Signals and Communication Technology

Bogdan Dumitrescu

Positive Trigonometric Polynomials and Signal Processing Applications

Second Edition

Springer

Signals and Communication Technology

More information about this series at http://www.springer.com/series/4748

Bogdan Dumitrescu

Positive Trigonometric Polynomials and Signal Processing Applications

Second Edition

Springer

Bogdan Dumitrescu
Department of Automatic Control
and Computers
University Politehnica of Bucharest
Bucharest
Romania

ISSN 1860-4862 ISSN 1860-4870 (electronic)
Signals and Communication Technology
ISBN 978-3-319-53687-3 ISBN 978-3-319-53688-0 (eBook)
DOI 10.1007/978-3-319-53688-0

Library of Congress Control Number: 2017932002

The use of general descriptive names, registered names, trademarks, service marks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.

The publisher, the authors and the editors are safe to assume that the advice and information in this book are believed to be true and accurate at the date of publication. Neither the publisher nor the authors or the editors give a warranty, express or implied, with respect to the material contained herein or for any errors or omissions that may have been made. The publisher remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Printed on acid-free paper

This Springer imprint is published by Springer Nature
The registered company is Springer International Publishing AG
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

To Adina

Preface

A few words on the second edition. By a nice coincidence, Springer's proposal to revise the book came when I was giving a serious thought to the idea. Ten years have passed and my research shifted mostly to other topics, which discouraged my attempt, but there always seemed to be a small community interested in the book, which gave me hope that the work is not useless. The proposal tipped the balance.

Besides correcting some errors and typos, the new version has a few additions and modifications. Chapter 9, dedicated to optimization problems using the atomic norm and to the related super-resolution problem, is completely new. The Bounded Real Lemma (BRL) for trigonometric polynomial is central to the solution; it was a great reward to see that this BRL, which is the contribution that I consider the most personal and of which I was very proud at the time, has been applied in all its forms in a topic that I never foresaw. To help reading this chapter, all theory regarding the BRL is now gathered in Chap. 4. Another new topic, mentioned mostly in passing, is that of hybrid polynomials, having both real and trigonometric variables. The convex optimization software has greatly evolved, especially toward user convenience; some of the programs shown in the book are now written for CVX, which attracted immediate popularity due to its simple and versatile language; other programs use Pos3Poly, which is a package built on top of CVX, especially for optimization with positive polynomials.

On the contents of the book. Although trigonometric polynomials appear naturally in discrete-time signal processing and their positivity characterizes many design problems, it was only in the late 1990s that an exact and computationally useful parameterization of nonnegative trigonometric polynomials was found. The idea of parameterizing the coefficients of the polynomial as a linear function of the elements of a positive semidefinite matrix was already present (somewhat in disguise) in the previous literature; however, its implementation needed the emergence of semidefinite programming (SDP) methods in the early 1990s and, shortly after, of freely available SDP libraries. The following result is the foundation of this book. Any trigonometric polynomial

$$R(z) = \sum_{k=-n}^{n} r_k z^{-k}, \quad r_{-k} = r_k^*, \tag{1}$$

that is nonnegative on the unit circle (for $|z| = 1$), can be parameterized with a positive semidefinite matrix $\boldsymbol{Q}$ by

$$r_k = \mathrm{tr}[\boldsymbol{\Theta}_k \boldsymbol{Q}], \quad k = -n : n, \tag{2}$$

where $\boldsymbol{\Theta}_k$ is an elementary Toeplitz matrix, with ones on diagonal k and zeros elsewhere, and tr is the trace operator. The matrix $\boldsymbol{Q}$ is named *Gram matrix*. The parameterization (2) allows the description of a nonnegative trigonometric polynomial through a linear matrix inequality (LMI). Hence, SDP is applicable.

The book has two parts. In a simplistic classification, the first four chapters contain the theory and the other five chapters deal with applications. Here is a description of their contents that could help orient the lecture. Although the book treats also (inevitably) polynomials of real variable, we discuss here only the results pertaining to trigonometric polynomials, which have the lion's share.

Chapter 1 is written only in terms of polynomials. It starts with the spectral factorization of polynomials (1) that are nonnegative on the unit circle and which can be written as

$$R(z) = H(z)H^*(z^{-1}), \tag{3}$$

where $H(z)$ is causal and the asterisk denotes complex conjugated coefficients. It also describes polynomials (1) that are nonnegative on an interval, as a simple function of two nonnegative polynomials.

Chapter 2 is built around the Gram matrix parameterization (2) and contains examples of use and several side results linking it to the Kalman–Yakubovich–Popov lemma and spectral factorization. More importantly, it gives alternative parameterizations that are more efficient, for example, the *Gram-pair* parameterization, in which the matrix $\boldsymbol{Q}$ from (2) is replaced by two smaller positive definite matrices.

In Chap. 3, the presentation goes to multivariate polynomials. The most prominent trigonometric polynomial becomes now the *sum-of-squares*

$$R(\boldsymbol{z}) = \sum_{\ell=1}^{\nu} H_\ell(\boldsymbol{z})H_\ell^*(\boldsymbol{z}^{-1}). \tag{4}$$

(We use bold letters, like $\boldsymbol{z} = (z_1, \ldots, z_d)$, to denote multidimensional entities.) The polynomials $H_\ell(\boldsymbol{z})$ have support on the positive orthant, while the support of $R(\boldsymbol{z})$ is symmetric with respect to the origin. It turns out that all trigonometric polynomials that are strictly positive on the unit d-circle (where $|z_1| = \cdots = |z_d| = 1$) are also sum-of-squares; note that sum-of-squares are by construction nonnegative on the unit d-circle. However, the degrees of the polynomials $H_\ell(\boldsymbol{z})$ from (4) can be arbitrarily high. A parameterization like (2) holds, this time for sum-of-squares polynomials. In a practical implementation, an optimization problem with nonnegative polynomials can be solved only in a relaxed way, with sum-of-squares whose factors $H_\ell(z)$ have the degrees bounded to a convenient value. Typically, a higher relaxation degree leads to a better approximation of the original problem, but with a higher complexity, due to the higher size of the Gram matrix $\boldsymbol{Q}$. Chapter 3 contains also the multivariate version of the Gram-pair parameterization and the means for reducing the size of the Gram matrix for sparse polynomials. The chapter ends with a short presentation of polynomials with matrix coefficients, for which, mutatis mutandis, all previous results hold true.

Chapter 4, dealing also with multivariate polynomials, is dedicated to the most general results, which are of three types.

Polynomials positive on domains. Let

$$\mathcal{D} = \left\{ \boldsymbol{\omega} \in [-\pi, \pi]^d \mid D_\ell(\boldsymbol{\omega}) \geq 0,\ \ell = 1 : L \right\} \tag{5}$$

be a frequency domain defined by the positivity of L given trigonometric polynomials $D_\ell(\mathbf{z})$. Then, any trigonometric polynomial $R(z)$ that is positive on $\mathcal{D}$ can be expressed as

$$R(z) = S_0(z) + \sum_{\ell=1}^{L} D_\ell(z) \cdot S_\ell(z), \tag{6}$$

where $S_\ell(z)$, $\ell = 0 : L$, are sum-of-squares. Using a Gram matrix (or a pair of matrices) to parameterize the sum-of-squares, we associate an LMI with $R(z)$.

Bounded Real Lemma. Let $H(z)$ be a polynomial with positive orthant support. Then, the inequality $|H(z)| < \gamma$, with $\gamma \in \mathbb{R}$, can be written in the form of an LMI; see Theorems 4.26 (basic general result), 4.32 (extension to matrix coefficients), and 4.35 (Gram-pair version). This LMI makes possible the formulation of some optimization problems in terms of $H(z)$; the lack of spectral factorization (3) for multivariate polynomials can be thus circumvented in some cases.

Positivstellensatz. We add equalities to the set (5), obtaining

$$\mathcal{D}_E = \left\{ \boldsymbol{\omega} \in [-\pi, \pi]^d \,\middle|\, \begin{array}{l} E_k(\boldsymbol{\omega}) = 0,\ k = 1 : K \\ D_\ell(\boldsymbol{\omega}) \geq 0,\ \ell = 1 : L \end{array} \right\}. \tag{7}$$

Sum-of-squares polynomials can be used to determine whether the set (7) is empty. This happens if and only if there exist polynomials $U_k(\mathbf{z})$ and sum-of-squares polynomials $S_\ell(\mathbf{z})$ such that

$$1 + \sum_{k=1}^{K} E_k(z) U_k(z) + S_0(z) + \sum_{\ell=1}^{L} D_\ell(z) S_\ell(z) = 0. \tag{8}$$

In all the results from Chap. 4 listed above, the degrees of the variable polynomials can be high. Hence, only relaxed versions (i.e., sufficient conditions) can be actually implemented.

In Chaps. 5–9, each basic theoretical result is applied at least once. With univariate polynomials, the typical optimization problems are obtained by replacing the unknown FIR filter $H(z)$, in which the problem is not convex, with its squared magnitude (3), in which the problem is convex (and SDP). After solving the SDP problem, the desired filter is obtained by spectral factorization, with algorithms discussed in Appendix B. The optimization problems are signal processing classics, ranging from the design of FIR and IIR filters to the design of filterbanks and wavelets.

With multivariate polynomials, the applications are the design of 2-D FIR and IIR filters, H_∞ deconvolution, and stability tests, including robust stability. One interesting conclusion is that the relaxations of minimal degree, obtained, e.g., by taking in (4) the degrees of the factors $H_\ell(z)$ equal to the degree of $R(z)$, give practically optimal solutions in almost all problems. So, the limitations of relaxations are mostly theoretical. This allows solving optimally some problems for which no other known algorithm could guarantee practical optimality.

The BRL is used in filter design, deconvolution, and especially in all optimization problems involving the atomic norm and deconvolution, such as line spectrum and direction of arrival estimation.

Each chapter ends with bibliographical notes and a number of problems, whose difficulty ranges from very simple to medium. There are no solutions given in the book, but some hints are provided for many of the "not-so-trivial" problems. The programs for solving the numerical examples are available at http://www.schur.pub.ro/postrigpol_book.htm; in case of trouble, e-mail to bogdan.dumitrescu@upb.ro.

Acknowledgements. Ten years is actually quite a long time, especially when you are ten years older. A lot of things happened since the first edition. My dual affiliation to Tampere University of Technology and University Politehnica of Bucharest, miraculously lasting for so many professionally happy years, came to its natural end. I work now only in Romania. My genial friend (now I dare to say it) Corneliu Popeea passed away; independently of other researchers, he made in August 1999 the breakthrough that led, on a tortuous primal-dual-primal way, to the parameterization of positive trigonometric polynomials. Good things happened also. I still love my family. I learned to play a bit the trombone, despite being somewhat tone deaf. I ran a marathon (in 3:47) and many half marathons. I still play football.

The first edition of this book was based mostly on work done in Tampere. I am and always be grateful to Ioan Tăbuş, long time friend and collaborator, and Jaakko Astola, for the opportunity to come and stay there. Besides them, I thank to my other coauthors (in alphabetical order) on topics related to this book: Ilker Bayram, Robert Bregović, B.C. Chang, Boris Jora, Riitta Niemistö, Tae Roh, Tapio Saramäki, Ivan Selesnick, Petre Stoica, and Lieven Vandenberghe. A special

mention goes to my graduated PhD student Bogdan C. Şicleru, the main author of Pos3Poly, whose unfortunate illness forced him stop research activities. Some other people gave me advice or found errors in the first edition: Michael Dritschel, Nicholas Foti, Didier Henrion, Raf Mertens, Victoria Powers, Bruce Reznick, Claus Scheiderer, Carsten Scherer, and Markus Schweighofer; I am thankful especially to the mathematicians, who know much more than me on positive polynomials. The most important non-scientific support for this book came from my family: Adina, Andrei, and Sebastian; they are my unsigned perpetual coauthors.

Bucharest, Romania
January 2017

Bogdan Dumitrescu

Contents

Chapter 1
Positive Polynomials

Abstract This short chapter presents the characterizations of nonnegative univariate polynomials, with an emphasis on trigonometric polynomials. The basic result (the well-known Riesz-Fejér theorem) is the existence of a spectral factorization for a globally nonnegative trigonometric polynomial. Polynomials that are nonnegative only on a specified interval can be parameterized as a function of two globally nonnegative polynomials. These first characterizations are mostly formulated in "polynomial language"; they will be helpful later, when they serve as a basis for translation into a linear matrix inequality (LMI) form that opens the way for semidefinite programming (SDP) optimization. For completeness, we have added an old characterization of positivity in terms of positive semidefinite Toeplitz matrices.

1.1 Types of Polynomials

The most important mathematical object in this book is the (Hermitian) trigonometric polynomial

$$R(z) = \sum_{k=-n}^{n} r_k z^{-k}, \quad r_{-k} = r_k^*, \tag{1.1}$$

defined for $z \in \mathbb{C}$. We denote $\mathbb{R}[z]$ the set of polynomials (1.1) with real coefficients; we name these polynomials *symmetric*, since $r_{-k} = r_k$, or even. If the coefficients are complex, the polynomials (1.1) are Hermitian, since $r_{-k} = r_k^*$; we denote $\mathbb{C}[z]$ the set of such polynomials. The degree of the polynomial is $\deg R = n$. When the degree is fixed to n, we denote the corresponding sets of polynomials by $\mathbb{R}_n[z]$ and $\mathbb{C}_n[z]$. We note that the sum and the product of symmetric (Hermitian) polynomials are also symmetric (Hermitian) polynomials; the sets $\mathbb{R}[z]$ and $\mathbb{C}[z]$ are rings.

A *causal* polynomial is $H(z) = \sum_{k=0}^{n} h_k z^{-k}$, and the set of causal polynomials is denoted $\mathbb{R}_+[z]$ or $\mathbb{C}_+[z]$, as the coefficients of $H(z)$ are real or complex, respectively. The causal part of the polynomial (1.1) is

B. Dumitrescu, *Positive Trigonometric Polynomials and Signal Processing Applications*, Signals and Communication Technology,
DOI 10.1007/978-3-319-53688-0_1

$$R_+(z) = \frac{r_0}{2} + \sum_{k=1}^{n} r_k z^{-k}. \tag{1.2}$$

We are interested especially by the values of $R(z)$ on the unit circle $\mathbb{T}$, i.e., when

$$z = \mathrm{e}^{j\omega}, \quad \omega \in [-\pi, \pi]. \tag{1.3}$$

If $R \in \mathbb{R}[z]$, then, on the unit circle, it has the form

$$R(\omega) \overset{\Delta}{=} R(\mathrm{e}^{j\omega}) = 2\mathrm{Re}[R_+(\mathrm{e}^{j\omega})] = r_0 + 2\sum_{k=1}^{n} r_k \cos k\omega \tag{1.4}$$

and has real values. Note that $R(\omega)$ is the Fourier transform of the sequence r_k, $k = -n : n$. (We use the notation $R(\omega)$ to enhance the idea of spectrum of the discrete-time "signal" r_k.) The symmetry relation $R(-\omega) = R(\omega)$ holds. The form (1.4) explains the name *trigonometric* polynomial attached to $R(z)$. Denoting

$$t = \cos\omega = \frac{z + z^{-1}}{2} \tag{1.5}$$

and $C_k(t) = \cos(k \arccos t)$ the k-th order Chebyshev polynomial, the polynomial (1.4) can be written in the form

$$R(\omega) = r_0 + 2\sum_{k=1}^{n} r_k C_k(t) = \sum_{k=0}^{n} p_k t^k = P(t), \quad t \in [-1, 1]. \tag{1.6}$$

The transformation between the coefficients r_k and p_k is linear and is given in Sect. 1.5.1.

If $R \in \mathbb{C}[z]$, we can write

$$R(z) = \sum_{k=-n}^{n} (u_k + jv_k) z^{-k} = U(z) + jV(z), \tag{1.7}$$

where $U(z)$ is a symmetric polynomial, while $V(z)$ is antisymmetric, i.e., $v_{-k} = -v_k$ (in particular, $v_0 = 0$). On the unit circle, the polynomial (1.7) becomes

$$R(\omega) = u_0 + 2\sum_{k=1}^{n} u_k \cos k\omega + 2\sum_{k=1}^{n} v_k \sin k\omega \tag{1.8}$$

and has real values.

Finally, let us define the set $\mathbb{R}[t]$ of polynomials

$$P(t) = \sum_{k=0}^{n} p_k t^k, \tag{1.9}$$

where t runs on the real axis $\mathbb{R}$, and the coefficients p_k are real. There is a one-to-one correspondence between the polynomials of the same degree in $\mathbb{R}[t]$ and $\mathbb{R}[z]$ (i.e., between $\mathbb{R}_n[t]$ and $\mathbb{R}_n[z]$), as suggested by (1.6). As shown in Sect. 1.5.1, the linear transformation (1.45) between their coefficients is invertible.

1.2 Positive Polynomials

Characterization of polynomials (1.1) that are nonnegative on the unit circle, i.e., $R(\omega) \geq 0$, $\forall \omega \in \mathbb{T}$, or positive ($R(\omega) > 0$) is of great interest in several signal processing problems, as we will see starting with this section. We denote $\overline{\mathbb{RP}}[z]$ and $\overline{\mathbb{CP}}[z]$ ($\mathbb{RP}[z]$ and $\mathbb{CP}[z]$) the sets of polynomials that are nonnegative (positive) on the unit circle, with real and complex coefficients, respectively. We treat mainly polynomials with complex coefficients, since particularization from $\mathbb{C}[z]$ to $\mathbb{R}[z]$ is trivial.

