\boldsymbol{T}}^T] + \mathrm{tr}[\hat{\boldsymbol{T}}^{-T}\hat{\boldsymbol{K}}_C\hat{\boldsymbol{T}}^{-1}] \tag{35a}$$

where $\boldsymbol{x} = (\boldsymbol{t}_1^T, \boldsymbol{t}_2^T, \cdots, \boldsymbol{t}_p^T)^T$ and

$$\begin{aligned}\hat{\boldsymbol{M}}_A(\hat{\boldsymbol{T}}) &= \sum_{i=0}^{\infty}\sum_{j=0}^{\infty}\begin{bmatrix}\mathbf{0} & \boldsymbol{K}^{\frac{1}{2}}\end{bmatrix}\hat{\boldsymbol{M}}_{ij}^A\begin{bmatrix}\mathbf{0} \\ \boldsymbol{K}^{\frac{1}{2}}\end{bmatrix} \\ \hat{\boldsymbol{M}}_{ij}^A &= \begin{bmatrix}\boldsymbol{A}_2 & \boldsymbol{B}_2\boldsymbol{R}_{ij}\boldsymbol{C}_2 \\ \mathbf{0} & \boldsymbol{A}_2\end{bmatrix}^T\hat{\boldsymbol{M}}_{ij}^A\begin{bmatrix}\boldsymbol{A}_2 & \boldsymbol{B}_2\boldsymbol{R}_{ij}\boldsymbol{C}_2 \\ \mathbf{0} & \boldsymbol{A}_2\end{bmatrix} \\ &\quad + \begin{bmatrix}\boldsymbol{K}^{-\frac{1}{2}}\hat{\boldsymbol{T}}^{-1}\hat{\boldsymbol{T}}^{-T}\boldsymbol{K}^{-\frac{1}{2}} & \mathbf{0} \\ \mathbf{0} & \mathbf{0}\end{bmatrix} \\ \hat{\boldsymbol{W}}_B &= \boldsymbol{K}^{\frac{1}{2}}\boldsymbol{W}_B\boldsymbol{K}^{\frac{1}{2}}, \quad \hat{\boldsymbol{K}}_C = \boldsymbol{K}^{-\frac{1}{2}}\boldsymbol{K}_C\boldsymbol{K}^{-\frac{1}{2}}.\end{aligned} \tag{35b}$$

In this way, the original constrained optimization problem can be converted into an unconstrained optimization problem of obtaining a $p^2 \times 1$ vector $\boldsymbol{x}$ which minimizes $J(\boldsymbol{x})$ in (35a).

Applying a quasi-Newton (known as the Broyden-Fletcher-Goldfarb-Shanno (BFGS)) algorithm [30] to minimize $J(\boldsymbol{x})$ in (35a), in the kth iteration the most recent point $\boldsymbol{x}_k$ is updated to point $\boldsymbol{x}_{k+1}$ as

$$\boldsymbol{x}_{k+1} = \boldsymbol{x} + \alpha_k\boldsymbol{d}_k \tag{36a}$$

where

$$\begin{aligned}&\boldsymbol{d}_k = -\boldsymbol{S}_k\nabla J(\boldsymbol{x}_k), \qquad \alpha_k = arg\min_{\alpha} J(\boldsymbol{x}_k + \alpha\boldsymbol{d}_k) \\ &\boldsymbol{S}_{k+1} = \boldsymbol{S}_k + \left(1 + \frac{\gamma_k^T\boldsymbol{S}_k\gamma_k}{\gamma_k^T\delta_k}\right)\frac{\delta_k\delta_k^T}{\gamma_k^T\delta_k} - \frac{\delta_k\gamma_k^T\boldsymbol{S}_k + \boldsymbol{S}_k\gamma_k\delta_k^T}{\gamma_k^T\delta_k} \\ &\boldsymbol{S}_0 = \boldsymbol{I}_{p^2}, \quad \delta_k = \boldsymbol{x}_{k+1} - \boldsymbol{x}_k, \quad \gamma_k = \nabla J(\boldsymbol{x}_{k+1}) - \nabla J(\boldsymbol{x}_k).\end{aligned} \tag{36b}$$

In the above, $\nabla J(\boldsymbol{x})$ is the gradient of $J(\boldsymbol{x})$ with respect to $\boldsymbol{x}$, and $\boldsymbol{S}_k$ is a positive-definite approximation of the inverse Hessian matrix of $J(\boldsymbol{x})$. The algorithm starts with a trivial initial point $\boldsymbol{x}_0$ obtained from an initial assignment $\boldsymbol{T} = \boldsymbol{I}_p$, and this iteration process continues until

$$|J(\boldsymbol{x}_{k+1}) - J(\boldsymbol{x}_k)| < \epsilon \tag{37}$$

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

The BFGS algorithm is a well-known descent algorithm meaning that a twice continuously differentiable objective function at the iterates generated by the BFGS algorithm is monotonically decreasing. In theory, it was shown [31] that if the starting point is sufficiently close to a local minimizer and the initial Hessian approximation is sufficiently close to the Hessian at that minimizer, then the BFGS iterates will converge to the minimizer.

