DESIGN OF TWO-DIMENSIONAL DIGITAL FILTERS USING SINGULAR-VALUE DECOMPOSITION

by

Hui Ping Wang

B.Eng., 1984 Beijing Institute of Telecommunications
M.A.Sc., 1987 University of Victoria

A DISSERTATION SUBMITTED IN PARTIAL FULFILLMENT
OF THE REQUIREMENTS FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY
in the Department of
Electrical and Computer Engineering

We accept this dissertation as conforming
to the required standard

DATE 9/12/9

Dr. W.-S. Lu, Co-supervisor, Dept. of Elect. & Comp. Eng.

Dr. A. Antoniou, Co-supervisor, Dept. of Elect. & Comp. Eng.

Dr. P. Agathoklis, Department member, Dept. of Elect. & Comp. Eng.

Dr. R. Vahldieck, Department member, Dept. of Elect. & Comp. Eng.

Dr. D. E. Hewgill, Outside Member, Dept. of Mathematics

Dr. H. A. Muller, Outside Member, Dept. of Comp. Science

Dr. M. Ahmadi, External Examiner, University of Windsor

Co-supervisors: Dr. W.-S. Lu and Dr. A. Antoniou

ABSTRACT

This thesis presents a study on the design of two-dimensional (2-D) digital filters by using the singular-value decomposition (SVD).

A new method for the design of 2-D quadrantally symmetric FIR filters with linear phase response is proposed. It is shown that three realizations are possible, namely, a direct realization, a modified version of the direct realization, and a realization based on the combined application of the SV and LU decompositions. Each of the three realizations consists of a parallel arrangement of cascaded pairs of 1-D filters; hence extensive parallel processing and pipelining can be applied. The three realizations are compared and it is shown that the realization based on the SV and LU decompositions leads to the lowest approximation error and involves the smallest number of multiplications.

It is shown that the SVD of the sampled amplitude response of a 2-D digital filter with real coefficients possesses a special structure: every singular vector is either mirror-image symmetric or antisymmetric with respect to its midpoint. Consequently, the SVD method can be applied along with 1-D FIR techniques for the design of linear-phase 2-D filters with arbitrary prescribed amplitude responses which are symmetrical with respect to the origin of the (ω_1, ω_2) plane.

A method for the design of 2-D IIR digital filters based on the combined application of the SVD and the balanced approximation (BA) is proposed. It is shown that the approximation error in the phase angle is bounded by the sum of the neglected Hankel singular values of the filter. Consequently, the phase response of the resulting filter is approximately linear over the passband region provided that only small Hankel singular values are neglected. It is also shown that the resulting 2-D filter is nearly balanced, which implies that the filter has low roundoff noise as well as low parameter sensitivity. Furthermore, the 2-D filter obtained is more economical and computationally more efficient than the

original 2-D FIR filter, and in the case where an IIR filter is obtained the stability of the filter is guaranteed.

Efficient general algorithms for the evaluation of the 1-D and 2-D gramians for 1-D and 2-D, causal, stable, recursive digital filters are proposed, which facilitate the application of the BA method in the design of digital filters. The algorithms obtained are based on a two-stage extension of the Åström-Jury-Agniel (ÅJA) algorithm. It is shown that the ÅJA algorithm can be modified to solve a 1-D Lyapunov equation in a recursive manner. The recursive algorithm is then extended to the case where the rational function vector involved depends on two complex variables. It is shown that the two algorithms obtained can be combined to evaluate the 2-D gramians. The proposed algorithms are also useful in obtaining optimal digital filter structures that minimize the output-noise power due to the roundoff of products.

Examiners:

Dr. W.-S. Lu, Co-supervisor, Dept. of Elect. & Comp. Eng.

Dr. A. Antoniou, Co-supervisor, Dept. of Elect. & Comp. Eng.

Dr. P. Agathoklis, Department Member, Dept. of Elect. & Comp. Eng.

Dr. R. Vahldieck, Department Member, Dept. of Elect. & Comp. Eng.

Dr. D. E. Hewgill, Outside Member, Dept. of Mathematics

Dr. H. A. Muller, Outside Member, Dept. of Comp. Science

Dr. M. Ahmadi, External Examiner, University of Windsor

Contents

List of Tables

List of Figures

List of Abbreviations

ÅJA Åström-Jury-Agniel

BA balanced approximation

DSP digital signal processing

FIR finite impulse response

IIR infinite impulse response

SVD singular-value decomposition

1-D one-dimensional

2-D two-dimensional

ACKNOWLEDGEMENTS

The author wishes to express her sincere gratitude to her supervisors, Dr. Wu-Sheng Lu and Dr Andreas Antoniou for their constant guidance, support, and encouragement during the course of this work and the writting of this manuscript. Without their help, this dissertation would not have been written.

The author also wishes to thank faculty and staff in the Department of Electrical and Computer Engineering for their assistance during the course of her study in this department.

Finally, the author would like to thank the Natural Sciences and Engineering Research Council of Canada for providing financial support in the form of a research assistantship and the University of Victoria for a fellowship.

To my parents and sister

Chapter 1

Introduction

1.1 Background

Multidimensional (M-D) digital signal processing (DSP) is primarily concerned with the representation, transformation and manipulation of signals that can be represented as M-D arrays. M-D DSP, particularly two-dimensional (2-D) DSP, finds extensive applications in acoustics, sonar, radar, seismology, geophysical exploration, robotics and many other areas. In many cases, the central part of a 2-D DSP system is a specific piece of software or a dedicated hardware board implementing a filtering algorithm that can process signals received, and is referred to in general as a 2-D digital filter. This thesis presents a study on the design of 2-D digital filters by using the singular-value decomposition (SVD).

Two-dimensional digital filters can be classified as recursive or nonrecursive depending on whether or not the output of the filter depends on previous values of the output. Alternatively, they can be classified as infinite-impulse response (IIR) or finite-impulsive response (FIR) filters depending on whether their impulse response is of infinite or finite duration. These types of 2-D digital filters are consistent with their 1-D counterparts and have analogous properties. FIR filters have the advantages that they are free of stability problems and that linear phase can easily be achieved. IIR filters, on the other hand, have the advantage that the amount of computation necessary for their operation and the required memory

resources are relatively low.

1.2 Existing Methods for the Design of 2-D Filters

Methods for the design of 2-D FIR and IIR filters have been investigated by a number of researchers during the past two decades. The main design approaches for 2-D FIR and IIR filters can be classified into the following four categories:

1.2.1 Window Method

In this method, an ideal frequency response $H(\omega_1, \omega_2)$ is approximated by an FIR filter by multiplying the ideal impulse response $h(n_1, n_2)$ by a finite window array $w(n_1, n_2)$ to produce the filter impulse response.

$$h_w(n_1, n_2) = h(n_1, n_2)w(n_1, n_2) \tag{1.1}$$

The frequency response of the resulting filter will be a good approximation to $H(\omega_1, \omega_2)$ if $W(\omega_1, \omega_2)$, the frequency spectrum of $w(n_1, n_2)$, is a good approximation to a 2-D impulse function.

Huang [1] first described the circularly symmetric window formulation. The window used has a circular region of support and is formed as

$$\omega_C(n_1, n_2) = \omega(\sqrt{n_1^2 + n_2^2}) \tag{1.2}$$

by sampling a rotated 1-D continuous window function in the 2-D plane. Another type of 2-D window function has a rectangular region of support and can be formed as the outer product of two 1-D windows [2] i.e.

$$\omega_R(n_1, n_2) = \omega_1(n_1)\omega_2(n_2) \tag{1.3}$$

Several 1-D windows can be used in (1.2) and (1.3). Among the most popular ones are the rectangular, von Hann, Hamming, and Kaiser windows [3]. The

advantage of using window functions is that they require less computation than optimization techniques and they are effective in reducing Gibbs' oscillations.

1.2.2 McClellan Transformation Method

The McClellan transformation method [4-10] is used to design 2-D linear phase FIR filters either with circularly symmetric amplitude responses or with fan filter specifications. The advantage of the method is that by using optimal 1-D filters it is possible to design optimal 2-D filters that can be implementated efficiently. The method starts with a high-order 1-D filter design that satisfies certain frequency response specifications. The 1-D filter is then transformed into a 2-D filter using the McClellan transformation. To be more specific, the frequency response of a 1-D zero-phase FIR filter of length $(2N+1)$ is written as

$$H(\omega) = h(0) + \sum_{n=1}^{N} h(n)[e^{-j\omega n} + e^{j\omega n}] \tag{1.4}$$

$$= \sum_{n=0}^{N} a(n) \cos(n\omega) \tag{1.5}$$

where

$$a(n) = \begin{cases} h(0), & n = 0 \\ 2h(n), & n > 0 \end{cases}$$

and $h(n)$ is the impulse response of the filter. The function $\cos(n\omega)$ can be expressed as a polynomial of degree n in the variable $\cos\omega$. The resulting polynomial is the nth Chebyshev polynomial $T_n[\cdot]$. Therefore

$$\cos n\omega = T_n[\cos\omega] \tag{1.6}$$

By substituting (1.6) into (1.5), $H(\omega)$ can be written as

$$H(\omega) = \sum_{n=0}^{N} a(n) T_n[\cos\omega] \tag{1.7}$$

The generalized McClellan transformation which converts 1-D filters into 2-D filters possessing quadrantal symmetry can be written as

$$\cos(\omega) = \sum_{p=0}^{P}\sum_{q=0}^{Q} t_{pq} \cos p\omega_1 \cos q\omega_2 \tag{1.8}$$

where t_{pq} are real constants.

The frequency response of the resulting 2-D FIR filter is

$$H(\omega_1, \omega_2) = \sum_{n=0}^{N} a(n) \left[\sum_{p=0}^{P}\sum_{q=0}^{Q} t_{pq} \cos p\omega_1 \cos q\omega_2\right]^n \tag{1.9}$$

The coefficients of the original McClellan transformation are chosen as $t_{11} = t_{10} = t_{01} = -t_{00} = 1/2$ for circular symmetry. These coefficients result in nearly circular contours for low values of ω and increasingly square contours for larger values of ω. Thus the McClellan transformation is quite useful for the design of lowpass or highpass filters with a low cutoff radius. The McClellan transformation is not quite suitable for designing either lowpass filters with large cutoff frequency or broadband bandpass filters since it cannot provide circular contours at high frequencies.

The coefficients of the McClellan transformation are computed using optimization techniques [7, 8, 9]. The design usually requires a large amount of computation. Thus an approximate solution is sometimes used in order to reduce the computation burden. An approximate technique which results in simple formulas for fast calculation of the McClellan transformation coefficients has been described in [10].

1.2.3 Optimization Methods

An objective function can be defined in terms of the error between the actual and desired frequency responses. The filter coefficients can then be obtained by various optimization techniques [11, 12, 13, 14, 15, 16, 17, 18]. Filters designed

using different error criteria can be quite different. The most commonly used error criteria are the L_2 and the L_∞ norms. The design of IIR filters is generally more complicated than that for FIR filters since stability constraints must also be imposed in the former case.

1.2.4 SVD Method

The main advantage of the SVD method is that 2-D filter designs can be accomplished by designing a set of 1-D subfilters and, therefore, the many well-established techniques for the design of 1-D filters can be employed. In the following two subsections, the theory of the SVD and its previous applications in the design of 2-D filters will be given.

SVD of a Matrix

Theorem 1.1 If $\mathbf{A} \in R^{L\times M}$ is a matrix, then there are orthogonal matrices $\mathbf{U}$ and $\mathbf{V}$ such that

$$\mathbf{U}^T\mathbf{A}\mathbf{V} = \begin{pmatrix} \mathbf{\Sigma} & 0 \\ 0 & 0 \end{pmatrix} \tag{1.10}$$

where $\mathbf{\Sigma} = \text{diag}\,(\sigma_1, \sigma_2, \cdots, \sigma_r)$, and $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_r > 0$. A proof of this theorem can be found in Appendix A [19].

The decomposition in Theorem 1.1 is unique. From (1.10) we have

$$\mathbf{V}^T\mathbf{A}^T\mathbf{A}\mathbf{V} = \text{diag}\,(\mathbf{\Sigma}^2, 0) \tag{1.11}$$

Thus the numbers $\sigma_1^2, \sigma_2^2, \cdots, \sigma_r^2$ must be the nonzero eigenvalues of $\mathbf{A}^T\mathbf{A}$ arranged in descending order. This, along with the requirement that the σ_i be nonnegative, completely determines the σ_i.

Theorem 1.1 shows that any nonzero matrix $\mathbf{A}$ of rank r may be written as the product of three factors

$$\mathbf{A} = \mathbf{U}\begin{pmatrix} \mathbf{\Sigma} & 0 \\ 0 & 0 \end{pmatrix}\mathbf{V}^T$$

$$= \sum_{i=1}^{r} \sigma_i \mathbf{u}_i \mathbf{v}_i^T \tag{1.12}$$

where σ_i for $i = 1, 2, \cdots, r$ are the singular values of $\mathbf{A}$ with $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_r > 0$, $\mathbf{u}_i$ is the ith eigenvector of $\mathbf{A}\mathbf{A}^T$ associated with the ith eigenvalue σ_i^2, and $\mathbf{v}_i$ is the ith eigenvector of $\mathbf{A}^T\mathbf{A}$ associated with the ith eigenvalue σ_i^2.

An important property of the SVD can be stated in terms of the following theorem.

Theorem 1.2 Let

$$\mathbf{A} = \mathbf{U} \begin{pmatrix} \mathbf{\Sigma} & 0 \\ 0 & 0 \end{pmatrix} \mathbf{V}^T \quad \in R^{L \times M}$$

with $L \geq M$ where $\mathbf{\Sigma} = \text{diag}\ (\sigma_1, \sigma_2, \cdots, \sigma_M)$ and $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_M$. If $\mathbf{\Sigma}' = \text{diag}\ (\sigma_1, \sigma_2, \cdots, \sigma_r, 0, \cdots, 0)$ and

$$\mathbf{A}' = \mathbf{U} \begin{pmatrix} \mathbf{\Sigma}' & 0 \\ 0 & 0 \end{pmatrix} \mathbf{V}^T$$

then

$$\| \mathbf{A} - \mathbf{A}' \|_F = \min_{\text{rank}(\mathbf{B})=r} \| \mathbf{A} - \mathbf{B} \|_F \tag{1.13}$$

where

$$\| \mathbf{A} \|_F = \left[\sum_{l=1}^{L} \sum_{m=1}^{M} a_{lm}^2 \right]^{1/2}$$

is the Frobenius norm of a matrix $\mathbf{A}$. A proof of Theorem 2 can be found in Appendix A [19].

Theorem 1.2 shows that for any fixed k ($1 \leq k \leq r$), $\sum_{i=1}^{k} \sigma_i \mathbf{u}_i \mathbf{v}_i^T$ is a minimal mean-square-error (MMSE) approximation to $\mathbf{A}$. A special case of Theorem 1.2 is that the MMSE separable approximation to $\mathbf{A}$ is given by

$$\mathbf{A} \approx \sigma_1 \mathbf{u}_1 \mathbf{v}_1^T \tag{1.14}$$

where σ_1 is the dominant singular value, and $\mathbf{u}_1$ and $\mathbf{v}_1$ are the corresponding dominant singular vectors [20].

For most digital signal and image processing applications, matrix $\mathbf{A}$ is nonnegative. It has been shown that for a nonnegative matrix $\mathbf{A}$, all entries of $\mathbf{u}_1$ and $\mathbf{v}_1$ are nonnegative [20, 21], and this property has been utilized in early developments of the SVD design method.

Previous Work on SVD Design Method

The SVD design method was first considered in [22] where Treitel and Shanks used the SVD of a matrix formed by a given ideal impulse response to approximate a spatial nonrecursive filter by a sum of separable 2-D filters. In [20] the SVD method was extended to the frequency-domain design of separable approximations to desired amplitude responses. Assume that matrix $\mathbf{A} = \{a_{l,m}\}$ represents a desired amplitude response and

$$a_{l,m} = A(\frac{\pi\mu_l}{T_1}, \frac{\pi\nu_m}{T_2}), \quad 1 \le l \le L \text{ and } 1 \le m \le M \tag{1.15}$$

where μ_l and ν_m are normalized frequencies such that

$$\mu_l = \frac{l-1}{L-1}, \quad \nu_m = \frac{m-1}{M-1}$$

and $0 \le \mu_l \le 1$, $0 \le \nu_m \le 1$. Since the elements of the first pair of singular vectors $\tilde{\mathbf{v}}_1 = \sigma_1^{1/2}\mathbf{v}_1$ and $\tilde{\mathbf{u}}_1 = \sigma_1^{1/2}\mathbf{u}_1$ are always positive, the problem of designing a 2-D filter characterized by $\mathbf{A}$ is acomplished by designing two 1-D filters whose sampled amplitude responses are given by

$$f_l = |F(e^{j\pi\mu_l/T_1})|, \quad l = 1, 2, \cdots, L \tag{1.16}$$

$$g_m = |G(e^{j\pi\nu_m/T_2})|, \quad m = 1, 2, \cdots, M \tag{1.17}$$

This task can be carried out by using one of the standard optimization algorithms such as the Fletcher-Powell algorithm. In [20] the Marquardt minimization algorithm was used with an L_p error criterion to design both FIR and IIR filters. In the FIR case, the user specifies p and the order N of the desired 1-D filter and the optimization program yields a filter characterized by a transfer function

$$H(z) = K(1 + \sum_{n=1}^{N} a_n z^{-n}) \tag{1.18}$$

such that the error

$$E_p = \sum_{n=1}^{M} \{|H(e^{j\omega_n})| - d_n\}^p \tag{1.19}$$

is minimized, where d_n is the desired response of the 1-D filter at $\omega = \omega_n$, p is an even integer, and M is the number of sampling points. For the IIR case, a filter characterized by

$$H(z) = K \prod_{n=1}^{N} \frac{(1 + a_{n1}z^{-1} + a_{n2}z^{-2})}{(1 + b_{n1}z^{-1} + b_{n2}z^{-2})} \tag{1.20}$$

is obtained in the same way. The technique described in [20] has the limitation that if **A** has more than one dominant singular value, the approximation achieved is poor since only the first set of the singular vectors is utilized for the 2-D filter design. Therefore, the approximation error is relatively large. In [23] the SVD method is modified to include more than one set of singular vectors for the design of corresponding 1-D subfilters. Thus the 2-D design accuracy is increased. If 1-D IIR subfilters are used, zero phase is required for each subfilter. This necessitates data transpositions at the inputs and outputs of subfilters and, as a result, the usefulness of these designs is limited to nonreal-time applications, where the delay introduced in the processing is unimportant.

Each method described in the above sections has its advantages and disadvantages and significant improvements and extensions are still possible.

1.3 Background on Computation of Gramians

In practice, digital filters are implemented using finite-precision arithmetic and hence it becomes necessary to quantize products and coefficients. As a result, roundoff noise is introduced which varies significantly from one realization to another. Previous work [24] has shown that an optimal local state-space 2-D digital-filter realization that minimizes the output noise power due to the roundoff of products can be obtained. The key step in finding such optimal realization is the evaluation of two positive definite matrices $\mathbf{K}_2$ and $\mathbf{W}_2$ which are known as the controllability and the observability gramians of the digital filter, respectively. These gramians have also been used to obtain balanced approximations of 1-D and 2-D systems [25, 26], which are very useful in the design of 1-D and 2-D digital filters [27, 28].

The most efficient method for the evaluation of the gramians for the 1-D case is to solve two relevant Lyapunov equations, and reliable algorithms for computing the gramians are available in the literature [29, 30]. For the 2-D case, the corresponding Lyapunov equations depend on a complex parameter, as is demonstrated in [25, 31], which varies on the unit circle of a complex plane. In other words, if the Lyapunov approach is chosen for the evaluation of the 2-D gramians, one needs to solve a family of 1-D Lyapunov equations as opposed to two constant Lyapunov equations in the 1-D case. A Lyapunov approach for the 2-D case was described in [32] but the transfer function of the digital filters under consideration must have separable denominators. For general, 2-D, causal, stable, recursive digital filters, the most commonly used method is the truncation method described in [31] which provides numerical approximations of the gramians in terms of truncated double summations to guarantee acceptable numerical error. This is particularly the case when the filter under consideration has small stability margin in which case the convergence of the infinite series is rather slow. Therefore, an efficient and general method for the evaluation of gramians for the

case of 2-D, causal, stable, recursive digital filters is needed.

1.4 Contributions of This Thesis

The main contributions of this thesis can be summarized as follows:

1. A new design method based on the SVD is proposed for the design of 2-D FIR filters with linear phase response. It is shown that three realizations are possible, namely, a direct realization, a modified version of the direct realization, and a realization based on the combined application of the SV and LU decompositions. Each of the three realizations consists of a parallel arrangement of cascaded pairs of 1-D filters; hence extensive parallel processing and pipelining can be applied. The three realizations are compared and it is shown that the realization based on the SV and LU decompositions, leads to the lowest approximation error and involves the smallest number of multiplications.

2. A new design method based on the balanced approximation (BA) for the design of 2-D IIR digital filter is proposed. It is shown that, the approximation error in the phase angle is bounded by the sum of the neglected Hankel singular values of the filter. Consequently, the phase response of the resulting filter is approximately linear over the passband region provided that only small Hankel singular values are neglected. It is also shown that the resulting 2-D filter is nearly balanced, which implies that the filter has low roundoff noise as well as low parameter sensitivity [32]. Furthermore, the 2-D filter obtained is more economical and computationally more efficient than the original 2-D FIR filter, and in the case where an IIR filter is obtained the stability of the filter is guaranteed.

3. An efficient method for the evaluation of the controllability and observability gramians of 2-D digital filters is proposed. The algorithms obtained are based on a two-stage extension of the Åstrom-Jury-Agniel (ÅJA) algorithm

which was originally used for the evaluation of the scalar loss function of a stationary random process with rational spectral density [33, 34]. It is shown that the ÅJA algorithm can be modified to solve a 1-D Lyapunov equation in a recursive manner. The recursive algorithm is then extended to the case where the vector rational function involved depends on two complex variables. It is shown that the two algorithms obtained can be combined to evaluate the 2-D gramians. The proposed algorithms are useful in obtaining optimal digital filter structures that minimize the output-noise power due to the roundoff of products [35, 36], and in obtaining a balanced approximation of a given discrete-time dynamical system [37] or digital filter [27, 28].

1.5 Organization of This Thesis

In Chapter 2, the new SVD method for the design of 2-D quadrantally symmetric FIR digital filters is presented. The SVD, the McClellan transformation, and the 2-D window methods are used to design a bandpass and a fan filter, and the results obtained are compared.

In Chapter 3, a method for the design of 2-D FIR filters by using the SVD is presented. It is shown that the SVD of the sampled amplitude response of a 2-D digital filter with real coefficients possesses a special structure: every singular vector is either mirror-image symmetric or anti-symmetric with respect to its midpoint. Consequently, the SVD method can be applied along with 1-D FIR techniques for the design of linear-phase 2-D filters with arbitrary prescribed amplitude responses which are symmetrical with respect to the origin of the (ω_1, ω_2) plane.

In Chapter 4, a design method using the well-known balanced approximation [26, 27, 28, 38] is applied to linear-phase 2-D FIR filters of the type that may be obtained by using the SVD methods presented in Chapter 2 and 3. The BA

method leads to a lower-order separable 2-D filter, usually an IIR filter. It is shown that the designs obtained are causal and locally quasi-balanced, and in cases where IIR designs are obtained stability is quaranteed.

In Chapter 5, an efficient general method for the evaluation of the 1-D and 2-D gramians for 1-D and 2-D, causal, stable, recursive digital filters is presented. The proposed method is compared with other known methods for the evaluation of the 2-D gramians with respect to accuracy and computational efficiency.

Chapter 2

Design of Quadrantally Symmetric 2-D FIR Digital Filters by Using the SVD

2.1 Introduction

The design of 2-D digital filters by using the SVD and other similar decompositions has been investigated by a number of researchers [20, 22, 23, 28, 39, 40] [41, 42]. This design approach has several advantages. First, the design can be accomplished by designing a set of 1-D subfilters and, therefore, the many well established techniques for the design of 1-D filters can be employed; second, the resulting 2-D filter is stable if the 1-D subfilters employed are stable; and third, the 1-D subfilters form a parallel structure which allows extensive parallel processing, hence the structure obtained is suitable for VLSI implementation. As pointed out in [23], the SVD approach can be used for the design of either infinite-impulse response (IIR) or finite-impulse response (FIR) 2-D filters. While high selectivity can be achieved by using low-order IIR designs for the parallel 1-D subfilters, zero phase is required for each subfilter. This necessitates data transpositions at the inputs and outputs of subfilters and, as a result, the usefulness of these designs is limited to nonreal-time applications, where the delay introduced in the processing is unimportant. On the other hand, by using higher

order FIR designs for the parallel 1-D subfilters, high selectivity and linear phase 2-D FIR filter can be achieved.

In this chapter, the SVD method is applied in conjunction with 1-D FIR design techniques for the design of 2-D quandrantally symmetric FIR filters. It is shown that by using linear-phase 1-D filters, linear-phase causal 2-D filters can be designed which are suitable for real-time or quasi-real-time applications.

In Section 2.2 some preliminary material regarding 2-D digital filters is presented. In Section 2.3 the design of 2-D quadrantally symmetric FIR filters by the SVD method in conjunction with 1-D FIR techniques is described. In Section 2.4 three realizations schemes are proposed for the design of 2-D quadrantally symmetric FIR filters. It will be shown that in all three realization methods, the outcome is a 2-D causal, linear-phase, parallel filter. In Section 2.5 an error analysis is presented for the SVD design method. This would facilitate the determination of the number of singular values that should be used in the design and the maximum approximation error that should be achieved in the design of the 1-D filters. In Section 2.6 two examples are included to illustrate the effectiveness of the proposed design method and the results obtained are compared with those obtained by using the 2-D window and the McClellan transformation methods.

2.2 Preliminaries

A 2-D linear FIR digital filter with support in the rectangle defined by $-N_i/2 \leq n_i \leq N_i/2$, $i = 1,\ 2$, can be characterized by the transfer function

$$H(z_1,\ z_2) = \sum_{n_1=-N_1/2}^{N_1/2} \sum_{n_2=-N_2/2}^{N_2/2} h(n_1,\ n_2) z_1^{-n_1} z_2^{-n_2} \tag{2.1}$$

where $h(n_1,\ n_2)$ is its impulse response. If the filter is causal, then we have

$$H(z_1,\ z_2) = \sum_{n_1=0}^{N_1} \sum_{n_2=0}^{N_2} h(n_1,\ n_2) z_1^{-n_1} z_2^{-n_2} \tag{2.2}$$

Similarly, a linear, causal, recursive 2-D IIR digital filter can be characterized by

$$H(z_1,\ z_2) = \frac{\sum_{n_1=0}^{N_1}\ \sum_{n_2=0}^{N_2} a(n_1,\ n_2) z_1^{-n_1} z_2^{-n_2}}{\sum_{n_1=0}^{N_1}\ \sum_{n_2=0}^{N_2} b(n_1,\ n_2) z_1^{-n_1} z_2^{-n_2}} \tag{2.3}$$

If $z_i = e^{j\omega_i T_i}$ where $T_i = 2\pi/\omega_{si}$ for $i = 1,\ 2$ are sampling periods and ω_{si} are the sampling frequencies, we have

$$H(e^{j\omega_1 T_1},\ e^{j\omega_2 T_2}) = M(\omega_1,\ \omega_2) e^{j\theta(\omega_1,\ \omega_2)} \tag{2.4}$$

where

$$M(\omega_1,\ \omega_2) = |H(e^{j\omega_1 T_1},\ e^{j\omega_2 T_2})| \tag{2.5}$$

and

$$\theta(\omega_1,\ \omega_2) = \arg\ [H(e^{j\omega_1 T_1},\ e^{j\omega_2 T_2})] \tag{2.6}$$

$M(\omega_1,\ \omega_2)$ and $\theta(\omega_1,\ \omega_2)$ are the amplitude and phase responses of the filter, respectively.

For some digital signal applications such as image processing, it is important to have distortionless transmission. If an input signal is represented by $x(n_1,\ n_2)$, distortionless signal transmission is achieved if

$$y(n_1,\ n_2) = \alpha x(n_1 - n_{01},\ n_2 - n_{02}) \tag{2.7}$$

where α is constant and n_{01} and n_{02} are integers. This means that the output of the linear system is a scaled and shifted version of the input signal. By using the z transform, (2.7) can be expressed as

$$Y(e^{j\omega_1 T_1},\ e^{j\omega_2 T_2}) = \alpha X(e^{j\omega_1 T_1},\ e^{j\omega_2 T_2}) e^{-j(\omega_1 T_1 n_{01} + \omega_2 T_2 n_{02})} \tag{2.8}$$

Therefore the amplitude and the phase responses of a distortionless system can be written as

$$|H(e^{j\omega_1 T_1},\ e^{j\omega_2 T_2})| = \alpha \tag{2.9}$$

$$\arg\ [H(e^{j\omega_1 T_1},\ e^{j\omega_2 T_2})] = -\omega_1 T_1 n_{01} - \omega_2 T_2 n_{02} \tag{2.10}$$

respectively. Equations (2.9) and (2.10) show that if the signal is to be transmitted without distortion the amplitude response of the linear system should be constant and the phase response should be linear over those frequencies where $X(e^{j\omega_1 T_1},\ e^{j\omega_2 T_2})$ is nonzero. The linear phase requirement can also be specified by group delay functions defined by

$$\tau_1(\omega_1) = -\frac{\partial \arg\,[H(e^{j\omega_1 T_1},\ e^{j\omega_2 T_2})]}{\partial \omega_1} \tag{2.11}$$

and

$$\tau_2(\omega_2) = -\frac{\partial \arg\,[H(e^{j\omega_1 T_1},\ e^{j\omega_2 T_2})]}{\partial \omega_2} \tag{2.12}$$

A linear system is distortionless with respect to phase response, if its associated group delays with respect to ω_1 and ω_2 are constants over frequencies where $X(e^{j\omega_1 T_1},\ e^{j\omega_2 T_2})$ is nonzero. Consequently, in the design of 2-D filters, the amplitude response and group delays are usually required to be constant in the filter's passband.

2.3 Design

A quandrantally symmetric 2-D FIR filter requires that the impulse response $h(n_1,\ n_2)$ of (2.1) satisfy

$$h(n_1,\ n_2) = h(n_1,\ -n_2) = h(-n_1,\ n_2) = h(-n_1,\ -n_2)$$

Further, if $h(n_1,\ n_2)$ is real, then the frequency response of the filter

$$\begin{aligned} H(e^{j\omega_1 T_1},\ e^{j\omega_2 T_2}) &= \sum_{n_1=-N_1/2}^{N_1/2} \sum_{n_2=-N_2/2}^{N_2/2} h(n_1,\ n_2) e^{-j\omega_1 n_1 T_1} e^{-j\omega_2 n_2 T_2} \\ &= X(\omega_1,\ \omega_2) \end{aligned} \tag{2.13}$$

is a real function which is even with respect to ω_1 and ω_2.

The transfer function given in (2.1) can be rewritten as

$$H(z_1,\ z_2) = \sum_{i=1}^{K} F_i(z_1)G_i(z_2) \tag{2.14}$$

where $F_i(z_1)$ and $G_i(z_2)$ are transfer functions of 1-D subfilters in the z_1 and z_2 domains, respectively. If these subfilters are FIR filters with support in the rectangle defined by $-N_i/2 \leq n_i \leq N_i/2$, $i = 1,\ 2$, we have

$$F_i(z_1) = \sum_{n_1=-N_1/2}^{N_1/2} f_i(n_1)z_1^{-n_1} \tag{2.15}$$

and

$$G_i(z_2) = \sum_{n_2=-N_2/2}^{N_2/2} g_i(n_2)z_2^{-n_2} \tag{2.16}$$

and if $F_i(z_1)$ and $G_i(z_2)$ are assumed to represent zero-phase filters, then their frequency responses are given by

$$\begin{aligned} F_i(e^{j\omega_1 T_1}) &= \sum_{n_1=-N_1/2}^{N_1/2} f_i(n_1)e^{-j\omega_1 n_1 T_1} \\ &= \Phi_i(\omega_1) \end{aligned} \tag{2.17}$$

and

$$\begin{aligned} G_i(e^{j\omega_2 T_2}) &= \sum_{n_2=-N_2/2}^{N_2/2} g_i(n_2)e^{-j\omega_2 n_2 T_2} \\ &= \Gamma_i(\omega_2) \end{aligned} \tag{2.18}$$

where $\Phi_i(\omega_1)$ and $\Gamma_i(\omega_2)$ are real functions which are even with respect to ω_1 and ω_2, respectively. Consequently, from (2.14)-(2.18) the frequency response for a

quadrantally symmetric 2-D filter can be written as

$$\begin{aligned} H(e^{j\omega_1 T_1},\ e^{j\omega_2 T_2}) &= \sum_{i=1}^{K} F_i(e^{j\omega_1 T_1}) G_i(e^{j\omega_2 T_2}) \\ &= \sum_{i=1}^{K} \Phi_i(\omega_1)\Gamma_i(\omega_2) \\ &= A(\omega_1,\ \omega_2) \end{aligned} \tag{2.19}$$

Now assume that matrix $\mathbf{A} = \{a_{l,m}\}$ represents a desired frequency response, i.e.

$$a_{l,m} = A(\frac{\pi\mu_l}{T_1},\ \frac{\pi\nu_m}{T_2}), \quad 1 \le l \le L \quad \text{and} \quad 1 \le m \le M \tag{2.20}$$

where μ_l and ν_m are normalized frequencies such that

$$\mu_l = \frac{l-1}{L-1}, \quad \nu_m = \frac{m-1}{M-1}$$

and $0 \le \mu_l \le 1$, $0 \le \nu_m \le 1$. The SVD of $\mathbf{A}$ gives

$$\begin{aligned} \mathbf{A} &= \sum_{i=1}^{r} \sigma_i \mathbf{u}_i \mathbf{v}_i^T \\ &= \sum_{i=1}^{r} \tilde{\mathbf{u}}_i \tilde{\mathbf{v}}_i^T \end{aligned} \tag{2.21}$$

where σ_i are the singular values of $\mathbf{A}$ such that $\sigma_1 \ge \sigma_2 \ge \cdots \ge \sigma_r > 0$, $\tilde{\mathbf{u}}_i = \sigma_i^{\frac{1}{2}} \mathbf{u}_i$, and $\tilde{\mathbf{v}}_i = \sigma_i^{\frac{1}{2}} \mathbf{v}_i$.