Theorem 1.1 (Riesz-Fejér, spectral factorization) *A polynomial $R \in \mathbb{C}[z]$, defined as in (1.1), is nonnegative on the unit circle if and only if a causal polynomial*

$$H(z) = \sum_{k=0}^{n} h_k z^{-k} \tag{1.10}$$

exists such that

$$R(z) = H(z)H^*(z^{-1}), \tag{1.11}$$

where

$$H^*(z) \overset{\Delta}{=} \sum_{k=0}^{n} h_k^* z^{-k}.$$

The equality (1.11) is called spectral factorization *of the nonnegative polynomial.*

Proof If (1.11) holds, then, for $z = \mathrm{e}^{j\omega}$, it becomes

$$R(\omega) = H(\omega)H(\omega)^* = |H(\omega)|^2 \geq 0. \tag{1.12}$$

To prove the converse implication, start by noticing that

$$R(1/z^*)^* = \sum_{k=-n}^{n} r_k^* z^k = R(z)$$

and so if ζ is a zero of $R(z)$, then $1/\zeta^*$ is a zero of $R(z)$. Note that $\zeta \neq 1/\zeta^*$ means $|\zeta| \neq 1$. Since $R(z)$ has $2n$ zeros, it can be expressed as the product $R(z) = aF(z)G(z)$, with $a \geq 0$. The factor

$$F(z) = \prod_{k=1}^{m} \frac{(z - z_k)(z^{-1} - z_k^*)}{1 + |z_k|^2} \tag{1.13}$$

contains the zeros z_k that are not on the unit circle, and thus come in pairs, or are on the unit circle and have double multiplicity. The factor $G(z)$ can be restricted to have distinct single zeros on the unit circle and has the form

$$G(z) = \prod_{k=m+1}^{n} \frac{(z - \mathrm{e}^{j\alpha_k})(z^{-1} - \mathrm{e}^{-j\beta_k})}{1 + \mathrm{e}^{j(\alpha_k - \beta_k)}}, \tag{1.14}$$

where $\alpha_k \neq \beta_k \pm \pi$ and α_k, β_k are distinct numbers in $(-\pi, \pi]$. For any $z = \mathrm{e}^{j\omega}$, we have

$$F(z) = \prod_{k=1}^{m} \frac{|\mathrm{e}^{j\omega} - z_k|^2}{1 + |z_k|^2} \geq 0.$$

We now prove that $m = n$, i.e., $R(z)$ has no zeros of multiplicity one (or odd, generally) on the unit circle. To do this, consider the Hermitian polynomial

$$G_1(z) = \frac{(z - \mathrm{e}^{j\alpha})(z^{-1} - \mathrm{e}^{-j\beta})}{1 + \mathrm{e}^{j(\alpha - \beta)}} \tag{1.15}$$

and assume without loss of generality that $\alpha < \beta$; assume also that $\beta \neq \alpha + \pi$. For $z = \mathrm{e}^{j\omega}$, it results that

$$G_1(\omega) = \frac{1 + \cos(\beta - \alpha) - \cos(\omega - \alpha) - \cos(\beta - \omega)}{1 + \cos(\alpha - \beta)}.$$

Denoting

$$\omega - \alpha = \frac{\beta - \alpha}{2} + \theta, \quad \beta - \omega = \frac{\beta - \alpha}{2} - \theta,$$

it results that

$$G_1(\omega) = \frac{2\cos\frac{\beta-\alpha}{2}(\cos\frac{\beta-\alpha}{2} - \cos\theta)}{1 + \cos(\alpha - \beta)}.$$

It can be seen immediately that $G_1(\omega)$ has different signs for $\theta \in [-\frac{\beta-\alpha}{2}, \frac{\beta-\alpha}{2}]$ and its complement, or equivalently, for $\omega \in [\alpha, \beta]$ and its complement over $[-\pi, \pi]$. If $\beta = \alpha + \pi$, the only possible form corresponding to (1.15) is $G_1(z) = (z + j)(z^{-1} - j)$, for which $G_1(\omega) = 2\sin\omega$, which again changes its sign over $[-\pi, \pi]$. Since $G(\omega)$ is a finite product of terms of the form $G_1(\omega)$, it results that $G(z)$ cannot be nonnegative on the whole unit circle. So $R(z) = aF(z)$ and (1.11) holds with

$$H(z) = b \prod_{k=1}^{n} (z - z_k), \tag{1.16}$$

obtained by taking half of the factors from (1.13) (b is a convenient scalar). ■

Remark 1.2 The proof above gives an algorithm for computing the spectral factorization of a nonnegative polynomial $R(z)$. First compute the $2n$ zeros of $R(z)$ and pair the zeros that are one the conjugated reciprocal of the other. (If single roots on the unit circle remain, then the polynomial is actually not nonnegative.) Then compute $H(z)$ by assigning it a zero from each pair. It is clear that the spectral factorization is not unique. It can be made unique, e.g., if the zeros of $H(z)$ are chosen to be inside or on the unit circle; thus, a minimum-phase spectral factor is obtained.

The above algorithm behaves poorly numerically and can be recommended only for rather short polynomials, in general. One difficulty is in identifying multiple roots on the unit circle. However, the worst effect is due to the ill conditioning of the operation of forming the coefficients of a polynomial from its roots. Other algorithms are discussed in Appendix B. ■

Remark 1.3 In the particular case of real coefficients, if $R \in \mathbb{R}[z]$ and $R(\omega) \geq 0$, $\forall \omega \in \mathbb{T}$, then there exists a causal polynomial $H(z)$ with real coefficients such that $R(z) = H(z)H(z^{-1})$. This follows from the fact that if ζ is a zero of $R(z)$, then ζ^*, $1/\zeta$ and $1/\zeta^*$ are also zeros of $R(z)$. Hence, we can assign ζ and ζ^* to the same spectral factor, obtaining a polynomial with real coefficients. In particular, the minimum-phase spectral factor has real coefficients. However, a nonnegative $R(z)$ may have also spectral factors with complex coefficients. ■

The spectral factorization relation (1.11) can be written in terms of the coefficients of $R(z)$ and $H(z)$ as follows:

$$r_k = \sum_{i=k}^{n} h_i h_{i-k}^*, \quad k \geq 0. \tag{1.17}$$

This expression tells that r_k is an *autocorrelation sequence* if $R(\omega) \geq 0$.

Example 1.4 (Autocorrelations of an MA process.) Consider the MA (moving average) process

$$y(\ell) = \sum_{k=0}^{n} h_k w(\ell - k), \tag{1.18}$$

where $w(\ell)$ is white noise of variance 1, i.e., $E\{w(\ell)w^*(\ell - k)\} = \delta_k$. The autocorrelation sequence of the MA process is

$$r_k = E\{y(\ell)y^*(\ell-k)\} = E\left\{\sum_{i=0}^{n} h_i w(\ell-i) \sum_{m=0}^{n} h_m^* w^*(\ell-k-m)\right\}. \quad (1.19)$$

Simple computation shows that r_k is given by (1.17), for $k \geq 0$ (and $r_{-k} = r_k^*$). Due to Theorem 1.1, any finite Hermitian sequence r_k for which $R(\omega) \geq 0$ is the autocorrelation sequence of an MA process. From now on, we will use the terms nonnegative polynomial and autocorrelation (or nonnegative) sequence as synonyms. ■

Problem (*MA_Estimation*) Assume that we know the order n of the MA process (1.18) and we want to estimate its parameters h_k, $k = 0 : n$, from a finite realization of the process $y(\ell)$, $\ell = 0 : L-1$. A possible solution is to compute an estimation $\hat{r}_k$ of the MA autocorrelation sequence from the given L samples, using, for instance, the biased estimation

$$\hat{r}_k = \frac{1}{L}\sum_{\ell=k}^{L-1} y(\ell)y^*(\ell-k), \quad k = 0 : n, \quad (1.20)$$

or other estimations [1, 2]. Then, perform the spectral factorization $\hat{R}(z) = \hat{H}(z)\hat{H}^*(z^{-1})$ to obtain estimations of h_k.

This algorithm cannot work if, due to the finite character of the estimation, the estimated autocorrelation sequence is not nonnegative, i.e., there are frequencies ω for which $\hat{R}(\omega) < 0$. In this case, spectral factorization is not possible. For the biased estimation (1.20), the sequence $\hat{r}_k$ is always nonnegative; however, this is not true for other estimations that may be more meaningful, especially for short data sets [2]. So, if $\hat{R}(\omega)$ is not nonnegative, we must replace $\hat{r}_k$ with a nonnegative sequence, possibly the nearest. We will show immediately that this is possible. ■

Remark 1.5 The set $\mathbb{CP}_n[z] \subset \mathbb{C}[z]$ of Hermitian polynomials of degree at most n that are positive on the unit circle is *convex*. Indeed, if $R_1(z)$ and $R_2(z)$ are positive, so is any convex combination $aR_1(z) + (1-a)R_2(z)$, for $a \in [0, 1]$; note that the degree of the convex combination might be smaller than the degree of the polynomials. Moreover, $\mathbb{CP}_n[z]$ is a *cone*, since if $R \in \mathbb{CP}_n[z]$, then $aR \in \mathbb{CP}_n[z]$, for any $a > 0$. Also, the set $\overline{\mathbb{CP}}_n[z]$ of polynomials nonnegative on the unit circle (which is the closure of $\mathbb{CP}_n[z]$) is a convex cone. ■

Problem (*Nearest_autocorrelation*) As we have seen above in Problem *MA_Estimation*, it may be useful to find the autocorrelation nearest from a given sequence. Suppose that the Hermitian sequence $\hat{r}_k$, $k = -n : n$, with $\hat{r}_{-k} = \hat{r}_k^*$, is given and we want to find the nonnegative sequence r_k that is nearest from $\hat{r}_k$. The distance between the sequences is measured via the norm

$$dist(\boldsymbol{r}, \hat{\boldsymbol{r}}) = \sum_{k=-n}^{n} |r_k - \hat{r}_k|^2 = (\boldsymbol{r} - \hat{\boldsymbol{r}})^H \boldsymbol{\Gamma} (\boldsymbol{r} - \hat{\boldsymbol{r}}), \tag{1.21}$$

where $\boldsymbol{r} = [r_0 \ r_1 \ \dots \ r_n]^T \in \mathbb{C}^{n+1}$ is the vector of the elements in the sequence and $\boldsymbol{\Gamma} = \operatorname{diag}(1, 2, \dots, 2)$. Other norms could be used as well; also, $\boldsymbol{\Gamma}$ may be an arbitrary positive definite matrix. Since the set of autocorrelation sequences is convex, the problem

$$\begin{array}{cl} \min\limits_{\boldsymbol{r}} & (\boldsymbol{r} - \hat{\boldsymbol{r}})^H \boldsymbol{\Gamma} (\boldsymbol{r} - \hat{\boldsymbol{r}}) \\ \text{subject to} & R(\omega) \geq 0, \ \ \forall \omega \in [-\pi, \pi] \end{array} \tag{1.22}$$

of finding the autocorrelation nearest from $\hat{\boldsymbol{r}}$ has a *unique* solution. Solving (1.22) is not trivial. As posed, it is a semi-infinite optimization problem, since the number of constraints is infinite. An approximated solution may be obtained by discretizing the constraint $R(\omega) \geq 0$ over a finite grid of frequencies; note that for a given ω, the constraint is linear in the coefficients r_k; however, such a solution may become negative between some grid points and is usually not optimal. As this is probably the simplest problem involving a signal processing application of positive polynomials, we will concentrate in the sequel (in this and next chapter) on presenting the tools necessary for its solution, restating the problem as soon as we advance. Other important applications will be presented further in Chaps. 5–9. ■

The spectral factorization relations (1.11) and (1.12) allow a generalization, which now may seem insignificant but later will prove important.

Definition 1.6 A trigonometric polynomial $R(z)$ defined as in (1.1) is *sum-of-squares* if it can be written in the form

$$R(z) = \sum_{\ell=1}^{\nu} H_\ell(z) H_\ell^*(z^{-1}), \tag{1.23}$$

for some $\nu \geq 0$ and causal polynomials $H_\ell(z)$. ■

On the unit circle, a sum-of-squares polynomial is

$$R(\omega) = \sum_{\ell=1}^{\nu} |H_\ell(\omega)|^2 \tag{1.24}$$

and so is nonnegative. Theorem 1.1 says that any polynomial that is nonnegative on the unit circle is sum-of-squares with a single term. Therefore, the sets of nonnegative and sum-of-squares polynomials coincide.

For polynomials of real variable, there is no equivalent of the spectral factorization theorem. Let $P \in \mathbb{R}[t]$ be nonnegative, i.e., $P(t) \geq 0, \forall t \in \mathbb{R}$. It is clear that its degree n must be even. Such a polynomial is sum-of-squares if it can be written as

$$P(t) = \sum_{\ell=1}^{\nu} H_\ell(t)^2, \tag{1.25}$$

for some $\nu \geq 0$ and polynomials $H_\ell(t)$ of degree at most $n/2$.

Theorem 1.7 *Any polynomial $P \in \mathbb{R}[t]$ that is nonnegative on the real axis can be expressed as a sum-of-squares with two terms.*

The proof is elementary and is presented in Sect. 1.5.2. So again the sets of positive and sum-of-squares polynomials are identical; however, a positive polynomial cannot be expressed as a square, as in the case of polynomials in $\mathbb{C}[z]$.

1.3 Toeplitz Positivity Conditions

Let $R \in \mathbb{C}_n[z]$ be nonnegative. We have shown in the previous section that its coefficients r_k, $k = -n : n$, form an autocorrelation sequence. Let us consider again the MA process (1.18), reminding that $r_k = E\{y(\ell)y^*(\ell - k)\}$. For an arbitrary positive integer m, denote

$$\boldsymbol{y}_m(\ell) = \begin{bmatrix} y(\ell) \\ y(\ell-1) \\ \vdots \\ y(\ell-m) \end{bmatrix}. \tag{1.26}$$

It is clear that the $(m+1) \times (m+1)$ matrix

$$\boldsymbol{R}_m = E\{\boldsymbol{y}_m \boldsymbol{y}_m^H\} \tag{1.27}$$

is positive semidefinite. (For any $\boldsymbol{x} \in \mathbb{C}^{m+1}$, we have $\boldsymbol{x}^H \boldsymbol{R}_m \boldsymbol{x} = E\{\boldsymbol{x}^H \boldsymbol{y}_m \boldsymbol{y}_m^H \boldsymbol{x}\} = E\{|\boldsymbol{x}^H \boldsymbol{y}_m|^2\} \geq 0$.) Using (1.26) and the definition of r_k, we can write

$$\boldsymbol{R}_m = \begin{bmatrix} & \vdots & \\ \ldots & E\{y(\ell-i)y^*(\ell-s)\} & \ldots \\ & \vdots & \end{bmatrix}_{is} = \begin{bmatrix} & \vdots & \\ \ldots & r_{s-i} & \ldots \\ & \vdots & \end{bmatrix}_{is},$$

where $r_k = 0$ if $|k| > n$. So, for $m > n$, the matrix $\boldsymbol{R}_m$ has the Toeplitz structure

$$\boldsymbol{R}_m = \begin{bmatrix} r_0 & r_1 & \dots & r_n & 0 & \dots & 0 \\ r_{-1} & r_0 & r_1 & \ddots & r_n & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & \ddots & \ddots & 0 \\ \vdots & \ddots & \ddots & \ddots & \ddots & \ddots & r_n \\ 0 & \ddots & \ddots & \ddots & \ddots & \ddots & \vdots \\ \vdots & \ddots & r_{-n} & \ddots & r_{-1} & r_0 & r_1 \\ 0 & \dots & 0 & r_{-n} & \dots & r_{-1} & r_0 \end{bmatrix}$$
$$\stackrel{\Delta}{=} \operatorname{Toep}(r_0, r_1, \dots, r_n, 0, \dots, 0). \tag{1.28}$$

If $m \le n$, the matrix $\boldsymbol{R}_m$ is just a principal submatrix of the matrix above, precisely $\boldsymbol{R}_m = \operatorname{Toep}(r_0, \dots, r_m)$. The considerations above lead to the following result.

Theorem 1.8 *The polynomial $R \in \mathbb{C}_n[z]$ is nonnegative if and only if the matrices $\boldsymbol{R}_m$ defined by (1.28) are positive semidefinite for any m.*

Proof The "only if" part has been proved above. The "if" part results by contradiction. Suppose that some $\omega \in [-\pi, \pi]$ exists such that $R(\omega) < 0$. Denote

$$\boldsymbol{x}_m = [1 \ \mathrm{e}^{j\omega} \ \dots \ \mathrm{e}^{jn\omega} \ 0 \ \dots \ 0]^T \in \mathbb{C}^{m+1},$$

define

$$\alpha_m = \frac{1}{m+1} \boldsymbol{x}_m^H \boldsymbol{R}_m \boldsymbol{x}_m = \sum_{k=-n}^{n} \frac{m+1-|k|}{m+1} r_k \mathrm{e}^{-jk\omega} \tag{1.29}$$

and remark that

$$|R(\omega) - \alpha_m| = \left| \sum_{k=-n}^{n} \frac{|k|}{m+1} r_k \mathrm{e}^{-jk\omega} \right| \le \frac{2n}{m+1} \sum_{k=1}^{n} |r_k|.$$

Taking $m > 2n(\sum_{k=1}^{n} |r_k|)/|R(\omega)|$, it results that

$$|R(\omega) - \alpha_m| < |R(\omega)|,$$

which implies that $\alpha_m < 0$ and so, by virtue of the definition (1.29), the matrix $\boldsymbol{R}_m$ is not positive semidefinite, which is the sought after contradiction. ■

Remark 1.9 Defining $\boldsymbol{R}_m$ as in (1.28), for a particular m, the condition $\boldsymbol{R}_m \succeq 0$ is necessary but not sufficient for the nonnegativity of the polynomial $R(z)$. Theorem 1.8 provides only the following possible nonnegativity check: If $\boldsymbol{R}_m$ is not positive semidefinite, then $R(z)$ is not nonnegative. ■

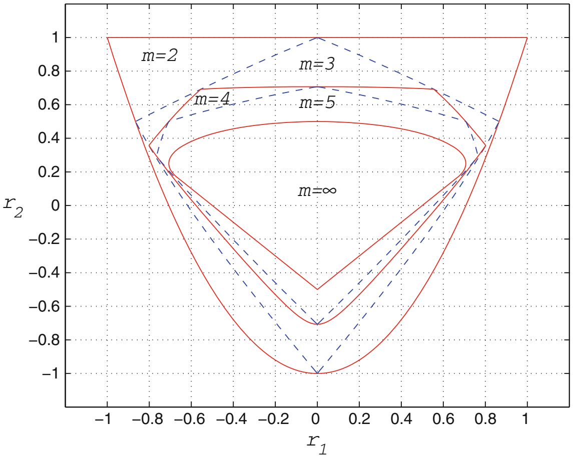


Fig. 1.1 From exterior to interior, the domains for which $\boldsymbol{R}_m \succeq 0$, drawn in the plane (r_1, r_2), for $m = 2, 3, 4, 5$ (with $n = 2$, $r_0 = 1$). The most interior domain corresponds to values of r_1, r_2 for which $1 + 2r_1 \cos\omega + 2r_2 \cos 2\omega \geq 0$

Example 1.10 Since the principal submatrices of a positive semidefinite matrix are positive semidefinite themselves, it is clear that if $\boldsymbol{R}_m \succeq 0$, then $\boldsymbol{R}_i \succeq 0$, for all $i \leq m$. On the other side, as $m \to \infty$, the condition $\boldsymbol{R}_m \succeq 0$ approximates better and better the nonnegativity condition $R(\omega) \geq 0$. One may wonder how good the approximation is for moderate values of m, i.e., values only slightly greater than n. To have an image of this phenomenon, let us take $n = 2$ and, without loss of generality, consider $r_0 = 1$. For a polynomial $R(z)$ with real coefficients, we draw, in the (r_1, r_2) plane, the domain for which $R(\omega) = 1 + 2r_1 \cos\omega + 2r_2 \cos 2\omega \geq 0$. In Fig. 1.1, this is the most interior domain, marked with $m = \infty$. Similarly, we draw domains for the values (r_1, r_2) for which $\boldsymbol{R}_m \succeq 0$. In Fig. 1.1, such domains are presented for values of m going from 2 (the most exterior domain) to 5 (the one but most interior). Obviously, all these domains are convex. It appears that for small values of m, the approximation is not very good and we may assume that this is happening also for larger n. So, probably, the practical importance of the conditions $\boldsymbol{R}_m \succeq 0$ is small, although theoretically they prove to be useful. ■

1.4 Positivity on an Interval

We seek now characterizations of polynomials that are nonnegative only on an interval, unlike those discussed in Sect. 1.2, which were nonnegative on the whole unit circle or real axis. For polynomials of real variable, $P \in \mathbb{R}[t]$, the relevant cases are those of the finite interval $[a, b]$ and of the half-infinite interval $[a, \infty)$. For $\mathbb{R}[z]$ or $\mathbb{C}[z]$, we are interested by trigonometric polynomials $R(z)$ for which $R(\omega) \geq 0$ for $\omega \in [\alpha, \beta] \subset [-\pi, \pi]$.