The implementation of (36a) requires the gradient of $J(\boldsymbol{x})$, which can be efficiently evaluated using closed-form expressions, see Appendix IV.

IV. A CASE STUDY

We now present a case study to demonstrate the effectiveness of the two algorithms developed in Section III. The case study was carried out using MATLAB on a PC with an Intel Core i-2500 CPU at 3.3 GHz.

Consider a stable 3-D separable-denominator digital filter in (2a) and (2b) specified by

$$\boldsymbol{\Delta}_0 = 10^{-2}\begin{bmatrix}0.00730 & 0.34297 & -0.09594 & 0.20541 \\ 3.33408 & -5.73707 & 3.94939 & -1.61598 \\ -1.46081 & 2.66051 & -1.68094 & 0.68022 \\ 1.12651 & -1.62192 & 1.24735 & -0.55781\end{bmatrix}$$

$$\boldsymbol{\Delta}_1 = 10^{-2}\begin{bmatrix}2.81318 & -5.00467 & 3.46926 & -0.84798 \\ -5.29980 & 9.24831 & -6.29206 & 2.80791 \\ 4.95232 & -8.39641 & 5.73329 & -1.62170 \\ 0.72029 & -1.34272 & 0.95941 & 0.54827\end{bmatrix}$$

$$\boldsymbol{\Delta}_2 = 10^{-2}\begin{bmatrix}-0.69409 & 1.54874 & -0.94779 & 0.39116 \\ 3.93785 & -6.79910 & 4.66564 & -1.96344 \\ -2.37995 & 4.20737 & -2.75482 & 0.95329 \\ 0.70545 & -0.90615 & 0.73168 & -0.55633\end{bmatrix}$$

$$\boldsymbol{\Delta}_3 = 10^{-2}\begin{bmatrix} 1.67681 & -2.69078 & 1.98218 & -0.33567 \\ -0.59937 & 1.11289 & -0.71981 & 0.43504 \\ 1.82472 & -2.93685 & 2.11591 & -0.43417 \\ 1.28875 & -2.01749 & 1.51782 & -0.09016 \end{bmatrix}$$

$$\begin{bmatrix} b_{11} & b_{12} & b_{13} \end{bmatrix} = \begin{bmatrix} b_{31} & b_{32} & b_{33} \end{bmatrix} = \begin{bmatrix} -1.81600 & 1.23756 & -0.31382 \end{bmatrix}$$

$$\begin{bmatrix} b_{21} & b_{22} & b_{23} \end{bmatrix} = \begin{bmatrix} -1.81611 & 1.23775 & -0.31391 \end{bmatrix}.$$

The 3-D filter can be realized by a minimal state-space model in (3) with

$$\boldsymbol{A}_2 = \begin{bmatrix} 0.00000 & -0.19089 & 0.29060 \\ 0.74393 & -86.40470 & 133.71075 \\ -0.27211 & -57.01643 & 88.22081 \end{bmatrix}$$

$$\boldsymbol{B}_2 = 10^3\begin{bmatrix} 0.00602 & -0.00921 & 0.00699 & -0.00095 \\ -1.10247 & 1.68622 & -1.27902 & 0.17267 \\ -0.71455 & 1.09291 & -0.82977 & 0.11192 \end{bmatrix}$$

$$\boldsymbol{C}_2 = \begin{bmatrix} 0.07236 & 0.06711 & -0.10298 \\ 0.01930 & 0.01789 & -0.02745 \\ 0.05887 & 0.05460 & -0.08378 \\ 0.07079 & 0.06565 & -0.10073 \end{bmatrix}.$$

Applying a coordinate transformation defined by

$$\boldsymbol{T} = 10^3\mathrm{diag}\{0.01077, 2.58588, 1.68384\}$$

and using (14b) with truncation $(0,0) \le (i,j) \le (100,100)$ and (18) to evaluate the Gramians $\boldsymbol{M}_A(\boldsymbol{I}_3)$, $\boldsymbol{W}_B$, $\boldsymbol{K}_C$, $\boldsymbol{N}_{\boldsymbol{\Delta}_0}$, $\boldsymbol{W}_1$, $\boldsymbol{K}_3$, $\boldsymbol{K}_0$ and $\boldsymbol{K}$ yields

$$\boldsymbol{M}_A(\boldsymbol{I}_3) = 10^7\begin{bmatrix} 0.00021 & 0.03478 & -0.03471 \\ 0.03478 & 5.83540 & -5.82400 \\ -0.03471 & -5.82400 & 5.81261 \end{bmatrix}$$

$$\boldsymbol{W}_B = 10^8\begin{bmatrix} 0.00014 & 0.02404 & -0.02399 \\ 0.02404 & 4.36159 & -4.35410 \\ -0.02399 & -4.35410 & 4.34663 \end{bmatrix}$$

$$\boldsymbol{K}_C = 10\begin{bmatrix} 5.70413 & -4.79529 & -4.78664 \\ -4.79529 & 5.70413 & 5.70410 \\ -4.78664 & 5.70410 & 5.70413 \end{bmatrix}$$