On comparing (2.19) with (2.21) and assuming that $\tilde{\mathbf{u}}_i$ and $\tilde{\mathbf{v}}_i$ are sampled versions of the frequency responses of the 1-D filters characterized by $F_i(z_1)$ and $G_i(z_2)$, respectively, a 2-D zero-phase FIR filter can be designed by designing two sets of 1-D zero-phase FIR subfilters characterized by $F_i(z_1)$ and $G_i(z_2)$, where $1 \le i \le r$. A 2-D, causal, linear-phase filter can readily be obtained by

shifting the impulse response by $N_1/2$ and $N_2/2$ with respect to the n_1 axis and n_2 axis, respectively. This can be accomplished by multiplying $F_i(z_1)$ and $G_i(z_2)$ by $z_1^{-N_1/2}$ and $z_2^{-N_2/2}$, respectively.

The design of the 2-D filter can be completed by using any one of the standard methods for the design of 1-D FIR filters. Using the Frequency sampling method in conjunction with window techniques [3], designs can be obtained very quickly with a small amount of computational effort. These designs are not optimal although the approximation error can be made arbitrarily small by increasing the order of the 1-D filters used. On the other hand, by using methods based on the Remez algorithm [43], it may be possible to obtain optimal designs although a large amount of computation would be required.

2.4 SVD Realizations

In this section, three distinct realizations are obtained through the application of the SVD. They are referred to as the direct-SVD realization, the modified-SVD realization, and the SVD-LUD realization.

2.4.1 Direct-SVD Realization

The direct-SVD realization is based on the direct application of the SVD to a matrix representing the required amplitude response and as in the IIR realizations reported in [23], the number of parallel filter sections is equal to the number of singular values used in the design. In practice, a designer would attempt to keep the error introduced by the application of the SVD as low as possible by using as many singular values as possible in the design. However, as more and more singular values are used, more and more parallel sections are required which increase the computational complexity or the cost of hardware involved in the implementation.

The direct-SVD realization is obtained by using the design method described

in Section 2.3. The steps involved are as follows:

1) Specify the desired amplitude response and thereby obtain the corresponding sampled amplitude-response matrix **A** given in (2.20).

2) Decompose matrix **A** using the SVD as in (2.21) to get $\tilde{\mathbf{u}}_i$ and $\tilde{\mathbf{v}}_i$, where $1 \leq i \leq r$.

3) Design K 2-D FIR filters, each of which is obtained by designing two 1-D zero-phase FIR filters characterized by transfer functions $F_i(z_1)$ and $G_i(z_2)$ corresponding to the desired amplitude responses $\tilde{\mathbf{u}}_i$ and $\tilde{\mathbf{v}}_i$, respectively, where $1 \leq i \leq K$ and $1 \leq K \leq r$.

4) Obtain the resulting 2-D zero-phase transfer function through (2.14).

5) Multiply the resulting 2-D zero-phase transfer function by $z_1^{-N_1/2}$ and $z_2^{-N_2/2}$ to obtain the linear-phase 2-D transfer function.

If $N_1 = N_2 = N$, the number of the multiplications needed in the direct SVD realization is $2K\bar{N}$ where

$$\bar{N} = \begin{cases} (N+1)/2 & \text{if } N \text{ is odd} \\ N/2+1 & \text{if } N \text{ is even} \end{cases}$$

is the number of multiplications needed in the implementation of a 1-D FIR filter. Integer K is the number of singular values that must be used in the design to reduce the approximation error introduced by the SVD to a satisfactory level.

2.4.2 Modified-SVD Realization

The second realization is a modified version of the first which takes advantage of the symmetry of the impulse response of a linear-phase 2-D FIR filter. The design is reformulated in terms of the singular values of a matrix **C** of rank r_c, where r_c does not exceed half of the order of the 1-D FIR filters used. This allows the

designer to eliminate the error introduced by the application of the SVD using a smaller number of parallel sections.

The modified-SVD realization can be obtained by manipulating the transfer function of the 2-D filter. From (2.14) to (2.18), we can write

$$\begin{aligned} H(z_1,\ z_2) &= \sum_{i=1}^{K}\Big[\sum_{n_1=-N/2}^{N/2} f_i(n_1)z_1^{-n_1}\Big]\Big[\sum_{n_2=-N/2}^{N/2} g_i(n_2)z_2^{-n_2}\Big] \\ &= \sum_{n_1=-N/2}^{N/2}\sum_{n_2=-N/2}^{N/2}\Big[\sum_{i=1}^{K} f_i(n_1)g_i(n_2)\Big]z_1^{-n_1}z_2^{-n_2} \\ &= \sum_{n_1=-N/2}^{N/2}\sum_{n_2=-N/2}^{N/2} c(n_1,\ n_2)z_1^{-n_1}z_2^{-n_2} \end{aligned} \tag{2.22}$$

where

$$c(n_1,\ n_2) = \sum_{i=1}^{K} f_i(n_1)g_i(n_2) \tag{2.23}$$

The SVD of matrix $\mathbf{C} = \{c(n_1,\ n_2)\}$ can be written as

$$\begin{aligned} \mathbf{C} &= \sum_{i=1}^{r_c} \sigma_{ci}\mathbf{u}_{ci}\mathbf{v}_{ci}^T \\ &= \sum_{i=1}^{r_c} \tilde{\mathbf{u}}_{ci}\tilde{\mathbf{v}}_{ci}^T \end{aligned} \tag{2.24}$$

where r_c is the rank of $\mathbf{C}$. Now if all the singular values of $\mathbf{C}$ for $i > K_c$ can be neglected, we can write

$$\mathbf{C} \approx \sum_{i=1}^{K_c} \tilde{\mathbf{u}}_{ci}\tilde{\mathbf{v}}_{ci}^T \tag{2.25}$$

and on combining (2.22) and (2.25), we have

$$H(z_1,\ z_2) \approx \sum_{n_1=-N/2}^{N/2}\sum_{n_2=-N/2}^{N/2}\sum_{i=1}^{K_c} \check{u}_{ci}(n_1)\tilde{v}_{ci}(n_2)z_1^{-n_1}z_2^{-n_2}$$

$$\approx \sum_{i=1}^{K_c} \tilde{F}_{ci}(z_1)\tilde{G}_{ci}(z_2) = \hat{H}(z_1,\ z_2) \tag{2.26}$$

where

$$\tilde{F}_{ci}(z_1) = \sum_{n_1=-N/2}^{N/2} \tilde{u}_{ci}(n_1) z_1^{-n_1}$$

and

$$\tilde{G}_{ci}(z_2) = \sum_{n_2=-N/2}^{N/2} \tilde{v}_{ci}(n_2) z_2^{-n_2}$$

Therefore, a realization can be obtained by connecting K_c $(1 \leq K_c \leq r_c)$ 2-D zero-phase filters in parallel, where each 2-D filter consists of two cascaded 1-D FIR filters represented by $\tilde{F}_{ci}(z_1)$ and $\tilde{G}_{ci}(z_2)$.

A step-by-step procedure for the modified-SVD realization starts with steps (1)-(4) given in section 2.4.1 and concludes with the following additional steps:

5) Form coefficient matrix $\mathbf{C}$ using (2.23).

6) Perform the SVD on matrix $\mathbf{C}$ to obtain vectors $\tilde{\mathbf{u}}_{ci}$ and $\tilde{\mathbf{v}}_{ci}$, $1 \leq i \leq r_c$, as in (2.24) and retain the vectors corresponding to the K_c most significant singular values.

7) Obtain the 2-D zero-phase transfer function using (2.26).

From (2.23)

$$\mathbf{C} = \sum_{i=1}^{K} \mathbf{f}_i \mathbf{g}_i^T = [\mathbf{f}_1 \cdots \mathbf{f}_K][\mathbf{g}_1 \cdots \mathbf{g}_K]^T$$

where

$$\mathbf{f}_i = [f_i(-N/2) \cdots f_i(N/2)]^T$$

and

$$\mathbf{g}_i = [g_i(-N/2) \cdots g_i(N/2)]^T$$

Hence it follows that the rank of $\mathbf{C}$ satisfies the inequality $r_c \leq K$. Moreover, since $\mathbf{f}_i$ and $\mathbf{g}_i$ are mirror-image symmetric, matrix $\mathbf{C}$ is quadrantally symmetric (see (A.1) in Appendix A). Consequently, there are at most N linearly independent row (or column) vectors in $\mathbf{C}$ and, therefore

$$r_c \leq \min\,(\bar{N},\ K) \tag{2.27}$$

In the modified-SVD realization, vectors $\tilde{\mathbf{u}}_{ci}$ and $\tilde{\mathbf{v}}_{ci}$ in (2.24) are all mirror-image symmetric, as shown in Appendix A. Consequently, $\tilde{F}_{ci}(z_1)$ and $\tilde{G}_{ci}(z_2)$ represent 1-D zero-phase FIR filters which can be readily designed using one of the standard methods. A 2-D, causal, linear-phase realization can be obtained by multiplying $\tilde{F}_{ci}(z_1)$ and $\tilde{G}_{ci}(z_2)$ by $z_1^{-N/2}$ and $z_2^{-N/2}$, respectively.

The number of multiplications required by the modified-SVD realization is $2K_c\bar{N}$. If $K < \bar{N}$ then from (2.27) we have $K_c \leq r_c \leq K$ and, therefore, the modified-SVD realization is more economical than the direct-SVD realization. If $K > \bar{N}$, we have $K_c \leq r_c \leq \bar{N}$ and, once again, the modified-SVD realization is more economical than the direct-SVD realization. In the later case, the modified-SVD realization has the additional advantage that the value of K can be increased to r, the number of singular values in $\mathbf{A}$, without increasing the number of multiplications. Consequently, the approximation error can be reduced further at no additional cost.

2.4.3 SVD-LUD Realization

The SVD-LUD realization is, in effect, a modified version of the second realization. As in the second method, the error introduced by the application of the SVD can be reduced by increasing the number of parallel sections and, in addition, the number of multiplications can be reduced further through the use of the LU decomposition.

Instead of decomposing $\mathbf{C}$ using the SVD, the LUD is used [44] to give

$$\mathbf{C} = \mathbf{L}_c \mathbf{U}_c \tag{2.28}$$

where $\mathbf{L}_c$ and $\mathbf{U}_c \in \mathbf{R}^{N \times N}$ are the lower- and upper-triangular matrices, respectively. Since matrix $\mathbf{C}$ given by (2.23) is quadrantally symmetric, it can be shown that $\mathbf{L}_c = \{l_{i,j}\}$ and $\mathbf{U}_c = \{u_{i,j}\}$ satisfy the relations

$$\begin{aligned} l_{i,j} &= l_{N-i+1,j} \quad \text{for } 1 \le i,\ j \le N \\ l_{i,j} &= 0 \quad \text{for } j > r_c \\ l_{i,j} &= 0 \quad \text{for } i < j \text{ and } j \le r_c \end{aligned}$$

and

$$\begin{aligned} u_{i,j} &= u_{i,N-j+1} \quad \text{for } 1 \le i,\ j \le N \\ u_{i,j} &= 0 \quad \text{for } i > r_c \\ u_{i,j} &= 0 \quad \text{for } j < i \text{ and } i \le r_c \end{aligned}$$

respectively, i.e. $\mathbf{L}_c$ and $\mathbf{U}_c$ have the forms of

$$\mathbf{L}_c = \begin{bmatrix} * & 0 & & & \cdots & 0 \\ * & * & 0 & & \cdots & 0 \\ * & \cdots & * & 0 & \cdots & 0 \\ \vdots & & \vdots & \vdots & & \vdots \\ * & \cdots & * & 0 & \cdots & 0 \\ * & * & 0 & & \cdots & 0 \\ * & 0 & & & \cdots & 0 \end{bmatrix}$$

and

$$\mathbf{U}_c = \begin{bmatrix} * & * & * & \cdots & * & * & * \\ 0 & * & \vdots & & \vdots & * & 0 \\ & 0 & * & \cdots & * & 0 & \\ & & 0 & \cdots & 0 & & \\ \vdots & \vdots & \vdots & & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & 0 & 0 & 0 \end{bmatrix}$$

resectively, (see Appendix B) where the nonzero columns (rows) in $\mathbf{L}_c$ ($\mathbf{U}_c$) are also mirror-image symmetric.

Now if we let $\mathbf{Z}_i = [z_i^{N/2}, \cdots, 1 \cdots, z_i^{-N/2}]$. $i = 1, 2$, then (2.26) can be written as

$$\begin{aligned} \hat{H}(z_1, z_2) &= \mathbf{Z}_1 \mathbf{L}_c \mathbf{U}_c \mathbf{Z}_2^T \\ &= \sum_{i=1}^{K_c} L_{ci}(z_1) U_{ci}(z_2) \end{aligned} \tag{2.29}$$

where $1 \leq K_c \leq r_c$ and

$$L_{ci}(z_1) = \sum_{n_1=i-(N/2+1)}^{(N/2+1)-i} L_c(n_1, i) z_1^{-n_1}$$

and

$$U_{ci}(z_2) = \sum_{n_2=i-(N/2+1)}^{(N/2+1)-i} U_c(i, n_2) z_2^{-n_2}$$

A 2-D, causal, linear-phase realization can be obtained by multiplying $L_{ci}(z_1)$ and $U_{ci}(z_2)$ by $z_1^{-N/2}$ and $z_2^{-N/2}$, respectively.

A step-by-step procedure for the SVD-LUD realization starts with steps (1)-(4) given in Section 2.4.1 and concludes with the following additional steps:

5) Form coefficient matrix $\mathbf{C}$ using (2.23).

6) Perform the LUD on matrix $\mathbf{C}$ to obtain matrices $\mathbf{L}_c$ and $\mathbf{U}_c$ by (2.28).

7) Obtain the 2-D zero-phase transfer function through (2.29).

As in the modified-SVD realization, full accuracy can be achieved by using r_c parallel sections along with $K = r$, according to (2.27). The number of multiplications required in this realization is $K_c(2\bar{N} - K_c + 1)$, where $1 \leq K_c \leq r_c$,

Table 2.1: Number of Multiplications Required by the Realization Schemes

Realizations	Direct SVD	Modified SVD	SVD-LUD
No. of Multipli.	$2K\bar{N}$	$2K_c\bar{N}$	$K_c(2\bar{N} - K_c + 1)$
Upper Bound	$2r\bar{N}$	$2\bar{N}^2$	$\bar{N}(\bar{N}+1)$

which is always less than $2K_c\bar{N}$, the number of multiplications required in the modified-SVD method. Also note that by (2.27), $K_c(2\bar{N} - K_c + 1)$ has an upper bound $\bar{N}(\bar{N}+1)$ which is less than $2\bar{N}^2$, the upper bound for the modified SVD method. Consequently, the SVD-LUD method usually leads to the most economical realization. The numbers of multiplications required by the three realization schemes are given in Table 2.1.

2.5 Error Analysis

As was demonstrated in Section 2.4, in the direct-SVD realization a number of singular values of **A** and in the modified-SVD and SVD-LUD realizations a number of singular values of **A** and/or **C** may be neglected in practice. In this section, a quantitative error analysis is undertaken which would facilitate the determination of the number of singular values that should be used in the design and the maximum approximation error that should be achieved in the design of the 1-D filters.

2.5.1 Direct-SVD Realization

Assume that the direct-SVD method has been used to design a 2-D FIR filter for a desired frequency response **A** and that its transfer function $H(z_1, z_2)$ is given by (2.14). Let $\tilde{\mathbf{f}}_i$ and $\tilde{\mathbf{g}}_i$ be the column vectors obtained by evaluating $\Phi_i(\omega_1)$ and $\Gamma_i(\omega_2)$ at frequencies $\omega_1 = \pi\mu_l/T_1$ and $\omega_2 = \pi\nu_m/T_2$, where $1 \leq l \leq L$ and

$1 \leq m \leq M$, respectively, i.e.

$$\tilde{\mathbf{f}}_i = [\Phi_i(\frac{\pi\mu_1}{T_1}) \cdots \Phi_i(\frac{\pi\mu_L}{T_1})]^T$$

$$\tilde{\mathbf{g}}_i = [\Gamma_i(\frac{\pi\nu_1}{T_2}) \cdots \Gamma_i(\frac{\pi\nu_M}{T_2})]^T$$

The amplitude response of the 2-D filter at frequency point $(\omega_1, \omega_2) = (\pi\mu_l/T_1, \pi\nu_m/T_2)$ is given by $\sum_1^K \tilde{\mathbf{f}}_i\tilde{\mathbf{g}}_i^T$ and, therefore, the approximation error at this frequency point is the (l,m) entry in the error matrix $\mathbf{E}$ defined by

$$\begin{aligned}
\mathbf{E} &= \{e_{l,m}\} \\
&= \sum_{i=1}^{K} \tilde{\mathbf{f}}_i\tilde{\mathbf{g}}_i^T - \mathbf{A} \\
&= \sum_{i=1}^{K} (\tilde{\mathbf{f}}_i\tilde{\mathbf{g}}_i^T - \tilde{\mathbf{u}}_i\tilde{\mathbf{v}}_i^T) - \sum_{i=K+1}^{r} \sigma_i\mathbf{u}_i\mathbf{v}_i^T
\end{aligned} \tag{2.30}$$

If we define

$$\Delta\tilde{\mathbf{u}}_i = \tilde{\mathbf{f}}_i - \tilde{\mathbf{u}}_i, \quad \Delta\tilde{\mathbf{v}}_i = \tilde{\mathbf{g}}_i - \tilde{\mathbf{v}}_i$$

then $\Delta\tilde{\mathbf{u}}_i$ and $\Delta\tilde{\mathbf{v}}_i$ represent the approximation errors in the 1-D frequency responses $\Phi_i(\omega_1)$ and $\Gamma_i(\omega_2)$, respectively. From (2.30), we can write

$$\begin{aligned}
\mathbf{E} &= \sum_{i=1}^{K} (\Delta\tilde{\mathbf{u}}_i\tilde{\mathbf{v}}_i^T + \tilde{\mathbf{u}}_i\Delta\tilde{\mathbf{v}}_i^T + \Delta\tilde{\mathbf{u}}_i\Delta\tilde{\mathbf{v}}_i^T) - \sum_{i=K+1}^{r} \sigma_i\mathbf{u}_i\mathbf{v}_i^T \\
&= \sum_{i=1}^{K} [\sigma_i^{1/2}(\Delta\tilde{\mathbf{u}}_i\mathbf{v}_i^T + \mathbf{u}_i\Delta\tilde{\mathbf{v}}_i^T) + \Delta\tilde{\mathbf{u}}_i\Delta\tilde{\mathbf{v}}_i^T] - \sum_{i=K+1}^{r} \sigma_i\mathbf{u}_i\mathbf{v}_i^T
\end{aligned}$$

The (l,m) entry of matrix $\mathbf{E}$ can be expressed as

$$e_{l,m} = [p_1\ p_2\ \cdots\ p_L]\ \mathbf{E}\ [q_1\ q_2\ \cdots\ q_M]^T$$

where

$$p_i = \begin{cases} 1 & \text{for } i = l \\ 0 & \text{otherwise} \end{cases}$$

and

$$q_i = \begin{cases} 1 & \text{for } i = m \\ 0 & \text{otherwise} \end{cases}$$

and hence

$$e_{l,m} = \sum_{i=1}^{K} [\sigma_i^{1/2} (\Delta\tilde{u}_{il} v_{im} + u_{il} \Delta\tilde{v}_{im}) + \Delta\tilde{u}_{il} \Delta\tilde{v}_{im}] - \sum_{i=K+1}^{r} \sigma_i u_{il} v_{im}$$

where u_{il} denotes the lth component of vector $\mathbf{u}_i$, etc. Since vectors $\mathbf{u}_i$ and $\mathbf{v}_i$ are all unit vectors for $1 \leq i \leq r$, an upper bound on the magnitude of $e_{l,m}$ can be obtained as

$$|e_{l,m}| \leq \sum_{i=1}^{K} [\sigma_i^{1/2} (\epsilon_{1i} + \epsilon_{2i}) + \epsilon_{1i}\epsilon_{2i}] + \sum_{i=K+1}^{r} \sigma_i \tag{2.31}$$

where ϵ_{1i} and ϵ_{2i} represent the maximum approximation errors in the frequency responses $\Phi_i(\omega_1)$ and $\Gamma_i(\omega_2)$, respectively.

If the approximation error in each 1-D filter design is reasonably small, then the high-order terms $\epsilon_{1i}\epsilon_{2i}$, $1 \leq i \leq K$, in inequality (2.31) can be neglected. We note also that the right-hand side in (2.31) is independent of l and m, and hence (2.31) holds also for the maximum of $|e_{l,m}|$. We have, therefore, obtained an upper bound for the maximum approximation error over the set of sampled frequency points as

$$e_\infty = \max_{1 \leq l \leq L,\ 1 \leq m \leq M} |e_{l,m}| \leq \epsilon_{pK} + \epsilon_{rK} \tag{2.32}$$

where the principal error ϵ_{pK} and the residual error ϵ_{rK} are defined by

$$\epsilon_{pK} = \sum_{i=1}^{K} \sigma_i^{1/2} (\epsilon_{1i} + \epsilon_{2i}) \tag{2.33}$$

and

$$\epsilon_{rK} = \sum_{i=K+1}^{r} \sigma_i \tag{2.34}$$

respectively.

The error bound given by (2.32) shows clearly how the choice of K and the approximation error introduced by a specific 1-D design technique affect the overall approximation error in the 2-D filter. In practice, K should be chosen to keep the number of parallel sections small and the residual error ϵ_{rK} acceptable. Having determined the value of K, the principal error ϵ_{pK} can also be made acceptable by controlling the 1-D design errors ϵ_{1i} and ϵ_{2i} for $1 \leq i \leq K$, by increassing the orders of the 1-D filters or by using a better design method for the 1-D filters. From (2.33) it follows that small approximation error should be obtained in the design of those 1-D filters that correspond to the large singular values.

2.5.2 Modified-SVD and SVD-LUD Realizations

In this section, an error analysis is given for the modified-SVD and SVD-LUD realizations. As in the direct-SVD realization, a trade-off exists between the number of parallel sections used and the approximation error. An error analysis is, therefore, useful to the designer.

Assume that a 2-D FIR filter has been designed using K singular values of the desired frequency response $\mathbf{A}$ and that its transfer function $H(z_1,\ z_2)$ is given by (2.22). If a modified-SVD or a SVD-LUD realization is obtained comprising K_c parallel sections, where $1 \leq K_c \leq r_c$, the error introduced is given from (2.22) and (2.26) as

$$H(z_1,\ z_2) - \hat{H}(z_1,\ z_2) = \mathbf{Z}_1\Big(\sum_{i=K_c+1}^{r_c} \sigma_{ci} u_{ci} v_{ci}^T\Big)\mathbf{Z}_2^T \tag{2.35}$$

where $\mathbf{Z_i}$, $i = 1,\ 2$, are row vectors as defined in Section 2.4.3 and

$$\begin{aligned}\hat{H}(z_1,\ z_2) &= \mathbf{Z}_1(\sum_{i=1}^{K_c}\sigma_{ci}u_{ci}v_{ci}^T)\mathbf{Z}_2^T \\ &= \sum_{i=1}^{K_c}\tilde{F}_{ci}(z_1)\tilde{G}_{ci}(z_2)\end{aligned}$$

Under these circumstances, the approximation error at frequency point $(\omega_1,\ \omega_2) = (\pi\mu_l/T_1,\ \pi\nu_m/T_2)$ is given by

$$\begin{aligned}|\hat{e}_{l,\ m}| &= |a_{l,m} - \hat{H}(e^{j\omega_1 T_1},\ e^{j\omega_2 T_2})| \\ &= |a_{l,m} - H(e^{j\omega_1 T_1},\ e^{j\omega_2 T_2})| + |H(e^{j\omega_1 T_1},\ e^{j\omega_2 T_2}) - \hat{H}(e^{j\omega_1 T_1},\ e^{j\omega_2 T_2})| \\ &= |e_{l,m}| + |H(e^{j\omega_1 T_1},\ e^{j\omega_2 T_2}) - \hat{H}(e^{j\omega_1 T_1},\ e^{j\omega_2 T_2})| \qquad (2.36)\end{aligned}$$

where $|e_{l,m}|$ has an upper bound given by (2.32), the second term on the right-hand side in (2.36) can be estimated as

$$\begin{aligned}|H(e^{j\omega_1 T_1},\ e^{j\omega_2 T_2}) - \hat{H}(e^{j\omega_1 T_1},\ e^{j\omega_2 T_2})| &= |\mathbf{Z}_1(\sum_{i=K_c+1}^{r_c}\sigma_{ci}u_{ci}v_{ci}^T)\mathbf{Z}_2^T| \\ &\leq \|\mathbf{Z}_1\|\ \|\mathbf{Z}_2\|\ \|(\sum_{i=K_c+1}^{r_c}\sigma_{ci}u_{ci}v_{ci}^T)\| \\ &= N\sum_{i=K_c+1}^{r_c}\sigma_{ci} \qquad (2.37)\end{aligned}$$

where $\|\cdot\|$ denotes the Euclidean norm of the matrix involved. The estimate in (2.36) in conjunction with (2.32) and (2.37) leads to

$$\hat{e}_\infty = \max_{1\leq l\leq L,\ 1\leq m\leq M}|\hat{e}_{l,m}| \leq \epsilon_{pK} + \epsilon_{rK} + \epsilon_s \qquad (2.38)$$

where ϵ_{pK} and ϵ_{rK} are given by (2.33) and (2.34), respectively, and ϵ_s given by

$$\epsilon_s = N\sum_{i=K_c+1}^{r_c}\sigma_{ci} \qquad (2.39)$$

is the error introduced by using a reduced number of sections.

With $K = r$, the residual error in (2.38) vanishes and (2.38) becomes

$$\hat{e}_\infty \leq \epsilon_{pr} + \epsilon_s \tag{2.40}$$

In other words, one can design a linear-phase 2-D FIR digital filter using only K_c parallel sections if all the 1-D filters involved are designed such that the principal error ϵ_{pr} is sufficiently small and, if K_c is chosen such that the error in (2.39) is also sufficiently small.

2.6 Examples

In this section, the design of a bandpass and a fan 2-D filter are presented to illustrate the effectiveness of the proposed design method.

2.6.1 Two-Dimensional Bandpass FIR Filter

The desired amplitude response of a circularly symmetric 2-D bandpass FIR filter is specified by

$$|H_1(e^{j\omega_1 T_1},\ e^{j\omega_2 T_2})| = \begin{cases} 0 & 0 \leq \sqrt{{\omega_1}^2 + {\omega_2}^2} \leq \omega_{a_1} \\ 1 & \omega_{p_1} \leq \sqrt{{\omega_1}^2 + {\omega_2}^2} \leq \omega_{p_2} \\ 0 & \omega_{a_2} \leq \sqrt{{\omega_1}^2 + {\omega_2}^2} \leq \pi \end{cases}$$

where $\omega_{a_1} = 0.24\pi$, $\omega_{p_1} = 0.36\pi$, $\omega_{p_2} = 0.64\pi$ and $\omega_{a_2} = 0.76\pi$ and is illustrated in Figure 2.1. The corresponding sampled amplitude response $\mathbf{A_1} = |H_1(e^{j\pi\mu_l},\ e^{j\pi\nu_m})|$ can be expresed as

$$|H_1(e^{j\pi\mu_l},\ e^{j\pi\nu_m})| = \begin{cases} 0 & 0 \leq \sqrt{{\mu_l}^2 + {\nu_m}^2} < \omega_{c_1}/\pi \\ 1 & \omega_{c_1}/\pi \leq \sqrt{{\mu_l}^2 + {\nu_m}^2} \leq \omega_{c_2}/\pi \\ 0 & \omega_{c_2}/\pi < \sqrt{{\mu_l}^2 + {\mu_m}^2} \leq 1 \end{cases}$$

where

$$\omega_{c_1} = \frac{1}{2}(\omega_{a_1} + \omega_{p_1}), \quad \omega_{c_2} = \frac{1}{2}(\omega_{a_2} + \omega_{p_2}).$$

for $l = 1, 2, \cdots, L$ and $m = 1, 2, \cdots, M$ with $L = M = 36$.

An easy-to-use numerically reliable software package called MATLAB has been used to perform SVD on matrix $\mathbf{A}_1$ in order to obtain the necessary data for the designs in the following steps.

The 1-D FIR filters were designed by using the Fourier series method along with the Kaiser window function, which is known for its simplicity and flexibility [3]. As may be expected, the higher the order of the 1-D filters, the lower the approximation error. By trial and error, a value of 29 for N was found to give satisfactory results.

As is shown in Figure 2.2, there are 19 nonzero singular values resulting from the SVD of matrix $\mathbf{A}_1$ but, as is shown in Table 2.2, if only the first 9 are used in the design, the approximation error is less than 4%. The 3-D plot and the contour plot of the amplitude response obtained for a direct-SVD realization with 9 sections are shown in Figures 2.3 and 2.4 (each contour plot used in this chapter has 12 levels), respectively.

If all 19 nonzero singular values of $\mathbf{A}_1$ are used in the design, then the rank of matrix $\mathbf{C}$ defined in (2.23) is 15. If singular values 10 to 15 are neglected, the digital filter can be realized using the modified-SVD or SVD-LUD scheme with $K_c = 9$. The 3-D plot and the contour plot of the amplitude response obtained are shown in Figures 2.5 and 2.6, respectively. The maximum passband and stopband errors for different values of K and K_c are shown in Table 2.2. The number of multiplications required by the three realization schemes are listed in Table 2.3. An examination of Tables 2.2 and 2.3 suggests that the best choice for the designer is to use $K = 19$ and $K_c = 9$, and realize the filter by the SVD-LUD scheme.

Table 2.2: Approximation Errors (Bandpass Filter)

	Direct SVD			Modified SVD or SVD-LUD
K	9	15	19	$K = 19, K_c = 9$
Passband	0.0332	0.0276	0.0275	0.0282
Stopband	0.0290	0.0287	0.0263	0.0274

Table 2.3: Number of Multiplications

K	Direct SVD	Modified SVD		SVD-LUD	
		$K_c = 9$	$K_c = 15$	$K_c = 9$	$K_c = 15$
9	270	270	N/A	198	N/A
15	450	270	450	198	240
19	570	270	450	198	240
22	660	270	450	198	240

Table 2.4: Approximation Errors (Fan Filter)

	Direct SVD			Modified SVD or SVD-LUD
K	9	15	22	$K = 22,\ K_c = 9$
Passband	0.0475	0.0391	0.0390	0.0411
Stopband	0.0331	0.0267	0.0250	0.0281

2.6.2 Two-Dimensional Fan FIR Filter

The above approach has also been applied for the design of a fan filter having an amplitude response

$$|H_2(e^{j\omega_1 T_1},\ e^{j\omega_2 T_2})| = \begin{cases} 1 & \omega_2 < 0.6\omega_1 - 0.02857\pi \\ 0 & \omega_2 > 0.6\omega_1 + 0.1143\pi \end{cases}$$

for $0 \leq \omega_1,\ \omega_2 \leq \pi$, as depicted in Figure 2.7. The corresponding sampled amplitude response $\mathbf{A_2} = |H_2(e^{j\pi\mu_l},\ e^{j\pi\nu_m})|$ can be written as

$$|H_2(e^{j\pi\mu_l},\ e^{j\pi\nu_m})| = \begin{cases} 1 & \nu_m < 0.6\mu_l + 0.0457 \\ 0 & \nu_m \geq 0.6\mu_l + 0.0457 \end{cases}$$

where $l = 1, 2, \cdots, L$ and $m = 1, 2, \cdots, M$ with $L = M = 36$.

There are 22 nonzero singular values resulting from the SVD of matrix $\mathbf{A_2}$ as depicted in Figure 2.8 but if the last 13 are neglected in the design, the approximation error is less than 5%. The amplitude response obtained for a direct-SVD realization with $K = 9$ is illustrated in Figures 2.9 and 2.10.

If all 22 nonzero singular values of $\mathbf{A_2}$ are used in the design, then the matrix $\mathbf{C}$ defined in (2.23) has 15 nonzero singular values, but the last 6 may be neglected. Therefore, the digital filter can be realized using the modified-SVD or SVD-LUD realization scheme with $K_c = 9$. The amplitude response obtained is illustrated in Figures 2.11 and 2.12. The maximum passband and stopband errors for different values of K and K_c are shown in Table 2.4. The number of multiplications required by the three realization schemes are given in Table 2.3. An examination of Tables 2.4 and 2.3 suggests that the best choice for the designer is to use $K = 22$, $K_c = 9$ and realize the filter by the SVD-LUD scheme.

2.6.3 Comparisons

To compare the proposed design method with other well-established design methods, the McClellan transformation method [2, 9] and the 2-D window method [1]

have been used to design the same bandpass and fan filters.

A 1-D bandpass FIR filter of order 29 was first designed and the linear McClellan transformation recommended in [9] was then used to obtain a 29×29 2-D, circularly symmetric, bandpass filter. The amplitude response obtained is illustrated in Figures 2.13 and 2.14. The application of the window method resulted in a 29×29 bandpass filter whose amplitude response is illustrated in Figures 2.15 and 2.16.

The maximum passband and stopband errors obtained are summarized in Table 2.5 along with the results obtained in the SVD-LUD realization with $K = 19$ and $K_c = 9$. As can be seen, the SVD approach gave the lowest approximation error with respect to the passband and stopbands and, in addition, it resulted in highly circular contours in the amplitude response. The window method gave a fairly good approximation but a relatively large number of multiplications is required in the implementation. The McClellan transformation, on the other hand, resulted in the most economical realization [65] but, unfortunately, the approximation error is quite large in the upper stopband. Furthermore, the circularity of the contours of the amplitude response deteriorates rapidly as ω_1 and ω_2 are increased.

The McClellan transformation and window method were then used to design the fan filter. The amplitude response obtained is illustrated in Figures 2.17 and 2.18. The application of the window method gives a 29×29 fan filter whose amplitude response is illustrated in Figures 2.19 and 2.20. The results obtained are summarized in Table 2.6 along with the results obtained in the SVD-LUD realization with $K = 22$ and $K_c = 9$. Evidently, these results are fairly consistent with those in Table 2.5 and one is, therefore, tempted to conclude that the aforementioned merits and demerits of the three approaches are applicable to other types of filters.

Table 2.5: Comparison with the McClellan Transformation Method and Window Method (Bandpass Filter)

	SVD-LUD realization with $K = 19$, $K_c = 9$	McClellan method	Window method
Passband	0.0262	0.0381	0.0281
Stopband	0.0274	0.3709†	0.0714
No. of Multipl.	198	169	240

Table 2.6: Comparison with the McClellan Transformation Method and Window Method (Fan Filter)

	SVD-LUD realization with $K = 19$, $K_c = 9$	McClellan method	Window method
Passband	0.0411	0.0325	0.2905
Stopband	0.0281	0.9634†	0.0220
No. of Multipl	198	169	240

2.7 Conclusions

The SVD has been applied in the design of 2-D FIR digital filters and three realizations have been proposed. The direct-SVD realization is based on the principles used in [23] for the design of 2-D IIR filters. However, the modified-SVD and SVD-LUD realizations are new.