In this section, we present only the main results. The proofs can be found at the end of the chapter; although, at least for polynomials of real variable, elementary proofs are available (see [3]), we have preferred to use the following derivation scheme: Theorem 1.1 ⇒ Theorem 1.11 ⇒ Theorem 1.13, Theorem 1.11 ⇒ Theorem 1.15, Theorem 1.17, Theorem 1.18. The proofs are based on transformations between different intervals, which may be interesting by themselves.

All the theorems below show the same type of result. A polynomial that is nonnegative on an interval can be expressed as a function of squared polynomials (which are globally nonnegative) and elementary polynomials (of degree one or two) that are positive on that interval.

Theorem 1.11 *Let $P \in \mathbb{R}[t]$ be such that $P(t) \geq 0$ for any $t \in [a, b]$. Then, if* $\deg P = 2n$*, the polynomial can be expressed as*

$$P(t) = F(t)^2 + (t-a)(b-t)G(t)^2, \tag{1.30}$$

with $F, G \in \mathbb{R}[t]$ and $\deg F \leq n$, $\deg G \leq n-1$.

If $\deg P = 2n+1$, *then the polynomial can be expressed as*

$$P(t) = (t-a)\tilde{F}(t)^2 + (b-t)\tilde{G}(t)^2, \tag{1.31}$$

with $\tilde{F}, \tilde{G} \in \mathbb{R}[t]$ and $\deg \tilde{F} \leq n$, $\deg \tilde{G} \leq n$.

Example 1.12 The polynomial $P(t) = 1 - t^6$ is nonnegative on $[-1, 1]$ (and negative elsewhere). It can be written as

$$1 - t^6 = 0.4641(1-t^2)^2 + (1-t^2)(0.73205 + t^2)^2,$$

and so the relation (1.30) holds with $F(t) = \sqrt{0.4641}(1-t^2)$ and $G(t) = 0.73205 + t^2$. ■

Theorem 1.13 *A polynomial $P \in \mathbb{R}_n[t]$ for which $P(t) \geq 0$, for any $t \in [a, \infty]$, can be expressed as*

$$P(t) = F(t)^2 + (t-a)G(t)^2, \tag{1.32}$$

with $F, G \in \mathbb{R}[t]$, $\deg F \leq \lfloor n/2 \rfloor$, $\deg G \leq \lfloor (n-1)/2 \rfloor$.

Example 1.14 The polynomial $P(t) = t^3 + t^2 + t$ is nonnegative on $[0, \infty]$ (and negative elsewhere). It can be written in the form (1.32) as $P(t) = 3t^2 + t(t-1)^2$. ■

Theorem 1.15 *A polynomial $R \in \mathbb{C}_n[z]$ for which $R(\omega) \geq 0$, for any $\omega \in [\alpha, \beta] \subset (-\pi, \pi)$, can be expressed as*

$$R(z) = F(z)F^*(z^{-1}) + D_{\alpha\beta}(z) \cdot G(z)G^*(z^{-1}), \tag{1.33}$$

where F, G are causal polynomials with complex coefficients, of degree at most n and $n-1$, respectively. The polynomial

$$D_{\alpha\beta}(z) = d_1 z^{-1} + d_0 + d_1^* z \tag{1.34}$$

is defined such that $D_{\alpha\beta}(\omega)$ is nonnegative for $\omega \in [\alpha, \beta]$ and negative on its complementary. Denoting

$$a = \tan\tfrac{\alpha}{2}, \quad b = \tan\tfrac{\beta}{2}, \tag{1.35}$$

the coefficients of $D_{\alpha\beta}(z)$ are

$$d_0 = -\tfrac{ab+1}{2}, \quad d_1 = \tfrac{1-ab}{4} + j\tfrac{a+b}{4}. \tag{1.36}$$

(These coefficients may be multiplied by any positive constant.)

Remark 1.16 The proof of the above theorem is based on the transformation $t = \tan(\omega/2)$ (explained in detail in Sect. 1.5.5) that maps $\omega \in [\alpha, \beta]$ to $t \in [a, b]$. Theorem 1.11 is used to express the transformed polynomial, which is nonnegative on $[a, b]$, as in relation (1.30). When $\alpha = -\pi$ or $\beta = \pi$, Theorem 1.13 should be applied, as either a or b are infinite; in this case, the polynomial $D_{\alpha\beta}$ has a slightly different form; the reader is invited to prove it in Problem 1.6.

A simple way of avoiding a different formula when $\alpha = -\pi$ or $\beta = \pi$ is to work with the polynomial $\tilde{R}(z) = R(e^{j\gamma} z)$ instead of $R(z)$, where γ is chosen such that $[\alpha + \gamma, \beta + \gamma] \subset (-\pi, \pi)$. It is obvious that $\tilde{R}(\omega)$ is nonnegative on $[\alpha + \gamma, \beta + \gamma]$ if $R(\omega)$ is nonnegative on $[\alpha, \beta]$. The transformation between the coefficients of $R(z)$ and $\tilde{R}(z)$ is linear.

Finally, note that if $\alpha = -\beta$, then the coefficients of $D_{\alpha\beta}$ are real. ■

For trigonometric polynomials with *real* coefficients, the analogous of Theorem 1.15 is as follows. Note that now $[\alpha, \beta] \subset [0, \pi]$, as $R(-\omega) = R(\omega)$ for $R \in \mathbb{R}[z]$.

Theorem 1.17 *Let $R \in \mathbb{R}_n[z]$ be such that $R(\omega) \geq 0$ for any $\omega \in [\alpha, \beta] \subset [0, \pi]$. If n is even, then the polynomial can be expressed as*

$$R(z) = F(z)F(z^{-1}) + (\tfrac{z+z^{-1}}{2} - \cos\beta)(\cos\alpha - \tfrac{z+z^{-1}}{2}) \cdot G(z)G(z^{-1}), \tag{1.37}$$

with $F, G \in \mathbb{R}_+[z]$, $\deg F \leq n$, $\deg G \leq n-2$.

If n is odd, then the polynomial can be expressed as

$$R(z) = (\tfrac{z+z^{-1}}{2} - \cos\beta) \cdot \tilde{F}(z)\tilde{F}(z^{-1}) + (\cos\alpha - \tfrac{z+z^{-1}}{2}) \cdot \tilde{G}(z)\tilde{G}(z^{-1}), \tag{1.38}$$

with $\tilde{F}, \tilde{G} \in \mathbb{R}_+[z]$, $\deg \tilde{F} \leq n-1$, $\deg \tilde{G} \leq n-1$.

On the unit circle, the relations (1.37) and (1.38) can be stated in another form, by using symmetric (not causal, as above) polynomials as parameters.

Theorem 1.18 *Let $R \in \mathbb{R}_n[z]$ be such that $R(\omega) \geq 0$ for any $\omega \in [\alpha, \beta] \subset [0, \pi]$. If n is even, then, on the unit circle, the polynomial can be expressed as*

$$R(\omega) = R_1(\omega)^2 + (\cos\omega - \cos\beta)(\cos\alpha - \cos\omega) \cdot R_2(\omega)^2, \tag{1.39}$$

with $R_1, R_2 \in \mathbb{R}[z]$*,* $\deg R_1 \le n/2$*,* $\deg R_2 \le n/2 - 1$*.*

If n *is odd, then the polynomial can be expressed as*

$$R(\omega) = (\cos\omega - \cos\beta) \cdot \tilde{R}_1(\omega)^2 + (\cos\alpha - \cos\omega) \cdot \tilde{R}_2(\omega)^2, \tag{1.40}$$

with $\tilde{R}_1, \tilde{R}_2 \in \mathbb{R}[z]$*,* $\deg \tilde{R}_1 \le (n-1)/2$*,* $\deg \tilde{R}_2 \le (n-1)/2$*.*

Example 1.19 The symmetric polynomial

$$R(z) = \tfrac{z^4+z^{-4}}{2} - 13\tfrac{z^3+z^{-3}}{2} - 30\tfrac{z^2+z^{-2}}{2} + 121\tfrac{z+z^{-1}}{2} - 79$$

has real coefficients and is nonnegative on the interval $[0, \pi/3]$ (and a little outside it; draw its graph !). Theorem 1.18 says that we can write

$$R(\omega) = R_1(\omega)^2 + D(\omega)R_2(\omega)^2,$$

where (we introduce a factor of 2, to get integer coefficients)

$$D(\omega) = 2(\cos\omega - \tfrac{1}{2})(1 - \cos\omega) = -\cos 2\omega + 3\cos\omega - 2.$$

It can be checked that

$$R_1(\omega) = 2(\cos 2\omega - 1), \quad R_2(\omega) = 2(\cos\omega + 4).$$

Using (1.5), it results that

$$\begin{aligned} R_1(z)^2 &= 4\left(\tfrac{z^2+z^{-2}}{2} - 1\right)^2 = (z^{-4} - 2z^{-2} + 1)(z^4 + 2z^2 + 1), \\ R_2(z)^2 &= 4\left(\tfrac{z+z^{-1}}{2} + 4\right)^2 = (z^{-2} + 8z^{-1} + 1)(z^2 + 8z + 1). \end{aligned}$$

We see immediately that relation (1.37) holds with

$$F(z) = z^{-4} - 2z^{-2} + 1, \quad G(z) = \tfrac{1}{\sqrt{2}}(z^{-2} + 8z^{-1} + 1).$$

The reader is invited to use the technique illustrated by this example for deriving Theorem 1.18 from Theorem 1.17. ■

Remark 1.20 (Positivity on a union of intervals) The results in this section have been all for polynomials that are positive on an interval. A natural question is whether the same type of result holds for an union of disjoint intervals $\mathcal{U} = \bigcup_{i=1}^{\nu}[a_i, b_i]$. The full answer will be given in a broader context in Chap. 4 . Here, we give only the generalization of Theorem 1.11. Let $D(t)$ be a polynomial that is nonnegative on $\mathcal{U}$ and positive elsewhere, for example

$$g_1(t) = (-1)^{\nu+1} \prod_{i=1}^{\nu} (t - a_i)(b_i - t).$$

Then, any polynomial $P \in \mathbb{R}[t]$ with $P(t) \geq 0$ for all $t \in \mathcal{U}$ can be expressed as

$$P(t) = s_0(t) + g_1(t)s_1(t), \tag{1.41}$$

where s_0 and s_1 are sum-of-squares. So, the difference with respect to Theorem 1.11 and (1.30) is that here we have sum-of-squares instead of simple squares and also that the degrees of these sum-of-squares may be higher than deg P.

1.5 Details and Other Facts

1.5.1 Chebyshev Polynomials

The (first kind) k-th order Chebyshev polynomial is defined as $C_k(t) = \cos(k \arccos t)$, for $t \in [-1, 1]$. With $t = \cos\omega$, the definition is $C_k(\cos\omega) = \cos(k\omega)$. It is immediate that $C_0(t) = 1$, $C_1(t) = t$, $C_2(t) = 2t^2 - 1$. Also, the recurrence relation

$$C_{k+1}(t) = 2tC_k(t) - C_{k-1}(t), \quad k \geq 1, \tag{1.42}$$

holds. Since $\deg C_k = k$, the polynomials $C_k(t)$, $k = 0 : n$, form a basis to $\mathbb{R}_n[t]$. Let us denote

$$C_k(t) = \sum_{i=0}^{k} c_{ki} t^i, \tag{1.43}$$

where the coefficients c_{ki} can be computed through the recurrence given by (1.42), i.e., $c_{k+1,i} = 2c_{k,i-1} - c_{k-1,i}$. The transformation between the canonical basis and the basis of Chebyshev polynomials is given by

$$\begin{bmatrix} C_0(t) \\ C_1(t) \\ \vdots \\ C_n(t) \end{bmatrix} = \begin{bmatrix} c_{00} & & & \\ c_{10} & c_{11} & & \\ \vdots & \vdots & \ddots & \\ c_{n0} & c_{n1} & \dots & c_{nn} \end{bmatrix} \begin{bmatrix} 1 \\ t \\ \vdots \\ t^n \end{bmatrix}. \tag{1.44}$$

Let us denote $\boldsymbol{C}$ the lower triangular matrix from (1.44). If $P \in \mathbb{R}[t]$ is an arbitrary polynomial and

$$P(t) = \sum_{k=0}^{n} p_k t^k = \sum_{k=0}^{n} r_k C_k(t),$$

then the vectors of coefficients $\boldsymbol{p}$ (in the canonical basis) and $\boldsymbol{r}$ (in the Chebyshev basis) are related by

$$\boldsymbol{p} = \boldsymbol{C}^T \boldsymbol{r}. \tag{1.45}$$

Finally, note that an equivalent definition of Chebyshev polynomials is based on the relation

$$\tfrac{z^{-k}+z^k}{2} = C_k(\tfrac{z^{-1}+z}{2}).$$

1.5.2 Positive Polynomials in $\mathbb{R}[t]$ as Sum-of-Squares

Theorem 1.7 states that if $P \in \mathbb{R}[t]$ and $P(t) \geq 0$, $\forall t \in \mathbb{R}$, then $P(t) = F(t)^2 + G(t)^2$. Here is the proof.

The degree of the polynomial is $2n$. Without loss of generality, we can assume that the coefficient of t^{2n} in $P(t)$ is 1. By expressing the polynomial function of its roots $a_i \pm jb_i$, $i = 1 : n$, we obtain

$$\begin{aligned} P(t) &= \prod_{i=1}^{n} (t - a_i - jb_i)(t - a_i + jb_i) \\ &= \prod_{i=1}^{n} (t - a_i - jb_i) \prod_{i=1}^{n} (t - a_i + jb_i) \\ &= [F(t) - jG(t)][F(t) + jG(t)] = F(t)^2 + G(t)^2, \end{aligned} \tag{1.46}$$

where $F, G \in \mathbb{R}[t]$ and their degrees are at most n (but at least one has degree equal to n).

1.5.3 Proof of Theorem 1.11

It is only necessary to prove the Theorem for the interval $[-1, 1]$. The linear change of variable

$$t = \frac{(b-a)\tau + a + b}{2} \tag{1.47}$$

transforms $[-1, 1]$ into $[a, b]$. So, if $P(t) \geq 0$ for any $t \in [a, b]$, then

$$\tilde{P}(\tau) = P(\tfrac{(b-a)\tau+a+b}{2}) \geq 0, \quad \forall \tau \in [-1, 1].$$

So, we prove that any $P \in \mathbb{R}[t]$ such that $P(t) \geq 0$ for $t \in [-1, 1]$ can be expressed as

$$P(t) = \begin{cases} F(t)^2 + (1-t^2)G(t)^2, & \text{if} \deg P = 2n, \\ (1-t)\tilde{F}(t)^2 + (1+t)\tilde{G}(t)^2, & \text{if} \deg P = 2n+1. \end{cases} \tag{1.48}$$

We discuss first the case where deg $P = 2n$. Let us replace $t = (z^{-1}+z)/2$ and denote $R(z) = P(t)$. For $z = \mathrm{e}^{j\omega}$, it results that $t = \cos\omega$, relation that links polynomials in $\mathbb{R}_n[t]$ and $\mathbb{R}_n[z]$ (and polynomials in $\mathbb{R}_n[t]$ nonnegative on $[-1, 1]$ to polynomials in $\mathbb{R}_n[z]$ nonnegative on the unit circle). Since $R(z)$ is nonnegative on the unit circle, by Theorem 1.1 it can be written as $R(z) = H(z)H(z^{-1})$. We can write

$$H(z) = \sum_{k=0}^{2n} h_k z^{-k} = z^{-n}[A(z) + B(z)], \tag{1.49}$$

where $A(z)$ is symmetric ($a_{-k} = a_k$), $B(z)$ is antisymmetric ($b_{-k} = -b_k$) and their degree is at most n. Symmetry means that $A(z^{-1}) = A(z)$, while antisymmetry means that $B(z^{-1}) = -B(z)$. It results that

$$R(z) = H(z)H(z^{-1}) = A(z)^2 - B(z)^2. \tag{1.50}$$

Since

$$\tfrac{z^{-k}+z^k}{2} = C_k(\tfrac{z^{-1}+z}{2}),$$

where $C_k(z)$ is the k-th order Chebyshev polynomial, the symmetric polynomial $A(z)$ can be written as

$$A(z) = a_0 + 2\sum_{k=1}^{n} a_k \frac{z^{-k}+z^k}{2} = a_0 + 2\sum_{k=1}^{n} a_k C_k(t) \overset{\Delta}{=} F(t). \tag{1.51}$$

The antisymmetric polynomial $B(z)$ can be expressed as

$$\begin{aligned} B(z) &= 2\sum_{k=1}^{n} b_k \frac{z^{-k}-z^k}{2} \\ &= \frac{z^{-1}-z}{2} \cdot 2\sum_{k=1}^{n} b_k (z^{-k+1} + \ldots + 1 + \ldots + z^{k-1}) \\ &\overset{\Delta}{=} \frac{z^{-1}-z}{2} G(t). \end{aligned} \tag{1.52}$$

Since

$$\left(\tfrac{z^{-1}-z}{2}\right)^2 = \left(\tfrac{z^{-1}+z}{2}\right)^2 - 1,$$

it results that

$$B(z)^2 = (t^2-1)G(t)^2. \tag{1.53}$$

Substituting (1.51) and (1.53) into (1.50), we obtain the first expression from (1.48).

The case with deg $P = 2n + 1$ is proved similarly, with the difference that (1.49) is replaced with

$$H(z) = z^{-n}[(1 + z^{-1})A(z) + (1 - z^{-1})B(z)],$$

where now both $A(z)$ and $B(z)$ are symmetric polynomials.