$$\boldsymbol{N}_{\boldsymbol{\Delta}_0} = 10\begin{bmatrix} 8.61482 & 8.61482 & 8.61482 & 8.61482 \\ 8.61482 & 8.61482 & 8.61482 & 8.61482 \\ 8.61482 & 8.61482 & 8.61482 & 8.61482 \\ 8.61482 & 8.61482 & 8.61482 & 8.61482 \end{bmatrix}$$

$$\boldsymbol{W}_1 = 10^3\begin{bmatrix} 2.68584 & 2.59589 & 2.34321 \\ 2.59589 & 2.68584 & 2.59589 \\ 2.34321 & 2.59589 & 2.68584 \end{bmatrix}$$

$$\boldsymbol{K}_3 = 10^3\begin{bmatrix} 2.52585 & 2.43584 & 2.18528 \\ 2.43584 & 2.52585 & 2.43584 \\ 2.18528 & 2.43584 & 2.52585 \end{bmatrix}$$

$$\boldsymbol{K}_0 = 10\begin{bmatrix} 1.42603 & 1.29740 & 0.99842 \\ 1.29740 & 1.42603 & 1.29740 \\ 0.99842 & 1.29740 & 1.42603 \end{bmatrix}$$

$$\boldsymbol{K} = \begin{bmatrix} 1.00000 & -0.84067 & -0.83915 \\ -0.84067 & 1.00000 & 0.99999 \\ -0.83915 & 0.99999 & 1.00000 \end{bmatrix}.$$

The total l_2-sensitivity defined by (14a) was found to be

$$S = J(\boldsymbol{I}_3) + 1.59797 \times 10^4$$

where $J(\boldsymbol{I}_3)$ defined by (16b) was found to be

$$J(\boldsymbol{I}_3) = 9.87319 \times 10^8.$$

We see that the sensitivity associated with $\boldsymbol{H}_2(z_2)$, namely $J(\boldsymbol{I}_3)$, was indeed dominating the filter's sensitivity.

A. Application of the Lagrange method

Choosing $\boldsymbol{P}^{(0)} = \boldsymbol{I}_3$ in (28) as an intial estimate and a tolerance $\epsilon = 10^{-8}$ in the bisection method of (29a) as well as in (30), it took the Lagrange-based algorithm 7 iterations to converge to the solution

$$\boldsymbol{P}^{opt} = \begin{bmatrix} 2.26693 & -2.29606 & -2.28843 \\ -2.29606 & 3.27352 & 3.26762 \\ -2.28843 & 3.26762 & 3.26176 \end{bmatrix}.$$

With $\boldsymbol{K}$ given above and $\boldsymbol{P} = \boldsymbol{P}^{opt}$, an orthogonal matrix $\boldsymbol{U}$ satisfying (22) was found to be

$$\boldsymbol{U} = \begin{bmatrix} 0.91281 & 0.01760 & -0.40801 \\ 0.37727 & 0.34617 & 0.85897 \\ -0.15636 & 0.93801 & -0.30934 \end{bmatrix}$$

which in conjunction with (21) yielded the optimal coordinate transformation matrix as

$$\boldsymbol{T}^{opt} = \begin{bmatrix} 0.95919 & -0.79515 & -0.84535 \\ -0.31867 & 1.52484 & 0.92024 \\ -0.31499 & 1.52445 & 0.91574 \end{bmatrix}$$

and the optimal state-space coefficient matrices in (15) were constructed as

$$\overline{\boldsymbol{A}}_2 = \begin{bmatrix} 0.62960 & -0.10557 & -0.29017 \\ -0.03842 & 0.56172 & -0.29573 \\ 0.40269 & 0.09232 & 0.62479 \end{bmatrix}$$

$$\overline{\boldsymbol{B}}_2 = \begin{bmatrix} 0.36268 & -0.55634 & 0.50511 & -0.05885 \\ -0.12255 & 0.18883 & -0.23407 & 0.02035 \\ -0.13464 & 0.20306 & 0.02528 & 0.01847 \end{bmatrix}$$

$$\overline{\boldsymbol{C}}_2 = \begin{bmatrix} 0.06690 & -0.34372 & 0.24644 \\ 0.01682 & -0.08626 & 0.06888 \\ 0.05258 & -0.27075 & 0.20585 \\ 0.06038 & -0.31154 & 0.25590 \end{bmatrix}.$$

The controllability Gramian associated with the optimized state-space model was found to be

$$\overline{\boldsymbol{K}} = \begin{bmatrix} 1.00041 & -0.63416 & 0.63416 \\ -0.63416 & 1.00041 & -0.89767 \\ 0.63416 & -0.89767 & 1.00041 \end{bmatrix}$$

showing that the l_2-scaling constraints in (19) were practically satisfied. Then the l_2-sensitivity measure in (24) was computed as

$$J_o(\boldsymbol{P}^{opt}, \lambda) = 3.24252 \times 10^3$$

where $\lambda = 8.42179 \times 10^2$. The profiles of the l_2-sensitivity measure $J_o(\mathbf{P}, \lambda)$ and the Lagrange multiplier λ during the first 7 iterations are shown in Fig. 2, from which it is observed that with a tolerance $\epsilon = 10^{-8}$ the algorithm converges in 7 iterations.