Each of the three realizations consists of a parallel arrangement of cascaded pairs of 1-D digital filters. Hence extensive parallel processing and pipelining can be applied. The 1-D FIR filters can be designed using well-known standard design methods and by using linear-phase causal 1-D filters, linear-phase causal 2-D filters can be obtained.

A quantitative error analysis has been carried out which can facilitate the determination of K and K_c, the number of singular values of matrices $\mathbf{A}$ and $\mathbf{C}$, that should be used in the design and the maximum approximation error that should be achieved in the design of the 1-D filters.

The three realizations were used to design a bandpass and a fan filter and

† These designs may not be optimal

the results obtained were compared. It was found that when the same number of parallel sections are used in the three realizations, the modified-SVD realization results in a smaller approximation error than the direct-SVD realization and the SVD-LUD realization results in the same approximation error as the modified-SVD realization but requires a reduced number of multiplications. Therefore, the SVD-LUD realization is the best of the three and should always be preferred. However, this does not obviate the need for the other two realizations since the direct-SVD realization must be obtained before the modified-SVD realization can be obtained and, in turn, the modified-SVD realization must be obtained before the SVD-LUD realization can be obtained.

The bandpass and fan filters were also designed using the McClellan transformation and the 2-D window method and the results obtained were compared with the results obtained using the SVD approach. The SVD-LUD realization resulted in very good approximations both with respect to the passband and stopband(s) while the computational complexity was moderate in both examples; in the case of the bandpass filter, excellent contour circularity was achieved in the passband. The 2-D window method resulted in fairly good approximations but the number of multiplications required in each design was very high. The McClellan transformation resulted in the most economical designs but the approximation error was found to be quite large in the upper stopband of the bandpass filter and the stopband of the fan filter; in addition, in the bandpass filter the circularity of the contours was found to deteriorate at higher frequencies.

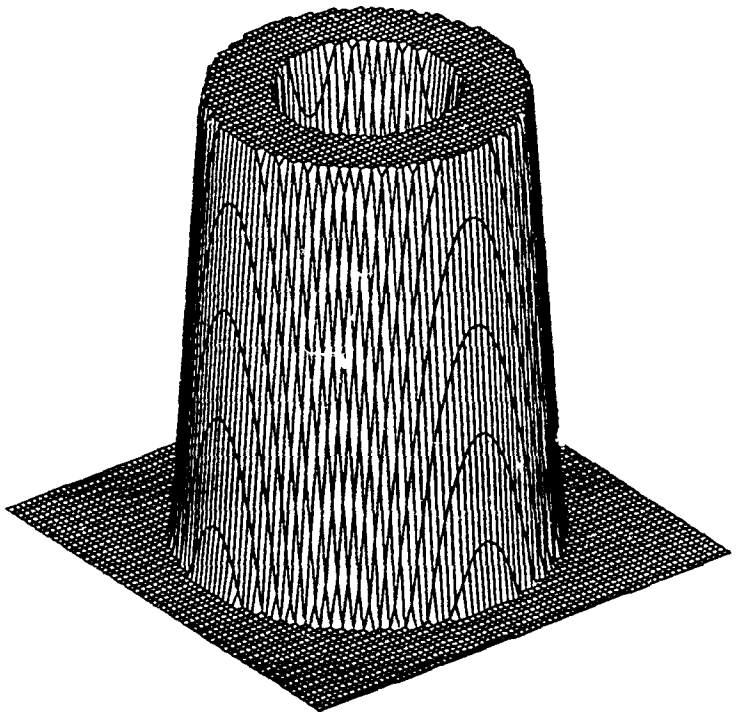

Figure 2.1: Ideal amplitude response of 2-D FIR bandpass filter.

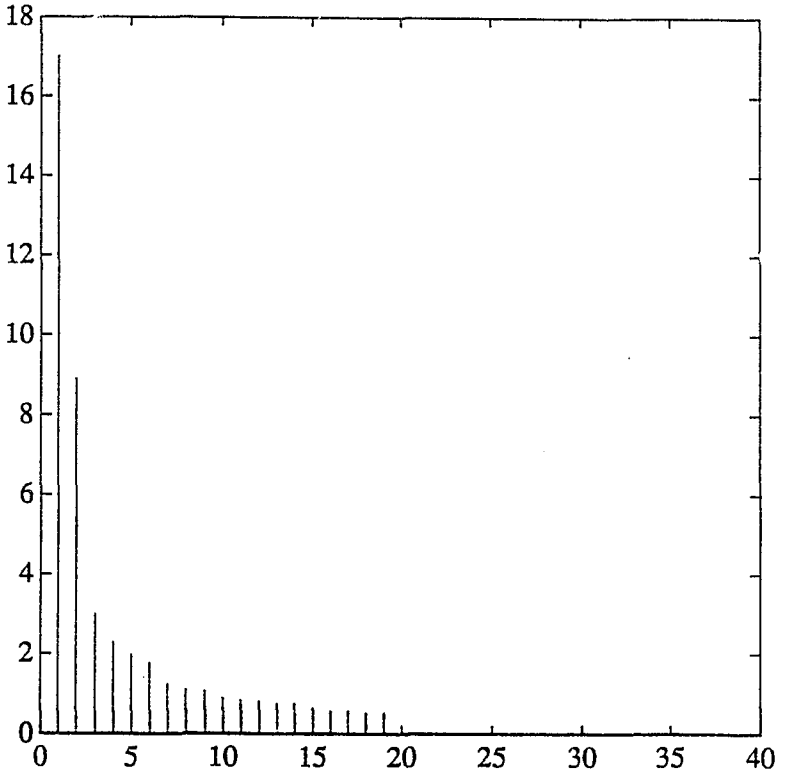

Figure 2.2: Singular-value distribution of matrix $\mathbf{A}_1$.

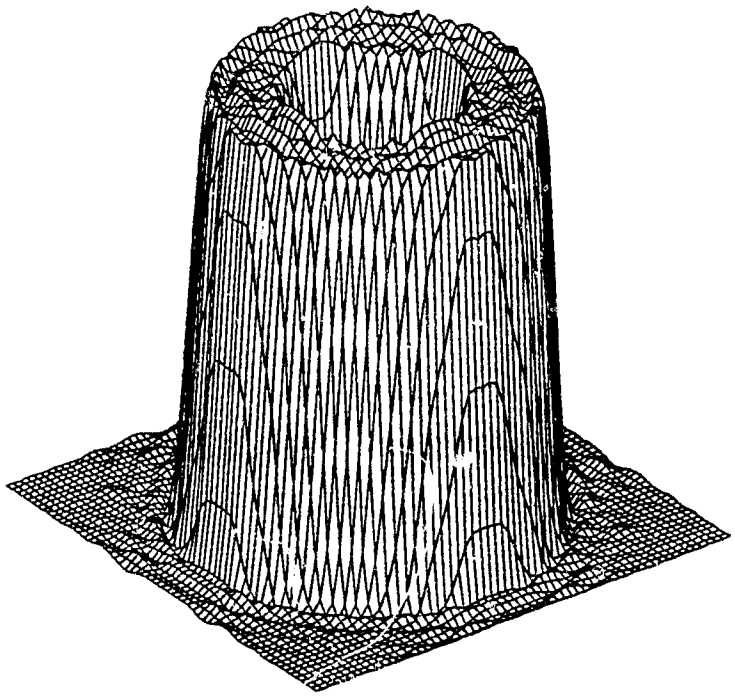

Figure 2.3: Three-dimensional plot of the amplitude response of the bandpass filter obtained by using the direct-SVD realization with $K = 9$.

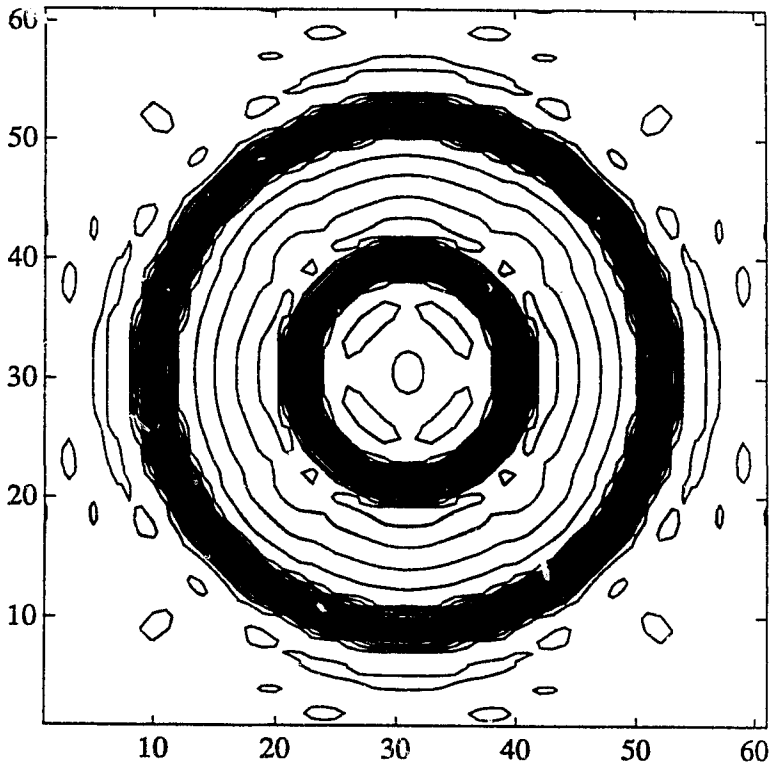

Figure 2.4: Contour plot of the amplitude response of bandpass filter obtained by using the direct-SVD realization with $K = 9$.

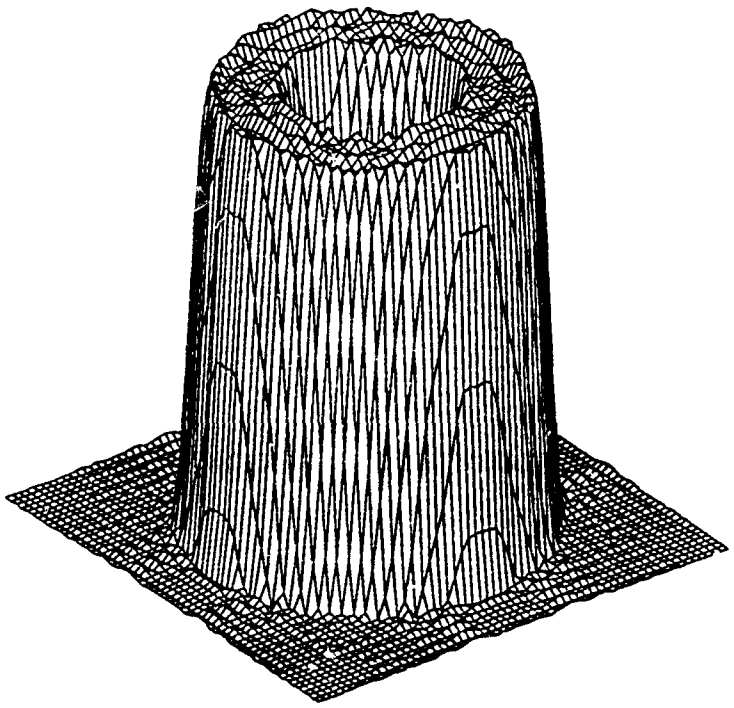

Figure 2.5: Three-dimensional plot of the amplitude response of bandpass filter obtained by using the modified-SVD or SVD-LUD realization with $K = 19$ and $K_c = 9$.

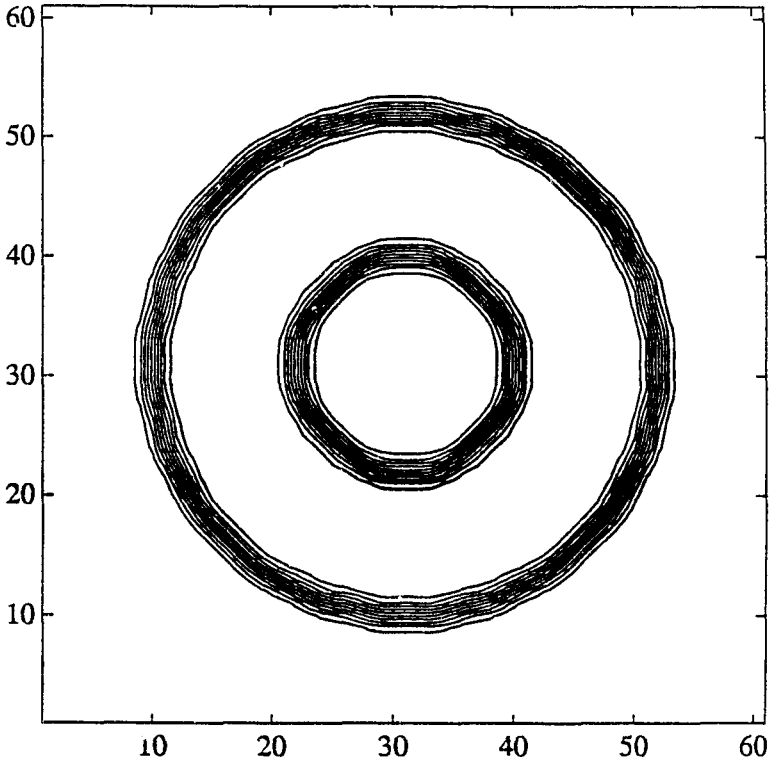

Figure 2.6: Contour plot of the amplitude response of bandpass filter obtained by using the modified-SVD or SVD-LUD realization with $K = 19$ and $K_c = 9$.

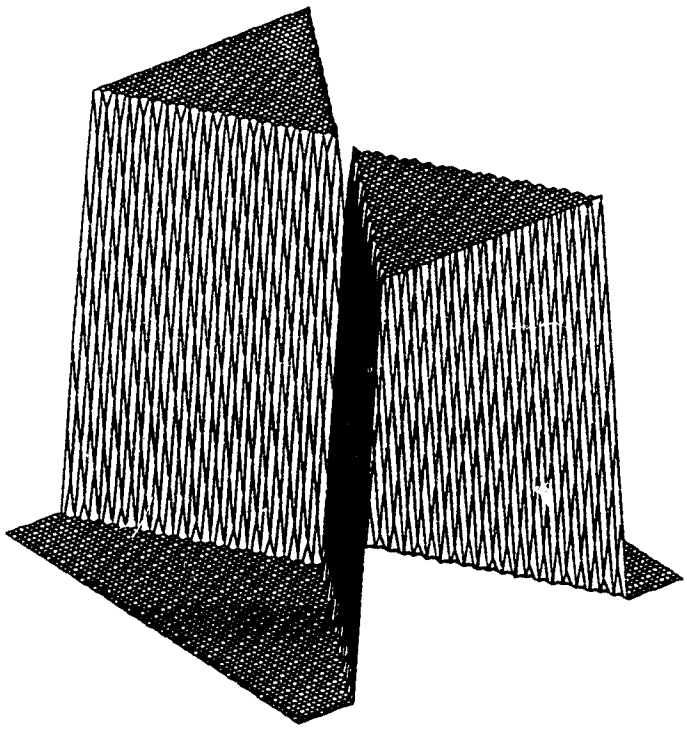

Figure 2.7: Ideal amplitude response of 2-D FIR fan filter.

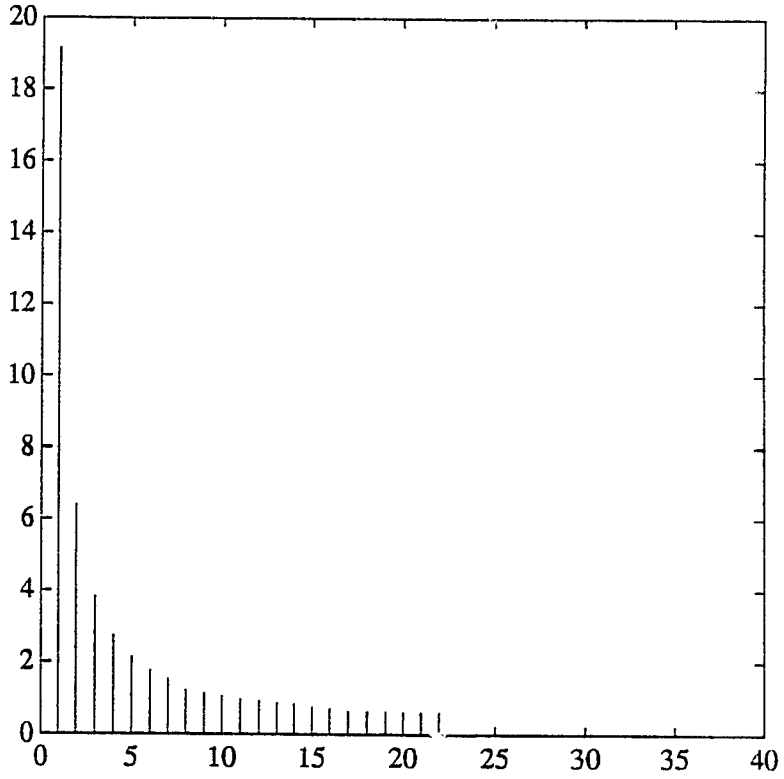

Figure 2.8: Singular-value distribution of matrix $\mathbf{A}_2$

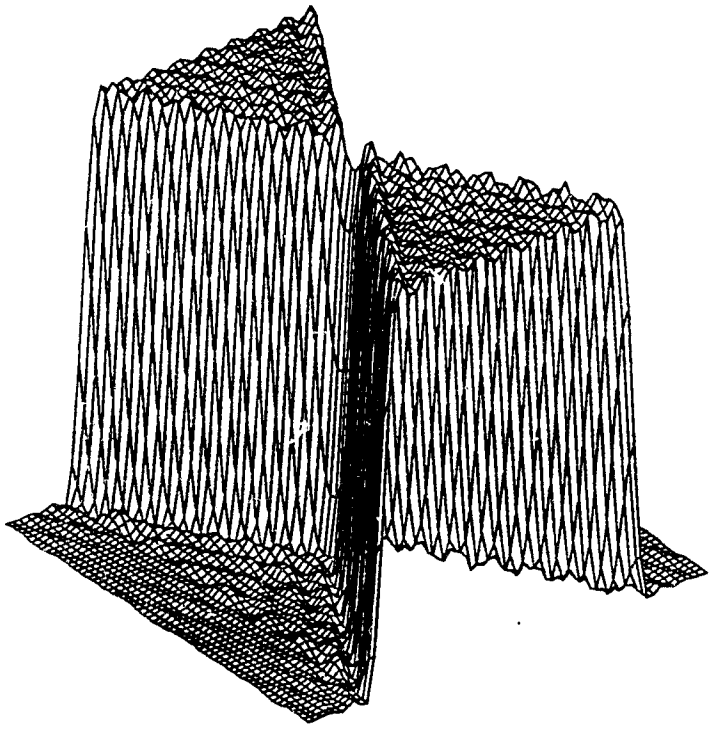

Figure 2.9: Three-dimensional plot of the amplitude response of fan filter obtained by using the direct-SVD realization with $K = 9$.

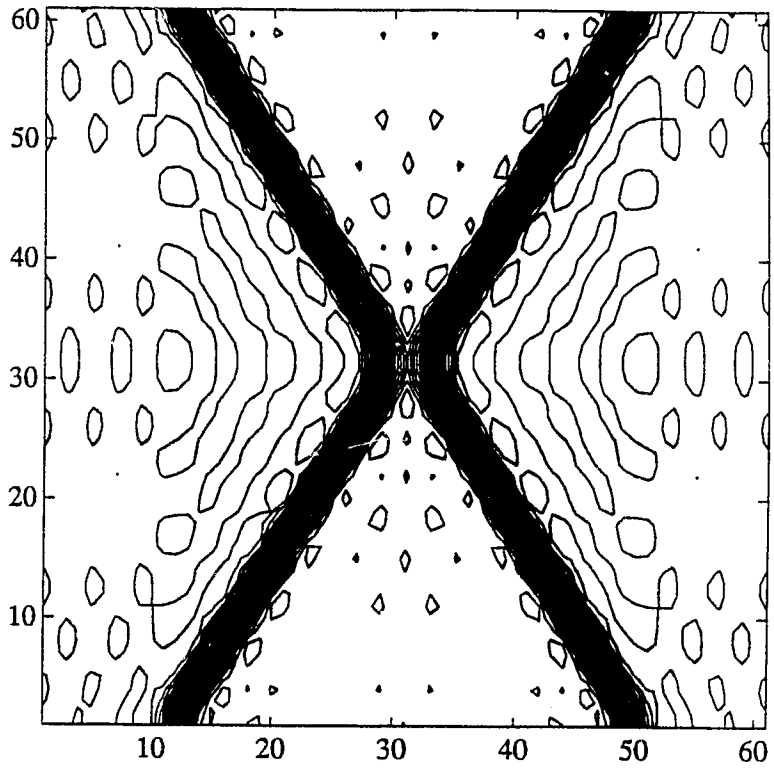

Figure 2.10: Contour plot of the amplitude response of fan filter obtained by using the direct-SVD realization with $K = 9$.

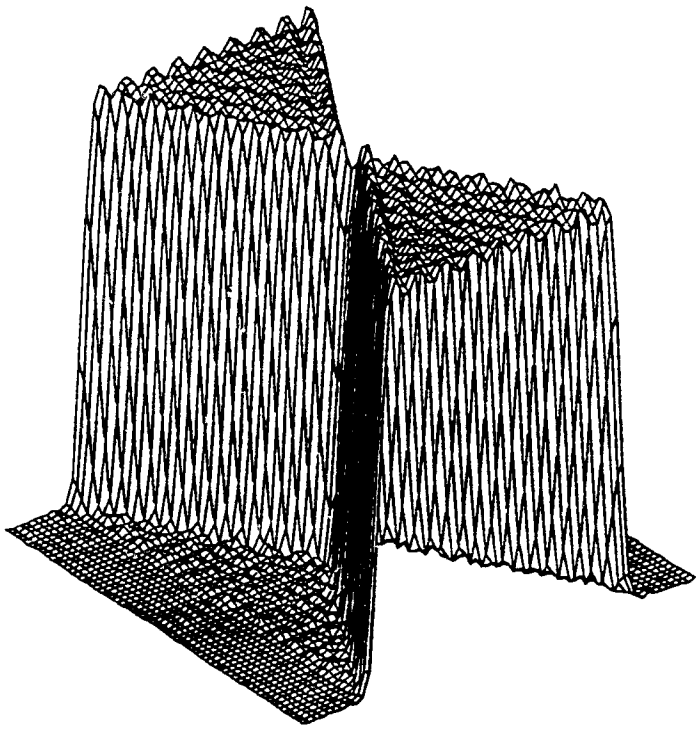

Figure 2.11: Three-dimensional plot of the amplitude response of fan filter obtained by using the modified-SVD or SVD-LUD realization with $K = 22$ and $K_c = 9$.

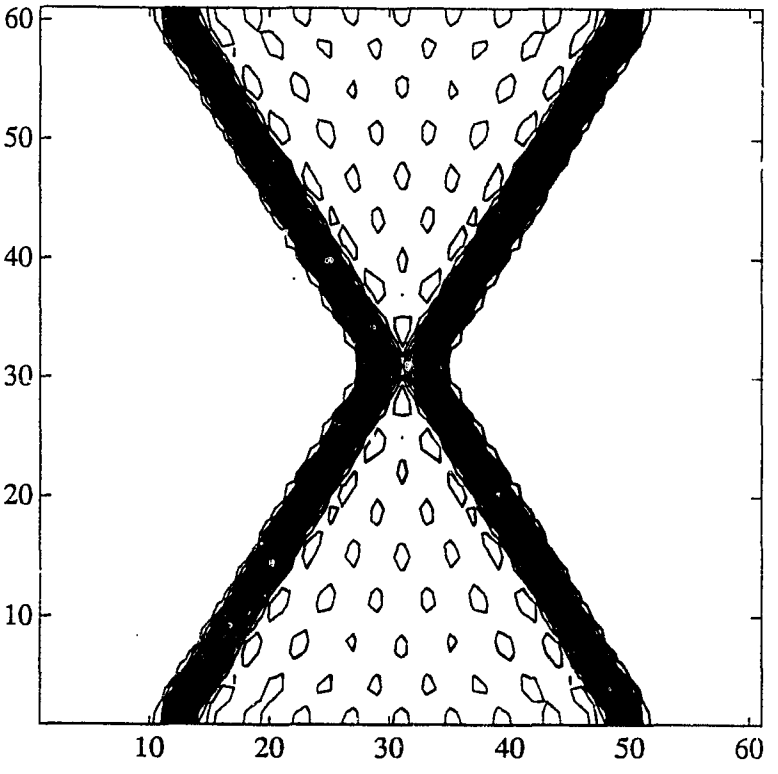

Figure 2.12: Contour plot of the amplitude response of fan filter obtained by using the modified-SVD or SVD-LUD realization with $K = 22$ and $K_c = 9$.

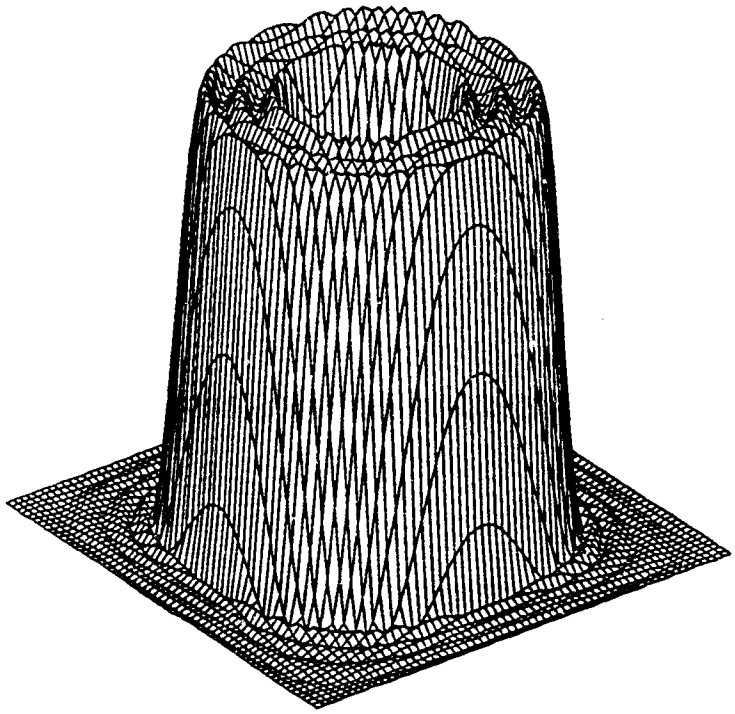

Figure 2.13: Three-dimensional plot of the amplitude response of bandpass filter obtained by using the McClellan transformation method.

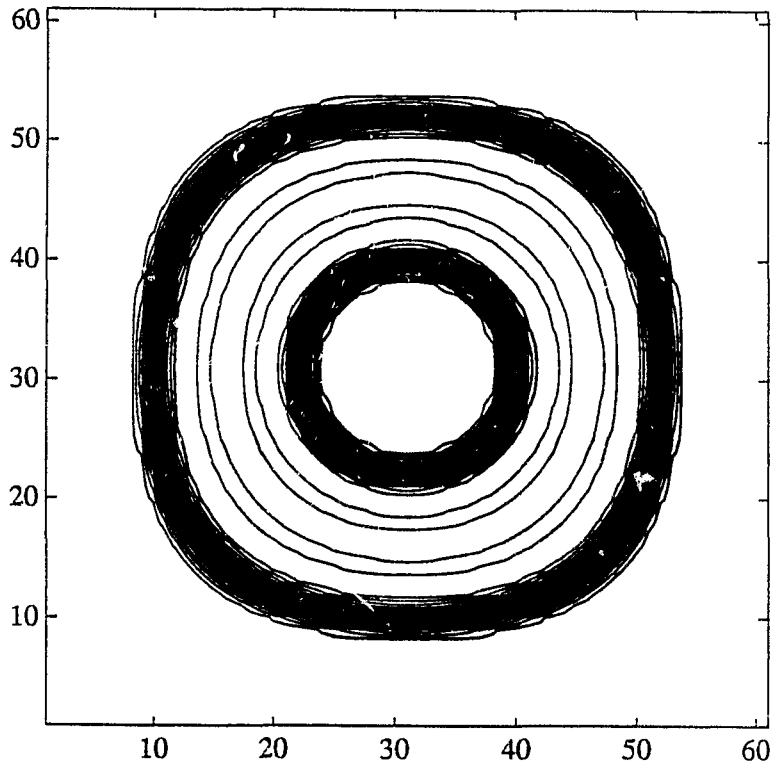

Figure 2.14: Contour plot of the amplitude response of bandpass filter obtained by using the McClellan transformation method.

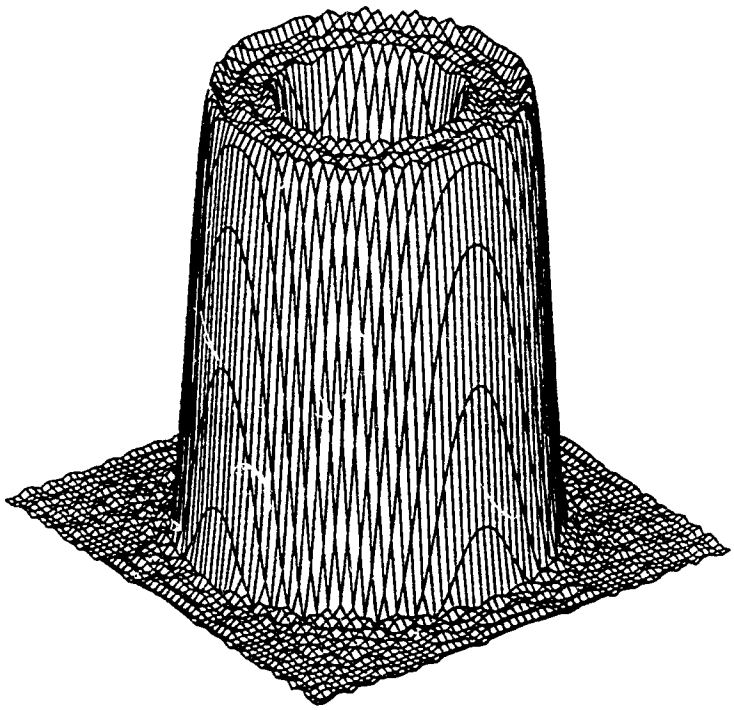

Figure 2.15: Three-dimensional plot of the amplitude response of bandpass filter obtained by using the window method.

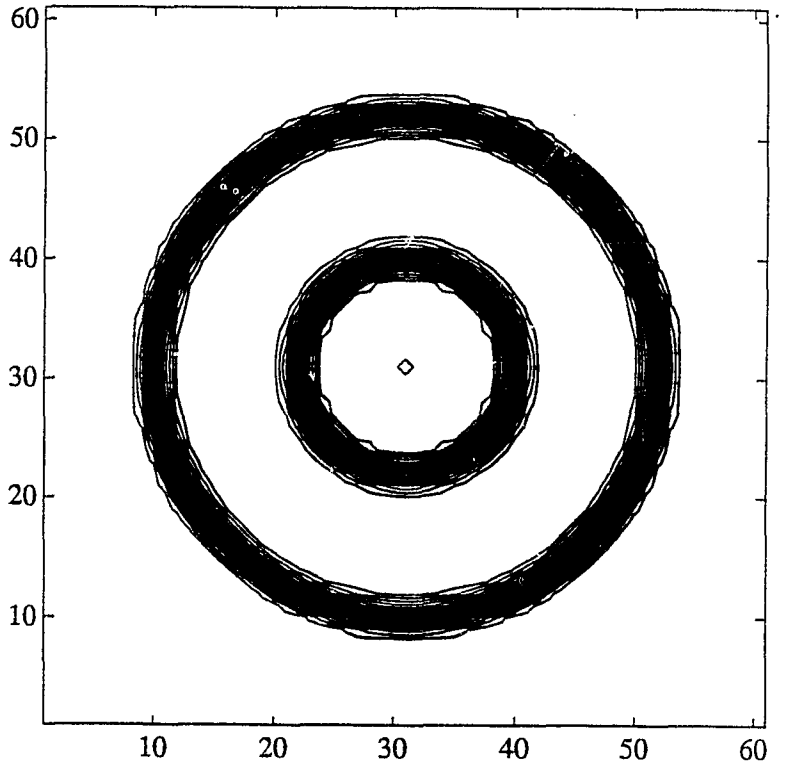

Figure 2.16: Contour plot of the amplitude response of bandpass filter obtained by using the window method.

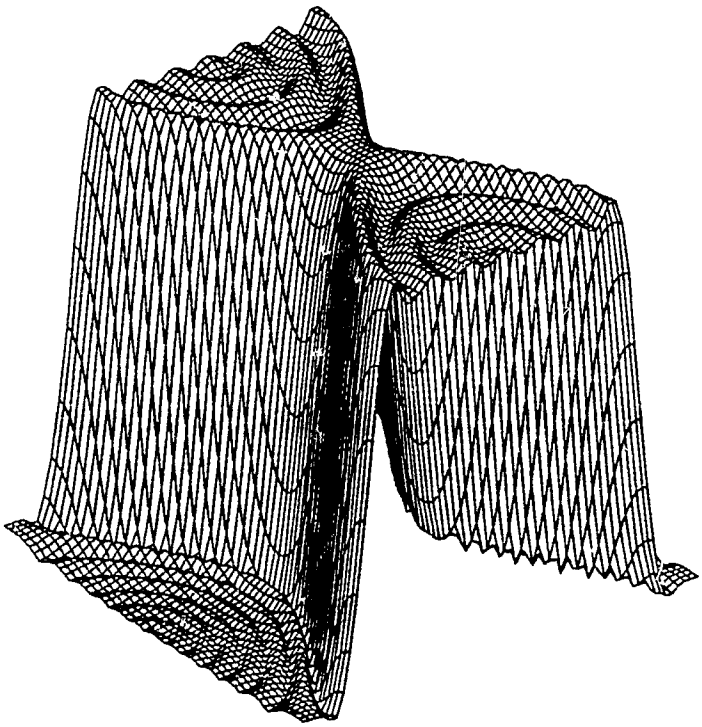

Figure 2.17: Three-dimensional plot of the amplitude response of fan filter obtained by using the McClellan transformation method.