1.5.4 Proof of Theorem 1.13

It is clear that if $P(t) \geq 0, \forall t \in [a, \infty)$, then $P(t - a) \geq 0, \forall t \in [0, \infty)$. So, instead of (1.32), we prove that any polynomial $P(t)$ that is nonnegative for $t \geq 0$ can be written as

$$P(t) = F(t)^2 + tG(t)^2. \tag{1.54}$$

For the proof, we use Theorem 1.11 and the *Goursat transform* [4], which transforms a polynomial positive on $[-1, 1]$ into a polynomial positive on $[0, \infty)$ (and viceversa). Denote as usual $n = \deg P$. The n-th order Goursat transform of $P(t)$ is

$$\breve{P}(t) = (1 + t)^n P\left(\tfrac{1-t}{1+t}\right).$$

Note that $\breve{n} = \deg \breve{P} \leq n$. The Goursat transform is its own inverse, modulo a constant factor:

$$\breve{\breve{P}}(t) = (1 + t)^n \breve{P}\left(\tfrac{1-t}{1+t}\right) = (1 + t)^n \left(1 + \tfrac{1-t}{1+t}\right)^n P\left(\tfrac{1-\frac{1-t}{1+t}}{1+\frac{1-t}{1+t}}\right) = 2^n P(t).$$

If $P(t) \geq 0$ for $t \geq 0$, then $\breve{P}(t) \geq 0$ for $t \in [-1, 1]$. From Theorem 1.11, it results that depending on the parity of $\breve{n}$, we can write either

$$\breve{P}(t) = A(t)^2 + (1 - t^2)B(t)^2 \tag{1.55}$$

or

$$\breve{P}(t) = (1 - t)A(t)^2 + (1 + t)B(t)^2. \tag{1.56}$$

In both cases, by applying the Goursat transform, we obtain

$$2^n P(t) = (1 + t)^m [F(t)^2 + tG(t)^2], \tag{1.57}$$

where $m = n - \breve{n}$. (If (1.55) holds, then $F(t) = A(t)$ and $G(t) = 2B(t)$. If (1.56) holds, then $F(t) = \sqrt{2}B(t)$ and $G(t) = \sqrt{2}A(t)$.) If m is even, then (1.57) has already the form (1.54). If m is odd, then we notice that

$$(1+t)[F(t)^2 + tG(t)^2] = [F(t) + tG(t)]^2 + t[F(t) - G(t)]^2$$

has the form (1.54).

1.5.5 Proof of Theorem 1.15

Consider the bilinear transform

$$z = \frac{1+jt}{1-jt} = \frac{(1+jt)^2}{1+t^2}. \tag{1.58}$$

For $t \in \mathbb{R}$ and $z = \mathrm{e}^{j\omega}$, it results that

$$\cos\omega = \frac{z+z^{-1}}{2} = \frac{1-t^2}{1+t^2}, \quad \sin\omega = \frac{z-z^{-1}}{2j} = \frac{2t}{1+t^2}, \tag{1.59}$$

It is clear that the transform (1.58) maps the real axis to the unit circle, since relations (1.59) are equivalent to

$$t = \tan(\omega/2). \tag{1.60}$$

A polynomial $R \in \mathbb{C}[z]$, with $\deg R = n$, can be written as in (1.7)

$$R(z) = U(z) + jV(z) = \tilde{U}(\tfrac{z+z^{-1}}{2}) + \tfrac{z-z^{-1}}{2j} \cdot \tilde{V}(\tfrac{z+z^{-1}}{2}),$$

with $\tilde{U}, \tilde{V} \in \mathbb{R}[z]$ and $\deg \tilde{U} = n$, $\deg \tilde{V} = n-1$ (one of the degrees could be smaller, but this is irrelevant). On the unit circle, this is

$$R(\omega) = \tilde{U}(\cos\omega) + \sin\omega \cdot \tilde{V}(\cos\omega). \tag{1.61}$$

Taking (1.59) into account, we obtain

$$R(\omega) = \tilde{U}(\tfrac{1-t^2}{1+t^2}) + \tfrac{2t}{1+t^2} \cdot \tilde{V}(\tfrac{1-t^2}{1+t^2}) = \frac{P(t)}{(1+t^2)^n}, \tag{1.62}$$

where $P \in \mathbb{R}[t]$ and $\deg P = 2n$. Reciprocally, for any polynomial $P \in \mathbb{R}[t]$ of degree $2n$, we can find $\tilde{U}, \tilde{V} \in \mathbb{R}[z]$, of degrees n and $n-1$, respectively, such that (1.62) holds; this is due to the fact that the polynomials $(1-t^2)^k(1+t^2)^{n-k}$, $k = 0:n$, and $t(1-t^2)^k(1+t^2)^{n-1-k}$, $k = 0:n-1$, form a basis to $\mathbb{R}_{2n}[t]$.

From (1.62), it results also that if $R(\omega) \geq 0$ for $\omega \in [\alpha, \beta]$, then $P(t) \geq 0$ for $t \in [a, b]$, where a and b are defined in (1.35). So, using Theorem 1.11, we can write

$$\frac{P(t)}{(1+t^2)^n} = \frac{\tilde{A}(t)^2}{(1+t^2)^n} + \frac{(t-a)(b-t)}{(1+t^2)} \cdot \frac{\tilde{B}(t)^2}{(1+t^2)^{n-1}}, \tag{1.63}$$

with $\deg \tilde{A} \le n$, $\deg \tilde{B} \le n-1$.

Now, using the transformations (1.59), the following relations similar to (1.62) hold:

$$\frac{\tilde{A}(t)^2}{(1+t^2)^n} = A(\omega), \quad \frac{\tilde{B}(t)^2}{(1+t^2)^{n-1}} = B(\omega), \tag{1.64}$$

for some $A, B \in \mathbb{C}[z]$. Moreover, $A(\omega) \ge 0$, $B(\omega) \ge 0$ for any ω. By the spectral factorization Theorem 1.1, it results that $A(z) = F(z)F^*(z^{-1})$, $B(z) = G(z)G^*(z^{-1})$, with $\deg F \le n$, $\deg G \le n-1$.

Finally, we note that

$$\frac{(t-a)(b-t)}{(1+t^2)} = \frac{\frac{1-ab}{2}(1-t^2) + (a+b)t - \frac{ab+1}{2}(1+t^2)}{(1+t^2)} = D_{\alpha\beta}(\omega),$$

for $D_{\alpha\beta}(z)$ defined as in (1.36). It results that $R(z)$ has the expression (1.33).

1.5.6 Proof of Theorem 1.17

We prove only the case of even n, the other being similar. Using (1.6), we note that $P(t) \ge 0$, $\forall t \in [\cos\beta, \cos\alpha]$. From Theorem 1.11, it results that

$$P(t) = \tilde{A}(t)^2 + (t - \cos\beta)(\cos\alpha - t)\tilde{B}(t)^2, \tag{1.65}$$

with $\tilde{A}, \tilde{B} \in \mathbb{R}[t]$, $\deg \tilde{A} \le n/2$, $\deg \tilde{B} \le n/2 - 1$. Going back to the Chebyshev polynomials basis, there exist $A, B \in \overline{\mathbb{RP}}[z]$ such that

$$\tilde{A}(t)^2 = A(\omega), \quad \tilde{B}(t)^2 = B(\omega).$$

Using the spectral factorization Theorem 1.1, there exist $F, G \in \mathbb{R}_+[z]$, with $\deg F \le n$, $\deg G \le n-2$, such that $A(z) = F(z)F(z^{-1})$, $B(z) = G(z)G(z^{-1})$. Replacing these relations in (1.65), together with (1.5), leads to (1.37).

1.5.7 Proof of Theorem 1.18

The relation (1.39) results directly from (1.65), by putting $R_1(\omega) = \tilde{A}(t)$, $R_2(\omega) = \tilde{B}(t)$ (via the usual transformation $t = \cos\omega$).

1.6 Bibliographical and Historical Notes

The characterizations of real polynomials that are nonnegative on an interval, as in Theorem 1.11 and Theorem1.13, are attributed to Lukács [3] and date from the first decades of the twentieth century. A more advanced treatment and other references can be found in [4]. The interest in trigonometric polynomials is more recent, although results similar to Theorems 1.15, 1.17 were presented in [5]; in [6–9], they have been rediscovered or rediscussed and interpreted in connection with SDP.

Problems

P 1.1 Prove that if $R(z^{-1}) = R^*(z)$, then $R(\mathrm{e}^{j\omega}) \in \mathbb{R}$. Moreover, if $R(z)$ is a finite support polynomial, i.e., $R(z) = \sum_{k=-n_1}^{n_2} r_k z^{-k}$, then it has the form (1.1).

P 1.2 (Spectral factorization with respect to the imaginary axis.) Let $P(s) = \sum_{k=0}^{n} p_k s^k$ be a polynomial of complex variable, with $p_k \in \mathbb{C}$, such that $P(-s) = P^*(s)$. Show that $P(jt) \in \mathbb{R}$, for any $t \in \mathbb{R}$; moreover, $p_k \in \mathbb{R}$ if k is even and $p_k \in j\mathbb{R}$ if k is odd. If $P(jt) \geq 0$, $\forall t \in \mathbb{R}$, then n is even and $P(s) = F(s)F(-s)$, with $F(s) = \sum_{k=0}^{n/2} f_k s^k$.

P 1.3 ("Spectral factorization" with respect to the real axis.) Let $P(s) = \sum_{k=0}^{n} p_k s^k$, be a polynomial of complex variable, with $p_k \in \mathbb{C}$, such that $P(s) = P^*(s)$. Show that $p_k \in \mathbb{R}$ and so $P(t) \in \mathbb{R}$, for any $t \in \mathbb{R}$. If $P(t) \geq 0$, $\forall t \in \mathbb{R}$, then n is even and $P(s) = F^*(s)F(s)$, with $F(s) = \sum_{k=0}^{n/2} f_k s^k$.

Use this result to solve the previous problem.

P 1.4 Let $P \in \mathbb{R}_n[t]$ be such that $P(t) \geq 0$ for $t \in (-\infty, a] \cup [b, \infty)$. Prove that the polynomial can be expressed as

$$P(t) = \begin{cases} F(t)^2 - (t-a)(b-t)G(t)^2, & \text{if } n \text{ even}, \\ (a-t)F(t)^2 + (t-b)G(t)^2, & \text{if } n \text{ odd}, \end{cases}$$

with $F, G \in \mathbb{R}[t]$.

P 1.5 Show that if $R \in \mathbb{C}[z]$ is such that $R(\omega) \geq 0$ for $\omega \in [-\alpha, \alpha]$, then

$$R(z) = F(z)F^*(z^{-1}) + (\tfrac{z+z^{-1}}{2} - \cos\alpha) \cdot G(z)G^*(z^{-1}),$$

where $F, G \in \mathbb{C}_+[z]$, $\deg F \leq n$, $\deg G \leq n-1$. (Hint: this is a particular case of Theorem 1.15.)

P 1.6 Show that if $R \in \mathbb{C}[z]$ is such that $R(\omega) \geq 0$ for $\omega \in [\alpha, \pi]$, then

$$R(z) = F(z)F^*(z^{-1}) + D_\alpha(z) \cdot G(z)G^*(z^{-1}),$$

where $F, G \in \mathbb{C}_+[z]$, $\deg F \le n$, $\deg G \le n-1$ and

$$D_\alpha(z) = (-\tfrac{a}{4} + \tfrac{1}{4}j)z^{-1} - \tfrac{a}{2} + (-\tfrac{a}{4} - \tfrac{1}{4}j)z,$$

with $a = \tan(\alpha/2)$. (Hint: The coefficients of $D_\alpha(z)$ result by dividing with b in (1.36) and then putting $b \to \infty$.)

P 1.7 Show that if $R \in \mathbb{C}[z]$ is such that $R(\omega) \ge 0$ for $\omega \in [-\pi, \pi] \setminus (\alpha, \beta)$, then

$$R(z) = F(z)F^*(z^{-1}) - D_{\alpha\beta}(z) \cdot G(z)G^*(z^{-1}),$$

where $F, G \in \mathbb{C}_+[z]$, $\deg F \le n$, $\deg G \le n-1$ and $D_{\alpha\beta}(z)$ is defined as in Theorem 1.15.

References

1. P. Stoica, R.L. Moses, *Introduction to Spectral Analysis* (Prentice Hall, Upper Saddle River, 1997)
2. P. Stoica, T. McKelvey, J. Mari, MA estimation in polynomial time. IEEE Trans. Signal Process. **48**(7), 1999–2012 (2000)
3. G. Pólya, G. Szegö, *Problems and Theorems in Analysis II* (Springer, New York, 1976)
4. V. Powers, B. Reznick, Polynomials that are positive on an interval. Trans. Am. Math. Soc. **352**, 4677–4692 (2000)
5. M.G. Krein, A.A. Nudelman, *The Markov Moment Problem and Extremal Problems* (American Mathematical Society, Providence, 1977)
6. T.N. Davidson, Z.Q. Luo, J.F. Sturm, Linear matrix inequality formulation of spectral mask constraints with applications to FIR filter design. IEEE Trans. Signal Proc. **50**(11), 2702–2715 (2002)
7. B. Alkire, L. Vandenberghe, Convex optimization problems involving finite autocorrelation sequences. Math. Progr. Ser. A **93**(3), 331–359 (2002)
8. L. Faybusovich, On Nesterov's approach to semi-infinite programming. Acta Appl. Math. **74**, 195–215 (2002)
9. J.W. McLean, H.J. Woerdeman, Spectral factorizations and sums of squares representations via semidefinite programming. SIAM J. Matrix Anal. Appl. **23**(3), 646–655 (2002)

Chapter 2
Gram Matrix Representation

Abstract There are several ways of characterizing nonnegative polynomials that may be interesting for a mathematician. However, not all of them are appropriate for computational purposes, by "computational" understanding primarily optimization methods. Nonnegative polynomials have a basic property extremely useful in optimization: They form a convex set. So, an optimization problem whose variables are the coefficients of a nonnegative polynomial has a unique solution (or, in the degenerate case, multiple solutions belonging to a convex set), if the objective and the other constraints besides positivity are also convex. Convexity is not enough for obtaining efficiently a reliable solution. Efficiency and reliability are specific only to some classes of convex optimization, such as linear programming (LP), second-order cone problems (SOCP), and semidefinite programming (SDP). SDP includes LP and SOCP and is probably the most important advance in optimization in the last decade of the previous century. See some basic information on SDP in Appendix A. In this chapter, we present a parameterization of nonnegative polynomials that is intimately related to SDP. Each polynomial can be associated with a set of matrices, called *Gram* matrices (Choi et al., Proc Symp Pure Math 58:103–126, 1995, [1]); if the polynomial is nonnegative, then there is at least a positive semidefinite Gram matrix associated with it. Solving optimization problems with nonnegative polynomials may thus be reduced, in many cases, to SDP. We give several examples of such problems and of programs that solve them. Spectral factorization is important in this context, and we present several techniques for its computation. Besides the standard, or trace, parameterization, we discuss several other possibilities that may have computational advantages.

2.1 Parameterization of Trigonometric Polynomials

Let us start with some notations. The vector

$$\boldsymbol{\psi}_n(z) = [1 \; z \; z^2 \; \ldots \; z^n]^T \tag{2.1}$$

B. Dumitrescu, *Positive Trigonometric Polynomials and Signal Processing Applications*, Signals and Communication Technology,
DOI 10.1007/978-3-319-53688-0_2

contains the canonical basis for polynomials of degree n in z. Whenever the degree results from the context, we denote $\boldsymbol{\psi}(z)$ the vector from (2.1). A causal polynomial (1.10) can be written in the form $H(z) = \boldsymbol{h}^T \boldsymbol{\psi}(z^{-1})$, where $\boldsymbol{h} = [h_0\ h_1\ \ldots\ h_n]^T \in \mathbb{R}^{n+1}$ (or $\mathbb{C}^{n+1}$) is the vector of its coefficients. We use the notation $\boldsymbol{\psi}(\omega)$ for $\boldsymbol{\psi}(\mathrm{e}^{j\omega})$; remark that $\boldsymbol{\psi}^T(-\omega) = \boldsymbol{\psi}^H(\omega)$. Also, we denote $n' = n + 1$.

Definition 2.1 Consider the trigonometric polynomial $R \in \mathbb{C}_n[z]$, defined as in (1.1). A Hermitian matrix $\boldsymbol{Q} \in \mathbb{C}^{n' \times n'}$ is called a *Gram* matrix associated with $R(z)$ if

$$R(z) = \boldsymbol{\psi}^T(z^{-1}) \cdot \boldsymbol{Q} \cdot \boldsymbol{\psi}(z). \tag{2.2}$$

We denote $\mathcal{G}(R)$ the set of Gram matrices associated with $R(z)$. ■

If $R \in \mathbb{R}_n[z]$, then the matrix $\boldsymbol{Q}$ obeying to (2.2) belongs to $\mathbb{R}^{n' \times n'}$ and is symmetric.

Example 2.2 Let us consider polynomials of degree two with real coefficients, $R(z) = r_2 z^{-2} + r_1 z^{-1} + r_0 + r_1 z + r_2 z^2$. A few computations show that if

$$R(z) = [1\ z^{-1}\ z^{-2}] \begin{bmatrix} q_{00} & q_{10} & q_{20} \\ q_{10} & q_{11} & q_{21} \\ q_{20} & q_{21} & q_{22} \end{bmatrix} \begin{bmatrix} 1 \\ z \\ z^2 \end{bmatrix},$$

where $\boldsymbol{Q} \in \mathbb{R}^{3 \times 3}$ is a Gram matrix associated with $R(z)$, then

$$\begin{aligned} r_0 &= q_{00} + q_{11} + q_{22}, \\ r_1 &= q_{10} + q_{21}, \\ r_2 &= q_{20}. \end{aligned} \tag{2.3}$$

Hence, any Gram matrix associated with $R(z)$ has the form

$$\begin{aligned} \boldsymbol{Q} &= \begin{bmatrix} r_0 - q_{11} - q_{22} & r_1 - q_{21} & r_2 \\ r_1 - q_{21} & q_{11} & q_{21} \\ r_2 & q_{21} & q_{22} \end{bmatrix} \\ &= \begin{bmatrix} r_0 & r_1 & r_2 \\ r_1 & 0 & 0 \\ r_2 & 0 & 0 \end{bmatrix} + \begin{bmatrix} -q_{11} - q_{22} & -q_{21} & 0 \\ -q_{21} & q_{11} & q_{21} \\ 0 & q_{21} & q_{22} \end{bmatrix}. \end{aligned}$$

It is clear that, in general, any Hermitian matrix in $\mathbb{C}^{n' \times n'}$ produces a Hermitian polynomial through the mapping (2.2), which is many-to-one. For instance, taking

$$R(z) = 2z^{-2} - 3z^{-1} + 6 - 3z + 2z^2 = (2 - z^{-1} + z^{-2})(2 - z + z^2), \tag{2.4}$$

the following three matrices

$$\boldsymbol{Q}_0 = \begin{bmatrix} 6 & -3 & 2 \\ -3 & 0 & 0 \\ 2 & 0 & 0 \end{bmatrix}, \quad \boldsymbol{Q}_1 = \begin{bmatrix} 4 & -2 & 2 \\ -2 & 1 & -1 \\ 2 & -1 & 1 \end{bmatrix}, \\ \boldsymbol{Q}_2 = \begin{bmatrix} 2.2 & -1.5 & 2.0 \\ -1.5 & 1.6 & -1.5 \\ 2.0 & -1.5 & 2.2 \end{bmatrix} \tag{2.5}$$

are Gram matrices associated with $R(z)$. ■

A natural (and simple) question regards the relation between the coefficients of $R(z)$ and the elements of $\boldsymbol{Q} \in \mathcal{G}(R)$. From (2.3), we may infer that r_k is the sum of elements of $\boldsymbol{Q}$ along diagonal $-k$ (the main diagonal has number 0, and the lower triangle diagonals have negative numbers, as in MATLAB). This is indeed the case.