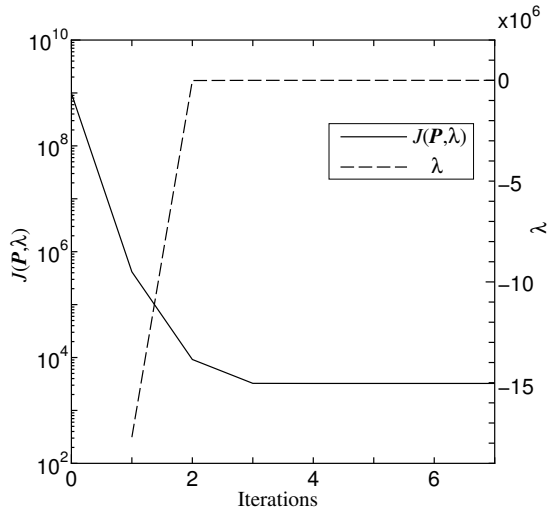

Fig. 2. Profiles of l_2-sensitivity and λ performances during the first 7 iterations of the constrained optimization method.

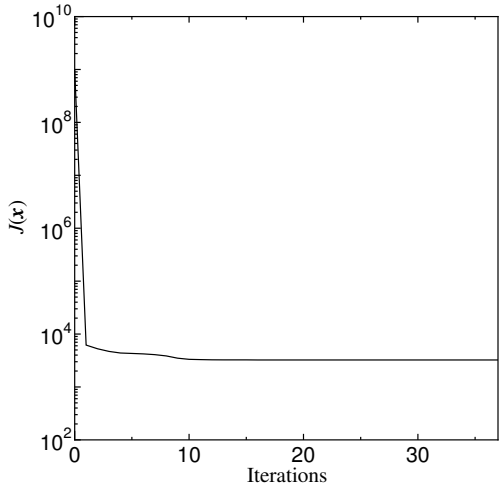

Fig. 3. Profile of l_2-sensitivity performance during the first 37 iterations of the unconstrained optimization method.

B. Application of the quasi-Newton method

By choosing $\hat{\boldsymbol{T}} = \boldsymbol{K}^{-\frac{1}{2}}$ with $\boldsymbol{T} = \boldsymbol{I}_3$ as an initial estimate in (36a) and a tolerance $\epsilon = 10^{-8}$ in (37), the quasi-Newton algorithm took 37 iterations to converge to the solution

$$\hat{\boldsymbol{T}}^{opt} = \begin{bmatrix} 1.03339 & 1.41754 & -0.13472 \\ -0.29813 & 2.46592 & -0.04995 \\ 0.90070 & -1.17579 & 0.96141 \end{bmatrix}$$

or equivalently,

$$\boldsymbol{T}^{opt} = \begin{bmatrix} 0.34976 & -1.20354 & 0.83486 \\ 0.42464 & 1.68766 & -0.49626 \\ 0.42613 & 1.68528 & -0.49116 \end{bmatrix}.$$

The l_2-sensitivity measure in (35a) was then computed as

$$J(\boldsymbol{x}) = 3.24356 \times 10^3$$

and the optimal state-space coefficient matrices in (15) were constructed as

$$\overline{\boldsymbol{A}}_2 = \begin{bmatrix} 0.61448 & -0.17729 & 0.32325 \\ -0.03747 & 0.57287 & 0.17307 \\ -0.43314 & -0.03545 & 0.62876 \end{bmatrix}$$

$$\overline{\boldsymbol{B}}_2 = \begin{bmatrix} 0.24441 & -0.37603 & 0.40596 & -0.04083 \\ -0.25567 & 0.39234 & -0.37591 & 0.04130 \\ 0.19877 & -0.30151 & 0.06567 & -0.02906 \end{bmatrix}$$

$$\overline{\boldsymbol{C}}_2 = \begin{bmatrix} 0.07181 & -0.29561 & -0.30148 \\ 0.02071 & -0.07290 & -0.08203 \\ 0.06091 & -0.23067 & -0.24815 \\ 0.07718 & -0.26188 & -0.30285 \end{bmatrix}.$$

The profile of the l_2-sensitivity measure $J(\boldsymbol{x})$ during the first 37 iterations is shown in Fig. 3, from which it is observed that with a tolerance $\epsilon = 10^{-8}$ the algorithm converges in 37 iterations.

We see that as far as the example discrete system is concerned, the two algorithms offer practically the same performance in terms of sensitivity minimization. Concerning the computational complexity, the 7 iterations required by the constrained algorithm consumed 46.01250 seconds of CPU time versus 7.15158×10^3 seconds for the unconstrained algorithm to finish its 37 iterations. Obviously the constrained method is considerably more efficient than its unconstrained counterpart. However, for discrete system 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 sensitivity minimization for a wide range of system scales.