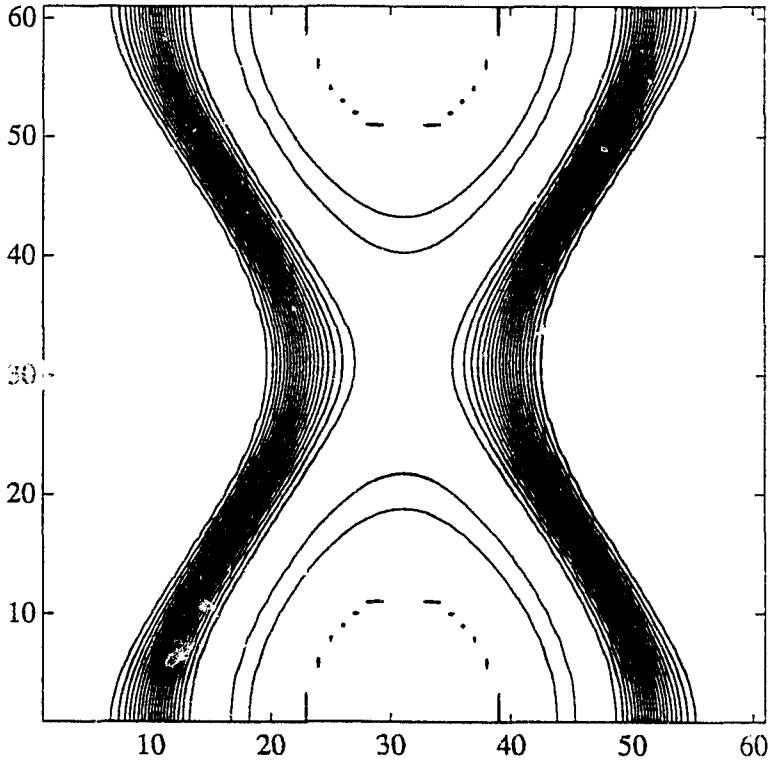

Figure 2.18: Contour plot of the amplitude response of fan filter obtained by using the McClellan transformation method.

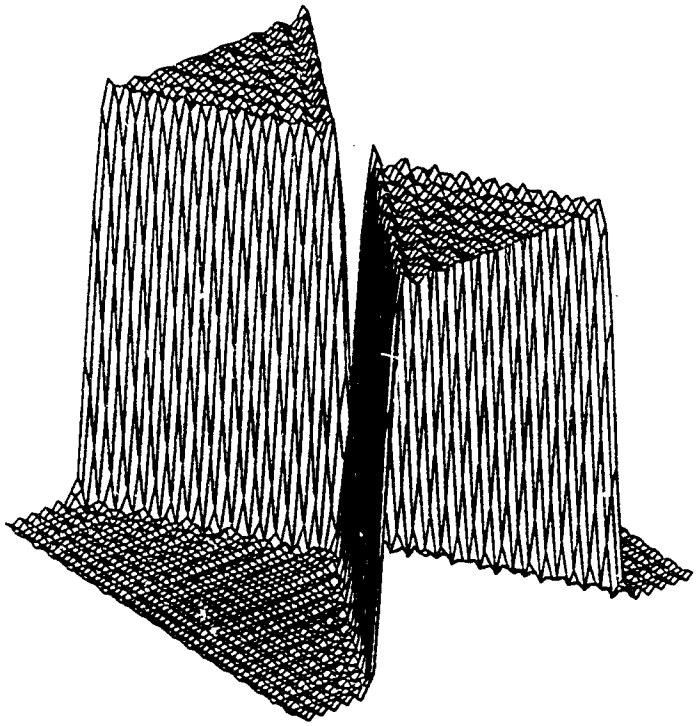

Figure 2.19: Three-dimensional plot of the amplitude response of fan filter obtained by using the window method.

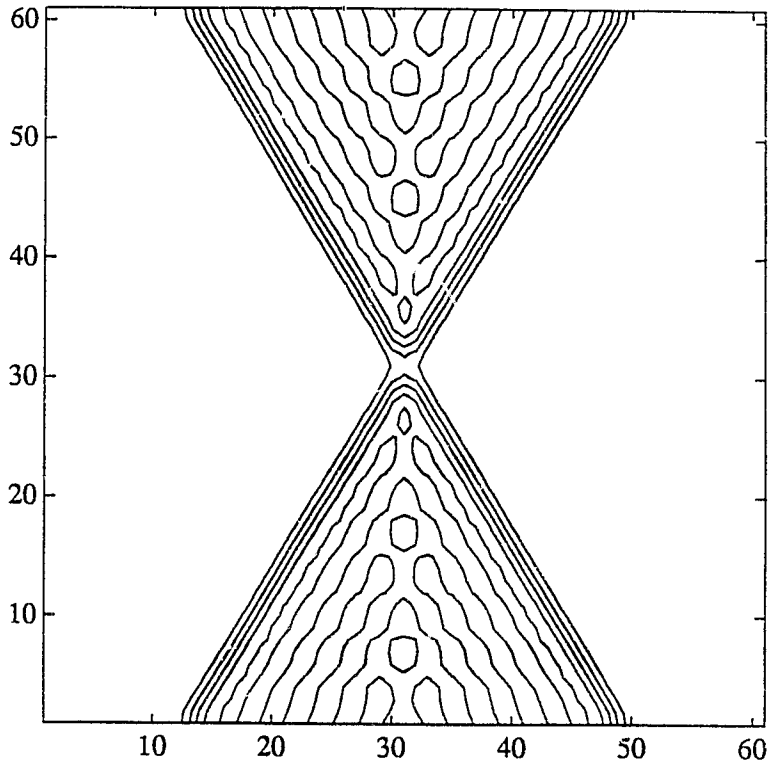

Figure 2.20: Contour plot of the amplitude response of fan filter obtained by using the window method.

Chapter 3

Design of General 2-D FIR Digital Filters by Using the SVD

3.1 Introduction

A limitation of the SVD method as presented in Chapter 2 is that the amplitude response of the filter is required to have quadrantal symmetry with respect to the (ω_1, ω_2) plane. This limits the application of the method.

In this chapter, it is shown that the SVD of the sampled amplitude response of a 2-D digital filter with real coefficients possesses a special structure: every singular vector is either mirror-image symmetric or antisymmetric about its midpoint. As a result, the SVD method can be applied along with 1-D FIR filter techniques for the design of linear-phase 2-D filters with arbitrary amplitude responses which are symmetrical with respect to the origin of the (ω_1, ω_2) plane. A quantitative error analysis is then described which can facilitate the determination of the number of subfilters to be used and the required design accuracy for each 1-D subfilters to achieve a desired approximation error.

In Section 3.2 the general SVD 2-D FIR design method is discussed and an important property of a class of 2-D FIR filters is presented. In Section 3.3 an error analysis is given for the general SVD design method. In Section 3.4, a 2-D lowpass FIR filter with rotated elliptical passband is designed to illustrate the effectiveness of the proposed method.

3.2 General SVD Design Method

3.2.1 Design

The impulse response $h(n_1,\ n_2)$ of a general 2-D linear-phase FIR filter must satisfy the relation

$$h(n_1,\ n_2) = h(-n_1,\ -n_2) \tag{3.1}$$

Furthermore, if $h(n_1,\ n_2)$ is real then the frequency response of the filter given by (2.13) is a real function which is symmetrical with respect to the origin of the $(\omega_1,\ \omega_2)$ plane such that

$$X(\omega_1,\ \omega_2) = X(-\omega_1,\ -\omega_2) \tag{3.2}$$

where $-\pi \leq \omega_1,\ \omega_2 \leq \pi$.

A 2-D FIR filter having an arbitrary amplitude response satisfying (3.2) can readily be designed by using a parallel arrangement of K 2-D FIR sections each comprising two 1-D subfilters in cascade as will now be demonstrated. Such an arrangement can be represented by the transfer function as in (2.13)-(2.18) and if $F_i(z_1)$ and $G_i(z_2)$ are assumed to represent zero-phase or $\frac{\pi}{2}$-phase filters, then their frequency responses are given by

$$\begin{aligned} F_i(e^{j\omega_1 T_1}) &= \sum_{n_1=-N_1/2}^{N_1/2} f_i(n_1) e^{-j\omega_1 n_1 T_1} \\ &= \Phi_i(\omega_1) e^{j\theta_i} \end{aligned} \tag{3.3}$$

and

$$\begin{aligned} G_i(e^{j\omega_2 T_2}) &= \sum_{n_2=-N_2/2}^{N_2/2} g_i(n_2) e^{-j\omega_2 n_2 T_2} \\ &= \Gamma_i(\omega_2) e^{j\theta_i} \end{aligned} \tag{3.4}$$

If $f_i(n_1)$ and $g_i(n_2)$ are mirror-image symmetric, then $\theta_i = 0$ in (3.3) and (3.1) and $\Phi_i(\omega_1)$ and $\Gamma_i(\omega_2)$ are real functions which are even with respect to ω_1 and ω_2, respectively; if $f_i(n_1)$ and $g_i(n_2)$ are mirror-image antisymmetric, then $\theta_i = \pi/2$ and $\Phi_i(\omega_1)$ and $\Gamma_i(\omega_2)$ are real functions which are odd with respect to ω_1 and ω_2, respectively. Under these circumstances, a zero-phase 2-D filter is obtained whose frequency response is given as

$$\begin{aligned} H(e^{j\omega_1 T_1},\ e^{j\omega_2 T_2}) &= \sum_{i=1}^{K} F_i(e^{j\omega_1 T_1}) G_i(e^{j\omega_2 T_2}) \\ &= \sum_{i=1}^{K} \pm \Phi_i(\omega_1)\Gamma_i(\omega_2) \end{aligned} \tag{3.5}$$

where the '+' sign corresponds to $\theta_i = 0$, and the '−' sign corresponds to $\theta_i = \pi/2$. On comparing (3.5) with (2.19), we obtain

$$A(\omega_1,\ \omega_2) = \sum_{i=1}^{K} \pm \Phi_i(\omega_1)\Gamma_i(\omega_2) \tag{3.6}$$

Now assume that matrix $\mathbf{A} = \{a_{l,m}\}$ represents a desired arbitrary frequency response such that (3.2) is satisfied, i.e.

$$a_{l,m} = A(\pi\mu_l,\ \pi\nu_m) = A(-\pi\mu_l,\ -\pi\nu_m) \tag{3.7}$$

where $1 \le l \le L$ and $1 \le m \le M$. The quantities μ_l and ν_m are normalized frequencies over the entire baseband such that

$$\mu_l = -1 + 2(\frac{l-1}{L-1}), \qquad \nu_m = -1 + 2(\frac{m-1}{M-1})$$

and $-1 \le \mu_l \le 1$, $-1 \le \nu_m \le 1$. The SVD of $\mathbf{A}$ is given by (2.21). In the next section, an important property of a class of FIR filters can be stated in terms of the following theorem.

3.2.2 Property

Theorem 3.1 If the frequency response of an FIR filter satisfies (3.2), then vectors $\mathbf{u}_i$ and $\mathbf{v}_i$ in (2.21) are either mirror-image symmetric or antisymmetric simultaneously for $i = 1, 2, \cdots, r$.

Proof: Let $\tilde{\mathbf{H}} = \{\tilde{h}_{i,j},\ 1 \le i,\ j \le 2N\}$ be an $2N \times 2N$ arbitrary matrix with real elements such that

$$\tilde{h}_{i,j} = \tilde{h}_{2N+1-i,2N+1-j} \tag{3.8}$$

and, for the sake of simplicity, assume that the matrix is square and of even size, and has distinct singular values. If matrices $\hat{\mathbf{I}}$, $\tilde{\mathbf{I}}$, and $\mathbf{H}$ are defined by

$$\hat{\mathbf{I}} = \begin{bmatrix} 0 & \dots & 0 & 1 \\ 0 & \dots & 1 & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 1 & \dots & 0 & 0 \end{bmatrix},\ \tilde{\mathbf{I}} = \begin{bmatrix} \mathbf{I}_N & \mathbf{0} \\ \mathbf{0} & \hat{\mathbf{I}}_N \end{bmatrix},\ \text{and } \mathbf{H} = \hat{\mathbf{I}}\tilde{\mathbf{H}}\hat{\mathbf{I}} = \begin{bmatrix} \mathbf{H}_1 & \mathbf{H}_2 \\ \mathbf{H}_2 & \mathbf{H}_1 \end{bmatrix}$$

respectively, where the sizes of $\hat{\mathbf{I}}_\mathbf{N}$, $\mathbf{I}_\mathbf{N}$, $\mathbf{H_1}$ and $\mathbf{H_2}$ are $N \times N$, and the sizes of $\tilde{\mathbf{I}}$ and $\mathbf{H}$ are $2N \times 2N$, respectively, then matrix $\tilde{\mathbf{H}}$ can be decomposed as

$$\tilde{\mathbf{H}} = \tilde{\mathbf{I}}\mathbf{H}\tilde{\mathbf{I}} = \mathbf{U}\boldsymbol{\Sigma}\mathbf{V}^\mathbf{T} = [\mathbf{u}_1\ \cdots\ \mathbf{u}_i\ \cdots\ \mathbf{u}_{2N}]\boldsymbol{\Sigma}[\mathbf{v}_1\ \cdots\ \mathbf{v}_i\ \cdots\ \mathbf{v}_{2N}]^T \tag{3.9}$$

where $\mathbf{u}_i$ and $\mathbf{v}_i$ are normalized eigenvectors of $\tilde{\mathbf{H}}\tilde{\mathbf{H}}^T$ and $\tilde{\mathbf{H}}^T\tilde{\mathbf{H}}$, respectively. Assume that $\mathbf{u}_i$ is a normalized eigenvector of $\tilde{\mathbf{H}}\tilde{\mathbf{H}}^T$, i.e. there exists a σ_i such that

$$\tilde{\mathbf{H}}\tilde{\mathbf{H}}^T\mathbf{u}_i = \sigma_i\ \mathbf{u}_i \tag{3.10}$$

Substituting (3.9) into (3.10), we have

$$\tilde{\mathbf{I}}\mathbf{H}\tilde{\mathbf{I}}(\tilde{\mathbf{I}}\mathbf{H}\tilde{\mathbf{I}})^T\mathbf{u}_i = (\tilde{\mathbf{I}}\mathbf{H}\mathbf{H}^T\tilde{\mathbf{I}})\mathbf{u}_i = \sigma_i\mathbf{u}_i \tag{3.11}$$

and

$$\tilde{\mathbf{I}}\mathbf{H}\mathbf{H}^T\tilde{\mathbf{I}}\ \mathbf{u}_i = \sigma_i\ \tilde{\mathbf{I}}\ \mathbf{u}_i \tag{3.12}$$

If we let

$$\mathbf{u}_i = \begin{bmatrix} \mathbf{u}_{i1} \\ \mathbf{u}_{i2} \end{bmatrix}$$

then

$$\tilde{\mathbf{I}}\mathbf{u}_i = \begin{bmatrix} \tilde{\mathbf{u}}_{i1} \\ \hat{\mathbf{I}}\mathbf{u}_{i2} \end{bmatrix} \equiv \begin{bmatrix} \mathbf{x}_{i1} \\ \mathbf{x}_{i2} \end{bmatrix}$$

and, therefore, (3.12) becomes

$$\mathbf{H}\mathbf{H}^T \begin{bmatrix} \mathbf{x}_{i1} \\ \mathbf{x}_{i2} \end{bmatrix} = \sigma_i \begin{bmatrix} \mathbf{x}_{i1} \\ \mathbf{x}_{i2} \end{bmatrix} \tag{3.13}$$

Note that

$$\begin{aligned} \mathbf{H}\mathbf{H}^T &\equiv \begin{bmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{B} & \mathbf{A} \end{bmatrix} \mathbf{H}\mathbf{H}^T = \begin{bmatrix} \mathbf{H}_1 & \mathbf{H}_2 \\ \mathbf{H}_2 & \mathbf{H}_1 \end{bmatrix} \begin{bmatrix} \mathbf{H}_1^T & \mathbf{H}_2^T \\ \mathbf{H}_2^T & \mathbf{H}_1^T \end{bmatrix} \\ &= \begin{bmatrix} \mathbf{H}_1\mathbf{H}_1^T + \mathbf{H}_2\mathbf{H}_2^T & \mathbf{H}_1\mathbf{H}_2^T + \mathbf{H}_2\mathbf{H}_1^T \\ \mathbf{H}_2\mathbf{H}_1^T + \mathbf{H}_1\mathbf{H}_2^T & \mathbf{H}_2\mathbf{H}_2^T + \mathbf{H}_1\mathbf{H}_1^T \end{bmatrix} \equiv \begin{bmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{B} & \mathbf{A} \end{bmatrix} \end{aligned}$$

where $\mathbf{A}$ is positive semidefinite and $\mathbf{B}$ is symmetric. Therefore, (3.13) can be expressed as

$$\begin{bmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{B} & \mathbf{A} \end{bmatrix} \begin{bmatrix} \mathbf{x}_{i1} \\ \mathbf{x}_{i2} \end{bmatrix} = \sigma_i \begin{bmatrix} \mathbf{x}_{i1} \\ \mathbf{x}_{i2} \end{bmatrix} \tag{3.14}$$

By writing (3.14) in another matrix notation, we have

$$\begin{bmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{B} & \mathbf{A} \end{bmatrix} \begin{bmatrix} \mathbf{x}_{i2} \\ \mathbf{x}_{i1} \end{bmatrix} = \sigma_i \begin{bmatrix} \mathbf{x}_{i2} \\ \mathbf{x}_{i1} \end{bmatrix} \tag{3.15}$$

Now on comparing (3.15) with (3.14), we note that both vectors

$$\begin{bmatrix} \mathbf{x}_{i1} \\ \mathbf{x}_{i2} \end{bmatrix} \text{ and } \begin{bmatrix} \mathbf{x}_{i2} \\ \mathbf{x}_{i1} \end{bmatrix}$$

are eigenvectors of matrix $\mathbf{HH}^T$ associated with the same eigenvalue σ_i and, therefore, the two vectors must be linearly dependent, i.e. they must satisfy

$$\begin{bmatrix} \mathbf{x}_{i1} \\ \mathbf{x}_{i2} \end{bmatrix} = \pm \begin{bmatrix} \mathbf{x}_{i2} \\ \mathbf{x}_{i1} \end{bmatrix}$$

which implies that

$$\mathbf{x}_{i1} = \mathbf{x}_{i2} \text{ or } \mathbf{x}_{i1} = -\mathbf{x}_{i2}$$

If $\mathbf{x}_{i1} = \mathbf{x}_{i2} \equiv \mathbf{x}_i$, we can write

$$\tilde{\mathbf{I}}\mathbf{u}_i = \begin{bmatrix} \mathbf{x}_i \\ \mathbf{x}_i \end{bmatrix} \text{ and } \mathbf{u}_i = \begin{bmatrix} \mathbf{x}_i \\ \hat{\mathbf{I}}\mathbf{x}_i \end{bmatrix}$$

which means that $\mathbf{u}_i$ is mirror-image symmetric. On the other hand, if $\mathbf{x}_{i1} = -\mathbf{x}_{i2} \equiv \mathbf{x}_i$, we have

$$\mathbf{u}_i = \begin{bmatrix} \mathbf{x}_i \\ -\hat{\mathbf{I}}\mathbf{x}_i \end{bmatrix}$$

which implies that $\mathbf{u}_i$ is mirror-image antisymmetric. Furthermore, by (3.9)

$$\mathbf{V} = \tilde{\mathbf{I}}\mathbf{H}^T\tilde{\mathbf{I}}(\mathbf{U}^T)^{-1}\mathbf{\Sigma}^{-1} \tag{3.16}$$

Since, $\tilde{\mathbf{I}}$ is symmetric and $\mathbf{U}$ is orthogonal, i.e. $(\mathbf{U}^T)^{-1} = \mathbf{U}$, matrix $\mathbf{V}$ can be expressed as

$$\mathbf{V} = \tilde{\mathbf{I}}\mathbf{H}^T\tilde{\mathbf{I}}\mathbf{U}\mathbf{\Sigma}^{-1} \tag{3.17}$$

If

$$\mathbf{u}_i = \begin{bmatrix} \mathbf{x}_i \\ \hat{\mathbf{I}}\mathbf{x}_i \end{bmatrix}$$

then (3.17) implies

$$\mathbf{v}_i = \sigma_i^{-1}\tilde{\mathbf{I}}\mathbf{H}^T\tilde{\mathbf{I}}\mathbf{u}_i = \sigma_i^{-1}\begin{bmatrix} (\mathbf{H}_1 + \mathbf{H}_2)^T\mathbf{x}_i \\ \hat{\mathbf{I}}(\mathbf{H}_1 - \mathbf{H}_2)^T\mathbf{x}_i \end{bmatrix} = \begin{bmatrix} \mathbf{y}_i \\ \hat{\mathbf{I}}\mathbf{y}_i \end{bmatrix}$$

where $\mathbf{y}_i = \sigma_i^{-1}(\mathbf{H}_1 + \mathbf{H}_2)^T\mathbf{x}_i$. If

$$\mathbf{u}_i = \begin{bmatrix} \mathbf{x}_i \\ -\hat{\mathbf{I}}\mathbf{x}_i \end{bmatrix}$$

then (3.17) implies

$$\mathbf{v}_i = \sigma_i^{-1}\tilde{\mathbf{I}}\mathbf{H}^T\tilde{\mathbf{I}}\mathbf{u}_i = \sigma_i^{-1}\begin{bmatrix} (\mathbf{H}_2 - \mathbf{H}_1)^T\mathbf{x}_i \\ -\hat{\mathbf{I}}(\mathbf{H}_2 - \mathbf{H}_1)^T\mathbf{x}_i \end{bmatrix} = \begin{bmatrix} \hat{\mathbf{y}}_i \\ -\hat{\mathbf{I}}\hat{\mathbf{y}}_i \end{bmatrix}$$

where $\hat{\mathbf{y}}_i = \sigma_i^{-1}(\mathbf{H}_1 - \mathbf{H}_2)^T\mathbf{x}_i$. This shows that the two vectors $\mathbf{u}_i$ and $\mathbf{v}_i$ have the same symmetry properties simultaneously, that is, they are either both mirror-image symmetric or antisymmetric. □

On comparing (3.5) with (2.21), $\tilde{\mathbf{u}}_i$ and $\tilde{\mathbf{v}}_i$ may be taken to be sampled versions of the frequency responses $\Phi_i(\omega_1)$ and $\Gamma_i(\omega_2)$, respectively. By designing the 1-D FIR filters characterized by $F_i(z_1)$ and $G_i(z_2)$ $(1 \leq i \leq K,\ 1 \leq K \leq r)$ as zero-phase or $\frac{\pi}{2}$-phase filters and then interconnecting the filters obtained as in Figure 3.1, a zero-phase 2-D filter is obtained whose amplitude response is a good approximation to the desired amplitude response in the minimal mean-square-error sense (see Theorem 1.2). The impulse response of the resulting filter is given by

$$h(n_1,\ n_2) \ = \ \sum_{i=1}^{K} f_i(n_1)g_i(n_2) \tag{3.18}$$

A 2-D causal linear-phase filter can readily be obtained by shifting the impulse response by $N_1/2$ and $N_2/2$ with respect to the n_1 axis and n_2 axis, respectively. This can be accomplished by multiplying $F_i(z_1)$ and $G_i(z_2)$ by $z_1^{-N_1/2}$ and $z_2^{-N_2/2}$, respectively. The design of the 2-D filter can be completed by using any one of the standard methods for the design of 1-D FIR filters. Using the Fourier series method in conjunction with window techniques [3], designs can be obtained very quickly with a small amount of computational effort. These designs are

not optimal although the approximation error can be made arbitrarily small by increasing the order of the 1-D filters used. On the other hand, by using methods based on the Remez algorithm [43], it may be possible to obtain optimal designs although a large amount of computation would be required.

In the above design approach, the sampling density must be high enough in order for matrix **A** to represent the ideal amplitude response well. However, since the dimension of **A**, namely $L \times M$, determines the dimensions of vectors $\tilde{\mathbf{u}}_i$ and $\tilde{\mathbf{v}}_i$, a high sampling density leads to high-order FIR filters. Furthermore, the SVD of matrix **A** is time-consuming and may entail numerical ill-conditioning, particularly if L or $M \geq 100$. Based on a variety of designs we carried so far, a sampling density $L \times M$ with $60 < L,\ M < 90$ is appropriate if the order of the FIR filter to be designed is in the range 25×25 to 41×41.

3.2.3 Design procedure

To summarize, the general SVD design can be accomplished through the following procedure:

1) Specify the desired amplitude response and thereby obtain the corresponding sampled amplitude response matrix **A**.

2) Decompose matrix **A** using (2.21) to get $\tilde{\mathbf{u}}_i$ and $\tilde{\mathbf{v}}_i$, where $1 \leq i \leq r$.

3) Obtain K $(1 \leq K \leq r)$ 2-D FIR filters by designing either two 1-D zero-phase or two $\frac{\pi}{2}$-phase FIR filters characterized by transfer functions $F_i(z_1)$ and $G_i(z_2)$ assuming sampled frequency responses $\tilde{\mathbf{u}}_i$ and $\tilde{\mathbf{v}}_i$, respectively.

4) Obtain the impulse response of the resulting zero-phase 2-D filter through (3.18).

5) Multiply the resulting 2-D zero-phase transfer function by $z_1^{-(N_1-1)/2}$ and $z_2^{-(N_2-1)/2}$ to obtain a causal linear-phase 2-D FIR filter.

3.3 Error Analysis

Assume that the SVD method has been used to design a 2-D FIR filter for a desired frequency response **A** and that the transfer function $H(z_1,\ z_2)$ is given by (2.14). Let $\tilde{\mathbf{f}}_i$ and $\tilde{\mathbf{g}}_i$ be the column vectors obtained by evaluating $\Phi_i(\omega_1)$ and $\Gamma_i(\omega_2)$ at frequencies $\omega_1 = \pi\mu_l/T_1$ and $\omega_2 = \pi\nu_m/T_2$, where $1 \leq l \leq L$, $1 \leq m \leq M$ and T_1, T_2 are the sampling periods, i.e.

$$\tilde{\mathbf{f}}_i = [\Phi_i(\frac{\pi\mu_1}{T_1}) \cdots \Phi_i(\frac{\pi\mu_L}{T_1})]^T$$

$$\tilde{\mathbf{g}}_i = [\Gamma_i(\frac{\pi\nu_1}{T_2}) \cdots \Gamma_i(\frac{\pi\nu_M}{T_2})]^T$$

The amplitude response of the 2-D filter at frequency point $(\omega_1,\ \omega_2) = (\pi\mu_l/T_1,\ \pi\nu_m/T_2)$ is the (l,m) entry of matrix $\sum_1^K \tilde{\mathbf{f}}_i\tilde{\mathbf{g}}_i^T$ and, therefore, the approximation error at this frequency point is the (l,m) entry in the error matrix **E** defined by

$$\begin{aligned}
\mathbf{E} &= \{\epsilon_{l,m}\} \\
&= \sum_{i=1}^{K} \tilde{\mathbf{f}}_i\tilde{\mathbf{g}}_i^T - \mathbf{A} \\
&= \sum_{i=1}^{K} (\tilde{\mathbf{f}}_i\tilde{\mathbf{g}}_i^T - \tilde{\mathbf{u}}_i\tilde{\mathbf{v}}_i^T) - \sum_{i=K+1}^{r} \sigma_i \mathbf{u}_i\mathbf{v}_i^T
\end{aligned} \tag{3.19}$$

If we define

$$\Delta\tilde{\mathbf{u}}_i = \tilde{\mathbf{f}}_i - \tilde{\mathbf{u}}_i, \qquad \Delta\tilde{\mathbf{v}}_i = \tilde{\mathbf{g}}_i - \tilde{\mathbf{v}}_i$$

then $\Delta\tilde{\mathbf{u}}_i$ and $\Delta\tilde{\mathbf{v}}_i$ represent the approximation errors in the 1-D frequency responses $\Phi_i(\omega_1)$ and $\Gamma_i(\omega_2)$, respectively. From (3.19) it follows that

$$\epsilon_{l,m} = \sum_{i=1}^{K} [\sigma_i^{1/2}(\Delta\tilde{u}_{il}v_{im} + u_{il}\Delta\tilde{v}_{im}) + \Delta\tilde{u}_{il}\Delta\tilde{v}_{im}] - \sum_{i=K+1}^{r} \sigma_i u_{il}v_{im}$$

where u_{il} denotes the lth component of vector $\mathbf{u}_i$, etc Since vectors $\mathbf{u}_i$ and $\mathbf{v}_i$ are unit vectors, an upper bound on the magnitude of $e_{l,m}$ can be obtained as

$$|e_{l,m}| \leq \sum_{i=1}^{K}[\sigma_i^{1/2}(\epsilon_{1i} + \epsilon_{2i}) + \epsilon_{1i}\epsilon_{2i}] + \sum_{i=K+1}^{r} \sigma_i \tag{3.20}$$

where ϵ_{1i} and ϵ_{2i} represent the maximum approximation errors in the frequency responses $\Phi_i(\omega_1)$ and $\Gamma_i(\omega_2)$, respectively.

If the approximation error in each 1-D filter design is reasonably small, then the high-order terms $\epsilon_{1i}\epsilon_{2i}$ in inequality (3.20) can be neglected. We note also that the right-hand side in (3.20) is independent of l and m, and hence (3.20) holds also for the maximum of $|e_{l,m}|$. We, therefore, have obtained an upper bound for the maximum approximation error over the set of sampled frequency points as

$$e_\infty = \max_{1\leq l\leq L,\ 1\leq m\leq M} |e_{l,m}| \leq \epsilon_{pK} + \epsilon_{rK} \tag{3.21}$$

where the principal error ϵ_{pK} and the residual error ϵ_{rK} are defined by

$$\epsilon_{pK} = \sum_{i=1}^{K} \sigma_i^{1/2}(\epsilon_{1i} + \epsilon_{2i}) \tag{3.22}$$

and

$$\epsilon_{rK} = \sum_{i=K+1}^{r} \sigma_i$$

respectively.

The error bound given by (3.21) shows clearly how the choice of K and the approximation error introduced by a specific 1-D design technique affect the overall approximation error in the 2-D filter. As $K \to r$, the residual error ϵ_{rK} will tend to zero and the approximation error in the design of the 2-D filter is reduced to that introduced in the design of the 1-D FIR filters but the number

of multiplications required becomes very large. Hence in practice, K should be chosen to keep the number of parallel sections small and the residual error ϵ_{rK} acceptable. Having determined the value of K, the principal error ϵ_{pK} can also be made acceptable by controlling the 1-D design errors ϵ_{1i} and ϵ_{2i} for $1 \leq i \leq K$, by increassing the orders of the 1-D filters or by using a better design method for the 1-D filters. From (3.22) it follows that small approximation error should be obtained in the design of those 1-D filters that correspond to the large singular values.

3.4 Example

In this section, a 2-D lowpass FIR filter with a rotated elliptical passband is designed to illustrate the effectiveness of the proposed method.

The desired amplitude response of the 2-D lowpass FIR filter, shown in Figure 3.2, is specified by

$$|H(e^{j\omega_1 T_1}, e^{j\omega_2 T_2})| = \begin{cases} 1 & (\tilde{\omega}_1/\omega_{p_1})^2 + (\tilde{\omega}_2/\omega_{p_2})^2 \leq 1 \\ 0 & (\tilde{\omega}_1/\omega_{a_1})^2 + (\tilde{\omega}_2/\omega_{a_2})^2 > 1 \end{cases}$$

where

$$\begin{aligned} \tilde{\omega}_1 &= \omega_1 \cos\alpha + \omega_2 \sin\alpha \\ \tilde{\omega}_2 &= -\omega_1 \sin\alpha + \omega_2 \cos\alpha \end{aligned}$$

and

$$\alpha = \pi/6,\ \omega_{p_1} = 0.32\pi,\ \omega_{p_2} = 0.52\pi,\ \omega_{a_1} = 0.48\pi,\ \omega_{a_2} = 0.68\pi.$$

With $L = M = 61$, the corresponding sampled amplitude response $\mathbf{A} = |H(e^{j\pi\mu_l}, e^{j\pi\nu_m})|$ can be expressed as

$$|H(e^{j\pi\mu_l}, e^{j\pi\nu_m})| = \begin{cases} 1 & (\tilde{\mu}_l/\omega_{c_1})^2 + (\tilde{\nu}_m/\omega_{c_2})^2 \leq 1/\pi^2 \\ 0 & \text{otherwise} \end{cases}$$

Table 3.1: Maximum Passband and Stopband Errors for FIR Filter Designed by Using the General SVD Method

K	Passband	Stopband
4	0.1360	0.1025
5	0.0851	0.0771
12	0.0393	0.0255
14	0.0365	0.0129
15	0.0295	0.0121
16	0.0150	0.0115
18	0.0124	0.0113
25	0.0117	0.0102

where

$$\begin{aligned} \tilde{\mu}_l &= \mu_l \cos\alpha + \nu_m \sin\alpha, \qquad 1 \le l \le 61 \\ \tilde{\nu}_m &= -\mu_l \sin\alpha + \nu_m \cos\alpha, \qquad 1 \le m \le 61 \end{aligned}$$

and

$$\omega_{c_1} = \frac{1}{2}(\omega_{a_1} + \omega_{p_1}), \qquad \omega_{c_2} = \frac{1}{2}(\omega_{a_2} + \omega_{p_2}).$$

The software package MATLAB has been used to obtain the SVD of matrix **A**. The 1-D FIR filters were designed by using the Fourier series method along with the Kaiser window function. As may be expected, the higher the order of the 1-D filters, the lower the approximation error. By trial and error, it has been found that a value of 29 for both N_1 and N_2 gives satisfactory results.

There were 25 nonzero singular values resulting from the SVD of matrix **A**. The resulting amplitude response of the lowpass 2-D FIR filter for $K = 12$ is shown in Figure 3.3. The maximum passband and stopband errors for $K = 4$, 5, 12, 14, 15 ,16, 18 and 25 are given in Table 3.1.

3.5 Conclusion

In this chapter, the SVD has been applied in conjunction with 1-D FIR filter techniques for the design of 2-D, causal, linear-phase FIR filters with *arbitrary* amplitude responses. The method is relatively simple to apply and leads to a parallel arrangement of pairs of cascade sections. This configuration is, therefore, suitable for parallel processing and is amenable to VLSI implementation.

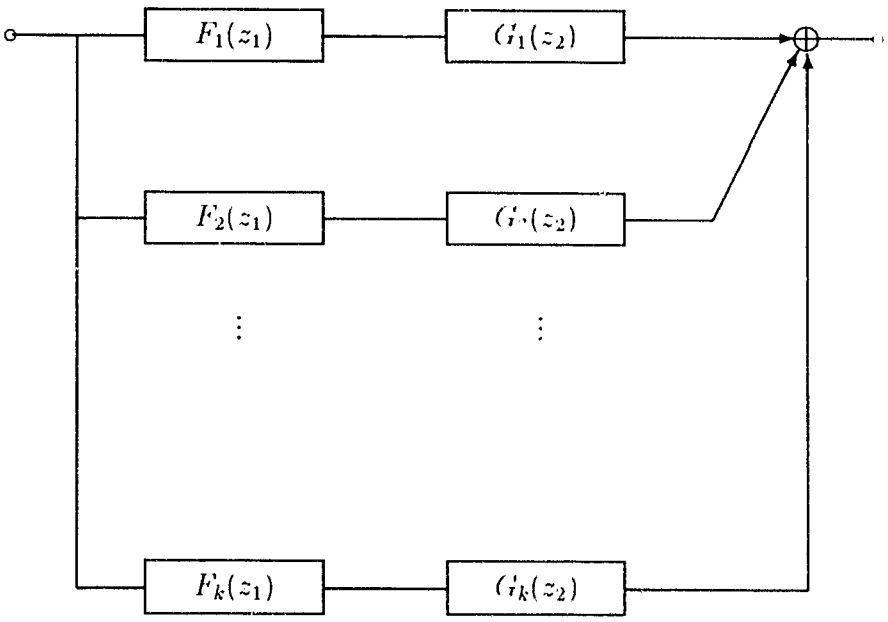

Figure 3.1: Parallel realization of 2-D digital filter.