Theorem 2.3 *If $R \in \mathbb{C}_n[z]$ and $\boldsymbol{Q} \in \mathcal{G}(R)$, then the relation*

$$r_k = \operatorname{tr}[\boldsymbol{\Theta}_k \boldsymbol{Q}] = \sum_{i=\max(0,k)}^{\min(n+k,n)} q_{i,i-k}, \quad k = -n : n, \tag{2.6}$$

holds, where $\boldsymbol{\Theta}_k$ is the elementary Toeplitz matrix with ones on the k-th diagonal and zeros elsewhere and tr$\boldsymbol{X}$ *is the trace of the matrix $\boldsymbol{X}$. We name (2.6) the* trace *parameterization of the trigonometric polynomial $R(z)$.*

Proof We recall that $\operatorname{tr}[\boldsymbol{ABC}] = \operatorname{tr}[\boldsymbol{CAB}]$, where $\boldsymbol{A}$, $\boldsymbol{B}$, and $\boldsymbol{C}$ are matrices of appropriate sizes and also that $a = \operatorname{tr}[a]$, if a is a scalar. The relation (2.2) can be written as

$$R(z) = \boldsymbol{\psi}^T(z^{-1}) \cdot \boldsymbol{Q} \cdot \boldsymbol{\psi}(z) = \operatorname{tr}[\boldsymbol{\psi}(z) \cdot \boldsymbol{\psi}^T(z^{-1}) \cdot \boldsymbol{Q}] = \operatorname{tr}[\boldsymbol{\Psi}(z) \cdot \boldsymbol{Q}],$$

where

$$\boldsymbol{\Psi}(z) = \begin{bmatrix} 1 \\ z \\ \vdots \\ z^n \end{bmatrix} [1\ z^{-1}\ \ldots\ z^{-n}] = \begin{bmatrix} 1 & z^{-1} & \ldots & z^{-n} \\ z & 1 & \ddots & z^{-n+1} \\ \vdots & \ddots & \ddots & \vdots \\ z^n & z^{n-1} & \ldots & 1 \end{bmatrix} = \sum_{k=-n}^{n} \boldsymbol{\Theta}_k z^{-k}. \tag{2.7}$$

Combining the above two relations, we obtain

$$R(z) = \sum_{k=-n}^{n} \operatorname{tr}[\boldsymbol{\Theta}_k \boldsymbol{Q}] z^{-k},$$

which proves (2.6) after identification with (1.1). ■

Example 2.4 For a polynomial of degree 2, as in Example 2.2, the trace parameterization (2.6) tells that

$$r_0 = \operatorname{tr} \boldsymbol{Q}, \quad r_1 = \operatorname{tr} \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{bmatrix} \boldsymbol{Q}, \quad r_2 = \operatorname{tr} \begin{bmatrix} 0 & 0 & 1 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} \boldsymbol{Q}.$$

We are now ready to state the main result, namely that the set $\mathbb{C}_n[z]$ of nonnegative polynomials (of order n) and the set of positive semidefinite matrices of size $n' \times n'$ are connected by the trace parameterization mapping (2.6).

Theorem 2.5 *A polynomial $R \in \mathbb{C}_n[z]$ is nonnegative (positive) on the unit circle if and only if there exists a positive semidefinite (definite) matrix $\boldsymbol{Q} \in \mathbb{C}^{n' \times n'}$ such that (2.6) holds.*

Proof If $\boldsymbol{Q} \succeq 0$ exists such that (2.6) holds, then, using the definition (2.2) of a Gram matrix, we can write

$$R(\omega) = [1 \ \mathrm{e}^{-j\omega} \ \ldots \mathrm{e}^{-jn\omega}] \cdot \boldsymbol{Q} \cdot \begin{bmatrix} 1 \\ \mathrm{e}^{j\omega} \\ \vdots \\ \mathrm{e}^{jn\omega} \end{bmatrix} = \boldsymbol{\psi}^H(\omega) \cdot \boldsymbol{Q} \cdot \boldsymbol{\psi}(\omega) \geq 0,$$

for all ω. The same reasoning shows that if $\boldsymbol{Q} \succ 0$, then $R(\omega) > 0$.

Reciprocally, if $R(\omega) \geq 0$, then the spectral factorization Theorem 1.1 says that

$$R(z) = H(z)H^*(z^{-1}) = \boldsymbol{h}^T \boldsymbol{\psi}(z^{-1}) \cdot \boldsymbol{h}^H \boldsymbol{\psi}(z) = \boldsymbol{\psi}^T(z^{-1}) \cdot \boldsymbol{h}\boldsymbol{h}^H \cdot \boldsymbol{\psi}(z).$$

It results that

$$\boldsymbol{Q}_1 = \boldsymbol{h}\boldsymbol{h}^H \succeq 0 \tag{2.8}$$

is a Gram matrix associated with $R(z)$. Note that $\operatorname{rank} \boldsymbol{Q}_1 = 1$.

If $R(\omega) > 0$, since $[-\pi, \pi]$ is compact, there exists $\varepsilon > 0$ such that $R_\varepsilon(z) = R(z) - \varepsilon$ is nonnegative. Denoting $H_\varepsilon(z)$ a spectral factor of $R_\varepsilon(z)$ and noticing that $\boldsymbol{\psi}^T(z^{-1}) \cdot \boldsymbol{\psi}(z) = n'$, it results as above that

$$R(z) = \boldsymbol{\psi}^T(z^{-1}) \cdot \left(\boldsymbol{h}\boldsymbol{h}^H + (\varepsilon/n')\boldsymbol{I}\right) \cdot \boldsymbol{\psi}(z)$$

and so $\boldsymbol{h}\boldsymbol{h}^H + (\varepsilon/n')\boldsymbol{I} \succ 0$ is a Gram matrix associated with $R(z)$. ■

Example 2.6 Returning to Example 2.2 and the three Gram matrices from (2.5), we notice that $\boldsymbol{Q}_1$ is defined as in (2.8) and so is positive semidefinite of rank 1 and also $\boldsymbol{Q}_2 \succ 0$. We conclude that $R(\omega) > 0$. The fact that a Gram matrix, in our case $\boldsymbol{Q}_0$, is not definite has no consequence on the positivity of $R(z)$. ■

Remark 2.7 Theorem 2.5 establishes a linear relation between the elements of two convex sets: nonnegative polynomials and positive semidefinite matrices. On one side, we have the usual parameterization of $\overline{\mathbb{CP}}_n[z]$ using the $n + 1$ coefficients of the polynomial $R(z)$. On the other side, we have an overparameterization, using the

$n(n+1)/2$ independent elements of the Gram matrix $\boldsymbol{Q}$. The high number of parameters of the latter is compensated by the reliability and efficiency of optimization algorithms dealing with linear combinations of positive semidefinite matrices, which belong to the class of semidefinite programming. ■

Remark 2.8 If the polynomial has complex coefficients, Theorem 2.5 can be formulated in terms of real matrices, as follows. The polynomial $R(z)$ is nonnegative if and only if there exist matrices $\boldsymbol{Q}_r, \boldsymbol{Q}_i \in \mathbb{R}^{n' \times n'}$ such that

$$\mathrm{Re} r_k = \mathrm{tr}[\boldsymbol{\Theta}_k \boldsymbol{Q}_r], \quad \mathrm{Im} r_k = \mathrm{tr}[\boldsymbol{\Theta}_k \boldsymbol{Q}_i] \tag{2.9}$$

and

$$\begin{bmatrix} \boldsymbol{Q}_r & -\boldsymbol{Q}_i \\ \boldsymbol{Q}_i & \boldsymbol{Q}_r \end{bmatrix} \succeq 0. \tag{2.10}$$

Indeed, putting $\boldsymbol{Q} = \boldsymbol{Q}_r + j\boldsymbol{Q}_i$, relation (2.10) is equivalent to $\boldsymbol{Q} \succeq 0$ and relation (2.9) is equivalent to (2.6). (Such a formulation might be useful when using SDP algorithms that work only with real matrices. However, all important SDP libraries are able to deal with complex matrices.) A more efficient way to parameterize complex polynomials with real Gram matrices will be presented in Sect. 2.8.1. ■

Remark 2.9 (Sum-of-squares decomposition) Let $R \in \mathbb{C}_n[z]$ be a nonnegative trigonometric polynomial, and let $\boldsymbol{Q} \succeq 0$ be a positive semidefinite Gram matrix associated with it. A distinct sum-of-squares decomposition (1.23) of $R(z)$ can be derived from each such Gram matrix. Let

$$\boldsymbol{Q} = \sum_{\ell=1}^{\nu} \lambda_\ell^2 \boldsymbol{x}_\ell \boldsymbol{x}_\ell^H, \tag{2.11}$$

be the eigendecomposition of $\boldsymbol{Q}$, in which ν is the rank, λ_ℓ^2 the eigenvalues, and $\boldsymbol{x}_\ell$ the eigenvectors, $\ell = 1 : \nu$. Inserting (2.11) into (2.2), we obtain the sum-of-squares decomposition

$$R(z) = \sum_{\ell=1}^{\nu} [\lambda_\ell \boldsymbol{\psi}^T(z^{-1}) \boldsymbol{x}_\ell] \cdot [\lambda_\ell \boldsymbol{x}_\ell^H \boldsymbol{\psi}(z)] = \sum_{\ell=1}^{\nu} H_\ell(z) H_\ell^*(z^{-1}), \tag{2.12}$$

where

$$H_\ell(z) = \lambda_\ell \boldsymbol{\psi}^T(z^{-1}) \boldsymbol{x}_\ell. \tag{2.13}$$

So, the sum-of-squares (2.12) has a number of terms equal to the rank of the Gram matrix $\boldsymbol{Q}$. If the Gram matrix is $\boldsymbol{Q}_1$ from (2.8), then the spectral factorization (1.11) is obtained. ■

Remark 2.10 (Toeplitz Gram matrices) There is a single Toeplitz Gram matrix of size $n' \times n'$ associated with a given polynomial $R(z)$ of degree n, namely

$\boldsymbol{Q} = \text{Toep}(r_0/(n+1), r_1/n, \ldots, r_n)$. If $R(\omega) \geq 0$, is this matrix positive semidefinite? For example, for the positive polynomial (2.4), this matrix is

$$\boldsymbol{Q} = \begin{bmatrix} 2.0 & -1.5 & 2.0 \\ -1.5 & 2.0 & -1.5 \\ 2.0 & -1.5 & 2.0 \end{bmatrix}$$

and is positive semidefinite (and singular). Modifying r_2 to e.g., 2.001 keeps the polynomial positive, but the Toeplitz Gram matrix is no more positive semidefinite. So, in general, there is no connection between the nonnegativity of the polynomial and the positive semidefiniteness of the Toeplitz Gram matrix.

However, we can show that, for any nonnegative $R(z)$, there is an arbitrarily close $\tilde{R}(z)$ for which the Toeplitz Gram matrix is positive semidefinite. The trick is to remove the size restrictions. We can artificially consider the degree of $R(z)$ to be $m > n$, by adding coefficients $r_k = 0$, $k = n+1 : m$. Remember now Theorem 1.8, which states that the Toeplitz matrices $\boldsymbol{R}_m$ defined in (1.28) are positive semidefinite. For any m, the polynomial $\tilde{R}(z)$, with coefficients defined by

$$\tilde{r}_k = \text{tr}\left[\boldsymbol{\Theta}_k \cdot \frac{1}{m+1}\boldsymbol{R}_m\right] = \left(1 - \frac{|k|}{m+1}\right) r_k,$$

is thus nonnegative and the Gram matrix $\boldsymbol{R}_m/(m+1)$ is Toeplitz and positive semidefinite. For large enough m, the polynomial $\tilde{R}(z)$ is arbitrarily close to $R(z)$. ■

2.2 Optimization Using the Trace Parameterization

We present now some simple problems that can be solved using the trace parameterization (2.6) and SDP.

Let us notice first that, given the polynomial $R(z)$, the set $\mathcal{G}(R)$ is convex. Indeed, for any $\alpha \in [0, 1]$ and $\boldsymbol{Q}$, $\tilde{\boldsymbol{Q}} \in \mathcal{G}(R)$, it is immediate from (2.2) that $\alpha\boldsymbol{Q} + (1-\alpha)\tilde{\boldsymbol{Q}} \in \mathcal{G}(R)$. Moreover, if $R(\omega) \geq 0$, then the set of positive semidefinite Gram matrices associated with $R(z)$ is also convex, as the intersection of two convex sets.

Problem (*Most_positive_Gram_matrix*) It is clear that, given $R \in \mathbb{C}_n[z]$ with $R(\omega) > 0$, there are an infinite number of positive definite Gram matrices in $\mathcal{G}(R)$. This results, for example, by taking all possible values for the parameter ε appearing at the end of the proof of Theorem 2.5. A distinguished member of $\mathcal{G}(R)$ is the most positive one, i.e., the most nonsingular. The distance to nonsingularity is measured by the smallest singular value, or, as we deal with positive definite matrices, by the smallest eigenvalue. So, we want the matrix in $\mathcal{G}(R)$ having the largest smallest eigenvalue, namely the solution of the optimization problem

$$\begin{aligned}\lambda^\star = \max_{\lambda, \boldsymbol{Q}} \ & \lambda \\ \text{s.t.} \ & \operatorname{tr}[\boldsymbol{\Theta}_k \boldsymbol{Q}] = r_k, \quad k = 0 : n \\ & \lambda \geq 0, \quad \boldsymbol{Q} \succeq \lambda \boldsymbol{I}\end{aligned} \tag{2.14}$$

The inequality $\boldsymbol{Q} \succeq \lambda \boldsymbol{I}$ ensures that maximization of λ is equivalent to maximization of the smallest eigenvalue of the Gram matrix $\boldsymbol{Q}$. (We always assume that the Gram matrices are Hermitian and will not specify it explicitly from now on.) The optimization problem (2.14) is a semidefinite program, since the variables are the positive semidefinite matrix $\boldsymbol{Q}$ and the positive scalar λ and the constraints are linear equalities in the elements of $\boldsymbol{Q}$.

Although in evolved optimization problem solvers such as CVX the problem (2.14) can be posed as it is, some popular SDP libraries need the problem in the standard equality form shown in Appendix A. The transformation to standard form can be made by denoting $\tilde{\boldsymbol{Q}} = \boldsymbol{Q} - \lambda \boldsymbol{I}$; as the matrix $\tilde{\boldsymbol{Q}}$ is positive semidefinite, it can serve as variable in the SDP problem. Since $\boldsymbol{\Theta}_0 = \boldsymbol{I}$, it follows that $\operatorname{tr}\boldsymbol{\Theta}_0 = n'$, while for $k \neq 0$ we have $\operatorname{tr}\boldsymbol{\Theta}_k = 0$. We conclude that

$$\operatorname{tr}[\boldsymbol{\Theta}_k \tilde{\boldsymbol{Q}}] = \begin{cases} r_0 - n'\lambda, & \text{if } k = 0, \\ r_k, & \text{otherwise.} \end{cases}$$

Thus, the problem (2.14) is equivalent to the standard SDP problem

$$\begin{aligned}\lambda^\star = \max_{\lambda, \tilde{\boldsymbol{Q}}} \ & \lambda \\ \text{s.t.} \ & n'\lambda + \operatorname{tr}\tilde{\boldsymbol{Q}} = r_0 \\ & \operatorname{tr}[\boldsymbol{\Theta}_k \tilde{\boldsymbol{Q}}] = r_k, \quad k = 1 : n \\ & \lambda \geq 0, \quad \tilde{\boldsymbol{Q}} \succeq 0\end{aligned} \tag{2.15}$$

We finally note that the SDP problem (2.14) or (2.15) gives no solution if the polynomial $R(z)$ is not nonnegative, as no positive semidefinite Gram matrix exists. However, the problem of finding the Gram matrix with maximum smallest eigenvalue is well defined; removing the constraint $\lambda \geq 0$ in (2.14) or (2.15) and thus leaving λ free is the single necessary modification. ■

Example 2.11 In Example 2.2, none of the three Gram matrices from (2.5) is the most positive one. Solving the SDP problem (2.14), we obtain

$$\boldsymbol{Q} = \begin{bmatrix} 2.2917 & -1.5000 & 2.0000 \\ -1.5000 & 1.4167 & -1.5000 \\ 2.0000 & -1.5000 & 2.2917 \end{bmatrix}.$$

The smallest eigenvalue of this matrix is $\lambda^\star = 0.2917$. For comparison, the smallest eigenvalue of the matrix $\boldsymbol{Q}_2$ from (2.5) is 0.2. ■

Problem (*Min_poly_value*) A problem related to that of finding the most positive Gram matrix is to compute the minimum value on the unit circle of a given trigonometric polynomial $R(z)$. Let $R \in \mathbb{C}_n[z]$ be a polynomial not necessarily nonnegative. We want to find

$$\mu^\star = \min_{\omega \in [-\pi,\pi]} R(\omega). \tag{2.16}$$

Certainly, this is a problem that can be solved using elementary tools, so the solution given below may not be the most efficient, although it is instructive in the current context. Since $\mu^\star$ is the maximum scalar for which $R(\omega) - \mu^\star$ is nonnegative, we can connect (2.16) to nonnegative polynomials by transforming it into

$$\begin{aligned} \mu^\star = \ & \max_{\mu} \ \mu \\ & \text{s.t.} \ \ R(\omega) - \mu \ge 0, \quad \forall \omega \in [-\pi, \pi] \end{aligned} \tag{2.17}$$

Denoting $\tilde{R}(z) = R(z) - \mu$, we note that $\tilde{r}_0 = r_0 - \mu$ and $\tilde{r}_k = r_k$, $k = 1 : n$. Using the trace parameterization (and reminding again that $\boldsymbol{\Theta}_0 = \boldsymbol{I}$), the problem (2.17) can brought to the following SDP form

$$\begin{aligned} \mu^\star = \ & \max_{\mu, \tilde{\boldsymbol{Q}}} \ \mu \\ & \text{s.t.} \ \ \mu + \operatorname{tr} \tilde{\boldsymbol{Q}} = r_0 \\ & \qquad \operatorname{tr}[\boldsymbol{\Theta}_k \tilde{\boldsymbol{Q}}] = r_k, \quad k = 1 : n \\ & \qquad \tilde{\boldsymbol{Q}} \succeq 0 \end{aligned} \tag{2.18}$$

If $R(\omega) \ge 0$, the SDP problems (2.14) and (2.18) are equivalent, expressing the connection between the most positive Gram matrix and the minimum value of a polynomial. The equivalence is shown by the relations

$$\mu = n'\lambda, \quad \tilde{\boldsymbol{Q}} = \boldsymbol{Q} - \lambda \boldsymbol{I} \tag{2.19}$$

between the variables of the two problems, which are obvious if we look at the (2.15), which is equivalent to (2.14) and becomes identical to (2.18) by taking $\mu = n'\lambda$. So, the optimal values of (2.14) and (2.18) are related by $\mu^\star = n'\lambda^\star$, and solving one problem leads immediately to the solution of the other through (2.19).