C. Comparisons with the work in [25]

The Lagrange method was also applied to the 3-D filter investigated in Section 4 of [25]. We shall omit the intermediate details but focus on comparisons of the results obtained. The total l_2-sensitivity defined by (14a) with (16b) after l_2-scaling was found to be $S = J(\boldsymbol{I}_3) + 21.89505 \times 10^2 = 112.71349 \times 10^2$ where $J(\boldsymbol{I}_3) = 90.81844 \times 10^2$. After 6 iterations, the Lagrange method converges to a solution $\boldsymbol{P}^{opt}$ with $J(\boldsymbol{P}^{opt}) = 9.14487 \times 10^2$, which implies a total l_2-sensitivity $S = 31.03992 \times 10^2$. At this solution, the dynamic-range constraints $(\overline{\boldsymbol{K}})_{ii} = 1$ for $i = 1, 2, 3$ are perfectly satisfied. The total number of nontrivial parameters in its realization was 55. Alternatively, the algorithm proposed in [25] was applied, which led to a minimized total l_2-sensitivity 3.12274×10^2. The Gramian $\overline{\boldsymbol{K}}$ associated with the optimal state-space filter was found to be

$$\overline{\boldsymbol{K}} = \begin{bmatrix} 1.76972 & 1.14927 & 0.77867 \\ 1.14927 & 1.21704 & 0.94657 \\ 0.77867 & 0.94657 & 2.25271 \end{bmatrix}.$$

Clearly, the dynamic-range constraints $(\overline{\boldsymbol{K}})_{ii} = 1$ are violated for all $i = 1, 2, 3$ in this case. The number of nontrivial parameters involved in this realization was 73. From above results, we see two options for 3-D state-space filters with separable

denominators: [25] offers techniques for minimum sensitivity but with more system parameters and possible violation of dynamic-range constraints, while the present paper provides techniques for low sensitivity, reduced number of parameters (hence faster implementation), and guaranteed freedom of internal overflow.

V. CONCLUSION

A minimal state-space realization technique for 3-D separable-denominator digital filters has been explored and the l_2-sensitivity of the minimal state-space model realized from a given 3-D separable-denominator digital filter has been analyzed more precisely by taking into account 0 and 1 elements contained in the model. Two iterative methods have been developed to minimize the l_2-sensitivity of the filter subject to a fixed number of nontrivial parameters and dynamic-range l_2-scaling constraints. Computer simulation results have demonstrated the validity and effectiveness of the proposed techniques.

A problem that might be worthwhile for future studies is to minimize coefficient sensitivity subject to filter's robust stability under l_2-scaling constraints. In such a problem, the results reported in [24] are expected to play a significant role.

APPENDIX I
MINIMAL REALIZATION OF $\boldsymbol{H}_2(z_2)$

We write $\boldsymbol{H}_2(z_2)$ in (2b) as

$$\boldsymbol{H}_2(z_2)=\boldsymbol{\Delta}_0+\frac{(\boldsymbol{\Delta}_1-b_{21}\boldsymbol{\Delta}_0)z_2^{-1}+\cdots+(\boldsymbol{\Delta}_{N_2}-b_{2N_2}\boldsymbol{\Delta}_0)z_2^{-N_2}}{D_2(z_2)} \tag{A.1}$$

which can be realized with a multivariable observable canonic form $(\boldsymbol{A}_0,\boldsymbol{B}_0,\boldsymbol{C}_0,\boldsymbol{\Delta}_0)_{N_2(N_1+1)}$ as

$$\boldsymbol{A}_0=\begin{bmatrix} \boldsymbol{0} & \cdots & \boldsymbol{0} & -b_{2N_2}\boldsymbol{I}_{N_1+1} \\ \boldsymbol{I}_{N_1+1} & \cdots & \boldsymbol{0} & \vdots \\ \vdots & \ddots & \vdots & -b_{22}\boldsymbol{I}_{N_1+1} \\ \boldsymbol{0} & \cdots & \boldsymbol{I}_{N_1+1} & -b_{21}\boldsymbol{I}_{N_1+1} \end{bmatrix}$$

$$\boldsymbol{B}_0=\begin{bmatrix} \boldsymbol{\Delta}_{N_2}-b_{2N_2}\boldsymbol{\Delta}_0 \\ \vdots \\ \boldsymbol{\Delta}_2-b_{22}\boldsymbol{\Delta}_0 \\ \boldsymbol{\Delta}_1-b_{21}\boldsymbol{\Delta}_0 \end{bmatrix}, \quad \boldsymbol{C}_0=[\boldsymbol{0} \quad \cdots \quad \boldsymbol{0} \quad \boldsymbol{I}_{N_1+1}]. \tag{A.2}$$