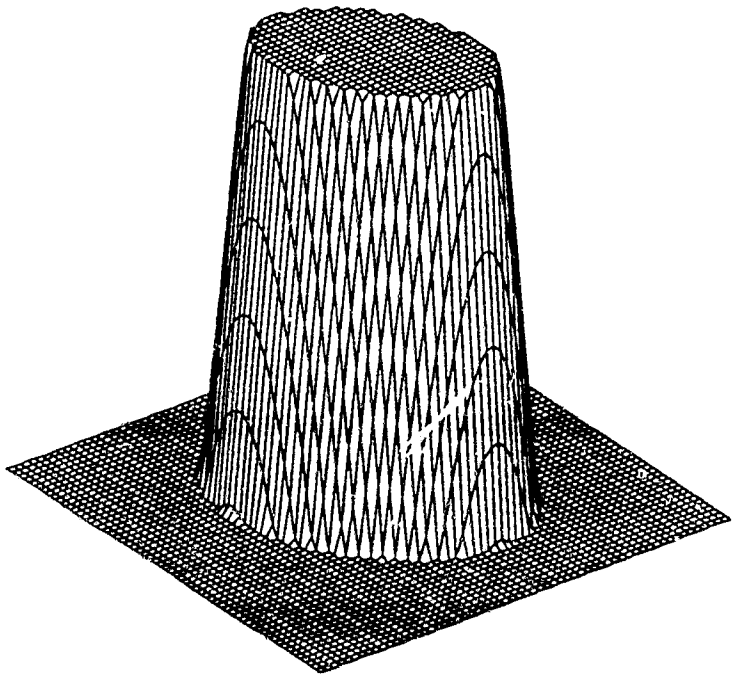

Figure 3.2: Ideal amplitude response of 2-D filter with rotated elliptical passband.

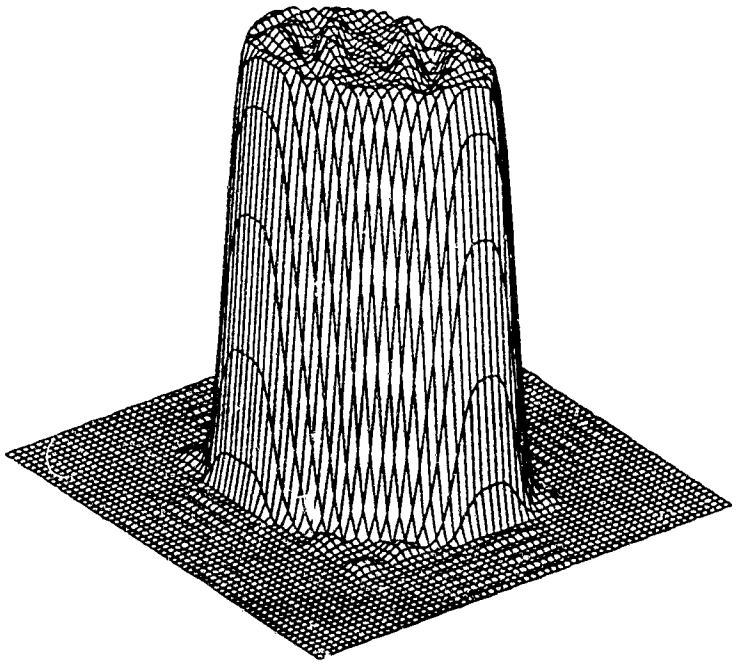

Figure 3.3: Amplitude response of 2-D FIR filter with rotated elliptical passband obtained by using SVD method ($N_1 = N_2 = 29, K = 12$).

Chapter 4

Design of 2-D IIR Digital Filters by Using Balanced Approximation Method

4.1 Introduction

Two dimensional IIR filters have advantages over 2-D FIR filters in terms of storage requirement and computation efficiency. The major drawback of 2-D IIR filters is that stability has to be considered during the design and that linear phase is generally difficult to obtain. In this chapter, a model reduction technique known as balanced approximation (BA) method [26, 25, 27, 38] is applied in the design of 2-D digital filters. The design starts with a linear-phase 2-D FIR filter of the type that may be obtained by using the proposed general SVD method and concludes with a lower-order 2-D digital filter, usually an IIR filter. The resulting 2-D filter has a specified amplitude response and its phase response is approximately linear over the passband. Furthermore, the 2-D filter obtained is more economical and computationally more efficient than the original 2-D FIR filter and the stability of the filter is quaranteed. Combining the general SVD and BA methods leads to excellent results as is demonstrated by designing a lowpass filter with a rotated elliptical passband.

In Section 4.2, the background information regarding model reduction using

the balanced approximation (BA) will be briefly reviewed, and some preliminary material regarding 1-D and 2-D gramians and balanced realizations is also presented. In Section 4.3 the 2-D IIR filter design using the BA method is discussed. In Section 4.4 an algorithm that implements the BA design procedure is given. In Section 4.5 properties related to the BA design are presented. It will be shown that the designs obtained are causal and locally quasi-balanced, and in cases where IIR designs are obtained, stability is guaranteed. In Section 4.6 the BA method is applied to the linear-phase 2-D FIR digital filter designed in Chapter 3 to illustrate the effectiveness of the method.

4 eliminaries

'ground Information

to approximate a model of a linear system by a lower order
s m of reduced order can be implemented more economically
an case of discrete systems, reduced computational complexity
can Model reduction techniques have been an active research area
for the last two decades. The BA technique initiated by Moore [26] is one of the most frequently used techniques. By using the BA technique, any weak subsystem which contributes little to the impulse response of the total system will be eliminated. More specifically, the least reachable and least observable states of the state-space representation of the original system will be eliminated. This is accomplished by first using a balanced transformation to transform the state space representation of the given system to a coordinate system where each state is equally reachable and observable; then the least reachable and least observable states are deleted.

The controllability and observability gramians carry useful information regarding the input-output behaviour of the system and they are, as a result, the two key factors used to determine the balanced transformation matrix. In the following section, some preliminary material dealing with 1-D and 2-D gramians

and the BA technique is presented.

4.2.2 One-Dimensional Balanced Realization

Consider a 1-D stable nth-order digital filter characterized by the state-space difference equations

$$\begin{aligned} \mathbf{x}(k+1) &= \mathbf{A}\mathbf{x}(k) + \mathbf{b}u(k) \\ y(k) &= \mathbf{c}\mathbf{x}(k) + du(k) \end{aligned} \tag{4.1}$$

where $\mathbf{A} \in R^{n\times n}$. The controllability and observability gramians are defined by

$$\mathbf{K}_1 = \sum_{k=0}^{\infty} \mathbf{A}^k \mathbf{b}\mathbf{b}^T (\mathbf{A}^T)^k \tag{4.2}$$

and

$$\mathbf{W}_1 = \sum_{k=0}^{\infty} (\mathbf{A}^T)^k \mathbf{c}^T \mathbf{c}\mathbf{A}^k \tag{4.3}$$

respectively. The two gramians can be obtained by solving the following two Lyapunov equations [51]

$$\mathbf{A}\mathbf{K}_1\mathbf{A}^T - \mathbf{K}_1 = -\mathbf{b}\mathbf{b}^T \tag{4.4}$$

$$\mathbf{A}^T\mathbf{W}_1\mathbf{A} - \mathbf{W}_1 = -\mathbf{c}^T\mathbf{c} \tag{4.5}$$

Alternatively by writing

$$\mathbf{f}_1(z) = (z\mathbf{I} - \mathbf{A})^{-1}\mathbf{b} \tag{4.6}$$

$$\mathbf{g}_1(z) = \mathbf{c}(z\mathbf{I} - \mathbf{A})^{-1} \tag{4.7}$$

and applying Parseval's relation, the two gramians $\mathbf{K}_1$ and $\mathbf{W}_1$ can also be expressed in terms of complex integrals as

$$\mathbf{K}_1 = \frac{1}{(2\pi j)} \oint_{|z|=1} \mathbf{f}_1(z)\mathbf{f}_1^H(z)z^{-1}\, dz \tag{4.8}$$

and

$$\mathbf{W}_1 = \frac{1}{(2\pi j)} \oint_{|z|=1} \mathbf{g}_1^H(z)\mathbf{g}_1(z)z^{-1}\, dz \tag{4.9}$$

where $\mathbf{f}_1^H(z)$ denotes the conjugate transpose of $\mathbf{f}_1(z)$. For a stable system, a nonsingular similarity transformation can be obained [51], which can transform the original system to a balanced system by simultaneously diagonalizing the two gramians such that

$$\mathbf{K}_d = \mathbf{W}_d = \Sigma = \text{diag}\,[\sigma_1 \cdots \sigma_n] \tag{4.10}$$

where $\sigma_1 \geq \cdots \geq \sigma_n > 0$. Since gramians contain measures of input to state and state-to-output couplings, and the balanced realization provides the system a coordinate setting where these couplings are equally weighted so that those state components which are weakly coupled may be discarded.

For example, if

$$\sigma_1 \geq \cdots \geq \sigma_r >> \sigma_{r+1} \geq \cdots \geq \sigma_n > 0$$

then $\mathbf{A}$, $\mathbf{b}$, and $\mathbf{c}$ of (4.1) can be partitioned as

$$\mathbf{A} = \begin{bmatrix} \mathbf{A}_{11} & \mathbf{A}_{12} \\ \mathbf{A}_{21} & \mathbf{A}_{22} \end{bmatrix}, \quad \mathbf{b} = \begin{bmatrix} \mathbf{b}_1 \\ \mathbf{b}_2 \end{bmatrix}, \quad \text{and } \mathbf{c} = \begin{bmatrix} \mathbf{c}_1 & \mathbf{c}_2 \end{bmatrix}$$

with $\mathbf{A}_{11} \in R^{r\times r}$, $\mathbf{b}_1 \in R^{r\times 1}$, and $\mathbf{c}_1 \in R^{1\times r}$. Therefore, a reduced system $(\mathbf{A}_{11}, \mathbf{b}_1, \mathbf{c}_1)$ of order r can be formed. It has been proven that such a lower order system represents a good approximation of the original system in terms of the L_∞ norm.

In the next section, it is shown that 1-D gramians can be extended to the 2-D case. In this way, a 2-D balanced realization and model reduction using the BA technique can be obtained.

4.2.3 Two-Dimensional Balanced Realization

In this section, some preliminary material regarding the 2-D balanced realization and model reduction using the BA technique is reviewed.

A 2-D digital filter of order $(N_1,\ N_2)$ can be represented by Roesser's local state-space equation

$$\begin{bmatrix} \mathbf{x}^h(i+1,\ j) \\ \mathbf{x}^v(i,\ j+1) \end{bmatrix} = \begin{bmatrix} \mathbf{A}_1 & \mathbf{A}_2 \\ \mathbf{A}_3 & \mathbf{A}_4 \end{bmatrix} \begin{bmatrix} \mathbf{x}^h(i,\ j) \\ \mathbf{x}^v(i,\ j) \end{bmatrix} + \begin{bmatrix} \mathbf{b}_1 \\ \mathbf{b}_2 \end{bmatrix} u(i,j) \equiv \mathbf{A}\mathbf{x} + \mathbf{b}u$$

$$y(i,j) = \begin{bmatrix} \mathbf{c}_1 & \mathbf{c}_2 \end{bmatrix} \begin{bmatrix} \mathbf{x}^h(i,\ j) \\ \mathbf{x}^v(i,\ j) \end{bmatrix} + d\ u(i,j) \equiv \mathbf{c}\mathbf{x} + du \tag{4.11}$$

where $\mathbf{A}_1 \in R^{N_1 \times N_2}$, $\mathbf{A}_4 \in R^{N_2 \times N_2}$, and $\mathbf{x}^h(i,\ j) \in R^{N_1}$, $\mathbf{x}^v(i,\ j) \in R^{N_2}$ form the local state for the system at (i,j). The 2-D z transform of the above state equations yields the transfer function

$$H(z_1,\ z_2) = \mathbf{c}[\mathbf{I}(z_1,\ z_2) - \mathbf{A}]^{-1}\mathbf{b} + d = \frac{n(z_1,\ z_2)}{p(z_1,\ z_2)} \tag{4.12}$$

where $\mathbf{I}(z_1,\ z_2) = z_1\mathbf{I}_{N_1} \oplus z_2\mathbf{I}_{N_2}$ and $p(z_1,\ z_2) = \det\ [\mathbf{I}(z_1,\ z_2) - \mathbf{A}]$. Throughout the chapter, we assume that

$$p(z_1,\ z_2) \neq 0 \ \text{ for } (z_1,\ z_2) \in \{(z_1,\ z_2):\ |z_1| \geq 1,\ |z_2| \geq 1\} \tag{4.13}$$

which guarantees the BIBO stability of the system.

If

$$\mathbf{f}_2(z_1,\ z_2) = [\mathbf{I}(z_1,\ z_2) - \mathbf{A}]^{-1}\mathbf{b}$$

and

$$\mathbf{g}_2(z_1,\ z_2) = \mathbf{c}[\mathbf{I}(z_1,\ z_2) - \mathbf{A}]^{-1}$$

then the generalized reachability and observability gramians of (4.11) [25] are defined as

$$\mathbf{K}_2 = \frac{1}{(2\pi j)^2} \oint_{|z_1|=1} \oint_{|z_2|=1} \mathbf{f}_2(z_1,\ z_2)\mathbf{f}_2{}^H(z_1,\ z_2) z_1^{-1} z_2^{-1}\ dz_1\ dz_2 \tag{4.14}$$

and

$$\mathbf{W}_2 = \frac{1}{(2\pi j)^2} \oint_{|z_1|=1} \oint_{|z_2|=1} \mathbf{g}_2^H(z_1,\ z_2)\mathbf{g}_2(z_1,\ z_2)z_1^{-1}z_2^{-1}\ dz_1\ dz_2 \qquad (4.15)$$

respectively. Further denote the N_1-dimensional upper left blocks and the N_2-dimensional lower left blocks of $\mathbf{K}_2$ and $\mathbf{W}_2$ by $\mathbf{K}_{11}$, $\mathbf{K}_{22}$ and $\mathbf{W}_{11}$, $\mathbf{W}_{22}$, respectively. Just as in the 1-D case, the generalized gramians provide certain system invariants which play an essential role in extending the concept of balanced realization to the 2-D case. It has been proven that [25] the eigenvalues of $\mathbf{K}_1\mathbf{W}_1$ and $\mathbf{K}_2\mathbf{W}_2$ are invariant under 2-D similarity transformations and, further, $\mathbf{K}_i$ ($\mathbf{W}_i$) for $i = 1, 2$ are positive definite if system (4.11) is locally reachable (observable) [25].

System (4.11) is said to be (locally) balanced if

$$\mathbf{K}_{11} = \mathbf{W}_{11} = \text{diag}\ (\sigma_{11},\ \sigma_{12},\ \cdots,\ \sigma_{1N_1})$$

and

$$\mathbf{K}_{22} = \mathbf{W}_{22} = \text{diag}\ (\sigma_{21},\ \sigma_{22},\ \cdots,\ \sigma_{2N_2})$$

where σ_{1i} $(1 \le i \le N_1)$ and σ_{2i} $(1 \le i \le N_2)$ are called the Hankel singular values of the system with $\sigma_{11} \ge \cdots \ge \sigma_{1N_1} \ge 0$, $\sigma_{21} \ge \cdots \ge \sigma_{2N_2} \ge 0$. If system (4.11) is not locally balanced, then a 2-D balancing transformation $\mathbf{T} = \mathbf{T}_1 \otimes \mathbf{T}_2$ can be found (assuming that (4.11) is locally reachable and observable) such that the system realization ($\mathbf{T}^{-1}\mathbf{A}\mathbf{T}$, $\mathbf{T}^{-1}\mathbf{b}$, $\mathbf{c}\mathbf{T}$, d) is locally balanced. A balancing transformation $\mathbf{T}$ can be computed by using $\mathbf{K}_{ii}$ and $\mathbf{W}_{ii}$ (i=1, 2) through any reliable algorithm for the 1-D balancing transformation [54] (see also Appendix D). Once a balanced realization, say ($\hat{\mathbf{A}}$, $\hat{\mathbf{b}}$, $\hat{\mathbf{c}}$, d), is found, a reduced state-space model ($\mathbf{A}_r$, $\mathbf{b}_r$, $\mathbf{c}_r$, d) of order $(r_1,\ r_2)$ can be obtained by subpartitioning $\hat{\mathbf{A}}$, $\hat{\mathbf{b}}$,

and $\hat{\mathbf{c}}$ as

$$\hat{\mathbf{A}} = \left[\begin{array}{c|c|c|c} \overbrace{\mathbf{A}_{1r}}^{r_1} & * & \overbrace{\mathbf{A}_{2r}}^{r_2} & * \\ \hline * & * & * & * \\ \hline \mathbf{A}_{3r} & * & \mathbf{A}_{4r} & * \\ \hline * & * & * & * \end{array}\right], \qquad \hat{\mathbf{b}} = \left[\begin{array}{c} \mathbf{b}_{1r} \\ * \\ \hline \mathbf{b}_{2r} \\ * \end{array}\right]$$

and

$$\hat{\mathbf{c}} = \left[\begin{array}{c|c} \overbrace{\mathbf{c}_{1r}}^{r_1} & * \end{array} \quad \begin{array}{c|c} \overbrace{\mathbf{c}_{2r}}^{r_2} & * \end{array}\right] \tag{4.16}$$

respectively, and then taking $(\mathbf{A}_r,\ \mathbf{b}_r,\ \mathbf{c}_r, d)$ where

$$\mathbf{A}_r = \left[\begin{array}{c|c} \mathbf{A}_{1r} & \mathbf{A}_{2r} \\ \hline \mathbf{A}_{3r} & \mathbf{A}_{4r} \end{array}\right], \quad \mathbf{b}_r = \left[\begin{array}{c} \mathbf{b}_{1r} \\ \hline \mathbf{b}_{2r} \end{array}\right] \text{ and } \mathbf{c}_r = \left[\begin{array}{c|c} \mathbf{c}_{1r} & \mathbf{c}_{2r} \end{array}\right] \tag{4.17}$$

to be an approximation for $(\hat{\mathbf{A}},\ \hat{\mathbf{b}},\ \hat{\mathbf{c}}, d)$. In the next section, particular attention will be given to the application of this reduction approach to 2-D FIR filters. It will be shown that the filter with reduced order is usually an IIR filter with separable denominator, which is always stable.

4.3 Design

The BA method has been applied extensively in the past in both 1-D as well as 2-D dynamical systems [47, 48, 49, 51, 52]. The method leads to more economical systems and in the case of discrete systems to reduced computational complexity.

Recently, it has been shown by Kimurd and Honoki that the BA method is also applicable in the design of 1-D digital filters [27]. These researchers have demonstrated that given a 1-D linear-phase FIR filter, a corresponding 1-D digital filter of reduced order can be obtained, which preserves approximately the

amplitude and phase responses of the original FIR filter. In this section, it is shown that the BA method can be put to good use in the design of 2-D digital filters as well.

Let

$$H(z_1,\ z_2) = \sum_{n_1=0}^{N_1} \sum_{n_2=0}^{N_2} h(n_1\ n_2) z_1^{-n_1} z_2^{-n_2} \tag{4.18}$$

be the transfer function of a 2-D FIR filter of order $(N_1,\ N_2)$. In Roesser's local state-space characterization, (4.18) can be represented by (4.11) with

$$\mathbf{A} = \left[\begin{array}{ccccc|cccccc} 0 & 1 & 0 & \dots & 0 & h_{1N_2} & \dots & & \dots & h_{11} \\ \vdots & \vdots & \vdots & \dots & \vdots & \vdots & \dots & & \dots & \vdots \\ 0 & 0 & 0 & \dots & 1 & \vdots & \dots & & \dots & \vdots \\ 0 & 0 & 0 & \dots & 0 & h_{N_1N_2} & \dots & & \dots & h_{N_11} \\ \hline & & & & & 0 & 1 & 0 & \dots & 0 \\ & & & & & \vdots & \vdots & \vdots & \dots & \vdots \\ & & \mathbf{0} & & & 0 & 0 & 0 & \dots & 1 \\ & & & & & 0 & 0 & 0 & \dots & 0 \end{array}\right],$$

$$\mathbf{b} = \left[\begin{array}{c} h_{10} \\ \vdots \\ h_{N_10} \\ \hline 0 \\ \vdots \\ 0 \\ 1 \end{array}\right],\quad \mathbf{c} = \left[\begin{array}{ccccc|ccc} 1 & 0 & \dots & 0 & | & h_{0N_2} & \dots & h_{01} \end{array}\right] \text{ and } d = h_{00} \tag{4.19}$$

Note that in $\mathbf{A}$, $\mathbf{A}_3 = 0$ means that the filter is separable. In addition, both $\mathbf{A}_1$ and $\mathbf{A}_4$ are nilpotent since $\mathbf{A}_1^{N_1} = 0$ and $\mathbf{A}_4^{N_2} = 0$. These properties considerably simplify the procedure of solving two Lyapunov equations which are described below.

To compute $\mathbf{K}_{ii}$ and $\mathbf{W}_{ii}$ $(i = 1,\ 2)$, we write $H(z_1,\ z_2)$ as

$$\begin{aligned} H(z_1, z_2) &= \sum_{n_1=0}^{N_1} [\sum_{n_2=0}^{N_2} h(n_1,\ n_2) z_2^{-n_2}] z_1^{-n_1} \equiv \sum_{n_1=0}^{N_1} b_{1n_1}(z_2) z_1^{-n_1} \\ &= \sum_{n_2=0}^{N_2} [\sum_{n_1=0}^{N_1} h(n_1,\ n_2) z_1^{-n_1}] z_2^{-n_2} \equiv \sum_{n_2=0}^{N_2} c_{2n_2}(z_1) z_2^{-n_2} \end{aligned} \tag{4.20}$$

and define

$$\begin{aligned} \tilde{\mathbf{A}}_1(z_2) &= \mathbf{A}_1 + \mathbf{A}_2(z_2\mathbf{I} - \mathbf{A}_4)^{-1}\mathbf{A}_3 \\ \tilde{\mathbf{b}}_1(z_2) &= \mathbf{b}_1 + \mathbf{A}_2(z_2\mathbf{I} - \mathbf{A}_4)^{-1}\mathbf{b}_2 \\ \mathbf{c}_1(z_2) &= \mathbf{c}_1 + \mathbf{c}_2(z_2\mathbf{I} - \mathbf{A}_4)^{-1}\mathbf{A}_3 \\ \tilde{\mathbf{A}}_2(z_1) &= \mathbf{A}_4 + \mathbf{A}_3(z_1\mathbf{I} - \mathbf{A}_1)^{-1}\mathbf{A}_2 \\ \tilde{\mathbf{b}}_2(z_1) &= \mathbf{b}_2 + \mathbf{A}_3(z_1\mathbf{I} - \mathbf{A}_1)^{-1}\mathbf{b}_1 \\ \tilde{\mathbf{c}}_2(z_1) &= \mathbf{c}_2 + \mathbf{c}_1(z_1\mathbf{I} - \mathbf{A}_1)^{-1}\mathbf{A}_2 \end{aligned} \tag{4.21}$$

It is known that $\mathbf{K}_{11}$ can be found through the integral [25]

$$\mathbf{K}_{11} = \frac{1}{(2\pi j)} \oint_{|z_2|=1} \mathbf{K}_1(z_2) z_2^{-1}\, dz_2 \tag{4.22}$$

where $\mathbf{K}_1(z_2)$ is the positive-definite Hermitian solution of the Lyapunov equation

$$\tilde{\mathbf{A}}_1(z_2)\mathbf{K}_1(z_2)\tilde{\mathbf{A}}_1^*(z_2) - \mathbf{K}_1(z_2) = -\tilde{\mathbf{b}}_1(z_2)\tilde{\mathbf{b}}_1^*(z_2) \tag{4.23}$$

For an FIR filter, $\tilde{\mathbf{A}}_1(z_2) = \mathbf{A}_1$ and

$$\tilde{\mathbf{b}}_1(z_2) = \begin{bmatrix} b_{11}(z_2) \\ \vdots \\ b_{1N_1}(z_2) \end{bmatrix}$$

where $b_{1n_1}(z_2)$, $1 \leq n_1 \leq N_1$, are defined by (4.20). Hence (4.23) becomes

$$\mathbf{A}_1\mathbf{K}_1(z_2)\mathbf{A}_1^T - \mathbf{K}_1(z_2) = -\hat{\mathbf{b}}_1(z_2)\tilde{\mathbf{b}}_1^*(z_2) \tag{4.24}$$

Since $\mathbf{A}_1$ is nilpotent, the solution $\mathbf{K}_1(z_2)$ is given by

$$\mathbf{K}_1(z_2) = \sum_{n_1=0}^{N_1-1} \mathbf{A}_1^{n_1}\tilde{\mathbf{b}}_1(z_2)\tilde{\mathbf{b}}_1^*(z_2)(\mathbf{A}_1^T)^{n_1} \tag{4.25}$$

Therefore

$$\mathbf{K}_{11} = \sum_{n_1=0}^{N_1-1} \mathbf{A}_1^{n_1}\mathbf{P}(\mathbf{A}_1^T)^{n_1} \tag{4.26}$$

with

$$\mathbf{P} = \frac{1}{(2\pi j)}\oint_{|z_2|=1} \tilde{\mathbf{b}}_1(z_2)\tilde{\mathbf{b}}_1^*(z_2)z_2^{-1}\, dz_2 \tag{4.27}$$

Evaluating the above integral shows that

$$\mathbf{P} = \mathbf{H}_b\mathbf{H}_b^T \tag{4.28}$$

where $\mathbf{H}_b$ is defined as

$$\mathbf{H}_b = \begin{bmatrix} h_{10} & h_{11} & \dots & h_{1N_2} \\ h_{20} & h_{21} & \dots & h_{2N_2} \\ \vdots & \vdots & \dots & \vdots \\ h_{N_10} & h_{N_11} & \dots & h_{N_1N_2} \end{bmatrix} \tag{4.29}$$

On the other hand, since both $\hat{\mathbf{A}}_2(z_1) = \mathbf{A}_4$ and $\tilde{\mathbf{b}}_2(z_1) = \mathbf{b}_2$ are independent of z_1 and of the filter coefficients, $\mathbf{K}_{22}$ is the positive-definite solution of the Lyapunov equation [38]

$$\mathbf{A}_4\mathbf{K}_{22}\mathbf{A}_4^T - \mathbf{K}_{22} = -\mathbf{b}_2\mathbf{b}_2^T \tag{4.30}$$

i.e.

$$\mathbf{K}_{22} = \sum_{n_2=0}^{N_2-1} \mathbf{A}_4^{n_2}\mathbf{b}_2\mathbf{b}_2^T(\mathbf{A}_4^T)^{n_2} = \mathbf{I}_{N_2} \tag{4.31}$$

Similarly, it can be shown that

$$\mathbf{W}_{11} = \mathbf{I}_{N_1} \tag{4.32}$$

and

$$\mathbf{W}_{22} = \sum_{n_2=0}^{N_2-1} (\mathbf{A}_4^T)^{n_2} \mathbf{Q} \mathbf{A}_4^{n_2} \tag{4.33}$$

with

$$\mathbf{Q} = \mathbf{H}_c^T \mathbf{H}_c \tag{4.34}$$

and

$$\mathbf{H}_c = \begin{bmatrix} h_{01} & h_{02} & \dots & h_{0N_2} \\ h_{11} & h_{12} & \dots & h_{1N_2} \\ \vdots & \vdots & \dots & \vdots \\ h_{N_1 1} & h_{N_1 2} & \dots & h_{N_1 N_2} \end{bmatrix} \tag{4.35}$$

Note that matrices $\mathbf{H}_b$, $\mathbf{H}_c$ are formed using subsets of the filter coefficients $h(n_1, n_2)$ for $0 \leq n_1 \leq N_1$, $0 \leq n_2 \leq N_2$ and can be obtained by properly segmenting the coefficient matrix $\mathbf{H} = \{h(n_1, n_2)\} \equiv \{h_{n_1 n_2}\}$, $0 \leq n_1 \leq N_1$, $0 \leq n_2 \leq N_2$ as follows:

$$\mathbf{H} = \left[\begin{array}{c|ccccc|c} & & & & & & \\ & h_{00} & h_{01} & h_{02} & \dots & h_{0N_2} & \\ \hline & h_{10} & h_{11} & h_{12} & \dots & h_{1N_2} & \\ & h_{20} & h_{21} & h_{22} & \dots & h_{2N_2} & \\ & \vdots & \vdots & \vdots & \dots & \vdots & \\ & H_{N_1 0} & h_{N_1 1} & h_{N_1 2} & \dots & h_{N_1 N_2} & \\ & & & & & & \\ \hline \end{array}\right] \tag{4.36}$$

Once $\mathbf{K}_{ii}$ and $\mathbf{W}_{ii}$ (i=1, 2) are found, a balancing transformation $\mathbf{T} = \mathbf{T}_1 \oplus \mathbf{T}_2$ with $\mathbf{T}_1$, $\mathbf{T}_2$ nonsingular can be computed such that the system realization $(\hat{\mathbf{A}}, \hat{\mathbf{b}}, \hat{\mathbf{c}}, \hat{d}) = (\mathbf{T}^{-1}\mathbf{A}\mathbf{T}, \mathbf{T}^{-1}\mathbf{b}, \mathbf{c}\mathbf{T}, d)$ is locally balanced. A reduced state-space

model ($\mathbf{A}_r$, $\mathbf{b}_r$, $\mathbf{c}_r$, d) of order (r_1, r_2) can be obtained by partitioning $\hat{\mathbf{A}}$, $\hat{\mathbf{b}}$, and $\hat{\mathbf{c}}$ as

$$\hat{\mathbf{A}} = \left[\begin{array}{c|c} \hat{\mathbf{A}}_1 & \hat{\mathbf{A}}_2 \\ \hline \hat{\mathbf{A}}_3 & \hat{\mathbf{A}}_4 \end{array}\right] = \left[\begin{array}{cc|cc} \mathbf{A}_{1r} & \mathbf{A}_{12} & \mathbf{A}_{2r} & \mathbf{A}_{22} \\ \mathbf{A}_{13} & \mathbf{A}_{14} & \mathbf{A}_{23} & \mathbf{A}_{24} \\ \hline \mathbf{0} & \mathbf{0} & \mathbf{A}_{4r} & \mathbf{A}_{42} \\ \mathbf{0} & \mathbf{0} & \mathbf{A}_{43} & \mathbf{A}_{44} \end{array}\right], \quad \hat{\mathbf{b}} = \left[\begin{array}{c} \hat{\mathbf{b}}_1 \\ \hline \hat{\mathbf{b}}_2 \end{array}\right] = \left[\begin{array}{c} \mathbf{b}_{1r} \\ * \\ \hline \mathbf{b}_{2r} \\ * \end{array}\right]$$

and

$$\hat{\mathbf{c}} = \left[\begin{array}{c|c} \hat{\mathbf{c}}_1 & \hat{\mathbf{c}}_2 \end{array}\right] = \left[\begin{array}{cc|cc} \mathbf{c}_{1r} & * & \mathbf{c}_{2r} & * \end{array}\right] \tag{4.37}$$

and then letting

$$\mathbf{A}_r = \left[\begin{array}{c|c} \mathbf{A}_{1r} & \mathbf{A}_{2r} \\ \hline \mathbf{0} & \mathbf{A}_{4r} \end{array}\right], \quad \mathbf{b}_r = \left[\begin{array}{c} \mathbf{b}_{1r} \\ \hline \mathbf{b}_{2r} \end{array}\right] \text{ and } \mathbf{c}_r = \left[\begin{array}{c|c} \mathbf{c}_{1r} & \mathbf{c}_{2r} \end{array}\right]$$

The transfer function of the reduced model is given by

$$H_r(z_1,\ z_2) = \mathbf{c}_r[\mathbf{I}_r(z_1,\ z_2) - \mathbf{A}_r]^{-1}\mathbf{b}_r + d$$

where $\mathbf{I}_r(z_1,\ z_2) = z_1\mathbf{I}_{r1} \oplus z_2\mathbf{I}_{r2}$.

The above method usually results in an IIR design which has approximately the same frequency and space-domain responses as the original FIR filter. The reduction in the filter order achieved in (4.37) leads to a more economical and computationally efficient design. Furthermore, if the original FIR filter has a linear phase response, an approximately linear phase response is achieved irrespective of whether an FIR or IIR filter is obtained.

4.4 Algorithm

An algorithm implementing the above design procedure is as follows:

1) Form $\mathbf{A}$, $\mathbf{b}$, $\mathbf{c}$ and calculate d according to (4.19) and obtain $\mathbf{A}_i$, $(1 \leq i \leq 4)$ $\mathbf{b}_1$, $\mathbf{b}_2$, $\mathbf{c}_1$ and $\mathbf{c}_2$ by subpartitioning $\mathbf{A}$, $\mathbf{b}$ and $\mathbf{c}$, respectively.

2) Obtain $\mathbf{H}_b$ and $\mathbf{H}_c$ by properly segmenting matrix $\mathbf{H}$ as in (4.36). Then compute matrices $\mathbf{P}$ and $\mathbf{Q}$ using (4.28) and (4.34), respectively.

3) Calculate $\mathbf{K}_{11}$ and $\mathbf{W}_{22}$ using equations (4.26) and (4.33), respectively. Then obtain $\mathbf{K}_{22}$ and $\mathbf{W}_{11}$ through (4.31) and (4.32).

4) From $\mathbf{K}_{11}\mathbf{W}_{11}$ and $\mathbf{K}_{22}\mathbf{W}_{22}$, compute the balancing transformations $\mathbf{T}_1$ and $\mathbf{T}_2$, respectively, by using Laub's algorithm (see Appendix D) and thereby obtain $\mathbf{T} = \mathbf{T}_1 \oplus \mathbf{T}_2$.