SeDuMi, CVX, and Pos3Poly programs for solving the SDP problems (2.14) and (2.18) are presented and commented in Sect. 2.12.1. ■

Example 2.12 The minimum value on the unit circle of the polynomial (2.4) (of degree $n = 2$) considered in Example 2.2 is $\mu^\star = 3\lambda^\star = 3 \cdot 0.2917 = 0.8750$. ■

Problem (*Nearest_autocorrelation*) We return to a problem discussed in the previous chapter: Given a symmetric (or Hermitian) sequence $\hat{r}_k$, $k = -n : n$, find the nonnegative sequence r_k that is nearest from $\hat{r}_k$. The optimization problem to be solved is (1.22); remind that $\boldsymbol{r} = [r_0 \ r_1 \ \ldots \ r_n]^T$. Expressing the nonnegativity condition with the trace parameterization, we obtain the problem

$$\begin{aligned} \min_{\boldsymbol{r},\boldsymbol{Q}} \ & (\boldsymbol{r}-\hat{\boldsymbol{r}})^H \boldsymbol{\Gamma} (\boldsymbol{r}-\hat{\boldsymbol{r}}) \\ \text{s.t.} \ & \operatorname{tr}[\boldsymbol{\Theta}_k \boldsymbol{Q}] = r_k, \quad k = 0:n \\ & \boldsymbol{Q} \succeq 0 \end{aligned} \tag{2.20}$$

where $\boldsymbol{\Gamma} \succ 0$. To bring (2.20) to a standard form, notice that

$$(\boldsymbol{r}-\hat{\boldsymbol{r}})^H \boldsymbol{\Gamma} (\boldsymbol{r}-\hat{\boldsymbol{r}}) = \|\boldsymbol{\Gamma}^{1/2}(\boldsymbol{r}-\hat{\boldsymbol{r}})\|^2, \tag{2.21}$$

where $\boldsymbol{\Gamma}^{1/2}$ is the square root of $\boldsymbol{\Gamma}$, i.e., the positive definite matrix $\boldsymbol{X}$ such that $\boldsymbol{X}^H\boldsymbol{X} = \boldsymbol{\Gamma}$. An alternative possibility in (2.21) is to use the Cholesky factor of $\boldsymbol{\Gamma}$ instead of $\boldsymbol{\Gamma}^{1/2}$. Using the same trick as in passing from (2.16) to (2.17), we obtain

$$\begin{aligned} \min_{\alpha,\boldsymbol{r},\boldsymbol{Q}} \ & \alpha \\ \text{s.t.} \ & \|\boldsymbol{\Gamma}^{1/2}(\boldsymbol{r}-\hat{\boldsymbol{r}})\| \le \alpha \\ & \operatorname{tr}[\boldsymbol{\Theta}_k \boldsymbol{Q}] = r_k, \quad k = 0:n \\ & \boldsymbol{Q} \succeq 0 \end{aligned} \tag{2.22}$$

The first constraint has a second-order cone form, so (2.22) is a semidefinite-quadratic-linear programming (SQLP) problem. We have only to bring it to one of the standard forms shown in Appendix A. To this purpose, denote

$$\boldsymbol{y} = \boldsymbol{\Gamma}^{1/2}(\boldsymbol{r}-\hat{\boldsymbol{r}})$$

and, in $\boldsymbol{r} - \boldsymbol{\Gamma}^{-1/2}\boldsymbol{y} = \hat{\boldsymbol{r}}$, replace $\boldsymbol{r}$ by its trace parameterization. So, the problem (2.22) is equivalent to

$$\begin{aligned} \min_{\alpha,\boldsymbol{y},\boldsymbol{Q}} \ & \alpha \\ \text{s.t.} \ & \begin{bmatrix} \vdots \\ \operatorname{tr}[\boldsymbol{\Theta}_k \boldsymbol{Q}] \\ \vdots \end{bmatrix} - \boldsymbol{\Gamma}^{-1/2}\boldsymbol{y} = \hat{\boldsymbol{r}} \\ & \boldsymbol{Q} \succeq 0, \quad \|\boldsymbol{y}\| \le \alpha \end{aligned} \tag{2.23}$$

This is a standard SQLP problem in equality form. ■

Remark 2.13 (Complexity issues) As discussed in Appendix A, the complexity of an SDP problem in equality form is $O(n^2m^2)$, where $n \times n$ is the size of the variable positive semidefinite matrix and m is the number of equality constraints. The scalar or SOC variables (from (2.14) and (2.23), respectively) do not change significantly the complexity. Since the size of the Gram matrix $\boldsymbol{Q}$ is $(n+1) \times (n+1)$ and the number of equality constraints is $n+1$, we can appreciate that the complexity of the three problems—*Most_positive_Gram_matrix*, *Min_poly_value* and *Nearest_autocorrelation*—formulated in SDP form in this section is $O(n^4)$. ■

2.3 Toeplitz Quadratic Optimization

In the previous section, we have presented several optimization problems in which the variable was genuinely a nonnegative polynomial. Here, we discuss a problem that can be transformed—in a general way—into one with nonnegative polynomials. The idea is to replace the variable causal polynomial $H(z)$ with $R(z) = H(z)H^*(z^{-1})$ (i.e., with its squared magnitude on the unit circle), solve the presumably easier problem with $R(z)$ as variable, and finally recover $H(z)$ by spectral factorization. We have already met a somewhat similar problem, namely *MA_Estimation* in Sect. 1.2.

Consider the quadratic optimization problem

$$\begin{aligned} \min_{\boldsymbol{h}} \ & \boldsymbol{h}^H \boldsymbol{A}_0 \boldsymbol{h} \\ \text{s.t.} \ & \boldsymbol{h}^H \boldsymbol{A}_\ell \boldsymbol{h} = b_\ell, \ \ell = 1 : L \end{aligned} \tag{2.24}$$

where the matrices $\boldsymbol{A}_\ell$, $\ell = 0 : L$, and the scalars b_ℓ, $\ell = 1 : L$, are given. The matrix $\boldsymbol{A}_0$ is positive semidefinite. The variable is the vector $\boldsymbol{h} \in \mathbb{C}^{n+1}$; we can interpret its elements as the coefficients of the causal filter (1.10). Although the objective function is convex, the problem (2.24) is not convex, in general; a notorious exception occurs when $L = 1$, $\boldsymbol{A}_1 = \boldsymbol{I}$, for which the solution is an eigenvector of $\boldsymbol{A}_0$ corresponding to the minimal eigenvalue. We treat here only the case where all the matrices $\boldsymbol{A}_\ell$ are *Toeplitz* and Hermitian; that the matrices are Hermitian is not a particularization, due to the quadratic form of the objective and constraints; for an anti-Hermitian matrix $\boldsymbol{A}$ (with $\boldsymbol{A}^H = -\boldsymbol{A}$), the quadratic function is $\boldsymbol{h}^H\boldsymbol{A}\boldsymbol{h} = 0$; so, if the matrices $\boldsymbol{A}_\ell$ were not Hermitian, they could be replaced with their Hermitian part $(\boldsymbol{A} + \boldsymbol{A}^H)/2$. We note also that if the matrices $\boldsymbol{A}_\ell$, $\ell = 1 : L$, are real, the equality constraints from (2.24) can be changed into inequalities without changing the character of the solution presented below.

We denote

$$\boldsymbol{A}_\ell = \text{Toep}(a_{\ell 0}, \ldots, a_{\ell n}) \tag{2.25}$$

and notice that

$$\boldsymbol{A}_\ell = a_{\ell 0}\boldsymbol{\Theta}_0 + \sum_{k=1}^{n}(a_{\ell k}\boldsymbol{\Theta}_k + a_{\ell k}^*\boldsymbol{\Theta}_{-k}). \tag{2.26}$$

If we consider $H(z)$ as the spectral factor of a nonnegative polynomial $R(z)$, i.e., relation (1.11) holds, then the coefficients of $H(z)$ and $R(z)$ are related through (1.17), which is equivalent to

$$r_k = \boldsymbol{h}^H\boldsymbol{\Theta}_k\boldsymbol{h}. \tag{2.27}$$

From (2.26) and (2.27), it results that

$$\boldsymbol{h}^H\boldsymbol{A}_\ell\boldsymbol{h} = r_0 + \sum_{k=1}^{n}(a_{\ell k}r_k + a_{\ell k}^* r_k^*) = r_0 + 2\sum_{k=1}^{n}\text{Re}(a_{\ell k}r_k). \tag{2.28}$$

So, the problem (2.24) can be transformed into

$$\begin{aligned} \min_{\boldsymbol{r}} \ & r_0 + 2\textstyle\sum_{k=1}^{n} \operatorname{Re}(a_{0k} r_k) \\ \text{s.t.} \ & r_0 + 2\textstyle\sum_{k=1}^{n} \operatorname{Re}(a_{\ell k} r_k) = b_\ell, \ \ell = 1 : L \\ & R(\omega) \geq 0, \ \forall \omega \in [-\pi, \pi] \end{aligned} \tag{2.29}$$

This is a convex optimization problem! The variables are the coefficients of a nonnegative polynomial, and the quadratic objective and constraints from (2.24) are now linear. We can use the trace parameterization (2.6) to transform (2.29) into an SDP problem. Inserting $r_k = \operatorname{tr}[\boldsymbol{\Theta}_k \boldsymbol{Q}]$ into (2.26), we obtain

$$\boldsymbol{h}^H \boldsymbol{A}_\ell \boldsymbol{h} = \operatorname{tr}\left[a_{\ell 0} \boldsymbol{\Theta}_0 \boldsymbol{Q} + \sum_{k=1}^{n} (a_{\ell k} \boldsymbol{\Theta}_k + a_{\ell k}^* \boldsymbol{\Theta}_{-k}) \boldsymbol{Q} \right] = \operatorname{tr}[\boldsymbol{A}_\ell \boldsymbol{Q}]. \tag{2.30}$$

Using this equality, the problem (2.29) is equivalent to the SDP problem

$$\begin{aligned} \min_{\boldsymbol{Q}} \ & \operatorname{tr}[\boldsymbol{A}_0 \boldsymbol{Q}] \\ \text{s.t.} \ & \operatorname{tr}[\boldsymbol{A}_\ell \boldsymbol{Q}] = b_\ell, \ \ell = 1 : L \\ & \boldsymbol{Q} \succeq 0 \end{aligned} \tag{2.31}$$

We conclude that the solution of the Toeplitz quadratic optimization problem (2.24) can be obtained as follows:

1. Solve the SDP problem (2.31) for the positive semidefinite matrix $\boldsymbol{Q}$.
2. Compute $R(z)$ with (2.6): $r_k = \operatorname{tr}[\boldsymbol{\Theta}_k \boldsymbol{Q}]$.
3. Obtain $\boldsymbol{h}$ from the spectral factorization of $R(z)$.

It is clear from the above method that any spectral factor of $R(z)$ is a solution to (2.24). Spectral factorization algorithms compute usually (and reliably) only the minimum-phase (or maximum-phase) factor. This might be the only drawback of the method; however, in signal processing applications, the minimum-phase spectral factor is often the desired one.

Examples of problems of the form (2.24) and interpretations of their solutions will be given in Chap. 6.

2.4 Duality

As mentioned in Remark 1.5, the set $\overline{\mathbb{RP}}_n[z]$ of nonnegative trigonometric polynomials of degree n is a cone. Due to the interest in optimization problems, we naturally look at the *dual* cone, defined by

$$\overline{\mathbb{RP}}_n^\star[z] = \{\boldsymbol{y} \in \mathbb{R}^{n+1} \mid \boldsymbol{y}^T \boldsymbol{r} \geq 0, \ \forall R \in \overline{\mathbb{RP}}_n[z]\}. \tag{2.32}$$

Theorem 2.14 *The dual cone (2.32) is the space of sequences $\boldsymbol{y} \in \mathbb{R}^{n+1}$ for which Toep$(2y_0, y_1, \ldots, y_n) \succeq 0$. (In other words, the dual cone can be identified with the space of positive semidefinite Toeplitz matrices.)*

Proof Since the polynomial $R(z)$ is nonnegative, it admits a spectral factorization (1.11), relation (1.17) holds and we can write

$$\boldsymbol{y}^T\boldsymbol{r} = \sum_{k=0}^{n} y_k \sum_{i=k}^{n} h_i h_{i-k} = \frac{1}{2}\boldsymbol{h}^T \begin{bmatrix} 2y_0 & y_1 & \ldots & y_n \\ y_1 & 2y_0 & \ddots & y_{n-1} \\ \vdots & \ddots & \ddots & \vdots \\ y_n & y_{n-1} & \ldots & 2y_0 \end{bmatrix} \boldsymbol{h}.$$

Since the above quadratic form is nonnegative for all $\boldsymbol{h} \in \mathbb{R}^{n+1}$, it follows that the matrix Toep$(2y_0, y_1, \ldots, y_n)$ is positive semidefinite. ■

Knowing the form of the dual cone, we can build easier the duals of optimization problems with nonnegative trigonometric polynomials. Let us consider the problem (1.22), where, for simplicity, we take $\boldsymbol{\Gamma} = \boldsymbol{I}$. The function dual to

$$f(\boldsymbol{r}) \overset{\Delta}{=} (\boldsymbol{r} - \hat{\boldsymbol{r}})^T(\boldsymbol{r} - \hat{\boldsymbol{r}}) \tag{2.33}$$

is

$$g(\boldsymbol{y}) = \inf_{\boldsymbol{r}} \left[f(\boldsymbol{r}) - \boldsymbol{y}^T\boldsymbol{r} \right],$$

where the Lagrangean multiplier $\boldsymbol{y}$ belongs to the dual cone. The minimum is obtained trivially for $\boldsymbol{y} = 2(\boldsymbol{r} - \hat{\boldsymbol{r}})$ and so

$$g(\boldsymbol{y}) = -\frac{1}{4}\boldsymbol{y}^T\boldsymbol{y} - \boldsymbol{y}^T\hat{\boldsymbol{r}}.$$

The optimization problem dual to (1.22) is

$$\begin{array}{lll} \max_{\boldsymbol{y}} \ g(\boldsymbol{y}) & \Longleftrightarrow & \min_{\boldsymbol{y}} \ \frac{1}{4}\boldsymbol{y}^T\boldsymbol{y} + \boldsymbol{y}^T\hat{\boldsymbol{r}} \\ \text{s.t.} \ \ \boldsymbol{y} \in \overline{\mathbb{RP}}_n^{\star}[z] & & \text{s.t.} \ \ \text{Toep}(2y_0, y_1, \ldots, y_n) \succeq 0 \end{array} \tag{2.34}$$

and is (as all duals are) a convex problem. Since (1.22) is convex and the Slater condition holds (which translates to the mere existence of strictly positive polynomials), it follows that the problems (1.22) and (2.34) have the same optimal value.

Moreover, we see immediately that (2.34) is an SDP problem (more precisely, it can be written as an SQLP one) since its constraint is the positive semidefinite matrix

$$\boldsymbol{Y} = \text{Toep}(2y_0, y_1, \ldots, y_n) = 2y_0\boldsymbol{I} + \sum_{k=1}^{n} y_k(\boldsymbol{\Theta}_k + \boldsymbol{\Theta}_k^T) \tag{2.35}$$

that depends linearly on the variables y_k.

We can now derive the dual of (2.34), using the scalar product specific to the space of positive semidefinite matrices, when building the Lagrangean function. The new primal (i.e., dual of the dual) function is

$$\tilde{f}(\boldsymbol{Q}) = \inf_{\boldsymbol{y}} \left(-g(\boldsymbol{y}) - \operatorname{tr}[\boldsymbol{QY}]\right), \tag{2.36}$$

where the Lagrangean multiplier is $\boldsymbol{Q} \succeq 0$. We note that, due to (2.35), we have

$$\frac{\partial \operatorname{tr}[\boldsymbol{QY}]}{\partial y_k} = \operatorname{tr}[(\boldsymbol{\Theta}_k + \boldsymbol{\Theta}_k^T)\boldsymbol{Q}] = 2\operatorname{tr}[\boldsymbol{\Theta}_k \boldsymbol{Q}].$$

Since the function to be minimized in (2.36) is quadratic, the minimum is obtained by equating its derivative with zero, giving

$$\frac{1}{2}\boldsymbol{y} = 2\begin{bmatrix} \vdots \\ \operatorname{tr}[\boldsymbol{\Theta}_k \boldsymbol{Q}] \\ \vdots \end{bmatrix} - \hat{\boldsymbol{r}}.$$

The dual of (2.34) is identical to (1.22), for $\boldsymbol{r} \in \mathbb{R}^{n+1}$ given by

$$r_k = 2\operatorname{tr}[\boldsymbol{\Theta}_k \boldsymbol{Q}].$$

Barring an insignificant factor of 2 that can be included in $\boldsymbol{Q}$, we have obtained again the trace parameterization of nonnegative polynomials, stated by Theorem 2.5. Although this is not a complete proof, it is an instructive result on how the Lagrangean duality mechanism can be used.