Although this realization is always observable, it is uncontrollable unless $rank\,\boldsymbol{V}_{N_2(N_1+1)-r}=N_2(N_1+1)$, i.e. full rank, where $\boldsymbol{V}_i=\left[\boldsymbol{B}_0,\boldsymbol{A}_0\boldsymbol{B}_0,\cdots,\boldsymbol{A}_0^i\boldsymbol{B}_0\right]$ and r is the rank of $\boldsymbol{B}_0$ [27]. Suppose that

$$rank\,\boldsymbol{V}_{N_2(N_1+1)-1}=p \tag{A.3}$$

and $\boldsymbol{v}_1,\boldsymbol{v}_2,\cdots,\boldsymbol{v}_p$ are p linearly independent column vectors from matrix $\boldsymbol{V}_{N_2(N_1+1)-1}$. By defining $\boldsymbol{M}=[\boldsymbol{v}_1,\boldsymbol{v}_2,\cdots,\boldsymbol{v}_p]$, it is shown [32] that $\boldsymbol{H}_2(z_2)$ can be realized with a minimal state-space model $(\boldsymbol{A}_2,\boldsymbol{B}_2,\boldsymbol{C}_2,\boldsymbol{\Delta}_0)_p$ as

$$\begin{aligned} &\boldsymbol{A}_2=(\boldsymbol{M}^T\boldsymbol{M})^{-1}\boldsymbol{M}^T\boldsymbol{A}_0\boldsymbol{M}, \quad \boldsymbol{B}_2=(\boldsymbol{M}^T\boldsymbol{M})^{-1}\boldsymbol{M}^T\boldsymbol{B}_0 \\ &\boldsymbol{C}_2=\boldsymbol{C}_0\boldsymbol{M}. \end{aligned} \tag{A.4}$$

APPENDIX II
ERROR IN TRANSFER FUNCTION DUE TO PARAMETER VARIATIONS

Consider a transfer function $H(z_1,z_2,z_3)$ 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 $\tilde{H}(z_1,z_2,z_3)$ be the transfer function associated with perturbed parameters $\{\tilde{p}_i\}$. The first-order approximation of $\tilde{H}(z_1,z_2,z_3)$ then gives

$$\tilde{H}(z_1,z_2,z_3)=H(z_1,z_2,z_3)+\Delta H(z_1,z_2,z_3) \tag{A.5a}$$

where $\Delta H(z_1,z_2,z_3)$ will be

$$\Delta H(z_1,z_2,z_3)=\sum_{i=1}^{N}\frac{\partial H(z_1,z_2,z_3)}{\partial p_i}\Delta p_i. \tag{A.5b}$$

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

$$\sigma^2_{\Delta H}=\frac{1}{(2\pi j)^3}\oint_{|z_1|=1}\oint_{|z_2|=1}\oint_{|z_3|=1}E[|\Delta H(z_1,z_2,z_3)|^2]\cdot\frac{dz_1}{z_1}\frac{dz_2}{z_2}\frac{dz_3}{z_3} \tag{A.6}$$

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

$$E[|\Delta H(z_1,z_2,z_3)|^2]=\sum_{i=1}^{N}\left|\frac{\partial H(z_1,z_2,z_3)}{\partial p_i}\right|^2\sigma^2 \tag{A.7}$$

where $\sigma^2=E[(\Delta p_i)^2]=2^{-2B}/12$. Eq. (A.7) establishes an analytic relationship between variations in the transfer function induced by an FWL realization and parameter sensitivity.

APPENDIX III
l_2-SCALING CONSTRAINTS

Suppose that the input and the output of a 2-D filter $\boldsymbol{f}(z_2,z_3)$ are denoted by $\boldsymbol{x}(k,r)$ and $u(k,r)$, respectively. It follows from (9b) and (5a) that

$$\boldsymbol{x}(k,r)=\sum_{p=1}^{k}\sum_{q=0}^{r}\boldsymbol{h}(p,q)u(k-p,r-q) \tag{A.8a}$$

where

$$\begin{aligned} &\boldsymbol{h}(p,q)=\boldsymbol{A}_2^{p-1}\boldsymbol{B}_2\boldsymbol{C}_3\boldsymbol{A}_3^{q-1}\hat{\boldsymbol{e}}_{N_3} \quad \text{for} \quad p,q\geq 1 \\ &\boldsymbol{h}(p,0)=\boldsymbol{A}_2^{p-1}\boldsymbol{B}_2\hat{\boldsymbol{e}}_1 \quad \text{for} \quad p\geq 1. \end{aligned} \tag{A.8b}$$

Let $\boldsymbol{e}_i$ be the ith column of the identity matrix $\boldsymbol{I}_p$, an upper bound of the ith component of the local state vector $\boldsymbol{x}(k,r)$

in (A.8a) can be obtained as

$$\begin{aligned}|\boldsymbol{e}_i^T\boldsymbol{x}(k,r)|^2 &= \left[\sum_{p=1}^{k}\sum_{q=0}^{r}\boldsymbol{e}_i^T\boldsymbol{h}(p,q)u(k-p,r-q)\right]^2\\ &\le \boldsymbol{e}_i^T\left(\sum_{p=1}^{k}\sum_{q=0}^{r}\boldsymbol{h}(p,q)\boldsymbol{h}^T(p,q)\right)\boldsymbol{e}_i\\ &\quad\cdot\sum_{p=1}^{k}\sum_{q=0}^{r}u^2(k-p,r-q)\le \boldsymbol{e}_i^T\boldsymbol{K}\boldsymbol{e}_i\,||u||^2\end{aligned} \tag{A.9a}$$

where

$$\boldsymbol{K}=\sum_{p=1}^{\infty}\sum_{q=0}^{\infty}\boldsymbol{h}(p,q)\boldsymbol{h}^T(p,q). \tag{A.9b}$$

Note that $\boldsymbol{e}_i^T\boldsymbol{K}\boldsymbol{e}_i$ is the ith diagonal element of $\boldsymbol{K}$, thus if all the diagonal elements of $\boldsymbol{K}$ are equal to unity, (A.9a) implies that the amplitude of each component of vector $\boldsymbol{x}(k,r)$ will not exceed $||u||$. Therefore, the dynamic-range constraints on the state-variables may be imposed as

$$(\boldsymbol{K})_{ii}=1 \quad \text{for} \quad i=1,2,\cdots,p. \tag{A.10}$$

The above constraints correspond to (19) in the new realization and matrix $\boldsymbol{K}$ in (A.9b) coincides with that derived from (18).