5) Form the balanced realization $(\hat{\mathbf{A}},\ \hat{\mathbf{b}},\ \hat{\mathbf{c}}, d)$ with $\hat{\mathbf{A}} = \mathbf{T}^{-1}\mathbf{A}\mathbf{T}$, $\hat{\mathbf{b}} = \mathbf{T}^{-1}\mathbf{b}$ and $\hat{\mathbf{c}} = \mathbf{c}\mathbf{T}$.

6) Determine the order $(r_1,\ r_2)$ of the reduced filter by neglecting the insignificant singular values calculated using Laub's algorithm.

7) Obtain the reduced state-space realization $(\mathbf{A}_r,\ \mathbf{b}_r,\ \mathbf{c}_r, d)$ of order $(r_1,\ r_2)$ by subpartitioning $\hat{\mathbf{A}}$, $\hat{\mathbf{b}}$, and $\hat{\mathbf{c}}$ as shown in (4.16) and (4.17).

4.5 Properties

The balanced approximation has been used extensively for the design of automatic controllers of 1-D linear systems owing to its good approximation accuracy. The maximum error introduced in 2-D separable discrete systems can be estimated as [50]

$$e_r = \max_{|z_1|=1,\ |z_2|=1} |H_r(z_1,\ z_2) - H(z_1,\ z_2)| \le \alpha_1 \sum_{i=r_1+1}^{N_1} \sigma_{1i} + \alpha_2 \sum_{i=r_2+1}^{N_2} \sigma_{2i} \tag{4.38}$$

where

$$H_r(z_1,\ z_2) = \mathbf{c}_r[\mathbf{I}_r(z_1,\ z_2) - \hat{\mathbf{A}}_r]^{-1}\mathbf{b}_r + d$$

σ_{1i} $(r_1 + 1 \leq i \leq N_1)$ and σ_{2i} $(r_2 + 1 \leq i \leq N_2)$ are the Hankel singular values of the original FIR filter, and α_1, α_2 are constants determined by $\{h_{ij}\}$. An important property of the BA method is that the stability of the reduced-order 2-D system is guaranteed provided that the original system is stable and has a separable denominator [50]. Since our design approach starts with an FIR filter, the above-mentioned properties apply. In what follows, we present two more properties of the BA method, which are especialy desirable for digital filters.

Theorem 4.1 If $\varphi_r(\omega_1,\ \omega_2)$ and $\varphi(\omega_1,\ \omega_2)$ are the phase responses of the reduced and the original filter, respectively, and Ω_p denotes the passband region of the original filter, then

$$\max_{\Omega_p} |\varphi_r(\omega_1,\ \omega_2) - \varphi(\omega_1,\ \omega_2)| \approx \delta_m \leq \frac{2\epsilon_r}{1 - \epsilon_p} \tag{4.39}$$

provided that $\epsilon_r << 1$, where ϵ_r is defined in (4.38), ϵ_p is the maximum error over the passband region for the original FIR filter,

$$\delta_m = \max_{\Omega_p} |1 - \frac{X(\omega_1,\ \omega_2)|X_r(\omega_1,\ \omega_2)|}{X_r(\omega_1,\ \omega_2)|X(\omega_1,\ \omega_2)|}|$$

$X(\omega_1,\ \omega_2)$ is the frequency response of the original filter given by (3.2), and

$$X_r(\omega_1,\ \omega_2) = H_r(e^{j\omega_1 T_1},\ e^{j\omega_2 T_2})$$

Proof: Writing

$$\begin{aligned} X(\omega_1,\ \omega_2) &= M(\omega_1,\ \omega_2)e^{j\varphi(\omega_1 T_1,\ \omega_2 T_2)} \\ X_r(\omega_1,\ \omega_2) &= M_r(\omega_1,\ \omega_2)e^{j\varphi_r(\omega_1 T_1,\ \omega_2 T_2)} \end{aligned}$$

the approximation error in the phase angle can be expressed as

$$\begin{aligned} |\varphi_r(\omega_1,\ \omega_2) - \varphi(\omega_1,\ \omega_2)| &= |\ln\ (\frac{X(\omega_1,\ \omega_2)M_r(\omega_1,\ \omega_2)}{X_r(\omega_1,\ \omega_2)M(\omega_1,\ \omega_2)})| \\ &= |\ln\ (1 + \delta)| \end{aligned}$$

where δ can be estimated as

$$\begin{aligned} |\delta| &= |1 - \frac{XM_r}{X_rM}| = \frac{|X_rM - X_rM_r + X_rM_r - XM_r|}{M_rM} \\ &\leq \frac{|X_r(M - M_r)| + |M_r(X_r - X)|}{M_rM} = \frac{||X| - |X_r|| + |X_r - X|}{M} \\ &\leq \frac{2|X_r - X|}{M} \end{aligned}$$

Hence

$$\delta_m = \max_{\Omega_p} |\delta| \leq \frac{N_{max}}{D_{min}} \leq \frac{2\epsilon_r}{1 - \epsilon_p} \tag{4.40}$$

where

$$N_{max} = 2 \max_{\Omega_p} |X_r - X|$$

and

$$D_{min} = \min_{\Omega_p} M$$

Since for a good FIR filter $\epsilon_p \approx 0$, (4.40) indicates that $|\delta| << 1$ if $\epsilon_r << 1$. Under these circumstances, we have

$$\begin{aligned} \max_{\Omega_p} |\varphi_r(\omega_1, \omega_2) - \varphi(\omega_1, \omega_2)| &= \max_{\Omega_p} |\ln(1 + \delta)| \\ &\approx \max_{\Omega_p} |\delta| \\ &\leq \frac{2\epsilon_r}{1 - \epsilon_p} \qquad \square \end{aligned}$$

If the BA method is used to approximate a high-order FIR filter, then by (4.38) the approximation error can be made very small if both

$$\sum_{i=r_1+1}^{N_1} \sigma_{1i} \quad \text{and} \quad \sum_{i=r_2+1}^{N_2} \sigma_{2i}$$

are very small. In such a case, the phase-response linearity of the original filter will be well preserved in the reduced filter by virtue of Theorem 4.1.

It is known that a locally balanced 2-D state-space digital filter has the lowest output roundoff noise and, further, its sensitivity to coefficient quantization is low [25, 32, 53]. The second property, stated in terms of Theorem 4.2 below ensures that like the original filter the reduced filter has low roundoff noise and low sensitivity to coefficient quantization since it is nearly balanced.

Theorem 4.2 If $\Sigma_{1r_1} = \text{diag}\,(\sigma_{11}, \cdots, \sigma_{1r_1})$, $\Sigma_{1c} = \text{diag}\,(\sigma_{1r_1+1}, \cdots, \sigma_{1N_1})$, $\Sigma_{2r_2} = \text{diag}\,(\sigma_{21}, \cdots, \sigma_{2r_2})$, $\Sigma_{2c} = \text{diag}\,(\sigma_{2r_2+1}, \cdots, \sigma_{2N_2})$, then

$$\begin{aligned}
\mathbf{A}_{1r}\Sigma_{1r_1}\mathbf{A}_{1r}^T - \Sigma_{1r_1} &\approx -(\mathbf{b}_{1r}\mathbf{b}_{1r}^T + \mathbf{A}_{2r}\Sigma_{2r_2}\mathbf{A}_{2r}^T) \\
\mathbf{A}_{1r}^T\Sigma_{1r_1}\mathbf{A}_{1r} - \Sigma_{1r_1} &\approx -\mathbf{c}_{1r}^T\mathbf{c}_{1r} \\
\mathbf{A}_{4r}\Sigma_{2r_2}\mathbf{A}_{4r}^T - \Sigma_{2r_2} &\approx -\mathbf{b}_{2r}\mathbf{b}_{2r}^T \\
\mathbf{A}_{4r}^T\Sigma_{2r_2}\mathbf{A}_{4r} - \Sigma_{2r_2} &\approx -(\mathbf{c}_{2r}^T\mathbf{c}_{2r} + \mathbf{A}_{2r}^T\Sigma_{1r_1}\mathbf{A}_{2r})
\end{aligned} \tag{4.41}$$

provided that

$$\sigma_{1r_1+1} << \sigma_{1r_1} \quad \text{and} \quad \sigma_{2r_2+1} << \sigma_{2r_2} \tag{4.42}$$

Proof: Since $(\hat{\mathbf{A}}, \hat{\mathbf{b}}, \hat{\mathbf{c}}, \hat{d})$ given by (4.37) represents a balanced realization of the original system, it satisfies [32] the relations

$$\hat{\mathbf{A}}_1\Sigma_1\hat{\mathbf{A}}_1^T - \Sigma_1 = -(\hat{\mathbf{b}}_1\hat{\mathbf{b}}_1^T + \hat{\mathbf{A}}_2\Sigma_2\hat{\mathbf{A}}_2^T) \tag{4.43}$$

$$\hat{\mathbf{A}}_1^T\Sigma_1\hat{\mathbf{A}}_1 - \Sigma_1 = -\hat{\mathbf{c}}_1^T\hat{\mathbf{c}}_1 \tag{4.44}$$

$$\hat{\mathbf{A}}_4\Sigma_2\hat{\mathbf{A}}_4^T - \Sigma_2 = -\hat{\mathbf{b}}_2\hat{\mathbf{b}}_2^T \tag{4.45}$$

$$\hat{\mathbf{A}}_4^T\Sigma_2\hat{\mathbf{A}}_4 - \Sigma_2 = -(\hat{\mathbf{c}}_2^T\hat{\mathbf{c}}_2 + \hat{\mathbf{A}}_2^T\Sigma_1\hat{\mathbf{A}}_2) \tag{4.46}$$

where

$$\Sigma_1 = \begin{bmatrix} \Sigma_{1r_1} & \mathbf{0} \\ \mathbf{0} & \Sigma_{1c} \end{bmatrix} \quad \text{and} \quad \Sigma_2 = \begin{bmatrix} \Sigma_{2r_2} & \mathbf{0} \\ \mathbf{0} & \Sigma_{2c} \end{bmatrix}$$

By taking the upper-left $r_1 \times r_1$ submatrices from (4.43) and (4.44) and the upper-left $r_2 \times r_2$ submatrices from (4.45) and (4.46), we obtain

$$\mathbf{A}_{1r}\boldsymbol{\Sigma}_{1r_1}\mathbf{A}_{1r}^T - \boldsymbol{\Sigma}_{1r_1} = -(\mathbf{b}_{1r}\mathbf{b}_{1r}^T + \mathbf{A}_{2r}\boldsymbol{\Sigma}_{2r_2}\mathbf{A}_{2r}^T) - \mathbf{E}_{1b} \tag{4.48a}$$

$$\mathbf{A}_{1r}^T\boldsymbol{\Sigma}_{1r_1}\mathbf{A}_{1r} - \boldsymbol{\Sigma}_{1r_1} = -\mathbf{c}_{1r}^T\mathbf{c}_{1r} - \mathbf{E}_{1c} \tag{4.48b}$$

$$\mathbf{A}_{4r}\boldsymbol{\Sigma}_{2r_2}\mathbf{A}_{4r}^T - \boldsymbol{\Sigma}_{2r_2} = -\mathbf{b}_{2r}\mathbf{b}_{2r}^T - \mathbf{E}_{2b} \tag{4.48c}$$

$$\mathbf{A}_{4r}^T\boldsymbol{\Sigma}_{2r_2}\mathbf{A}_{4r} - \boldsymbol{\Sigma}_{2r_2} = -(\mathbf{c}_{2r}^T\mathbf{c}_{2r} + \mathbf{A}_{2r}^T\boldsymbol{\Sigma}_{1r_1}\mathbf{A}_{2r}) - \mathbf{E}_{2c} \tag{4.48d}$$

Since the balanced realization $(\hat{\mathbf{A}},\ \hat{\mathbf{b}},\ \hat{\mathbf{c}},\ \hat{d})$ is obtained from a canonic realization of the FIR filter which is locally reachable and observable [28], it is also locally reachable and observable. Moreover it can be shown the $||\hat{\mathbf{A}}_1|| \leq 1$, $||\hat{\mathbf{A}}_4|| \leq 1$ and $||\hat{\mathbf{A}}_2|| \leq 1$ (see Theorem 5.2.3 and Appendix 5C of [50]) which imply that $||\mathbf{A}_{12}|| \leq 1$, $||\mathbf{A}_{42}|| \leq 1$, and $||\mathbf{A}_{22}|| \leq 1$, respectively. Therefore, matrices $\mathbf{E}_{ib}$ and $\mathbf{E}_{ic}$ (i = 1, 2) can be estimated as

$$\begin{aligned}
\| \mathbf{E}_{1b} \| &= \| \mathbf{A}_{12}\boldsymbol{\Sigma}_{1c}\mathbf{A}_{12}^T + \mathbf{A}_{22}\boldsymbol{\Sigma}_{2c}\mathbf{A}_{22}^T \| \leq \| \mathbf{A}_{12} \|^2 \| \boldsymbol{\Sigma}_{1c} \| + \| \mathbf{A}_{22} \|^2 \| \boldsymbol{\Sigma}_{2c} \| \\
&\leq \sigma_{1r_1+1} + \sigma_{2r_2+1} \\
\| \mathbf{E}_{1c} \| &= \| \mathbf{A}_{12}^T\boldsymbol{\Sigma}_{1c}\mathbf{A}_{12} \| \leq \| \mathbf{A}_{12} \|^2 \| \boldsymbol{\Sigma}_{1c} \| \leq \sigma_{1r_1+1} \\
\| \mathbf{E}_{2b} \| &= \| \mathbf{A}_{42}\boldsymbol{\Sigma}_{2c}\mathbf{A}_{42}^T \| \leq \| \mathbf{A}_{42} \|^2 \| \boldsymbol{\Sigma}_{2c} \| \leq \sigma_{2r_2+1} \\
\| \mathbf{E}_{2c} \| &= \| \mathbf{A}_{42}^T\boldsymbol{\Sigma}_{2c}\mathbf{A}_{42} + \mathbf{A}_{22}^T\boldsymbol{\Sigma}_{1c}\mathbf{A}_{22} \| \leq \sigma_{1r_1+1} + \sigma_{2r_2+1}
\end{aligned}$$

Consequently, matrices $\mathbf{E}_{ib}$ and $\mathbf{E}_{ic}$ $(i = 1, 2)$ can be neglected from equations (4.48) if the conditions in (4.42) are satisfied. This modifies (4.43) to (4.11). □

4.6 Example

In order to demonstrate the effectiveness of the BA method, it was applied to the linear-phase 2-D FIR digital filter designed with $K = 25$ in Chapter 3. Following

the steps of the algorithm described above, the balanced realization was first obtained. Then by examining the two sets of singular values and neglecting the insignificant ones, reduced IIR realizations of orders $(r_1,\ r_2) = (10,\ 12)$ and $(13,\ 15)$ were obtained. In the first realization, all σ_{1i} and σ_{2i} that are less than 0.4% of σ_{11} and σ_{21}, respectively, were ignored and in the second realization all σ_{1i} and σ_{2i} that are less than 0.02% of σ_{11} and σ_{21}, respectively, were ignored. For the sake of comparison, the BA method was also applied to an FIR filter of order (31, 31) to obtain an IIR filter of order (15, 17). The maximum passband and stopband errors for the three designs are given in Table 4.1. The amplitude response of the reduced filter for the case $(r_1,\ r_2) = (13,\ 15)$ is depicted in Figure 4.1.

The group delays of the 2-D IIR filter, namely

$$\tau_1 = -\frac{\partial\ \{\arg\ [H(e^{j\omega_1 T_1},\ e^{j\omega_2 T_2})]\}}{\partial \omega_1}$$

and

$$\tau_2 = -\frac{\partial\ \{\arg\ [H(e^{j\omega_1 T_1},\ e^{j\omega_2 T_2})]\}}{\partial \omega_2}$$

were used to construct the contour plots of Figures 4.2 and 4.3, respectively, for the case $(r_1,\ r_2) = (13,\ 15)$, assuming fifty layers. It is clearly seen that the linear-phase characteristic of the original FIR filter is well preserved over the entire passband of the 2-D filter. The maximum relative errors in the group delays are given in Table 4.2.

We conclude this section with a brief discussion on the efficiency of the filters obtained by the BA method. We observe in Tables 4.1 and 4.2 that if the errors in the amplitude response and the group delays are required to be less than 3.0% and 1.5%, respectively, an IIR filter of order (13, 15) is required. The same amplitude-response accuracy can be achieved by using an FIR filter of order (29, 29) comprising 15 parallel sections, as can be seen in Table 3.1. The number

Table 4.1: Maximum Passband and Stopband Errors for Reduced Realization

(r_1, r_2)	Passband	Stopband
(10, 12)	0.0955	0.0716
(13, 15)	0.0230	0.0202
(15, 17)	0.0102	0.0087

Table 4.2: Maximum Relative Errors in Group Delays for Reduced Realization

(r_1, r_2)	τ_1	τ_2
(10, 12)	6.30%	5.52%
(13, 15)	1.48%	1.49%
(15, 17)	0.79%	0.93%

of multiplications needed for a direct realization of the FIR filter is $(15+15) \times 15 = 450$ as compared to $(13\times15)+(13+15) = 223$ multiplications required by a direct realization of the reduced IIR filter. If the FIR filter is implemented using 1-D fast convolution with block size 512×512, then the total number of multiplications required to process an array of size 512×512 is 166×512^2 [2]. On the other hand, by writing the transfer function of the (13×15) IIR filter as

$$H_r(z_1, z_2) = \frac{N_r(z_1, z_2)}{D_{r_1}(z_1)D_{r_2}(z_2)}$$

the filter can be implemented using the scheme of Figure 4.4. Through the SVD of the 13×15 coefficient matrix of $N_r(z_1, z_2)$, implementation can be achieved by using at most 13 parallel 1-D sections. Consequently, the application of 1-D fast convolution to the FIR filters in both feedforward and feedback paths leads to an implementation of the IIR filter which requires a total of 156×512^2

multiplications in the worst case. Experience with a variety of practically useful 2-D filters has shown that some of the singular values of the coefficient matrix of $N_r(z_1, z_2)$ are often negligible (less than 10^{-4}) and, therefore, the number of multiplications required is usually less than that of the worst case.

Now if the errors in the amplitude response and the group delays are required to be less than 1.2% and 1.0%, respectively, an IIR filter of order (15, 17) is required according to Tables 4.2 and 4.3. The same amplitude-response accuracy can be achieved by using an FIR filter of order (29, 29) comprising 25 parallel sections as seen in Table 3.1. In this case, the implementation of the FIR filter by 1-D fast convolution requires 266×512^2 multiplications while its IIR counterpart requires only 176×512^2 multiplications.

At the other extreme, if the error in the amplitude response is allowed to be as high as 9%, an IIR filter of order (10, 12) is required. The same accuracy can be achieved by using an FIR filter of order (29, 29) comprising only five parallel sections. In this case, the FIR implementation is more efficient than its IIR counterpart.

In effect, if the amplitude-response approximation error is required to be low, say less than 5%, and a small variation in the group delays of the order of 1 to 2% can be tolerated, the BA method is very likely to yield a more efficient IIR design. Furthermore, a nearly balanced state-space realization is achieved which has low roundoff output noise as well as low sensitivities to coefficient quantization, according to Theorem 4.2.

4.7 Conclusions

In this chapter, the BA method has been applied for the design of 2-D digital filters. In this approach, the design starts with a 2-D causal, linear-phase, FIR filter of the type that can be obtained using the SVD and concludes with a design of reduced order, which is almost always an IIR design. The method

leads to computationally efficient designs and tends to preserve the linear phase response of the original FIR filter irrespective of whether an FIR or an IIR design is obtained. Furthermore, the designs obtained are causal and locally quasi-balanced, and in cases where IIR designs are obtained, stability is guaranteed. In other words, the method is highly suitable for the design of 2-D, causal, linear-phase IIR filters, which are very difficult to design by other methods.

The design approach has two important advantages. First, it leads to more efficient implementations in applications where the amplitude-response approximation error is required to be low. Second, nearly balanced state-space realizations are achieved which have low roundoff output noise as well as low sensitivity to coefficient quantization.

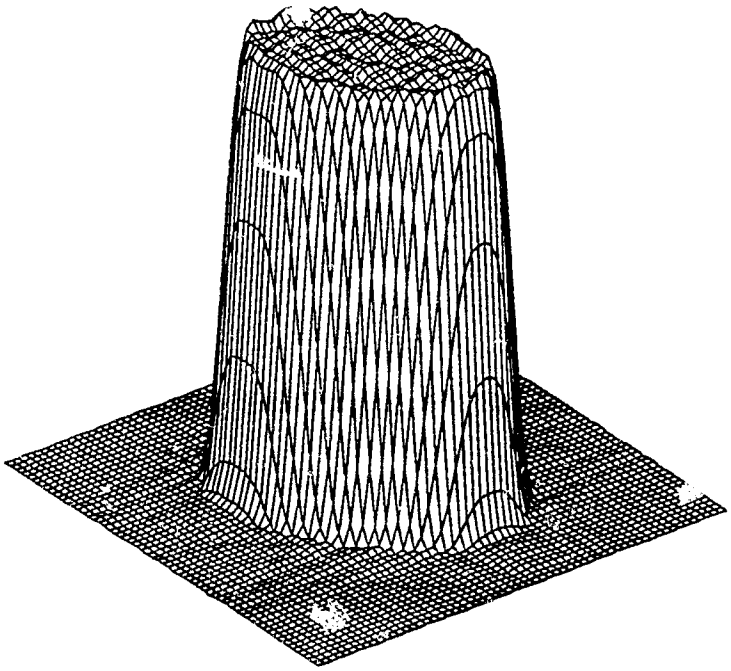

Figure 4.1: Amplitude response of 2-D IIR filter with rotated elliptical passband obtained by using SVD and BA methods ($N_1 = 13$, $N_2 = 15$).

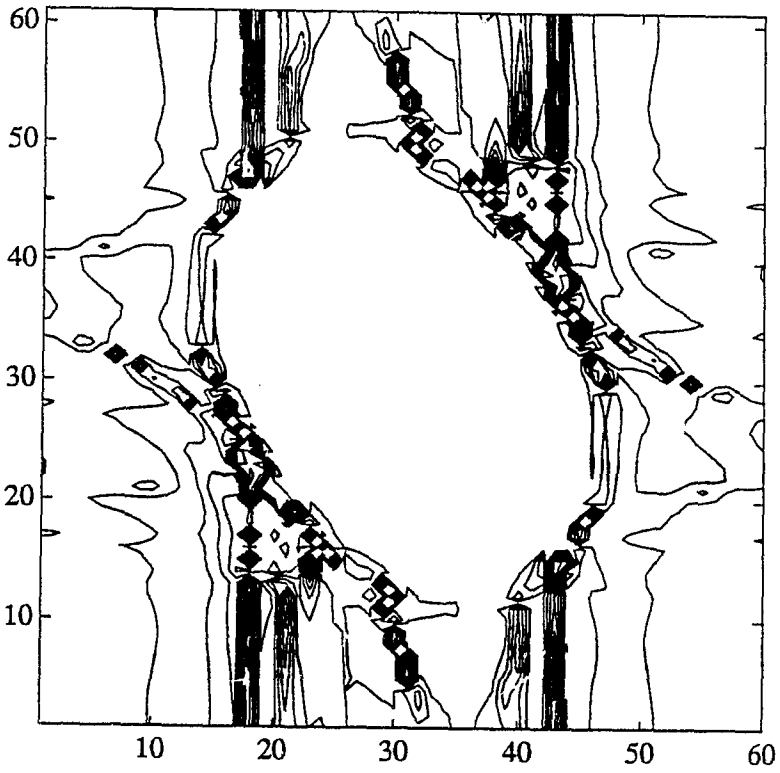

Figure 4.2: Contour plot of group delays of 2-D IIR filter with respect to ω_1.

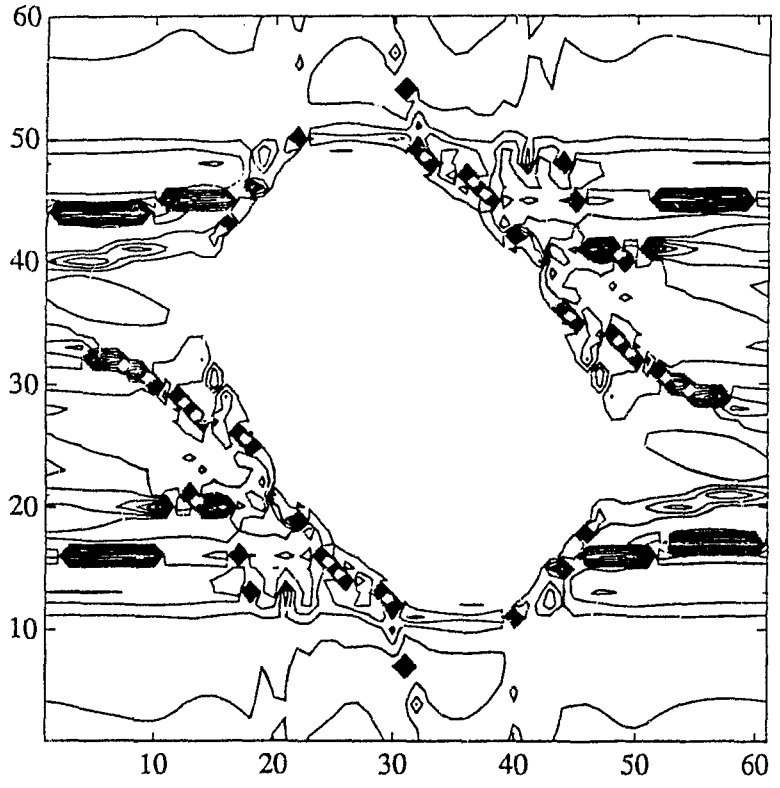

Figure 4.3: Contour plot of group delays of 2-D IIR filter with respect to ω_2.

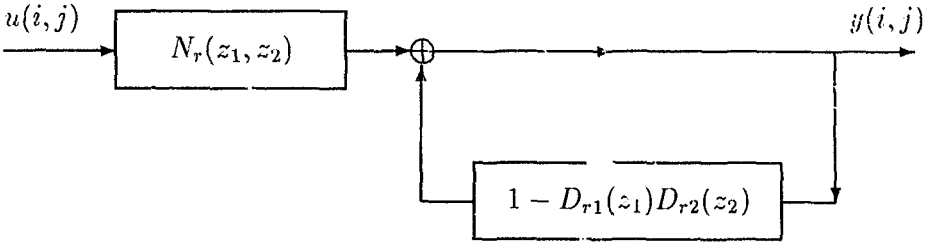

Figure 4.4: A realization of $H_r(z_1,\ z_2)$.

Chapter 5

Evaluation of the Controllability and Observability Gramians of 2-D Digital Filters

5.1 Introduction

As has been demonstrated in Chapter 4, the computation of 1-D and 2-D gramians is an important step in obtaining balanced approximations of 1-D and 2-D systems and also in finding optimal digital-filter structures that minimize the output-noise power due to the roundoff of products [24, 31, 35, 36].

The most efficient method for the evaluation of the gramians for the 1-D case is to solve two relevant Lyapunov equations, and reliable algorithms for this purpose are available in the literature [29, 30]. For the 2-D case, the Lyapunov equations depend on a complex parameter, as is demonstrated in [31, 25], which varies on the unit circle of a complex plane. In other words, should the Lyapunov approach be chosen for the evaluation of the 2-D gramians, one would need to solve a family of 1-D Lyapunov equations as opposed to two constant Lyapunov equations in the 1-D case. A Lyapunov approach for the 2-D case was described in [32] but the transfer function of the digital filters under consideration must have separable denominators. For general, 2-D, causal, stable, recursive digital filters, the most commonly used method is the truncation method described in [31] which

provides numerical approximations of the gramians in terms of truncated double summations. In practice, a large number of terms must be used to guarantee acceptable numerical error. This is particularly the case when the filter under consideration has small stability margin in which case the convergence of the infinite series is rather slow.

In this chapter, a general and efficient method for the evaluation of the gramians for the case of 2-D, causal, stable, recursive digital filters is presented. The method is based on a two-stage extension of the Åström-Jury-Agniel (ÅJA) algorithm which was originally used in [33, 34] for the evaluation of the scalar loss function of a stationary random process with rational spectral density. In Section 5.2 some preliminary material regarding the ÅJA algorithm is reviewed. In Section 5.3 the ÅJA algorithm is first extended to the vector case. It is shown that the modified ÅJA algorithm can be used to solve a 1-D Lyapunov equation in a recursive manner. In Section 5.4, the recursive algorithm obtained is further extended to the case where the vector rational function involved depends on two complex variables. It is shown that the two algorithms obtained can be combined to evaluate the 2-D gramians. In Section 5.5 the proposed method is compared with other known methods for the evaluation of the 2-D gramians with respect to accuracy and computational efficiency and is illustrated by a numerical example.

5.2 Preliminaries

It is known that for a stationary stochastic process with a rational spectral density $\phi(\omega)$, there exists a proper rational function $H(z)$ with poles inside the unit circle such that

$$\phi(\omega) = \mid H(e^{j\omega}) \mid^2$$

To evaluate the performance of a stochastic system and to optimize it with respect to system parameters, it is often required to evaluate the loss function L defined

by

$$L = \frac{1}{2\pi}\int_0^{2\pi} \phi(\omega)\, d\omega = \frac{1}{2\pi j}\oint_{|z|=1} H(z)H(z^{-1})z^{-1}\, dz \tag{5.1}$$

Let $H(z) = n(z)/d(z)$ where

$$\begin{aligned} n(z) &= n_0^{(n)}z^n + n_1^{(n)}z^{n-1} + \cdots + n_{n-1}^{(n)}z + n_n^{(n)} \\ d(z) &= d_0^{(n)}z^n + d_1^{(n)}z^{n-1} + \cdots + d_{n-1}^{(n)}z + d_n^{(n)} \end{aligned} \tag{5.2}$$

are scalar polynomials with real coefficients. A recursive formula that enables one to evaluate the loss function (5.1) in n steps was proposed by Åström [33, 34]. To describe the formula, let $\hat{d}(z)$ be the polynomial defined by

$$\hat{d}(z) = z^n[d(1/z)] \tag{5.3}$$

$$= d_0^{(n)} + d_1^{(n)}z + \cdots + d_n^{(n)}z^n \tag{5.4}$$

Further, let $n_k(z)$ and $d_k(z)$ $(k = n,\ n-1, \cdots,\ 1,\ 0)$ be a set of recursive polynomials given by

$$\begin{aligned} n_n(z) &= n(z) \\ d_n(z) &= d(z) \end{aligned} \tag{5.5}$$

$$\begin{aligned} n_{k-1}(z) &= z^{-1}[n_k(z) - \xi_k\hat{d}_k(z)] \\ d_{k-1}(z) &= z^{-1}[d_k(z) - \eta_k\hat{d}_k(z)] \end{aligned} \tag{5.6}$$

where

$$n_k(z) = n_0^{(k)}z^k + n_1^{(k)}z^{k-1} + \cdots + n_k^{(k)} \tag{5.7}$$

$$d_k(z) = d_0^{(k)}z^k + d_1^{(k)}z^{k-1} + \cdots + d_k^{(k)} \tag{5.8}$$

$$\xi_k = n_k^{(k)}/d_0^{(k)} \tag{5.9}$$

$$\eta_k = d_k^{(k)}/d_0^{(k)} \tag{5.10}$$

The coefficients of the polynomials $n_k^{(k)}$ and $d_k^{(k)}$ are given by the recursive equations

$$n_i^{(k-1)} = n_i^{(k)} - \xi_k d_{k-i}^{(k)} \tag{5.11}$$

$$d_i^{(k-1)} = d_i^{(k)} - \eta_k d_{k-i}^{(k)} \tag{5.12}$$

Åström's algorithm computes the complex integral L_k defined by

$$L_k = \frac{1}{2\pi j} \oint_{|z|=1} \frac{n_k(z) n_k(z^{-1})}{d_k(z) d_k(z^{-1})} z^{-1} \, dz \tag{5.13}$$

by using the following recursive equation

$$L_k = (1 - \eta_k^2) L_{k-1} + \xi_k^2 \quad (k = 1, 2, \cdots n) \tag{5.14}$$

$$L_0 = \xi_0^2 \tag{5.15}$$

Since integral L defined by (5.1) is equal to L_n, the evaluation of (5.1) can be carried out by applying (5.6)-(5.15) n times recursively. It is also shown in [33] that

$$L_k = \frac{1}{d_0^{(k)}} \sum_{i=0}^{k} \frac{\left(n_i^{(i)}\right)^2}{d_0^{(i)}} \tag{5.16}$$

Thus the integral L in (5.1) is given by

$$L = \frac{1}{d_0^{(n)}} \sum_{i=0}^{n} \frac{\left(n_i^{(i)}\right)^2}{d_0^{(i)}} \tag{5.17}$$

The derivation of formula (5.16) is quite lengthy and the interestered reader is referred to references [33, 34]. In the next section, it is shown that the above algorithm can be modified to include the complex matrix case so that the modified algorithm can be used to solve the 1-D Lyapunov equations in (4.4) and (4.5).

5.3 New Recursive Algorithm for the Solution of 1-D Lyapunov Equations

It follows from (4.2)-(4.3) that matrices $\mathbf{K}_1$ and $\mathbf{W}_1$ can be evaluated by solving the 1-D discrete Lyapunov equations in (4.4) and (4.5). Note that equations (4.4) and (4.5) are linear in $\mathbf{K}_1$ and $\mathbf{W}_1$ and a unique positive-definite solution $\mathbf{K}_1$ ($\mathbf{W}_1$) exists if the filter represented by (4.1) is a stable and controllable (observable) system.

Several algorithms are available for the solution of equations (4.4) and (4.5) [29, 30]. In this section, a new recursive algorithm for the solution of these equations based on the ÅJA algorithm is presented. As will be demonstrated, the new algorithm is much more efficient than other known algorithms and leads to an accurate solution of the problem at hand. Furthermore, it can readily be extended to the 2-D case and can be applied for the evaluation of the 2-D gramians given in (4.14) and (4.15).

In what follows, attention is focused on the evaluation of matrix $\mathbf{K}_1$ but the algorithm obtained can also be used for the evaluation of matrix $\mathbf{W}_1$. Let

$$\mathbf{f}_1(z) = (z\mathbf{I} - \mathbf{A})^{-1}\mathbf{b} = \frac{\mathbf{n}(z)}{d(z)}$$

where

$$\mathbf{n}(z) = \text{adj}\ (z\mathbf{I} - \mathbf{A})\mathbf{b} = \mathbf{n}_0^{(n)} z^{n-1} + \cdots + \mathbf{n}_n^{(n)} \tag{5.18}$$

$$d(z) = \det\ (z\mathbf{I} - \mathbf{A}) = d_0^{(n)} z^n + d_1^{(n)} z^{n-1} + \cdots + d_n^{(n)} \tag{5.19}$$

with $\mathbf{n}_i^{(n)} \in R^{n\times 1}$ for $i = 0, 1, \cdots, n$ and $d_0^{(n)} = 1$, then equation (4.8) becomes

$$\begin{aligned}\mathbf{K}_1 &= \frac{1}{2\pi j} \oint_{|z|=1} \left[\frac{\mathbf{n}(z)}{d(z)}\right] \left[\frac{\mathbf{n}(z)}{d(z)}\right]^H z^{-1}\, dz \\ &= \frac{1}{2\pi j} \oint_{|z|=1} \left[\frac{\mathbf{n}(z)}{d(z)}\right] \left[\frac{\mathbf{n}(z^{-1})}{d(z^{-1})}\right]^T z^{-1}\, dz \end{aligned} \tag{5.20}$$

where $\bar{z} = z^{-1}$ for $z \in T = \{z : |z| = 1\}$.