2.5 Kalman–Yakubovich–Popov Lemma

We show here that the trace parameterization (2.6) can be derived from the Kalman–Yakubovich–Popov (KYP) lemma. Consider a discrete-time system with transfer function $G(z) = \boldsymbol{C}(z\boldsymbol{I} - \boldsymbol{A})^{-1}\boldsymbol{B}$, where $(\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C}, \boldsymbol{D})$ is a state-space model. Assume that the state-space representation is minimal (or $(\boldsymbol{A}, \boldsymbol{B})$ is controllable, $(\boldsymbol{C}, \boldsymbol{A})$ is observable). The KYP lemma states that the system is positive real, i.e.,

$$\operatorname{Re}[G(\omega)] \geq 0, \ \forall \omega \in [-\pi, \pi],$$

if and only if there exists a matrix $\boldsymbol{P} \succeq 0$ such that

$$\boldsymbol{Q} = \begin{bmatrix} \boldsymbol{P} - \boldsymbol{A}^T \boldsymbol{P} \boldsymbol{A} & \text{sym} \\ \boldsymbol{C} - \boldsymbol{B}^T \boldsymbol{P} \boldsymbol{A} & (\boldsymbol{D} + \boldsymbol{D}^T) - \boldsymbol{B}^T \boldsymbol{P} \boldsymbol{B} \end{bmatrix} \succeq 0. \tag{2.37}$$

The causal part $R_+(z)$ of a nonnegative trigonometric polynomial, defined in (1.2), is positive real. Its controllable state-space realization is

$$\begin{gathered} \boldsymbol{A} = \boldsymbol{\Theta}_1 = \begin{bmatrix} 0 & 1 & \ldots & 0 \\ \vdots & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & 1 \\ 0 & \ldots & \ldots & 0 \end{bmatrix}, \ \boldsymbol{B} = \begin{bmatrix} 0 \\ \vdots \\ 0 \\ 1 \end{bmatrix}, \\ \boldsymbol{C} = \begin{bmatrix} r_n & \ldots & r_2 & r_1 \end{bmatrix}, \quad \boldsymbol{D} = r_0/2. \end{gathered} \tag{2.38}$$

Replacing these matrices in (2.37), we obtain

$$\begin{aligned} \boldsymbol{Q} &= \begin{bmatrix} \boldsymbol{P} & \boldsymbol{C}^T \\ \boldsymbol{C} & \boldsymbol{D} + \boldsymbol{D}^T \end{bmatrix} - \begin{bmatrix} \boldsymbol{A}^T \\ \boldsymbol{B}^T \end{bmatrix} \boldsymbol{P} \begin{bmatrix} \boldsymbol{A} & \boldsymbol{B} \end{bmatrix} \\ &= \left[\begin{array}{c|c} & r_n \\ \boldsymbol{P} & \vdots \\ & r_1 \\ \hline r_n \ \ldots \ r_1 & r_0 \end{array}\right] - \left[\begin{array}{c|c} 0 & 0 \ldots 0 \\ \hline 0 & \\ \vdots & \boldsymbol{P} \\ 0 & \end{array}\right] \succeq 0. \end{aligned} \tag{2.39}$$

Remarking that

$$\operatorname{tr} \boldsymbol{\Theta}_k \left(\begin{bmatrix} \boldsymbol{P} & \boldsymbol{0} \\ \boldsymbol{0} & 0 \end{bmatrix} - \begin{bmatrix} 0 & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{P} \end{bmatrix} \right) = 0, \tag{2.40}$$

for all $k = 0 : n$, the relation (2.39) is equivalent to $\operatorname{tr}[\boldsymbol{\Theta}_k \boldsymbol{Q}] = r_k$, i.e., the trace parameterization (2.6).

Despite the equivalence, it is not efficient to solve problems with nonnegative polynomials, as those presented in Sect. 2.2, by using the constraint (2.39), $\boldsymbol{P} \succeq 0$, instead of the trace parameterization. The LMI (2.39) has size $(n + 1) \times (n + 1)$ (as in the trace parameterization), but it contains $O(n^2)$ scalar variables in the matrix $\boldsymbol{P}$ (compared to only $n + 1$ equalities for the trace parameterization, which is an LMI in equality form). Consequently, the use of the KYP lemma leads to a complexity of $O(n^6)$, much higher than the $O(n^4)$ needed by the trace parameterization.

2.6 Spectral Factorization from a Gram Matrix

Let $R(z)$ be a nonnegative trigonometric polynomial and $H(z)$ a spectral factor respecting (1.11). As we have seen, we can express certain optimization problems involving $R(z)$ in terms of its associated Gram matrices. In this section, we explore how the spectral factor $H(z)$ can be computed directly from a Gram matrix $\boldsymbol{Q}$, and not by using (2.6) to get $R(z)$ and then obtaining $H(z)$ with one of the spectral factorization algorithms described in Appendix B.

2.6.1 SDP Computation of a Rank-1 Gram Matrix

We have remarked that the positive semidefinite matrix $\boldsymbol{Q}_1 = \boldsymbol{h}\boldsymbol{h}^H$ is a Gram matrix associated with $R(z)$ (see the lines preceding (2.8)). So, if

$$\boldsymbol{Q} = \begin{bmatrix} q_{00} & \boldsymbol{q}^H \\ \boldsymbol{q} & \hat{\boldsymbol{Q}} \end{bmatrix} \succeq 0 \tag{2.41}$$

is a rank-1 Gram matrix, then the spectral factor can be readily obtained from its first column as

$$\boldsymbol{h} = \frac{1}{\sqrt{q_{00}}} \begin{bmatrix} q_{00} \\ \boldsymbol{q} \end{bmatrix}. \tag{2.42}$$

The following theorem gives the conditions to obtain such a Gram matrix, more precisely the one giving the minimum-phase spectral factor.

Theorem 2.15 *Let $R \in \mathbb{C}_n[z]$ be a nonnegative trigonometric polynomial. Let $\boldsymbol{Q} \in \mathbb{C}^{n' \times n'}$ be the positive semidefinite Gram matrix (2.41) associated with $R(z)$ which has the largest element q_{00}. Then, the rank of the matrix $\boldsymbol{Q}$ is equal to 1 and $\boldsymbol{Q} = \boldsymbol{h}\boldsymbol{h}^H$, where the vector $\boldsymbol{h}$ contains the coefficients of the minimum-phase spectral factor of $R(z)$.*

Proof The set $\mathcal{G}(R)$ of positive semidefinite Gram matrices $\boldsymbol{Q}$ associated with $R(z)$ is convex and closed and the function $f(\boldsymbol{Q}) = q_{00}$ is linear and so is convex. It results that the maximum of $f(\boldsymbol{Q})$ is attained for some $\boldsymbol{Q} \in \mathcal{G}(R)$, $\boldsymbol{Q} \succeq 0$, and thus, the Gram matrix asserted in the theorem indeed exists. Let us assume that its rank is not one. Then, writing $\boldsymbol{Q}$ as in (2.41), there exists a nonzero matrix $\boldsymbol{P} \succeq 0$ such that

$$\begin{bmatrix} q_{00} & \boldsymbol{q}^H \\ \boldsymbol{q} & \hat{\boldsymbol{Q}} - \boldsymbol{P} \end{bmatrix} \succeq 0. \tag{2.43}$$

For example, we can take $\boldsymbol{P} = \hat{\boldsymbol{Q}} - \boldsymbol{q}\boldsymbol{q}^H / q_{00}$, i.e., the Schur complement of q_{00} in $\boldsymbol{Q}$. For this $\boldsymbol{P}$, it is clear that $\boldsymbol{P} = \boldsymbol{0}$ only if the rank $\boldsymbol{Q} = 1$, which we have assumed not true. Let us write

$$\boldsymbol{P} = \begin{bmatrix} \boldsymbol{0} & \boldsymbol{0} \\ \boldsymbol{0} & \hat{\boldsymbol{P}} \end{bmatrix},$$

where $\hat{p}_{00} > 0$. So, we put in evidence the first nonzero (and positive, since $\boldsymbol{P} \succeq 0$ and $\boldsymbol{P}$ is nonzero) diagonal element of $\boldsymbol{P}$ as upper-left element of the block $\hat{\boldsymbol{P}}$. Define the matrix

$$\boldsymbol{X} = \boldsymbol{Q} - \begin{bmatrix} 0 & 0 \\ 0 & \hat{\boldsymbol{P}} \end{bmatrix} + \begin{bmatrix} \hat{\boldsymbol{P}} & 0 \\ 0 & 0 \end{bmatrix}. \tag{2.44}$$

Since $\boldsymbol{X}$ is obtained by adding a positive semidefinite matrix to (2.43), it follows that $\boldsymbol{X} \succeq 0$. Moreover, taking (2.40) into account, it results that $\mathrm{tr}[\boldsymbol{\Theta}_k \boldsymbol{X}] = \mathrm{tr}[\boldsymbol{\Theta}_k \boldsymbol{Q}]$, for any $k = 0 : n$, and so $\boldsymbol{X} \in \mathcal{G}(R)$. Finally, we note that $x_{00} = q_{00} + \hat{p}_{00} > q_{00}$. We have thus built a Gram matrix associated with $R(z)$, whose upper-left element is greater than q_{00}, which is impossible. We conclude that the rank of $\boldsymbol{Q}$ is one and so $\boldsymbol{Q} = \boldsymbol{h}\boldsymbol{h}^H$, with $\boldsymbol{h}$ defined by (2.42).

That $\boldsymbol{h}$ is minimum-phase follows from the well-known Robinson's energy delay property, stating that the minimum-phase filter has the most energy concentrated in its first coefficients. Let $\boldsymbol{g}$ be a spectral factor of $R(z)$ having at least one zero outside the unit circle; if $\boldsymbol{h}$ is minimum-phase, then

$$\sum_{i=0}^{k} |h_i|^2 \geq \sum_{i=0}^{k} |g_i|^2, \quad \forall k = 0 : n-1,$$

and reciprocally. Moreover, for $k = 0$, the inequality is strict, i.e., $|h_0|^2 > |g_0|^2$. Since for the vector (2.42) we have $q_{00} = |h_0|^2 > |g_0|^2$ and q_{00} is maximum, it results that $\boldsymbol{h}$ is minimum-phase. ■

We conclude that the spectral factorization of a polynomial $R(z)$ can be computed with the following algorithm.

1. Solve the SDP problem

$$\begin{aligned} &\max_{\boldsymbol{Q}} \; q_{00} \\ &\text{s.t.} \;\; \mathrm{tr}[\boldsymbol{\Theta}_k \boldsymbol{Q}] = r_k, \quad k = 0 : n \\ &\qquad \boldsymbol{Q} \succeq 0 \end{aligned} \tag{2.45}$$

2. Writing the solution $\boldsymbol{Q}$ as in (2.41), compute the minimum-phase spectral factor $\boldsymbol{h}$ with (2.42).

This spectral factorization algorithm has generally a higher complexity than many of those presented in Appendix B. However, it has two advantages. With appropriate modifications described in Sect. B.5, it can be used for polynomials with matrix coefficients (a topic discussed later in Sect. 3.10). Also, it can be combined with certain optimization problems in order to avoid spectral factorization as a separate operation. Consider for example the Toeplitz quadratic optimization problem (2.24). Instead of solving (2.31), computing $R(z)$, and then its spectral factor, we can solve

$$\begin{aligned} &\min_{\boldsymbol{Q}} \; \mathrm{tr}[\boldsymbol{A}_0 \boldsymbol{Q}] - \alpha q_{00} \\ &\text{s.t.} \;\; \mathrm{tr}[\boldsymbol{A}_\ell \boldsymbol{Q}] = b_\ell, \; \ell = 1 : L \\ &\qquad \boldsymbol{Q} \succeq 0 \end{aligned} \tag{2.46}$$

where α is a constant. This constant should be small enough such that (2.46) gives (approximately) the same $R(z)$ as the original problem (2.31). However, the Gram matrix $\boldsymbol{Q}$ given by (2.46) will have rank equal to 1. (That there is a rank-1 solution to (2.31) is ensured by its equivalence to (2.24)!)

2.6.2 Spectral Factorization Using a Riccati Equation

Let $(\boldsymbol{\Theta}_1^T, \tilde{\boldsymbol{h}}, \boldsymbol{c}^T, h_0)$ be the observable state-space realization of $H(z)$, where

$$\tilde{\boldsymbol{h}} = [h_n \ \dots \ h_1]^T, \quad \boldsymbol{c} = [0 \ \dots \ 0 \ 1]^T. \tag{2.47}$$

Given $R(z)$, the state-space formalism can be used to obtain a spectral factorization algorithm, based on solving a Riccati equation, as follows. Note also that in the spectral factorization relation (1.11), we can always take $H(z)$ such that h_0 is real.

Theorem 2.16 *Let $R \in \mathbb{C}_n[z]$ be a nonnegative polynomial and denote*

$$\tilde{\boldsymbol{r}} = [r_n \ \dots \ r_1]^T.$$

Let $\boldsymbol{\Xi}$ be the (positive semidefinite) solution of the discrete-time matrix Riccati equation

$$\boldsymbol{\Xi} = \boldsymbol{\Theta}_1^T \boldsymbol{\Xi} \boldsymbol{\Theta}_1 + (\tilde{\boldsymbol{r}} - \boldsymbol{\Theta}_1^T \boldsymbol{\Xi} \boldsymbol{c})(r_0 - \boldsymbol{c}^T \boldsymbol{\Xi} \boldsymbol{c})^{-1} (\tilde{\boldsymbol{r}} - \boldsymbol{\Theta}_1^T \boldsymbol{\Xi} \boldsymbol{c})^H. \tag{2.48}$$

The minimum-phase spectral factor of $R(z)$ is given by

$$\begin{aligned} h_0 &= (r_0 - \boldsymbol{c}^T \boldsymbol{\Xi} \boldsymbol{c})^{1/2}, \\ \tilde{\boldsymbol{h}} &= (\tilde{\boldsymbol{r}} - \boldsymbol{\Theta}_1^T \boldsymbol{\Xi} \boldsymbol{c}) / h_0, \end{aligned} \tag{2.49}$$

where $\tilde{\boldsymbol{h}}$ and $\boldsymbol{c}$ are like in (2.47).

This is a relatively well-known result; for completeness, the proof is given in Sect. 2.12.2. Note that the matrix $\boldsymbol{\Theta}_1$ is stable (has all eigenvalues inside the unit circle) and thus makes possible the existence of a positive semidefinite solution $\boldsymbol{\Xi}$ of the Riccati equation (2.48); this ensures the minimum-phase property of the spectral factor.

So, the minimum-phase spectral factor is computed simply with (2.49), after solving the Riccati equation (2.48); due to the special form of $\boldsymbol{\Theta}_1$ and $\boldsymbol{c}$, some computations are trivial; for example, $\boldsymbol{c}^T \boldsymbol{\Xi} \boldsymbol{c}$ is the element of $\boldsymbol{\Xi}$ from the lower-right corner, etc. Note that relations (2.48), (2.49) can be written in the equivalent form

$$\begin{aligned} h_0^2 &= r_0 - \boldsymbol{c}^T \boldsymbol{\Xi} \boldsymbol{c}, \\ h_0 \tilde{\boldsymbol{h}} &= \tilde{\boldsymbol{r}} - \boldsymbol{\Theta}_1^T \boldsymbol{\Xi} \boldsymbol{c}, \\ \tilde{\boldsymbol{h}} \tilde{\boldsymbol{h}}^H &= \boldsymbol{\Xi} - \boldsymbol{\Theta}_1^T \boldsymbol{\Xi} \boldsymbol{\Theta}_1. \end{aligned} \tag{2.50}$$

Now, let assume that we have a positive semidefinite Gram matrix $\boldsymbol{Q}$ associated with $R(z)$, split as follows:

$$\boldsymbol{Q} = \begin{bmatrix} \tilde{\boldsymbol{Q}} & \boldsymbol{s} \\ \boldsymbol{s}^H & \rho \end{bmatrix}, \tag{2.51}$$

where ρ is a scalar. Writing $\boldsymbol{Q}$ as in (2.39) and identifying the blocks with (2.51), we obtain

$$\begin{aligned} \rho &= r_0 - \boldsymbol{c}^T \boldsymbol{P} \boldsymbol{c}, \\ \boldsymbol{s} &= \tilde{\boldsymbol{r}} - \boldsymbol{\Theta}_1^T \boldsymbol{P} \boldsymbol{c}, \\ \tilde{\boldsymbol{Q}} &= \boldsymbol{P} - \boldsymbol{\Theta}_1^T \boldsymbol{P} \boldsymbol{\Theta}_1. \end{aligned} \tag{2.52}$$

Subtracting (2.52) from (2.50) and denoting $\boldsymbol{\Pi} = \boldsymbol{P} - \boldsymbol{\Xi}$, we obtain

$$\begin{aligned} h_0^2 &= \rho + \boldsymbol{c}^T \boldsymbol{\Pi} \boldsymbol{c}, \\ h_0 \tilde{\boldsymbol{h}} &= \boldsymbol{s} + \boldsymbol{\Theta}_1^T \boldsymbol{\Pi} \boldsymbol{c}, \\ \tilde{\boldsymbol{h}} \tilde{\boldsymbol{h}}^H &= \tilde{\boldsymbol{Q}} - \boldsymbol{\Pi} + \boldsymbol{\Theta}_1^T \boldsymbol{\Pi} \boldsymbol{\Theta}_1. \end{aligned} \tag{2.53}$$

These relations give the spectral factorization algorithm working directly with the Gram matrix $\boldsymbol{Q}$, split as in (2.51):

1. Compute the matrix $\boldsymbol{\Pi}$ by solving the Riccati equation

$$\boldsymbol{\Pi} = \tilde{\boldsymbol{Q}} + \boldsymbol{\Theta}_1^T \boldsymbol{\Pi} \boldsymbol{\Theta}_1 - (\boldsymbol{s} + \boldsymbol{\Theta}_1^T \boldsymbol{\Pi} \boldsymbol{c})(\rho + \boldsymbol{c}^T \boldsymbol{\Pi} \boldsymbol{c})^{-1} (\boldsymbol{s} + \boldsymbol{\Theta}_1^T \boldsymbol{\Pi} \boldsymbol{c})^H. \tag{2.54}$$

2. Compute the minimum-phase spectral factor $H(z)$ with

$$\begin{aligned} h_0 &= (\rho + \boldsymbol{c}^T \boldsymbol{\Pi} \boldsymbol{c})^{1/2}, \\ \tilde{\boldsymbol{h}} &= (\boldsymbol{s} + \boldsymbol{\Theta}_1^T \boldsymbol{\Pi} \boldsymbol{c}) / h_0. \end{aligned} \tag{2.55}$$

In principle, the algorithm for solving the Riccati equation may fail if the polynomial $R(z)$ has zeros on the unit circle (and so a symplectic matrix built with the parameters of the equation has eigenvalues on the unit circle). However, in practice, the algorithm based on solving (2.54) works very well; this appears to be due to the presence of small numerical errors in the Gram matrix, and so in $\tilde{\boldsymbol{Q}}$. (On the contrary, the algorithm based on solving (2.48) was observed to fail!) Although this algorithm is rather slow and can be used only for degrees up to 200–300, the author's experience recommends it as very safe.

2.7 Parameterization of Real Polynomials

The presentation of the Gram matrix concept given in Sect. 2.1 can be followed with few modifications for the case of polynomials of real variable. Since we are interested by positive polynomials, we consider only even degrees. Most of the proofs are given at the end of the chapter.