APPENDIX IV
GRADIENT OF $J(\boldsymbol{x})$

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

$$\frac{\partial J(\hat{\boldsymbol{T}})}{\partial t_{ij}}=\lim_{\Delta\to 0}\frac{J(\hat{\boldsymbol{T}}_{ij})-J(\hat{\boldsymbol{T}})}{\Delta}=2\beta_1-2\beta_2+2\beta_3-2\beta_4 \tag{A.11a}$$

where $\hat{\boldsymbol{T}}_{ij}$ is the matrix obtained from $\hat{\boldsymbol{T}}$ with a perturbed (i,j)th component, which is given by [33, p.655]

$$\begin{aligned}\hat{\boldsymbol{T}}_{ij}&=\hat{\boldsymbol{T}}+\frac{\Delta\hat{\boldsymbol{T}}\boldsymbol{g}_{ij}\boldsymbol{e}_j^T\hat{\boldsymbol{T}}}{1-\Delta\boldsymbol{e}_j^T\hat{\boldsymbol{T}}\boldsymbol{g}_{ij}},\quad \hat{\boldsymbol{T}}_{ij}^{-1}=\hat{\boldsymbol{T}}^{-1}-\Delta\boldsymbol{g}_{ij}\boldsymbol{e}_j^T\\ \boldsymbol{g}_{ij}&=\partial\left\{\frac{\boldsymbol{t}_j}{||\boldsymbol{t}_j||}\right\}/\partial t_{ij}=\tfrac{1}{||\boldsymbol{t}_j||^3}(t_{ij}\boldsymbol{t}_j-||\boldsymbol{t}_j||^2\boldsymbol{e}_i)\\ \beta_1&=\boldsymbol{e}_j^T\hat{\boldsymbol{T}}\hat{\boldsymbol{M}}_A(\hat{\boldsymbol{T}})\hat{\boldsymbol{T}}^T\hat{\boldsymbol{T}}\boldsymbol{g}_{ij},\ \beta_3=\boldsymbol{e}_j^T\hat{\boldsymbol{T}}\hat{\boldsymbol{W}}_B\hat{\boldsymbol{T}}^T\hat{\boldsymbol{T}}\boldsymbol{g}_{ij}\\ \beta_2&=\boldsymbol{e}_j^T\hat{\boldsymbol{T}}^{-T}\hat{\boldsymbol{N}}_A(\hat{\boldsymbol{T}})\boldsymbol{g}_{ij},\qquad \beta_4=\boldsymbol{e}_j^T\hat{\boldsymbol{T}}^{-T}\hat{\boldsymbol{K}}_C\boldsymbol{g}_{ij}\\ \hat{\boldsymbol{N}}_A(\hat{\boldsymbol{T}})&=\sum_{i=0}^{\infty}\sum_{j=0}^{\infty}\begin{bmatrix}\boldsymbol{K}^{-\frac{1}{2}} & \boldsymbol{0}\end{bmatrix}\hat{\boldsymbol{N}}_{ij}^A\begin{bmatrix}\boldsymbol{K}^{-\frac{1}{2}}\\ \boldsymbol{0}\end{bmatrix}\\ \hat{\boldsymbol{N}}_{ij}^A&=\begin{bmatrix}\boldsymbol{A}_2 & \boldsymbol{B}_2\boldsymbol{R}_{ij}\boldsymbol{C}_2\\ \boldsymbol{0} & \boldsymbol{A}_2\end{bmatrix}\hat{\boldsymbol{N}}_{ij}^A\begin{bmatrix}\boldsymbol{A}_2 & \boldsymbol{B}_2\boldsymbol{R}_{ij}\boldsymbol{C}_2\\ \boldsymbol{0} & \boldsymbol{A}_2\end{bmatrix}^T\\ &\quad+\begin{bmatrix}\boldsymbol{0} & \boldsymbol{0}\\ \boldsymbol{0} & \boldsymbol{K}^{\frac{1}{2}}\hat{\boldsymbol{T}}^T\hat{\boldsymbol{T}}\boldsymbol{K}^{\frac{1}{2}}\end{bmatrix}.\end{aligned} \tag{A.11b}$$

REFERENCES

[1] 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.

[2] 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.

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

[4] C. Xiao, "Improved L_2-sensitivity for state-space digital system," *IEEE Trans. Signal Processing*, vol. 45, pp. 837-840, Apr. 1997.