The original ÅJA algorithm provides a recursive formula for the evaluation of the scalar loss function of a stationary stochastic process with a rational spectral density $\phi = |H(e^{j\omega})|^2$. More specifically, it can be used to compute the loss function L defined by (5.1). On the other hand, the new algorithm evaluates the series of integrals

$$\mathbf{Y}_k = \frac{1}{2\pi j} \oint_{|z|=1} \left[\frac{\mathbf{n}_k(z)}{d_k(z)}\right] \left[\frac{\mathbf{n}_k(z)}{d_k(z)}\right]^H z^{-1}\, dz \tag{5.21}$$

for $k = 1, 2, \cdots, n$ recursively. Functions $\mathbf{n}_k(z)$ and $d_k(z)$ for $k = n, n-1, \cdots, 1, 0$ are vector and scalar polynomials, respectively, defined as

$$\mathbf{n}_n(z) = \mathbf{n}(z) \tag{5.22}$$

$$d_n(z) = d(z) \tag{5.23}$$

$$\mathbf{n}_{k-1}(z) = z^{-1}[\mathbf{n}_k(z) - \xi_k \hat{d}_k(z)] \tag{5.24}$$

$$d_{k-1}(z) = z^{-1}[d_k(z) - \eta_k \hat{d}_k(z)] \tag{5.25}$$

where

$$\mathbf{n}_k(z) = \mathbf{n}_0^{(k)} z^k + \mathbf{n}_1^{(k)} z^{k-1} + \cdots + \mathbf{n}_k^{(k)} \tag{5.26}$$

$$d_k(z) = d_0^{(k)} z^k + d_1^{(k)} z^{k-1} + \cdots + d_k^{(k)} \tag{5.27}$$

$$\hat{d}_k(z) = \bar{d}_0^{(k)} + \bar{d}_1^{(k)} z + \cdots + \bar{d}_k^{(k)} z^k \tag{5.28}$$

$$\xi_k = \mathbf{n}_k^{(k)} / d_0^{(k)} \tag{5.29}$$

$$\eta_k = d_k^{(k)} / d_0^{(k)} \tag{5.30}$$

and $\bar{d}$ denotes the complex conjugate of d.

From (5.20) and (5.21), it is noted that $\mathbf{K}_1 = \mathbf{Y}_n$ and, therefore, $\mathbf{K}_1$ can be evaluated in a recursive manner through the following steps:

Algorithm 1 One-Dimensional Modified ÅJA Algorithm

1) Compute $\mathbf{n}_n(z) = \mathbf{n}(z)$ and $d_n(z) = d(z)$ using (5.18) to (5.19).

2) Compute $\mathbf{n}_{k-1}(z)$ and $d_{k-1}(z)$ for $k = n, \cdots, 1$ and ξ_k, η_k for $k = n, \cdots, 0$ using (5.24) to (5.30).

3) Form

$$\mathbf{Y}_0 = \xi_0 \xi_0^H \tag{5.31}$$

4) Compute

$$\mathbf{Y}_k = (1 - |\eta_k|^2)\mathbf{Y}_{k-1} + \xi_k \xi_k^H \tag{5.32}$$

for $k = 1, 2, \cdots, n$.

A closed-form solution of the Lyapunov equation (4.4) can be obtained as

$$\mathbf{K}_1 = \frac{1}{d_0^{(n)}} \sum_{i=0}^{n} [\mathbf{n}_i^{(i)} (\mathbf{n}_i^{(i)})^H] / d_0^{(i)} \tag{5.33}$$

by using (5.26) to (5.32). This is the counterpart of Corollary 2.1 of Chapter 5, Theorem 2.3 in [33] when $k = n$.

It follows from (5 28) and (5.31) that Algorithm 1 is applicable in the case where $\mathbf{A}$ and $\mathbf{b}$ are complex. Moreover, by repeating the arguments presented in [33] it can be shown that Algorithm 1 is valid even for the case where $\mathbf{b}$ is a complex matrix of dimension $n \times m$ with $m > 1$. These properties will prove of significant importance when we attempt to extend Algorithm 1 to the 2-D case in the following section.

5.4 A Recursive Algorithm for Evaluating 2-D Gramians

The important parts of the 2-D gramians $\mathbf{K}_2$ and $\mathbf{W}_2$ defined by (4.14) and (4.15) are their diagonal blocks of dimension $n_1 \times n_1$ and $n_2 \times n_2$, namely matrices $\mathbf{K}_{11}$, $\mathbf{K}_{22}$, $\mathbf{W}_{11}$ and $\mathbf{W}_{22}$ matrices specified below [28, 25, 24]:

$$\mathbf{K}_2 = \begin{bmatrix} \mathbf{K}_{11} & \mathbf{K}_{12} \\ \mathbf{K}_{12}^H & \mathbf{K}_{22} \end{bmatrix}, \quad \mathbf{W}_2 = \begin{bmatrix} \mathbf{W}_{11} & \mathbf{W}_{12} \\ \mathbf{W}_{12}^H & \mathbf{W}_{22} \end{bmatrix}$$

In what follows attention is focused on the evaluation of matrix $\mathbf{K}_{11}$, but with some straightforward modifications, the recursive algorithm obtained can also be used to evaluate matrices $\mathbf{K}_{22}$, $\mathbf{W}_{11}$ and $\mathbf{W}_{22}$.

From Section 5 of [25], matrix $\mathbf{K}_{11}$ can be written as

$$\mathbf{K}_{11} = \frac{1}{2\pi j} \oint_{|z_2|=1} \mathbf{K}_p(z_2) z_2^{-1} \, dz_2 \tag{5.34}$$

where $z_2 \in T_2 = \{z_2 : |z_2| = 1\}$ and $\mathbf{K}_p(z_2)$ is the unique positive-definite Hermitian solution of the parametric Lyapunov equation

$$\mathbf{A}_1(z_2)\mathbf{K}_p(z_2)\mathbf{A}_1^H(z_2) - \mathbf{K}_p(z_2) = -\mathbf{b}_1(z_2)\mathbf{b}_1^H(z_2), \quad z_2 \in T_2 \tag{5.35}$$

with

$$\mathbf{A}_1(z_2) = \mathbf{A}_1 + \mathbf{A}_2(z_2\mathbf{I} - \mathbf{A}_4)^{-1}\mathbf{A}_3 \tag{5.36}$$

$$\mathbf{b}_1(z_2) = \mathbf{b}_1 + \mathbf{A}_2(z_2\mathbf{I} - \mathbf{A}_4)^{-1}\mathbf{b}_2 \tag{5.37}$$

It follows that the evaluation of $\mathbf{K}_{11}$ can be carried out by first solving the parametric Lyapunov equation in (5.35) and then computing the complex integral in (5.34). Since the resulting $\mathbf{K}_p(z_2)$ has the form

$$\mathbf{K}_p(z_2) = \sum_{k=0}^{n} \tilde{\beta}_k(z_2)\tilde{\beta}_k^H(z_2) \tag{5.38}$$

where each $\tilde{\beta}_k(z_2)$ is a stable rational function of dimension $n_1 \times 1$ (i.e. its poles are located inside the unit circle $|z_2| = 1$), the evaluation of the complex integral in (5.34) can readily be carried out by applying Algorithm 1 described in Section 5.3, assuming that $\mathbf{K}_p(z_2)$ is known. We shall now develop an algorithm that can be used for the evaluation of $\mathbf{K}_p(z_2)$.

The Hermitian solution of equation (5.35) can be expressed as the complex integral

$$\begin{aligned}\mathbf{K}_p(z_2) &= \frac{1}{2\pi j}\oint_{|z_1|=1}[z_1\mathbf{I}-\mathbf{A}_1(z_2)]^{-1}\mathbf{b}_1(z_2)\mathbf{b}_1^H(z_2) \\ &\quad \times[z_1\mathbf{I}-\mathbf{A}_1(z_2)]^{-H}z_1^{-1}\,dz_1 \qquad (5.39)\end{aligned}$$

where $[z_1\mathbf{I}-\mathbf{A}_1(z_2)]^{-H}$ denotes the complex-conjugate transpose of $[z_1\mathbf{I}-\mathbf{A}_1(z_2)]^{-1}$. We can write $[z_1\mathbf{I}-\mathbf{A}_1(z_2)]^{-1}\mathbf{b}_1(z_2)$ in the form

$$[z_1\mathbf{I}-\mathbf{A}_1(z_2)]^{-1}\mathbf{b}_1(z_2) = \frac{\mathbf{n}(z_1,\ z_2)}{d(z_1,\ z_2)} \qquad (5.40)$$

where

$$\begin{aligned}\mathbf{n}(z_1,\ z_2) &= \text{adj}\ [z_1\mathbf{I}-\mathbf{A}_1(z_2)]\mathbf{b}_1(z_2) \\ &= \mathbf{n}_1^{(n)}(z_2)z_1^{n-1}+\cdots+\mathbf{n}_{n-1}^{(n)}(z_2)z_1+\mathbf{n}_n^{(n)}(z_2) \qquad (5.41)\end{aligned}$$

$$\begin{aligned}d(z_1,\ z_2) &= \det\ [z_1\mathbf{I}-\mathbf{A}_1(z_2)] \\ &= d_0^{(n)}(z_2)z_1^n+\cdots+d_{n-1}^{(n)}(z_2)z_1+d_n^{(n)}(z_2) \qquad (5.42)\end{aligned}$$

and $d_0^{(n)}(z_2) \equiv 1$. Further, we can define vector and scalar polynomials $\mathbf{n}_k(z_1,\ z_2)$ and $d_k(z_1,\ z_2)$, respectively, for $k = n, n-1, \cdots, 0$ in terms of the recursive relations

$$\mathbf{n}_n(z_1,\ z_2) = \mathbf{n}(z_1,\ z_2) \qquad (5.43)$$

$$d_n(z_1,\ z_2) = d(z_1,\ z_2) \tag{5.44}$$

$$\mathbf{n}_{k-1}(z_1,\ z_2) = z_1^{-1}[\mathbf{n}_k(z_1,\ z_2) - \beta_k(z_2)\hat{d}_k(z_1,\ z_2)] \tag{5.45}$$

$$d_{k-1}(z_1,\ z_2) = z_1^{-1}[d_k(z_1,\ z_2) - \alpha_k(z_2)\hat{d}_k(z_1,\ z_2)] \tag{5.46}$$

where

$$\mathbf{n}_k(z_1,\ z_2) = \mathbf{n}_0^{(k)}(z_2)z_1^k + \cdots + \mathbf{n}_{k-1}^{(k)}(z_2)z_1 + \mathbf{n}_k^{(k)}(z_2) \tag{5.47}$$

$$d_k(z_1,\ z_2) = d_0^{(k)}(z_2)z_1^k + \cdots + d_{k-1}^{(k)}(z_2)z_1 + d_k^{(k)}(z_2) \tag{5.48}$$

$$\alpha_k(z_2) = d_k^{(k)}(z_2)/d_0^{(k)}(z_2) \tag{5.49}$$

$$\beta_k(z_2) = \mathbf{n}_k^{(k)}(z_2)/d_0^{(k)}(z_2) \tag{5.50}$$

and

$$\hat{d}_k(z_1,\ z_2) = z_1^n d_k(z_1^{-1},\ \bar{z}_2) \tag{5.51}$$

$$= z_1^n d_k(z_1^{-1},\ z_2^{-1}) \text{ for } z_2 \in T_2 \tag{5.52}$$

Function $\mathbf{K}_p(z_2)$ can be determined by evaluating the series of integrals

$$\mathbf{J}_k(z_2) = \frac{1}{2\pi j}\oint_{|z_1|=1}\left[\frac{\mathbf{n}_k(z_1,\ z_2)}{d_k(z_1,\ z_2)}\right]\left[\frac{\mathbf{n}_k(z_1,\ z_2)}{d_k(z_1,\ z_2)}\right]^H z_1^{-1}\,dz_1 \tag{5.53}$$

for $k = 0, 1, \cdots, n$ and then noting from (5.39)-(5.44), and (5.53) that $\mathbf{K}_p(z_2) = \mathbf{J}_n(z_2)$. The steps involved are detailed in Algorithm 2 below.

Algorith 2 Two-Dimensional ÅJA Algorithm

1) Compute $\mathbf{n}_n(z_1,\ z_2) = \mathbf{n}(z_1,\ z_2)$ and $d_n(z_1,\ z_2) = d(z_1,\ z_2)$ using (5.41)-(5.42).

2) Compute $\mathbf{n}_{k-1}(z_1,\ z_2)$ and $d_{k-1}(z_1,\ n_2)$ for $k = n, \cdots, 1$ and $\alpha_k(z_2)$, $\beta_k(z_2)$ for $k = n, \cdots, 0$ using (5.43)-(5.52).

3) Form

$$\mathbf{J}_0(z_2) = \beta_0(z_2)\beta_0^H(z_2) \tag{5.54}$$

4) Compute

$$\mathbf{J}_k(z_2) = [1 - |\alpha_k(z_2)|^2]\mathbf{J}_{k-1}(z_2) + \beta_k(z_2)\beta_k^H(z_2) \tag{5.55}$$

for $k = 1, 2, \cdots, n$.

The derivation of formulas (5.54) and (5.55) is based on the theorems given in the following section.

5.5 Theorems

The following theorems are very important for the iterative computation of 2-D gramians.

Theorem 5.1

If $d_0^{(k)}(z_2) > 0$ for $z_2 \in T_2$ where $T_2 = \{z_2 : |z_2| = 1\}$, then polynomial $d_k(z_1, z_2)$ is stable with respect to (*w.r.t.*) z_1 for $z_2 \in T_2$ if and only if $d_{k-1}(z_1, z_2)$ is stable *w.r.t.* z_1 for $z_2 \in T_2$ and $d_0^{(k-1)}(z_2) > 0$ for $z_2 \in T_2$.

Theorem 5.2

If $d_0^{(n)}(z_2) > 0$ for $z_2 \in T_2$, then $d_n(z_1, z_2)$ is stable *w.r.t.* z_1 for $z_2 \in T_2$ if and only if $d_0^{(k)}(z_2) > 0$ for $z_2 \in T_2$ and for $k = 0, 1, \cdots, n-1$.

With the 2-D reciprocal polynomial $\hat{d}_k(z_1, z_2)$ given by (5.51) properly defined, Theorems 5.1 and 5.2 can be proved by using the arguments adopted in the proofs of Theorems 2.1 and 2.2 of [33, Chap. 5, Sec 2]. Now if we regard the complex variable z_2 in $\mathbf{J}_k(z_2)$ as an arbitrarily fixed parameter on the unit circle, then $\mathbf{J}_k(z_2)$ is very much the same as the integral $\mathbf{Y}_k$ defined by (5.32). Consequently, the argument presented in the proof of Theorem 2.3 of [33] in conjunction with Theorems 5.1 and 5.2 leads to formulas in (5.54) and (5.55).

The following analysis is of usefulness in the computation of the integral in (5.34). From (5.54) and (5.55), we have

$$\begin{aligned}
\mathbf{K}_{11} &= \frac{1}{2\pi j}\oint_{|z_2|=1} \mathbf{J}_n(z_2) z_2^{-1}\, dz_2 \\
&= \frac{1}{2\pi j}\oint_{|z_2|=1} [1-|\alpha_n(z_2)|]^2 \mathbf{J}_{n-1}(z_2) z_2^{-1}\, dz_2 \qquad (5.56) \\
&\quad + \frac{1}{2\pi j}\oint_{|z_2|=1} \beta_n(z_2)\beta_n^H(z_2) z_2^{-1}\, dz_2 \\
&= \frac{1}{2\pi j}\oint_{|z_2|=1} [1-|\alpha_{n-1}(z_2)|]^2 [1-|\alpha_n(z_2)|]^2 \mathbf{J}_{n-2}(z_2) z_2^{-1}\, dz_2 \\
&\quad + \frac{1}{2\pi j}\oint_{|z_2|=1} [1-|\alpha_n(z_2)|]^2 \beta_{n-1}(z_2)\beta_{n-1}^H(z_2) z_2^{-1}\, dz_2 \\
&\quad + \frac{1}{2\pi j}\oint_{|z_2|=1} \beta_n(z_2)\beta_n^H(z_2) z_2^{-1}\, dz_2 \\
&= \frac{1}{2\pi j}\sum_{k=0}^{n}\oint_{|z_2|=1} \left[\prod_{l=k+1}^{n+1} (1-|\alpha_l(z_2)|^2)\right] \beta_k(z_2)\beta_k^H(z_2) z_2^{-1}\, dz_2 \qquad (5.57)
\end{aligned}$$

where $\alpha_{n+1}(z_2) = 0$ is assumed. Further notice that as in the proof of Theorem 2.1 of [33], it can be shown that

$$|\alpha_k(z_2)| = \left|\frac{d_k^{(k)}(z_2)}{d_0^{(k)}(z_2)}\right| < 1 \quad \text{for all } z_2 \in T_2 \quad \text{and } k = 0, 1, \cdots, n \qquad (5.58)$$

Consequently, the scalar factor

$$\prod_{l=k+1}^{n+1} [1-|\alpha_l(z_2)|^2] \qquad (5.59)$$

in (5.57) is strictly positive for $z_2 \in T_2$ and $k = 0, 1, \cdots, n$, and, therefore, has the spectral factorization [33]

$$\prod_{l=k+1}^{n+1} [1-|\alpha_l(z_2)|^2] = r_k(z_2) r_k(z_2^{-1}) \quad \text{for } k = 0, 1, \cdots, n \qquad (5.60)$$

where $r_k(z_2)$ is a stable rational function. Since $z_2^{-1} = \bar{z}_2$ for $z_2 \in T_2$, (5.57) and (5.59) imply that

$$\mathbf{K}_{11} = \sum_{k=0}^{n} \mathbf{K}_1^{(k)} \tag{5.61}$$

where

$$\mathbf{K}_1^{(k)} = \frac{1}{2\pi j} \oint_{|z_2|=1} \tilde{\beta}_k(z_2)\tilde{\beta}_k^H(z_2)z_2^{-1}\, dz_2 \tag{5.62}$$

$$\tilde{\beta}_k(z_2) = r_k(z_2)\beta_k(z_2) \tag{5.63}$$

Note that $\tilde{\beta}_k(z_2)$ defined by (5.63) is a stable rational function of dimension $n_1 \times 1$ and, therefore, Algorithm 1 can be applied to the integral in (5.62).

In summary, a two-stage method for the evaluation of the controllability and observability gramians of 2-D digital filters and systems has been developed. In the first stage, Algorithm 2 is applied to obtain the positive Hermitian solution of the parametric Lyapunov equation (5.35), and in the second stage the 1-D spectral factorization technique is used to express the resulting $\mathbf{K}_p(z_2)$ in the form of (5.38) and Algorithm 1 is applied for the evaluation of the integral in (5.34).

5.6 Computational Issues

It is known that the $n_1 \times n_1$ and $n_2 \times n_2$ diagonal blocks in $\mathbf{K}_2$ and $\mathbf{W}_2$ can be expresed as [25]

$$\mathbf{K}_{11} = [\mathbf{I}_{n_1}\ \mathbf{0}] \left(\sum_{i=0}^{\infty}\sum_{j=0}^{\infty} q(i,\ j)q^T(i,\ j) \right))[\mathbf{I}_{n_1}\ \mathbf{0}]^T \tag{5.64}$$

$$\mathbf{K}_{22} = [\mathbf{0}\ \mathbf{I}_{n_2}] \left(\sum_{i=0}^{\infty}\sum_{j=0}^{\infty} q(i,\ j)q^T(i,\ j) \right) [\mathbf{0}\ \mathbf{I}_{n_2}]^T \tag{5.65}$$

$$\mathbf{W}_{11} = [\mathbf{I}_{n_1}\ \mathbf{0}] \left(\sum_{i=0}^{\infty} \sum_{j=0}^{\infty} (\mathbf{A}_{ij}^T \mathbf{c}^T \mathbf{c} \mathbf{A}_{ij}) \right) [\mathbf{I}_{n_1}\ \mathbf{0}]^T \tag{5.66}$$

$$\mathbf{W}_{22} = [\mathbf{0}\ \mathbf{I}_{n_2}] \left(\sum_{i=0}^{\infty} \sum_{j=0}^{\infty} (\mathbf{A}_{ij}^T \mathbf{c}^T \mathbf{c} \mathbf{A}_{ij}) \right) [\mathbf{0}\ \mathbf{I}_{n_2}]^T \tag{5.67}$$

where

$$q(i,\ j) = \mathbf{A}_{i-1,j} \begin{bmatrix} \mathbf{b}_1 \\ \mathbf{0} \end{bmatrix} + \mathbf{A}_{i,j-1} \begin{bmatrix} \mathbf{0} \\ \mathbf{b}_2 \end{bmatrix} \tag{5.68}$$

with $\mathbf{A}_{ij}$ defined by

$$\begin{aligned} \mathbf{A}_{ij} &= \mathbf{A}_{10}\mathbf{A}_{i-1,j} + \mathbf{A}_{01}\mathbf{A}_{i,j-1} \quad \text{for } (i,\ j) > (0,\ 0) \\ \mathbf{A}_{-i,j} &= \mathbf{A}_{i,-j} = 0 \quad \text{for } i \geq 1,\ j \geq 1 \\ \mathbf{A}_{00} &= \mathbf{I}_{n_1+n_2},\ \mathbf{A}_{10} = \begin{bmatrix} \mathbf{A}_1 & \mathbf{A}_2 \\ \mathbf{0} & \mathbf{0} \end{bmatrix},\ \mathbf{A}_{01} = \begin{bmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{A}_3 & \mathbf{A}_4 \end{bmatrix} \end{aligned} \tag{5.69}$$

The most commonly used approach to the numerical evaluation of $\mathbf{K}_{ii}$ and $\mathbf{W}_{ii}$ $(i = 1,\ 2)$ has been the truncation method [25, 24, 62, 63]. In this approach, (5.64)-(5.67) are approximated by truncating the double summation as

$$\mathbf{K}_{11} \approx [\mathbf{I}_n\ \mathbf{0}] \left(\sum_{i=0}^{N_1} \sum_{j=0}^{N_2} q(i,\ j) q^T(i,\ j) \right) [\mathbf{I}_n\ \mathbf{0}]^T$$

where N_1 and N_2 are sufficiently large positive integers to guarantee a small approximation error. A problem with the truncation method is its low computation efficiency, in particular, when the 2-D digital filter under consideration has a small stability margin, since $\mathbf{A}_{ij}$ tends to zero (as $i \to \infty,\ j \to \infty$) rather slowly. In practice, the indices N_1 and N_2 need to be large to obtain a satisfactory approximation of $\mathbf{K}_{11}$. Consequently, a very large amount of computation is required.

Another approach is to express $\mathbf{K}_2$ and $\mathbf{W}_2$ as double integrals over the rectangle $R = \{(\theta_1,\ \theta_2) :\ -\pi \leq \theta_1 \leq \pi, -\pi \leq \theta_2 \leq \pi\}$ i.e.

$$\mathbf{K}_2 = \frac{1}{4\pi^2}\int_0^{2\pi}\int_0^{2\pi} \mathbf{f}_2(e^{j\theta_1},\ e^{j\theta_2})\mathbf{f}_2^T(e^{-j\theta_1},\ e^{-j\theta_2})\ d\theta_1 d\theta_2 \tag{5.70}$$

$$\mathbf{W}_2 = \frac{1}{4\pi^2}\int_0^{2\pi}\int_0^{2\pi} \mathbf{g}_2^T(e^{-j\theta_1},\ e^{-j\theta_2})\mathbf{g}_2(e^{j\theta_1},\ e^{j\theta_2})\ d\theta_1 d\theta_2 \tag{5.71}$$

and then approximate them by the double summations

$$\mathbf{K}_2 \approx \frac{1}{N_1 N_2}\sum_{l=0}^{N_1}\sum_{m=0}^{N_2} \mathbf{f}_2(e^{jl\delta\theta_1},\ e^{jm\delta\theta_2})\mathbf{f}_2^T(e^{-jl\delta\theta_1},\ e^{-jm\delta\ell_2}) \tag{5.72}$$

$$\mathbf{W}_2 \approx \frac{1}{N_1 N_2}\sum_{l=0}^{N_1}\sum_{m=0}^{N_2} \mathbf{g}_2^T(e^{-jl\delta\theta_1},\ e^{-jm\delta\theta_2})\mathbf{g}_2(e^{jl\delta\theta_1},\ e^{jm\delta\theta_2}) \tag{5.73}$$

where $\delta\theta_1 = 2\pi/N_1$, $\delta\theta_2 = 2\pi/N_2$. This method is less sensitive to the stability margin of the filter than the truncation method but it also requires a very large amount of computation since each term at the right-hand side of (5.72) and (5.73) involves the inversion a complex matrix of dimension $(n_1 + n_2) \times (n_1 + n_2)$.

By contrast, the new approach provides an *exact* solution to the problem of evaluating matrices $\mathbf{K}_2$ and $\mathbf{W}_2$ with a much improved computation efficiency as compared to the truncation and numerical integration methods. As will be demonstrated by an illustrative example below, the basic types of operation used in Algorithms 1 and 2 and the 1-D spectral factorization include constant matrix-matrix addition and multiplication, polynomial addition and multiplication, computation of the characteristic polynomial of a square matrix and computation of the roots of an algebraic equation containing one unknown. Furthermore, all these operations can be programmed in terms of numerically reliable subroutines. Multiplication of two 1-D polynomials can be performed by computing the convolution of two finite sequences formed by the coefficients of the polynomials involved, the roots of an algebraic equation can be obtained by computing the eigenvalues of the corresponding companion matrix, and the product of an adjoint matrix and a vector as required in equations (5.18) and (5.41) can be

formed by using the formula

$$\mathbf{c}\ \text{adj}\ (z\mathbf{I}-\mathbf{A})\mathbf{b} = \det\ [z\mathbf{I}-(\mathbf{A}-\mathbf{bc})] - \det\ (z\mathbf{I}-\mathbf{A}) \tag{5.74}$$

which requires the computation of two characteristic polynomials.

5.7 Example

The effectiveness of the new 2-D gramian computation method can be illustrated by the following example. Consider a 2-D, stable, state-space digital filter of order $(n_1,\ n_2) = (2,2)$ characterized by (4.11) with

$$\mathbf{A}_1 = \begin{bmatrix} -0.5583 & 0.5825 \\ -0.0558 & 0.0583 \end{bmatrix},\quad \mathbf{A}_2 = \begin{bmatrix} -0.3744 & 0.7525 \\ -0.0374 & 0.0753 \end{bmatrix}$$

$$\mathbf{A}_3 = \begin{bmatrix} -0.1185 & -0.0356 \\ -0.0047 & -0.0014 \end{bmatrix},\quad \mathbf{A}_4 = \begin{bmatrix} -0.4527 & -0.1665 \\ 0.1037 & -0.0723 \end{bmatrix}$$

$$\mathbf{b}_1 = \begin{bmatrix} 1.0 \\ 1.2 \end{bmatrix},\quad \mathbf{b}_2 = \begin{bmatrix} -1.1 \\ 2.0 \end{bmatrix},\quad \mathbf{c}_1^T = \begin{bmatrix} 0.5 \\ -1.0 \end{bmatrix},\quad \mathbf{c}_2^T = \begin{bmatrix} 0.5 \\ 2.0 \end{bmatrix}$$

Following Algorithm 2, $\mathbf{n}(z_1, z_2)$ and $d(z_1, z_2)$ defined by (5.41) and (5.42), respectively, are first computed. Matrix $\mathbf{A}_1(z_2)$ and vector $\mathbf{b}_1(z_2)$ are given by (5.36) and (5.37). By repeatedly applying (5.74), we find that

$$\begin{aligned} \mathbf{n}(z_1,\ z_2) &= \mathbf{n}_1^{(2)}(z_2)z_1 + \mathbf{n}_2^{(2)}(z_2) \\ d(z_1,\ z_2) &= d_0^{(2)}(z_2)z_1^2 + d_1^{(2)}(z_2)z_1 + d_2^{(2)}(z_2) \end{aligned}$$

with

$$\mathbf{n}_1^{(2)}(z_2) = \begin{bmatrix} 1.0000 & 2.4418 & 0.7999 \\ 1.2000 & 0.8217 & 0.1350 \end{bmatrix} \frac{\mathbf{z}_2}{\Delta(z_2)} \equiv \begin{bmatrix} q_1(z_2)/\Delta(z_2) \\ q_2(z_2)/\Delta(z_2) \end{bmatrix}$$

$$\mathbf{n}_2^{(2)}(z_2) = \begin{bmatrix} 0.6408 & 0.6863 & 0.2451 & 0.0329 & 0.0015 \\ 0.6141 & 0.5998 & 0.2157 & 0.0346 & 0.0020 \end{bmatrix} \frac{\hat{\mathbf{z}}_2}{\Delta^2(z_2)} \equiv \begin{bmatrix} q_3(z_2)/\Delta^2(z_2) \\ q_4(z_2)/\Delta^2(z_2) \end{bmatrix}$$

$$
\begin{aligned}
d_0^{(2)}(z_2) &= 1 \\
d_1^{(2)}(z_2) &= \left[\begin{array}{ccc} 0.5000 & 0.2204 & 0.0332 \end{array}\right] \frac{\mathbf{z}_2}{\Delta(z_2)} \equiv \frac{p(z_2)}{\Delta(z_2)} \\
d_2^{(2)}(z_2) &= 0 \\
\Delta(z_2) &= \left[\begin{array}{ccc} 1.0000 & 0.5250 & 0.0500 \end{array}\right] \mathbf{z}_2 \\
\mathbf{z}_2 &= \left[\begin{array}{ccc} z_2^2 & z_2 & 1 \end{array}\right]^T \\
\hat{\mathbf{z}}_2 &= \left[\begin{array}{ccccc} z_2^4 & z_2^3 & z_2^2 & z_2 & 1 \end{array}\right]^T
\end{aligned}
$$

Using Eqns. (5.43)-(5.52), we compute

$$
\begin{aligned}
\alpha_2(z_2) &= 0 \\
\beta_2(z_2) &= \mathbf{n}_2^{(2)}(z_2) \\
\mathbf{n}_1(z_1, z_2) &= \mathbf{n}_0^{(1)}(z_2) z_1 + \mathbf{n}_1^{(1)}(z_2) \\
d_1(z_1, z_2) &= d_0^{(1)}(z_2) z_1 + d_1^{(1)}(z_2)
\end{aligned}
$$

where

$$
\begin{aligned}
\mathbf{n}_0^{(1)}(z_2) &= 0 \\
\mathbf{n}_1^{(1)}(z_2) &= \mathbf{n}_1^{(2)}(z_2) - \mathbf{n}_2^{(2)}(z_2) d_1^{(2)}(\bar{z}_2) \\
&= \left[\begin{array}{c} |\Delta(z_2)|^2 q_1(z_2) - q_3(z_2) p(\bar{z}_2) \\ |\Delta(z_2)|^2 q_2(z_2) - q_4(z_2) p(\bar{z}_2) \end{array}\right] \frac{1}{\Delta(z_2)|\Delta(z_2)|^2} \\
d_0^{(1)}(z_2) &= 1 \\
d_1^{(1)}(z_2) &= p(z_2)/\Delta(z_2)
\end{aligned}
$$

Hence

$$
\begin{aligned}
\alpha_1(z_2) &= p(z_2)/\Delta(z_2) \\
\beta_1(z_2) &= \mathbf{n}_1^{(1)}(z_2)
\end{aligned}
$$

$$\begin{aligned} \mathbf{n}_0(z_1, z_2) &= \mathbf{n}_0^{(0)}(z_2) = -d_1^{(1)}(\bar{z}_2)\mathbf{n}_1^{(1)}(z_2) \\ d_0(z_1, z_2) &= d_0^{(0)}(z_2) = \frac{|\Delta(z_2)|^2 - |p(z_2)|^2}{|\Delta(z_2)|^2} \\ \alpha_0(z_2) &= 1 \end{aligned}$$

and

$$\beta_0(z_2) = \mathbf{n}_0^{(0)}(z_2)/d_0^{(0)}(z_2)$$

By applying (5.49), (5.50), (5.54) and (5.55), we have

$$\mathbf{J}_2(z_2) = \frac{|\Delta(z_2)|^2}{|\Delta(z_2)|^2 - |p(z_2)|^2}\beta_1(z_2)\beta_1^H(z_2) + \mathbf{n}_2^{(2)}(z_2)[\mathbf{n}_2^{(2)}(z_2)]^H$$

where the second term can be written as

$$\mathbf{n}_2^{(2)}(z_2)[\mathbf{n}_2^{(2)}(z_2)]^H = \left(\begin{bmatrix} q_3(z_2) \\ q_4(z_2) \end{bmatrix} \frac{1}{\Delta^2(z_2)}\right)\left(\begin{bmatrix} q_3(z_2) \\ q_4(z_2) \end{bmatrix} \frac{1}{\Delta^2(z_2)}\right)^H \tag{5.75}$$

On applying Algorithm 1 to (5.75), we obtain

$$\frac{1}{2\pi j}\oint_{|z_2|=1} \mathbf{n}_2^{(2)}(z_2)[\mathbf{n}_2^{(2)}(z_2)]^H z_2^{-1}\, dz_2 = \begin{bmatrix} 0.4109 & 0.3925 \\ 0.3925 & 0.3804 \end{bmatrix} \tag{5.76}$$

Note that on the unit circle T_2, the spectral factorization of $|\Delta(z_2)|^2 - |p(z_2)|^2$ leads to

$$|\Delta(z_2)|^2 - |p(z_2)|^2 = \Delta(z_2)\Delta(z_2^{-1}) - p(z_2)p(z_2^{-1}) = v(z_2)v(z_2^{-1})$$

where

$$v(z_2) = 0.8637z_2^2 + 0.4807z_2 + 0.0387$$

Hence on T_2 the first term on the right-hand side of equation (5.75) becomes

$$\frac{|\Delta(z_2)|^2}{|\Delta(z_2)|^2 - |p(z_2)|^2}\beta_1(z_2)\beta_1^H(z_2) = \left[\frac{\mathbf{q}_{12}(z_2)}{v(z_2)\Delta^2(z_2)}\right]\left[\frac{\mathbf{q}_{12}(z_2)}{v(z_2)\Delta^2(z_2)}\right]^H \tag{5.77}$$

where

$$\mathbf{q}_{12}(z_2) = \begin{bmatrix} 0.0287 & 0.5093 & 2.1844 & 3.7150 & 2.2886 & 0.5463 & 0.0393 \\ 0.0396 & 0.5473 & 1.5470 & 1.4374 & 0.5699 & 0.0978 & 0.0058 \end{bmatrix} \tilde{\mathbf{Z}}_2$$

$$\tilde{\mathbf{Z}}_2 = \begin{bmatrix} z_2^6 & z_2^5 & z_2^4 & z_2^3 & z_2^2 & z_2^1 & 1 \end{bmatrix}^T$$

and the application of Algorithm 1 to (5.77) gives

$$\frac{1}{2\pi j}\oint_{|z_2|=1} \frac{|\Delta(z_2)|^2}{|\Delta(z_2)|^2 - |p(z_2)|^2}\beta_1(z_2)\beta_1^H(z_2)z_2^{-1}\, dz_2 = \begin{bmatrix} 5.5320 & 1.3018 \\ 1.3018 & 1.4894 \end{bmatrix} \quad (5.78)$$

Equations (5.75), (5.77), and (5.78) now imply that

$$\mathbf{K}_{11} = \begin{bmatrix} 5.9427 & 1.6943 \\ 1.6943 & 1.4894 \end{bmatrix}$$

With several straightforward modifications, Algorithms 1 and 2 can be used to obtain $\mathbf{K}_{22}$, $\mathbf{W}_{11}$, and $\mathbf{W}_{22}$ as follows:

$$\mathbf{K}_{22} = \begin{bmatrix} 1.3363 & -2.2425 \\ -2.2425 & 4.0696 \end{bmatrix}$$

$$\mathbf{W}_{11} = \begin{bmatrix} 0.3233 & -0.5691 \\ -0.5691 & 1.0758 \end{bmatrix}$$

$$\mathbf{W}_{22} = \begin{bmatrix} 0.5282 & -1.0647 \\ -1.0647 & 4.1374 \end{bmatrix}$$

In order to compare the proposed method with the truncation and integration methods with respect to computational efficiency and accuracy, the three methods were programmed in MATLAB (version 3.5e) on a Sun workstation using double-precision floating-point arithmetic. The total number of floating-point operations (flops) required to compute $\mathbf{K}_{ii}$ and $\mathbf{W}_{ii}$ $(i = 1, 2)$ by the proposed method was found to be 3.82×10^3. In order to obtain the solution to an accuracy of 4 significant digits, the truncation method required $N_1 = N_2 = 15$ and a total of 7.2×10^4 flops while the integration method required $N_1 = N_2 = 20$ and a total of 2.96×10^5 flops.

5.8 Conclusions

A general and computationally efficient method for the evaluation of the controllability and observability gramians of 2-D digital filters and systems has been proposed. The new method yields a high-accuracy closed-form solution of the problem at hand and, in addition, it requires only a fraction of the computation required by existing methods. In the case of a 2-D digital filter of order (2, 2), the new method required only 5.3% of the computation required by the truncation method and only 1.3% of that required by the numerical integration method. The theoretical foundation of the method is the ÅJA algorithm which was originally proposed for the evaluation of scalar loss functions of stationary random processes with rational spectral density. Through a two-stage extension of the ÅJA algorithm, the evaluation of 2-D gramians can be carried out by recursively computing the positive Hermitian solution of the parametric 1-D Lyapunov equation given in (5.35) and then recursively computing the complex integral in (5.34) by means of the 1-D spectral factorization technique.

Chapter 6

Conclusions and Recommendations for Further Work

6.1 Conclusions

This thesis presented a study on the design of 2-D digital filters. In the first part of the work, new methods for the design of 2-D quadrantally symmetric and general FIR and IIR filters using the SVD and BA have been presented. The advantages of the SVD methods are:

- Two-dimensional designs can be achieved by using well-known 1-D design methods. By using 1-D linear-phase FIR filters, linear-phase causal 2-D FIR filters can be obtained which are suitable for real-time or quasireal-time applications.

- The resulting filters consist of parallel arrangements of cascaded pairs of 1-D filters, hence extensive parallel processing and pipelining can be applied.

A new design method using the BA has been applied to linear-phase 2-D FIR filters of the type that may be obtained by using the SVD method. It has been shown that the BA method leads to a lower-order separable 2-D filter, usually an IIR filter, and the phase response of the resulting filter is approximately linear over the passband region provided that the approximation error in the phase

angle is bounded by the sum of the neglected Hankel singular values of the filter. Consequently, the phase response of the resulting filter is approximately linear over the passband region provided that only small Hankel singular values are neglected. The resulting 2-D filter is nearly balanced which implies that the filter has low roundoff noise as well as low parameter sensitivity. Furthermore, it has been shown that the 2-D filter obtained is more economical and computationally more efficient than the original 2-D FIR filter and, in the case where an IIR filter is obtained, the stability of the filter is guaranteed. In the second part of the work, new efficient and general methods for the evaluation of the 1-D and 2-D gramians for the case of 1-D and 2-D, causal, stable, recursive digital filters have been presented. The algorithms are based on extensions of the ÅJA algorithm which was originally used for the evaluation of the scalar loss function of a stationary random process with rational spectral density. It has been shown that the ÅJA algorithm can be modified to solve a 1-D Lyapunov equation in a recursive manner. Furthermore, the recursive algorithm can be extended to the case where the vector rational function involved depends on two complex variables. It has been shown that the two algorithms obtained can be combined to evaluate the 2-D gramians and the new method yields an accurate closed-form solution. In addition, it has been shown that the new method requires only a fraction of the computation required by existing methods. The proposed algorithms are useful in many applications, e.g. in obtaining optimal digital filter structures that minimize the output-noise power due to the roundoff of products, and in obtaining a balanced approximation of a given discrete-time dynamical system or filter.

6.2 Further Work

In Chapters 2 and 3, the error bound for the 2-D filter is given as a summation of the approximation error introduced by a specific 1-D design technique and the residual error. A method for choosing the approximation error in each 1-D

subfilter in such a way as to achieve the most economical 2-D design should be developed.

Presently, a fixed order is chosen for all 1-D subfilters. However, it may be possible to use a different order for each 1-D subfilter so as to achieve minimum error while using the least amount of computations.

In the present work, for simplicity, the Fourier series and window method were used for the design of the 1-D subfilters. However, other available 1-D FIR design methods may also be used. In particular, if the Remez method is employed, it is expected that reduced approximation error can be achieved.

In Chapter 4, the BA method was developed for the design of 2-D IIR filters with linear phase characteristics. It will be worthwhile to compare the BA approach with other methods for the design of linear phase filters, for example, the equalization method.

Presently, only the design of 2-D FIR filters with quadrantal symmetry has been studied. Preliminary work has shown that the SVD method may also be applied for the design of 3-D FIR filters. The SVD may be obtained by using an iterative least-square method. Further research is needed to prove that the decomposed vectors posess the symmetry property in Section 3.2.2. This property will allow the design of linear-phase 3-D FIR filters.

The SVD design method with compensation has been studied previously [23]. If the number of parallel sections is restricted, the design of 2-D FIR filters using the compensation scheme may under certain circumstances give better results. To investigate this possibility, a comparison of the SVD design method with and without compensation should be undertaken.

Bibliography

[1] T. S. Huang, "Two-dimensional windows," *IEEE Trans. on Audio Electroacoust.*, vol. AU-20, pp. 88-89, Mar. 1972.

[2] D. E. Dudgeon, R. M. Mersereau, *Multidimensional Digital Signal Processing,* Prentice-Hall, Inc., New Jersey, 1984.

[3] A. Antoniou, *Digital Filters: Analysis and Design,* New York: McGraw-Hill, 1979.

[4] J. H. McClellan, "The design of two-dimensional filters by transformations," *Proc. 7th Annual Princeton Conf. Information Sciences and Systems*, pp. 247-251, 1973.

[5] J. H. McClellan and T. W. Parks, "Equiripple approximation of fan filters," *Geophysics*, vol. 7, pp. 573-583, 1972.

[6] W. F. G. Mecklenbrauker and R. M. Mersereau, "McClellan Transformations for two-dimensional digital filtering: II - Implementation," *IEEE Trans. Circuits Syst.*, vol. CAS-23, pp. 414-422, Jul. 1976.

[7] R. M. Mersereau, W. F. G. Mecklenbrauker, and T. F. Jr. Quatieri, "McClellan Transformations for two-dimensional digital filtering: I - Design," *IEEE Trans. Circuits Syst.*, vol. CAS-23, pp. 405-413, Jul. 1976.

[8] P. K. Rajan and M. N. S. Swamy, "Design of circularly symmetric two-dimensional FIR filters employing transformations with variable parame-

ters," *IEEE Trans. Acoust., Speech, Signal Processing,* vol. ASSP-31, pp. 637-642, Jun. 1983.

[9] R. M. Mersereau, "The design of 2-D zero phase FIR filters using transformations," *IEEE Trans. Circuits Syst.,* vol. CAS-27, pp. 142-144, Feb. 1980.

[10] D. T. Nguyen and M. N. S. Swamy, "Approximation design of 2-D digital filters with elliptical magnitude response of arbitrary orientation," *IEEE Trans. Circuits Syst* , vol. CAS-33, pp. 597-603, Jun. 1986.

[11] G. A. Maria and M. M. Fahmy, "An L_p design technique for two-dimensional digital recursive filters," *IEEE Trans. Acoust., Speech, Signal Processing,* vol. ASSP-22, pp. 15-21, Feb. 1974.

[12] C. Charalambous, "Design of two-dimensional circularly symmetric digital filters," *Proc. Inst. Elec. Eng.,* vol. 129, pt. G, no. 2, pp. 47-54, Apr. 1982.

[13] C. Charalambous, "The performance of an algorithm for minimax design of two-dimensional linear phase FIR digital filters," *IEEE Trans. Circuits and Syst.,* vol. CAS-32, No. 10, Oct. 1985.

[14] B. G. Mertzios and A. N. Venetsanopoulos, "Design of two-dimensional half-plane recursive digital filters with octagonal symmetry," *Circuits Syst. Signal Process.,* vol. 4, no. 4, pp. 459-483, 1985.

[15] M. P. Ekstrom, R. E. Twogood, and J. W. Woods, "Two-dimensional recursive filter design – A spectral factorization approach," *IEEE Trans. Acoust., Speech, Signal Processing,* vol. ASSP-28, pp. 16-26, Feb. 1980.

[16] J. W. Woods, J. H. Lee, and I. Paul, "Two-dimensional IIR filter design with magnitude and phase error criteria," *IEEE Trans. Acoust., Speech, Signal Processing,* vol. ASSP-31, pp. 886-894, Aug. 1983.

[17] G. A. Lampropoulos and M. M. Fahmy, "A new technique for the design of two-dimensional FIR and IIR filters," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-33, pp. 268-280, Feb. 1985.

[18] J.-H. Lee and Y.-M. Chen, "A new method for the design of two-dimensional recursive digital filters," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-36, pp. 589-598, Apr. 1988.

[19] G. W. Stewart, *Introduction to Matrix Computations*, Orlando: Academic Press, 1973.

[20] R. E. Twogood and S. K. Mitra, "Computer-aided design of separable two-dimensional digital filters," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-25, pp. 165-169, Apr. 1977.

[21] P. Lancaster and M. Tismenetsky *The Theory of Matrices*, 2nd ed., New York: Academic Press, 1985.

[22] S. Treitel and J. L. Shanks, "The design of multistage separable planar filters," *IEEE Trans. Geosci. Electron.*, vol. GE-9, pp. 10-27, Jan. 1971.

[23] A. Antoniou and W.-S. Lu, "Design of two-dimensional digital filters by using the singular value decomposition," *IEEE Trans. Circuits and Syst.*, vol. CAS-34, pp. 1191-1198, Oct. 1987.

[24] W.-S. Lu and A. Antoniou, "Synthesis of 2-D state-space fixed-point digital filter structures with minimum roundoff noise," *IEEE Trans. Circuits and Syst.*, vol. CAS-33, pp. 965-973, Oct. 1986.

[25] W.-S. Lu, E. B. Lee, and Q.-T. Zhang, "Balanced approximation of two-dimensional and delay-differential systems," *Int. J. Control*, vol. 46, no. 6, pp. 2199-2218, 1987.

[26] B. C. Moore, "Principal component analysis in linear systems: controllability, observability, and model reduction," *IEEE Trans. Automat. Contr.*, vol. AC-26, pp. 17-32, Feb. 1981.

[27] H. Kimurd and Y. Honoki, "Balanced approximation of digital FIR filter with linear phase characteristic," *Proc. IEEE Inter. Symp. on Circuits and Systems*, pp. 283-286, 1985.

[28] W.-S. Lu, H.-P. Wang, and A. Antoniou, "Design of two-dimensional digital filters using the singular value decomposition and balanced approximation method," *IEEE Trans. Signal Processing,* vol. 39, pp. 2253-2262, Oct. 1991.

[29] S. J. Hammarling, "Numerical solution of the stable non-negative definite Lyapunov equation," *IMA J. Numer. Anal.,* vol. 2, pp. 303-323, 1982.

[30] A. Laub, M. T. Heath, C. C. Paige, and R. C. Ward, "Computation of system balancing transformations and other applications of simultaneous diagonalization algorithm," *IEEE Trans. Automat. Contr.*, vol AC-32, pp. 115-122, Feb. 1987.

[31] T. Lin, M. Kawamata, and T. Higuchi, "A unified study on the roundoff noise in 2-D state space digital filters," *IEEE Trans. Circuits and Syst.,* vol. CAS-33, pp. 724-730, Jul. 1986.

[32] M. Kawamata and T. Higuchi, "Synthesis of 2-D separable denominator digital filters with minimum roundoff noise and overflow oscillations," *Proc. IEEE Inter. Symp. on Circuits and Syst.,* pp. 1087-1091, Jun. 1985.

[33] K. J. Åström, *Introduction to Stochastic Control Theory*, (Chapter 5) New York: Academic Press, 1970.

[34] K. J. Åström, E. I. Jury, and R. G. Agniel, "A numerical method for the evaluation of complex integrals," *IEEE Trans. Automat, Contr.,* vol. AC-15, pp. 468-471, Aug. 1970.

[35] C. T. Mullis and R. A. Roberts, "Synthesis of minimum roundoff noise fixed point digital filters," *IEEE Trans. Circuits Syst.*, vol. CAS-23, pp. 531-562, Sept. 1976.

[36] S. Y. Hwang, "Minimum uncorrelated unit noise in state-space digital filtering," *IEEE Trans. Acoust. Speech, Signal Process.*, vol. ASSP-25, pp. 256-262, Aug. 1977.

[37] S. Y. Kung, "A new identification and model reduction algorithm via singular value decompositions," *Proc. 12th Asilomar Conf.*, pp. 704-714, 1978.

[38] W.-S. Lu, E. B. Lee, and Q.-T. Zhang, "Model reduction for 2-D systems," *Proc. IEEE Inter. Symp. on Circuits and Systems*, pp. 79-82, May 1986.

[39] W.-S. Lu, H.-P. Wang, and A. Antoniou, "Design of two-dimensional digital filters by using the singular value decomposition," *IEEE Trans. Circuits and Syst.*, vol. CAS-37, pp. 35-46, Jan. 1990.

[40] B. R. Suresh and B. A. Shenoi, "Exact realization of 2-dimensional digital filters by separable filters," *Electronics Letters*, vol. 12, no.10, pp. 242-244, 1976.

[41] P. Karivaratharajan and M. N. S. Swamy, "Realization of a 2-dimensional FIR digital filter using separable filters," *Electronics Letters*, vol. 14, no.8, pp. 294-251, Apr. 1978.

[42] A. N. Venetsanopoulos and C. L. Nikias, "Realization of two-dimensional digital filters by LU decomposition of their transfer function," *Proc. IEEE Inter. Conf. Acoust., Speech, Signal Processing*, pp. 20.4.1-20.4.4, Mar. 1984.

[43] T. W. Park and J. H. McClellan, "Chebyshev approximation for nonrecursive digital filters with linear phase," *IEEE Trans. Circuits Theory*, vol. CT-19, pp. 189-194, Mar. 1972.

[44] G. H. Golub and C. F. Van Loan, *Matrix Computations,* Baltimore: The Johns Hopkins University Press, 1983.

[45] T. S. Huang, J. W. Burne', and A. G. Deczky, "The importance of phase in image processing filters," *IEEE Trans. Acoust. Speech, Signal Process.,* vol. ASSP-23, pp. 529-542, Dec. 1975.

[46] T. C. Speake and R. M. Mersereau, "A note on the use of windows for two-dimensional FIR filter design," *IEEE Trans. Acoust. Speech, Signal Process.,* vol. ASSP-29, pp. 125-127, Feb. 1981.

[47] A. Kumar, F. W. Fairman, and J. R. Sveinsson, "Separately balanced realization and model reduction of 2-D separable denominator transfer functions from input data," *IEEE Trans. Circuits Syst.,* vol. CAS-34, pp. 233-239, Mar. 1987.

[48] E. I. Jury and K. Premaratne, "Model reduction of two dimensional discrete systems," *IEEE Trans. Circuits Syst.,* vol. CAS-33, pp. 558-562, May 1986.

[49] K. Premaratne and E. I. Jury, " Model reduction of two dimensional discrete systems via balanced realization," *Proc. 21st Asilomar Conf. on Signal, Systems and Computers,* Nov. 1987.

[50] K. Premaratne, " Model reduction of two dimensional discrete time and delay system," *Ph.D Thesis,* University of Miami, Aug. 1988.

[51] K. Glover, R. F. Curtain, and J. R. Partington, "Realization and approximation of linear infinite dimensional systems with error bounds," *Univ. of Cambridge, Dept. Engineering, Report CUED/F - CAMS / TR258,* 1986.

[52] S. Y. Kung, B. C. Levy, M. Morf, and T. Kailath, "New results in 2-D system theory, Part II: 2-D state-space model - Realization and notions of controllability, observability, and minimality," *Proc. IEEE,* vol. 65, pp. 945-961, 1977.

[53] A. Zilouchian and R. L. Carroll, "A coefficient sensitivity bound in 2-D state space digital filtering," *IEEE Trans. Circuits Syst.*, vol. CAS-33, pp. 665-667, Jul. 1986.

[54] A. Laub, "Computation of ' balancing ' transformations," *Proc. 1980 Joint Amer. Contr. Conf.*, FA8-E, 1980.

[55] W.-S. Lu and E. B. Lee, "Stability analysis of two-dimensional systems via a Lyapunov approach," *IEEE Trans. Circuits and Syst.*, vol. CAS-32, pp. 61-68, Jan. 1985.

[56] M. Sendaula, "On the frequency dependent Lyapunov equation," *IEEE Trans. Circuits Syst.*, vol. CAS-33, pp. 428-430, Apr. 1986.

[57] P. Agathoklis, E. I. Jury, and M. Mansour, "On the various forms and methods of solution of the Lyapunov equation for 2-D discrete systems," *IEEE Conf. on Decision and Control*, pp. 1573-1578, Dec. 1985.

[58] P. Agathoklis, E. I. Jury, and M. Mansour, "The discrete-time strictly bounded-real lemma and its computation of positive definite solutions to the 2-D Lyapunov equations," *IEEE Trans. Circuits and Syst.*, vol. CAS-36, pp. 830-837, Jun. 1989.

[59] R. A. Roberts and C. T. Mullis, *Digital Signal Processing*, Reading: Addison-Wesley, 1987.

[60] L. M. Smith and B. W. Momar, "An algorithm for constrained roundoff noise minimization in digital filters with application to two-dimensional filters," *IEEE Trans. Circuits Syst.*, vol. CAS-35, no. 11, pp. 1359-1368, Nov. 1988

[61] T. Kailath, *Linear Systems*, Englewood Cliffs: Prentice Hall, 1980.

[62] T. Hinamoto, T. Hamanaka, and S. Maekawa, "A generalized study on the synthesis of 2-D state-space digital filters with minimum roundoff noise,"

IEEE Trans. Circuits Syst., vol. CAS-35, pp. 1037-1042, Aug. 1988.

[63] T. Hinamoto, T. Hamanaka, and S. Maekawa, "Synthesis of 2-D state-space digital filters with low sensitivity based on the Fornasini-Marchesini model," *IEEE Trans. Acoust., Speech., Signal Processing,* vol. ASSP-38, pp. 1587-1594, Sept. 1990.

[64] W.-S. Lu, H.-P. Wang, and A. Antoniou, "On the Evaluation of the Controllability and Observability Gramians of 2-D Digital Filters," *Proc. IEEE Inter. Symp. on Circuits and Systems*, pp. 602-605, Jun. 1991.

[65] J. H. McClellan and D. S. K. Chen, "A 2-D filter structure derived from the Chebyshev recursion," *IEEE Trans. on Circuits and Syst.,* vol. CAS-24, pp. 372-378, Jul. 1977.

Appendix A

Proofs of Theorems 1.1 and 1.2

A.1 Proof of Theorem 1.1:

Since $\mathbf{A}^T\mathbf{A}$ is positive semidefinite, its eigenvalues are nonnegative. Let them be $\sigma_1^2, \sigma_2^2, \cdots, \sigma_M^2$, where $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_r > 0 = \sigma_{r+1} = \sigma_{r+2} = \cdots = \sigma_M$. Let $\mathbf{v}_1, \mathbf{v}_2, \cdots, \mathbf{v}_M$ be a set of orthonormal eigenvectors for $\sigma_1^2, \sigma_2^2, \cdots, \sigma_M^2$, and let $\mathbf{V}_1 = (\mathbf{v}_1, \mathbf{v}_2, \cdots, \mathbf{v}_r)$ and $\mathbf{V}_2 = (\mathbf{v}_{r+1}, \mathbf{v}_{r+2}, \cdots, \mathbf{v}_M)$. Then if $\mathbf{\Sigma} = \text{diag}\,(\sigma_1, \sigma_2, \cdots, \sigma_r)$, we have $\mathbf{V}_1^T\mathbf{A}^T\mathbf{A}\mathbf{V}_1 = \mathbf{\Sigma}^2$ and consequently

$$\mathbf{\Sigma}^{-1}\mathbf{V}_1^T\mathbf{A}^T\mathbf{A}\mathbf{V}_1\mathbf{\Sigma}^{-1} = \mathbf{I} \tag{A.1}$$

Also $\mathbf{V}_2^T\mathbf{A}^T\mathbf{A}\mathbf{V}_2 = 0$, whence

$$\mathbf{A}\mathbf{V}_2 = 0$$

Now let

$$\mathbf{U}_1 = \mathbf{A}\mathbf{V}_1\mathbf{\Sigma}^{-1}$$

Then from (A.1) $\mathbf{U}_1^T\mathbf{U}_1 = \mathbf{I}$; that is the columns of $\mathbf{U}_1$ are orthonormal. Let $\mathbf{U}_2$ be chosen so that $\mathbf{U} = (\mathbf{U}_1,\ \mathbf{U}_2)$ is orthogonal. Then

$$\begin{aligned}
\mathbf{U}^T\mathbf{A}\mathbf{V} &= \begin{pmatrix} (\mathbf{U}_1^T)\mathbf{A}\mathbf{V}_1 & \mathbf{U}_1^T(\mathbf{A}\mathbf{V}_2) \\ \mathbf{U}_2^T(\mathbf{A}\mathbf{V}_1) & \mathbf{U}_2^T(\mathbf{A}\mathbf{V}_2) \end{pmatrix} \\
&= \begin{pmatrix} (\mathbf{\Sigma}^{-1}\mathbf{V}_1^T\mathbf{A}^T)\mathbf{A}\mathbf{V}_1 & \mathbf{U}_1^T(0) \\ \mathbf{U}_2^T(\mathbf{U}_1\mathbf{\Sigma}) & \mathbf{U}_2^T(0) \end{pmatrix} \\
&= \begin{pmatrix} \mathbf{\Sigma} & 0 \\ 0 & 0 \end{pmatrix} \quad \square
\end{aligned}$$

Proof of Theorem 1.2:

Let $\mathbf{B}$ be a matrix of rank r such that $\| \mathbf{A} - \mathbf{B} \|_F$ is minimal. Let the SVD of $\mathbf{B}$ be

$$\mathbf{C} = \mathbf{Q}^T\mathbf{B}\mathbf{P} = \begin{pmatrix} \mathbf{C}_{11} & 0 \\ 0 & 0 \end{pmatrix}$$

where $\mathbf{C}_{11} = \text{diag}\ (\gamma_1, \gamma_2, \cdots, \gamma_r)$. Let

$$\mathbf{D} = \mathbf{Q}^T\mathbf{A}\mathbf{P} = \begin{pmatrix} \mathbf{D}_{11} & \mathbf{D}_{12} \\ \mathbf{D}_{21} & \mathbf{D}_{22} \end{pmatrix}$$

be partitioned conformally with $\mathbf{C}$.

We claim that $\mathbf{D}_{11} = \mathbf{C}_{11}$, $\mathbf{D}_{12} = 0$, and $\mathbf{D}_{21} = 0$. Suppose, for example, that $\mathbf{D}_{12} \neq 0$. Then the matrix

$$\mathbf{C}' = \begin{pmatrix} \mathbf{C}_{11} & \mathbf{D}_{12} \\ 0 & 0 \end{pmatrix}$$

has rank r and $\| \mathbf{D} - \mathbf{C}' \|_F < \| \mathbf{D} - \mathbf{C} \|_F$. However, if we set $\mathbf{B}' = \mathbf{Q}\mathbf{C}'\mathbf{P}^T$, then $\mathbf{B}'$ is also of rank r and

$$\| \mathbf{A} - \mathbf{B}' \|_F = \| \mathbf{D} - \mathbf{C}' \|_F < \| \mathbf{D} - \mathbf{C} \|_F = \| \mathbf{A} - \mathbf{B} \|_F$$

contradicting the minimality of $\mathbf{B}$. Similar arguments show that $\mathbf{D}_{21} = 0$ and $\mathbf{D}_{11} = \mathbf{C}_{11}$.

It follows that $\mathbf{D}$ has the form

$$\mathbf{D} = \begin{pmatrix} \mathbf{C}_{11} & 0 \\ 0 & \mathbf{D}_{22} \end{pmatrix}$$

and $\| \mathbf{A} - \mathbf{B} \|_F = \| \mathbf{D} - \mathbf{C} \|_F = \| \mathbf{D}_{22} \|_F$. Since $\mathbf{C}_{11}$ is diagonal, it consists of singular values of $\mathbf{A}$, and $\| \mathbf{D}_{22} \|_F^2$ will be the sum of squares of those left over. Obviously, this is a minimum when

$$\| \mathbf{D}_{22} \|_F^2 = \sigma_{r+1}^2 + \sigma_{r+2}^2 + \cdots + \sigma_M^2 = \| \mathbf{A} - \mathbf{A}' \|_F^2 \qquad \square$$

Appendix B

SVD of a Quadrantally Symmetric Matrix

Let $\mathbf{C} = \{c_{i,j},\ 0 \leq i,\ j \leq N-1,\ N = \text{odd}\}$ be an $N \times N$ quadrantally-symmetric matrix, i.e.

$$c_{i,j} = c_{i,N-1-j} = c_{N-1-i,j} = c_{N-1-i,N-1-j} \tag{B.1}$$

If a matrix $\hat{\mathbf{I}}$ is defined by

$$\hat{\mathbf{I}} = \begin{bmatrix} 0 & \dots & 0 & 1 & 0 \\ 0 & \dots & 1 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & \dots & 0 & 0 & 0 \end{bmatrix} \tag{B.2}$$

where the size of $\hat{\mathbf{I}}$ is $l \times k$, and $k = (N+1)/2$, $l = (N-1)/2$, then matrix $\mathbf{C}$ can be decomposed as

$$\mathbf{C} = \begin{bmatrix} \mathbf{I}_k & \mathbf{0} \\ \hat{\mathbf{I}} & \mathbf{I}_l \end{bmatrix} \begin{bmatrix} \mathbf{C}_1 & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} \begin{bmatrix} \mathbf{I}_k & \hat{\mathbf{I}}^T \\ \mathbf{0} & \mathbf{I}_l \end{bmatrix} \tag{B.3}$$

where $\mathbf{C}_1$ is the $k \times k$ principal minor of $\mathbf{C}$. Assume that the SVD of matrix $\mathbf{C}_1$ is given by

$$\mathbf{C}_1 = \mathbf{U}_1 \mathbf{S}_1 \mathbf{V}_1^T \tag{B.4}$$

From (B.3) and (B.4), matrix $\mathbf{C}$ can be expressed as

$$\begin{aligned}
\mathbf{C} &= \begin{bmatrix} \mathbf{U}_1 & \mathbf{0} \\ \hat{\mathbf{I}}\mathbf{U}_1 & \mathbf{0} \end{bmatrix} \begin{bmatrix} \mathbf{S}_1 & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} \begin{bmatrix} \mathbf{V}_1 & \mathbf{0} \\ \hat{\mathbf{I}}\mathbf{V}_1 & \mathbf{0} \end{bmatrix}^T \\
&= \left(\frac{1}{\sqrt{2}} \begin{bmatrix} \mathbf{U}_1 & \mathbf{0} \\ \hat{\mathbf{I}}\mathbf{U}_1 & \mathbf{0} \end{bmatrix}\right) \begin{bmatrix} 2\mathbf{S}_1 & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} \left(\frac{1}{\sqrt{2}} \begin{bmatrix} \mathbf{V}_1 & \mathbf{0} \\ \hat{\mathbf{I}}\mathbf{V}_1 & \mathbf{0} \end{bmatrix}^T\right) \\
&= \begin{bmatrix} \bar{\mathbf{U}}_1 & \mathbf{0} \end{bmatrix} \begin{bmatrix} 2\mathbf{S}_1 & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} \begin{bmatrix} \bar{\mathbf{V}}_1 & \mathbf{0} \end{bmatrix} \qquad (B.5)
\end{aligned}$$

where

$$\bar{\mathbf{U}}_1 = \frac{1}{\sqrt{2}} \begin{bmatrix} \mathbf{U}_1 \\ \hat{\mathbf{I}}\mathbf{U}_1 \end{bmatrix} \quad \text{and} \quad \bar{\mathbf{V}}_1 = \frac{1}{\sqrt{2}} \begin{bmatrix} \mathbf{V}_1 \\ \hat{\mathbf{I}}\mathbf{V}_1 \end{bmatrix} \qquad (B.6)$$

If $\bar{\mathbf{U}}_2$ and $\bar{\mathbf{V}}_2$ are the orthonormal complements of $\bar{\mathbf{U}}_1$ and $\bar{\mathbf{V}}_1$, respectively, then the SVD of matrix $\mathbf{C}$ can be obtained from (B.5) as

$$\begin{aligned}
\mathbf{C} &= \mathbf{U}\mathbf{S}\mathbf{V}^T \\
&= \sum_{i=1}^{r_c} \sigma_i \mathbf{u}_i \mathbf{v}_i^T \\
&= \sum_{i=1}^{\bar{r}_c} \tilde{\mathbf{u}}_i \tilde{\mathbf{v}}_i^T \qquad (B.7)
\end{aligned}$$

where $\mathbf{U} = [\bar{\mathbf{U}}_1\ \bar{\mathbf{U}}_2]$ and $\mathbf{V} = [\bar{\mathbf{V}}_1\ \bar{\mathbf{V}}_2]$ are orthogonal matrices, $\mathbf{u}_i$ and $\mathbf{v}_i$ represent the ith column of $\mathbf{U}$ and $\mathbf{V}$, respectively, $\tilde{\mathbf{u}}_i = \sigma_i^{\frac{1}{2}} \mathbf{u}_i$, $\tilde{\mathbf{v}}_i = \sigma_i^{\frac{1}{2}} \mathbf{v}_i$ and

$$\mathbf{S} = \begin{bmatrix} 2\mathbf{S}_1 & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} = \begin{bmatrix} \sigma_1 & 0 & \dots & & 0 \\ 0 & \sigma_2 & 0 & \dots & 0 \\ \vdots & \dots & \sigma_{r_c} & 0 & \dots & \vdots \\ & \dots & \dots & 0 & \dots & \\ & \dots & & \vdots & & 0 \\ 0 & \dots & & & & 0 \end{bmatrix}$$

From (B.2) and (B.6), we now observe that the first k columns of $\mathbf{U}$ and $\mathbf{V}$ are all mirror-image symmetric and, consequently, the vectors $\tilde{\mathbf{u}}_i$ and $\tilde{\mathbf{v}}_i$ in (B.7) are all mirror-image symmetric.

Appendix C

LUD of a Quadrantally Symmetric Matrix

If the LUD of matrix $\mathbf{C}_1$ is given by

$$\mathbf{C}_1 = \mathbf{L}_1\mathbf{U}_1 \tag{C.1}$$

where $\mathbf{L}_1$ and $\mathbf{U}_1$ are lower- and upper-triangular matrices of size N_1, then (B.1) implies that

$$\mathbf{C} = \begin{bmatrix} \mathbf{I}_k & \mathbf{0} \\ \hat{\mathbf{I}} & \mathbf{I}_l \end{bmatrix} \begin{bmatrix} \mathbf{L}_1 & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} \begin{bmatrix} \mathbf{U}_1 & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} \begin{bmatrix} \mathbf{I}_k & \hat{\mathbf{I}}^T \\ \mathbf{0} & \mathbf{I}_l \end{bmatrix}$$

$$= \begin{bmatrix} \mathbf{L}_1 & \mathbf{0} \\ \hat{\mathbf{I}}\mathbf{L}_1 & \mathbf{0} \end{bmatrix} \begin{bmatrix} \mathbf{U}_1 & \mathbf{U}_1\hat{\mathbf{I}}^T \\ \mathbf{0} & \mathbf{0} \end{bmatrix}$$

and, therefore, the LUD of $\mathbf{C}$ is given by

$$\mathbf{C} = \mathbf{L}_c\mathbf{U}_c$$

where $\mathbf{L}_c$ and $\mathbf{U}_c$ are of the form given in Section 2.4.3.

Appendix D

Laub's Algorithm

Step 1 Obtain the Cholesky factorization of $\mathbf{K}$ as

$$\mathbf{K} = \mathbf{L}\mathbf{L}^T$$

where $\mathbf{L}$ is lower triangular.

Step 2 Form $\mathbf{L}^T\mathbf{W}\mathbf{L}$

Step 3 Solve the symmetric eigenvalue/eigenvector problem

$$\mathbf{U}^T(\mathbf{L}^T\mathbf{W}\mathbf{L})\mathbf{U} = \Lambda^2$$

Step 4 Obtain matrix $\mathbf{T}$ as

$$\mathbf{T} = \mathbf{L}\mathbf{U}\Lambda^{-1/2}$$