[5] 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.

[6] S. Yamaki, M. Abe and M. Kawamata, "A closed form solution to L_2-sensitivity minimization of second-order state-space digital filters," in *Proc. 2006 IEEE Int. Symp. Circuits Syst.*, pp. 5223-5226.

[7] G. Li, "On frequency weighted minimal L_2 sensitivity of 2-D systems using Fornasini-Marchesini LSS model", *IEEE Trans. Circuits Syst. I*, vol. 44, pp. 642-646, July 1997.

[8] G. Li, "Two-dimensional system optimal realizations with L_2-sensitivity minimization," *IEEE Trans. Signal Processing*, vol. 46, pp. 809-813, Mar. 1998.

[9] T. Hinamoto and Y. Sugie, "L_2-sensitivity analysis and minimization of 2-D separable-denominator state-space digital filters," *IEEE Trans. Signal Processing*, vol. 50, pp. 3107-3114, Dec. 2002.

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

[11] 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 Processing*, vol. 54, pp. 1804-1812, May 2006.

[12] S. Yamaki, M. Abe and M. Kawamata, "A novel approach to L_2-sensitivity minimization of digital filters subject to L_2-scaling constraints," in *Proc. 2006 IEEE Int. Symp. Circuits Syst.*, pp. 5219-5222.

[13] T. Hinamoto, T. Oumi, O. I. Omoifo, and W.-S. Lu, "Minimization of frequency-weighted l_2-sensitivity subject to l_2-scaling constraints for two-dimensional state-space filters," *IEEE Trans. Signal Processing*, vol. 56, pp. 5157-5168, Oct. 2008.

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

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

[16] K. Hirano, M. Sakane and M. Z. Mulk, "Design of three-dimensional recursive digital filters," *IEEE Trans. Circuits Syst.*, vol. CAS-31, pp. 550-561, June 1984.

[17] H. Mutluay and M. M. Fahmy, "Frequency-domain design of N-D digital filters," *IEEE Trans. Circuits Syst.*, vol. CAS-32, pp. 1226-1233, Dec. 1985.

[18] M. E. Zervakis and A. N. Venetsanopoulos, "Design of three-dimensional digital filters using two-dimensional rotated filters," *IEEE Trans. Circuits Syst.*, vol. CAS-34, pp. 1452-1469, Dec. 1987.

[19] H. K. Kwan and C. L. Chan, "Design of multidimensional spherically symmetric and constant group delay recursive digital filters with sum of power-of-two coefficients," *IEEE Trans. Circuits Syst.*, vol. CAS-37, pp. 1027-1035, Aug. 1990.

[20] P. K. Rajan and M. N. S. Swamy, "Quadrantal symmetry associated with two-dimensional digital transfer functions," *IEEE Trans. Circuits Syst.*, vol. 25, pp. 340-343, June 1983.

[21] L. Xu, H. Fan, Z. Lin and N. K. Bose, "A direct-construction approach to multidimensional realization and LFR uncertainty modeling," *Multidimensional Systems and Signal Processing*, vol. 19, pp. 323-359, Dec. 2008.

[22] B. Sumanasena and P. H. Bauer, "Realization using the Roesser model for implementations in distributed grid sensor networks," *Multidimensional Systems and Signal Processing*, vol. 22, pp. 131-146, Mar. 2011.

[23] T. Zhou, "Boundedness of multidimensional filters over a prescribed frequency domain," *IEEE Trans. Signal Processing*, vol.56, pp. 5487-5499, Nov. 2008

[24] T. Zhou, "Stability and stability margin for a two-dimensional system," *IEEE Trans. Signal Processing*, vol. 54, pp. 3483-3488, Sept. 2006.

[25] T. Hinamoto, Y. Sugie, A. Doi and M. Muneyasu, "Synthesis of 3-D separable-denominator state-space digital filters with minimum L_2-sensitivity," *Multidimensional Systems and Signal Processing*, vol. 15, pp. 147-167, Apr. 2004.

[26] R. J. Ober, X. Lai, Z. Lin and E. S. Ward, "State-space realization of a three-dimensional image set with application to nise reduction of fluo-rescent microscopy images of cells," *Multidimensional Systems and Signal Processing*, vol. 16, pp. 7-48, Jan. 2005.
[27] C. T. Chen, Introduction to Linear System Theory, Holt, Rinehart and Winston, Inc., 1970.
[28] L. L. Scharf, Statistical Signal Processing, Reading, MA: Addison-Wesley, 1991.
[29] H. Togawa, Handbook of Numerical Methods, Saiensu-sha, Tokyo, 1992.
[30] R. Fletcher, Practical Methods of Optimization, 2nd ed., Wiley, New York, 1987.
[31] J. E. Dennis and J. J. More, "Quasi-Newton methods, motivation and theory," *SIAM Review*, vol. 19, pp. 46-89, 1977.
[32] R. E. Kalman, "Mathematical description of linear dynamical systems," *SIAM J. Contrl (ser. A)*, vol. 1, pp. 152-192, 1963.
[33] T. Kailath, *Linear System*, Englewood Cliffs, N.J.: Prentice-Hall, 1980.