Definition 2.17 Consider the polynomial $P \in \mathbb{R}_{2n}[t]$. A symmetric matrix $\boldsymbol{Q} \in \mathbb{R}^{n' \times n'}$, where $n' = n + 1$, is called a *Gram* matrix associated with $P(t)$ if

$$P(t) = \boldsymbol{\psi}_n^T(t) \cdot \boldsymbol{Q} \cdot \boldsymbol{\psi}_n(t). \tag{2.56}$$

We denote $\mathcal{G}(P)$ the set of Gram matrices associated with $P(t)$. ■

Example 2.18 Consider polynomials of degree four, $P(t) = p_0 + p_1 t + p_2 t^2 + p_3 t^3 + p_4 t^4$. A Gram matrix $\boldsymbol{Q} \in \mathbb{R}^{3\times 3}$ satisfies the relation

$$P(t) = [1 \;\; t \;\; t^2] \begin{bmatrix} q_{00} & q_{10} & q_{20} \\ q_{10} & q_{11} & q_{21} \\ q_{20} & q_{21} & q_{22} \end{bmatrix} \begin{bmatrix} 1 \\ t \\ t^2 \end{bmatrix}.$$

It results that

$$\begin{aligned} p_0 &= q_{00}, \\ p_1 &= 2q_{10}, \\ p_2 &= q_{11} + 2q_{20}, \\ p_3 &= 2q_{21}, \\ p_4 &= q_{22}. \end{aligned} \tag{2.57}$$

Unlike the 3 × 3 Gram matrix in Example 2.2, here there is only one degree of liberty left to the Gram matrices associated with $P(t)$, which have the form

$$\boldsymbol{Q} = \begin{bmatrix} p_0 & \frac{p_1}{2} & \frac{p_2}{2} \\ \frac{p_1}{2} & 0 & \frac{p_3}{2} \\ \frac{p_2}{2} & \frac{p_3}{2} & p_4 \end{bmatrix} + \begin{bmatrix} 0 & 0 & -\frac{q_{11}}{2} \\ 0 & q_{11} & 0 \\ -\frac{q_{11}}{2} & 0 & 0 \end{bmatrix}.$$

In general, the mapping (2.56) that associates symmetric matrices in $\mathbb{R}^{n' \times n'}$ to real polynomials is many-to-one. For instance, taking

$$P(t) = 2 + 2t + 7t^2 - 2t^3 + t^4, \tag{2.58}$$

the following two matrices

$$\boldsymbol{Q}_0 = \begin{bmatrix} 2 & 1 & 0 \\ 1 & 7 & -1 \\ 0 & -1 & 1 \end{bmatrix}, \quad \boldsymbol{Q}_1 = \begin{bmatrix} 2 & 1 & 2 \\ 1 & 3 & -1 \\ 2 & -1 & 1 \end{bmatrix} \tag{2.59}$$

are Gram matrices associated with $P(t)$. ■

Relations (2.57) suggest that the coefficients of $P(t)$ are obtained as sums along the antidiagonals of the Gram matrix.

Theorem 2.19 *If $P \in \mathbb{R}_{2n}[t]$ and $\boldsymbol{Q} \in \mathcal{G}(P)$, then the relation*

$$p_k = \operatorname{tr}[\boldsymbol{\Upsilon}_k \boldsymbol{Q}] = \sum_{i=\max(0,k-n)}^{\min(k,n)} q_{i,k-i}, \quad k = 0 : 2n, \tag{2.60}$$

holds, where $\boldsymbol{\Upsilon}_k$ is the elementary Hankel matrix with ones on the k-th antidiagonal and zeros elsewhere (antidiagonals are numbered from zero, starting with the upper left corner of the matrix).

Example 2.20 For a polynomial of degree 4, as in Example 2.18, the first three coefficients are given through (2.60) by

$$p_0 = \operatorname{tr}\begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} \boldsymbol{Q}, \quad p_1 = \operatorname{tr}\begin{bmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} \boldsymbol{Q}, \quad p_2 = \operatorname{tr}\begin{bmatrix} 0 & 0 & 1 \\ 0 & 1 & 0 \\ 1 & 0 & 0 \end{bmatrix} \boldsymbol{Q}.$$

Theorem 2.21 *A polynomial $P \in \mathbb{R}_{2n}[t]$ is nonnegative (positive) on the real axis if and only if there exists a positive semidefinite (definite) matrix $\boldsymbol{Q} \in \mathbb{R}^{n' \times n'}$ such that (2.60) holds.*

Example 2.22 In Example 2.18, the Gram matrix $\boldsymbol{Q}_0$ from (2.59) is positive definite, which shows that the polynomial (2.58) is positive. However, the Gram matrix $\boldsymbol{Q}_1$ is not positive semidefinite. ■

Remark 2.23 (Sum-of-squares decomposition) Let $P \in \mathbb{R}_{2n}[t]$ be a nonnegative polynomial. Let $\boldsymbol{Q}$ be an associated positive semidefinite Gram matrix, whose eigenvalue decomposition is (2.11); as $\boldsymbol{Q}$ is real, the eigenvectors $\boldsymbol{x}_\ell$ are also real. Inserting (2.11) into (2.56), we obtain the sum-of-squares decomposition

$$P(t) = \sum_{\ell=1}^{\nu} \lambda_\ell^2 \cdot [\boldsymbol{\psi}^T(t)\boldsymbol{x}_\ell] \cdot [\boldsymbol{x}_\ell^T \boldsymbol{\psi}(t)] = \sum_{\ell=1}^{\nu} F_\ell(t)^2, \tag{2.61}$$

where $F_\ell(t) = \lambda_\ell \boldsymbol{\psi}^T(t)\boldsymbol{x}_\ell$. ■

As examples of SDP programs solving some simple problems involving nonnegative polynomials of real variable, we will give below short descriptions of the problems *Min_poly_value* and *Most_positive_Gram_matrix* for real polynomials. In contrast with the trigonometric polynomials case, it will result that these problems are not equivalent.

Problem (*Min_poly_value*) Let $P \in \mathbb{R}_{2n}[t]$, with $p_{2n} > 0$. We compute $\mu^\star = \min_{t \in \mathbb{R}} P(t)$ by finding the maximum $\mu \in \mathbb{R}$ for which $\tilde{P}(t) = P(t) - \mu$ is a nonnegative polynomial. Using the parameterization (2.60), the following SDP problem,

$$\begin{aligned} \mu^\star = \max_{\mu, \tilde{\boldsymbol{Q}}} \ & \mu \\ \text{s.t.} \ & \mu + \operatorname{tr}[\boldsymbol{\Upsilon}_0 \tilde{\boldsymbol{Q}}] = p_0 \\ & \operatorname{tr}[\boldsymbol{\Upsilon}_k \tilde{\boldsymbol{Q}}] = p_k, \quad k = 1 : 2n \\ & \tilde{\boldsymbol{Q}} \succeq 0 \end{aligned} \tag{2.62}$$

similar to (2.18), provides the solution.

Problem (*Most_positive_Gram_matrix*) Let $P \in \mathbb{R}_{2n}[t]$ be a positive polynomial. To compute the most positive Gram matrix associated with $P(t)$, we have to solve an SDP problem similar to (2.14), namely

$$\begin{aligned} \lambda^\star = \max_{\lambda, \boldsymbol{Q}} \ & \lambda \\ \text{s.t.} \ & \operatorname{tr}[\boldsymbol{\Upsilon}_k \boldsymbol{Q}] = p_k, \quad k = 0 : 2n \\ & \lambda \geq 0, \quad \boldsymbol{Q} \succeq \lambda \boldsymbol{I} \end{aligned} \tag{2.63}$$

To bring the above problem to standard form, we use, as in Sect. 2.2, the positive definite matrix $\tilde{\boldsymbol{Q}} = \boldsymbol{Q} - \lambda \boldsymbol{I}$. The difference is that now we have

$$\operatorname{tr} \boldsymbol{\Upsilon}_k = \begin{cases} 1, & \text{if } k \text{ is even} \\ 0, & \text{if } k \text{ is odd.} \end{cases}$$

Thus, the standard form of (2.63) is

$$\begin{aligned} \lambda^\star = \max_{\lambda, \tilde{\boldsymbol{Q}}} \ & \lambda \\ \text{s.t.} \ & \lambda + \operatorname{tr}[\boldsymbol{\Upsilon}_k \tilde{\boldsymbol{Q}}] = p_k, \quad k = 0 : 2 : 2n, \\ & \operatorname{tr}[\boldsymbol{\Upsilon}_k \tilde{\boldsymbol{Q}}] = p_k, \quad k = 1 : 2 : 2n \\ & \lambda \geq 0, \quad \tilde{\boldsymbol{Q}} \succeq 0 \end{aligned} \tag{2.64}$$

It is obvious that the constraints of (2.64) and (2.62) are different. ■

Example 2.24 Consider again the polynomial (2.58). Its minimum value for $t \in \mathbb{R}$ is 1.8628. The most positive Gram matrix associated with $P(t)$ is

$$\boldsymbol{Q} = \begin{bmatrix} 2.0000 & 1.0000 & -0.1763 \\ 1.0000 & 7.3525 & -1.0000 \\ -0.1763 & -1.0000 & 1.0000 \end{bmatrix}.$$

The smallest eigenvalue of $\boldsymbol{Q}$ is 0.8458. ■

2.8 Choosing the Right Basis

In defining the Gram matrices for trigonometric polynomials, we have used the natural basis (2.1). However, there are other possibilities. The technically simplest way is to replace the vector $\boldsymbol{\psi}(z)$ with

$$\boldsymbol{\phi}(z) = \boldsymbol{C}\boldsymbol{\psi}(z), \tag{2.65}$$

where $\boldsymbol{C} \in \mathbb{C}^{(n+1)\times(n+1)}$ is a nonsingular matrix. The relation (2.2) becomes

$$R(z) = \boldsymbol{\phi}^H(z^{-1}) \cdot \boldsymbol{C}^{-H}\boldsymbol{Q}\boldsymbol{C}^{-1} \cdot \boldsymbol{\phi}(z). \tag{2.66}$$

From Theorem 2.5, we immediately conclude that $R(z)$ is nonnegative if and only if there exist $\widehat{\boldsymbol{Q}} \succeq 0$ such that

$$R(z) = \boldsymbol{\phi}^H(z^{-1}) \cdot \widehat{\boldsymbol{Q}} \cdot \boldsymbol{\phi}(z). \tag{2.67}$$

(Since $\boldsymbol{C}$ is nonsingular, any $\widehat{\boldsymbol{Q}} \succeq 0$ can be written as $\widehat{\boldsymbol{Q}} = \boldsymbol{C}^{-H}\boldsymbol{Q}\boldsymbol{C}^{-1}$, for some $\boldsymbol{Q} \succeq 0$.) We can name $\widehat{\boldsymbol{Q}}$ a Gram matrix associated with $R(z)$, for the basis $\boldsymbol{\phi}(z)$. The parameterization (2.6) takes the form

$$r_k = \mathrm{tr}[\boldsymbol{\Theta}_k\boldsymbol{C}^H\widehat{\boldsymbol{Q}}\boldsymbol{C}] = \mathrm{tr}[\boldsymbol{C}\boldsymbol{\Theta}_k\boldsymbol{C}^H \cdot \widehat{\boldsymbol{Q}}]. \tag{2.68}$$

This general approach may be not so useful, especially as it may produce complex Gram matrices $\widehat{\boldsymbol{Q}}$ even for polynomials with real coefficients. Since $R(\omega)$ has real values, we would be more interested in associating real Gram matrices even with polynomials with complex coefficients. We introduce in the sequel several new parameterizations.

2.8.1 *Basis of Trigonometric Polynomials*

Let us consider a nonnegative trigonometric polynomial $R(z)$ of degree $n = 2\tilde{n}$. The spectral factorization Theorem 1.1 says that $R(\omega) = |H(\omega)|^2$, where $H(z)$ is a causal polynomial; since $H(z)$ may be multiplied with any unit-norm constant and is still a spectral factor, we can take $h_{\tilde{n}}$ real. We can also write $R(\omega) = |\tilde{H}(\omega)|^2$, with

$$\tilde{H}(z) = z^{\tilde{n}}H(z) = \sum_{k=-\tilde{n}}^{\tilde{n}} h_{k+\tilde{n}}z^{-k}. \tag{2.69}$$

It results that

$$\tilde{H}(\omega) = A(\omega) + jB(\omega), \tag{2.70}$$

where

$$\begin{aligned} A(\omega) &= h_{\tilde{n}} + \textstyle\sum_{k=1}^{\tilde{n}} [\mathrm{Re}(h_{\tilde{n}-k} + h_{\tilde{n}+k}) \cos k\omega + \mathrm{Im}(-h_{\tilde{n}-k} + h_{\tilde{n}+k}) \sin k\omega], \\ B(\omega) &= \textstyle\sum_{k=1}^{\tilde{n}} [\mathrm{Im}(h_{\tilde{n}-k} + h_{\tilde{n}+k}) \cos k\omega + \mathrm{Re}(h_{\tilde{n}-k} - h_{\tilde{n}+k}) \sin k\omega] \end{aligned}$$

are trigonometric polynomials of degree $\tilde{n}$, with real coefficients. Introducing the basis vector (of length $n+1$)

$$\boldsymbol{\chi}(\omega) = [1 \;\; \cos\omega \;\; \sin\omega \;\; \ldots \;\; \cos\tilde{n}\omega \;\; \sin\tilde{n}\omega]^T, \tag{2.71}$$

we can write

$$A(\omega) = \boldsymbol{a}^T \boldsymbol{\chi}(\omega), \quad B(\omega) = \boldsymbol{b}^T \boldsymbol{\chi}(\omega),$$

with $\boldsymbol{a}, \boldsymbol{b} \in \mathbb{R}^{n+1}$, and so we obtain

$$|\tilde{H}(\omega)|^2 = A(\omega)^2 + B(\omega)^2 = \boldsymbol{\chi}^T(\omega)(\boldsymbol{a}\boldsymbol{a}^T + \boldsymbol{b}\boldsymbol{b}^T)\boldsymbol{\chi}(\omega). \tag{2.72}$$

This expression leads to a result similar with Theorems 2.5 and 2.21.

Theorem 2.25 *A polynomial $R \in \mathbb{C}_{2\tilde{n}}[z]$ is nonnegative on the unit circle if and only if there exists a positive semidefinite matrix $\boldsymbol{Q} \in \mathbb{R}^{(2\tilde{n}+1)\times(2\tilde{n}+1)}$ such that*

$$R(\omega) = \boldsymbol{\chi}^T(\omega) \cdot \boldsymbol{Q} \cdot \boldsymbol{\chi}(\omega), \tag{2.73}$$

where $\boldsymbol{\chi}(\omega)$ is defined in (2.71).

Proof If there exists $\boldsymbol{Q} \succeq 0$ such that (2.73) holds, then it is clear that $R(\omega) \geq 0$ for all ω. Reciprocally, if $R(\omega) \geq 0$, then the matrix $\boldsymbol{Q} \overset{\Delta}{=} \boldsymbol{a}\boldsymbol{a}^T + \boldsymbol{b}\boldsymbol{b}^T \succeq 0$ from (2.72) satisfies (2.73). ■

Example 2.26 Let us take a polynomial of degree $n = 2$ (i.e., $\tilde{n} = 1$), with complex coefficients. On the unit circle, according to (2.73), we have

$$R(\omega) = [1 \;\; \cos\omega \;\; \sin\omega] \begin{bmatrix} q_{00} & q_{10} & q_{20} \\ q_{10} & q_{11} & q_{21} \\ q_{20} & q_{21} & q_{22} \end{bmatrix} \begin{bmatrix} 1 \\ \cos\omega \\ \sin\omega \end{bmatrix}.$$

By using simple trigonometric expressions such as $(\cos\omega)^2 = (1+\cos 2\omega)/2$, we obtain

$$R(\omega) = (q_{00} + \tfrac{q_{11}}{2} + \tfrac{q_{22}}{2}) + 2q_{10}\cos\omega + 2q_{20}\sin\omega + (\tfrac{q_{11}}{2} - \tfrac{q_{22}}{2})\cos 2\omega + q_{21}\sin 2\omega. \tag{2.74}$$

In the particular case where the polynomial is

$$R(z) = (2+j)z^{-2} + (3-j)z^{-1} + 9 + (3+j)z + (2-j)z^2,$$

on the unit circle we have

$$R(\omega) = 9 + 6\cos\omega - 2\sin\omega + 4\cos 2\omega + 2\sin 2\omega. \tag{2.75}$$

Identifying with (2.74), we obtain the general form of the Gram matrix

$$\boldsymbol{Q} = \begin{bmatrix} q_{00} & 3 & -1 \\ 3 & 13 - q_{00} & 2 \\ -1 & 2 & 5 - q_{00} \end{bmatrix}. \tag{2.76}$$

Taking $q_{00} = 2$, we get the matrix

$$\boldsymbol{Q} = \begin{bmatrix} 2 & 3 & -1 \\ 3 & 11 & 2 \\ -1 & 2 & 3 \end{bmatrix} \succ 0.$$

Its positivity ensures that the polynomial $R(z)$ is positive on the unit circle. Indeed, the minimum value of the polynomial is 0.5224, as obtained by solving (2.18) with the programs shown in Sect. 2.12.1. ■

Using the standard trigonometric identities

$$\begin{aligned} \cos i\omega \cos \ell\omega &= \tfrac{1}{2}[\cos(i+\ell)\omega + \cos(i-\ell)\omega], \\ \sin i\omega \sin \ell\omega &= \tfrac{1}{2}[-\cos(i+\ell)\omega + \cos(i-\ell)\omega], \end{aligned} \tag{2.77}$$

and

$$\sin i\omega \cos \ell\omega = \frac{1}{2}[\sin(i+\ell)\omega + \sin(i-\ell)\omega], \tag{2.78}$$

it can be easily shown that the relation (2.73) can be written as a linear dependence between the coefficients of $R(z)$ and the elements of the matrix $\boldsymbol{Q}$. This proves the following.

Theorem 2.27 *A polynomial $R \in \mathbb{C}_{2\tilde{n}}[z]$ is nonnegative on the unit circle if and only if there exists a positive semidefinite matrix $\boldsymbol{Q} \in \mathbb{R}^{(2\tilde{n}+1)\times(2\tilde{n}+1)}$ such that*

$$r_k = tr[\boldsymbol{\Gamma}_k \boldsymbol{Q}], \quad k = 0 : 2\tilde{n}, \tag{2.79}$$

where $\boldsymbol{\Gamma}_k$ are constant matrices.

The expressions of the matrices $\boldsymbol{\Gamma}_k$ are not derived here. Note that these matrices are in general complex; in this sense, the parameterization (2.79) is opposed to the trace parameterization (2.6), where the constant matrices $\boldsymbol{\Theta}_k$ are real, while the parameter matrix $\boldsymbol{Q}$ is complex. However, the relation (2.79) can be immediately split into $\text{Re} r_k = \text{tr}[(\text{Re}\boldsymbol{\Gamma}_k)\boldsymbol{Q}]$, $\text{Im} r_k = \text{tr}[(\text{Im}\boldsymbol{\Gamma}_k)\boldsymbol{Q}]$; the total number of real equalities is $2n + 1$ (remind that $n = 2\tilde{n}$). In contrast, the trace parameterization (2.6) has $n + 1$ complex equalities (amounting also to $2n + 1$ real equalities).

For example, we derive from (2.73) that the (always real) free term is

$$r_0 = q_{00} + \tfrac{1}{2}\sum_{i=1}^{n} q_{ii}$$

and so $\boldsymbol{\Gamma}_0 = \operatorname{diag}(1, 1/2, \ldots, 1/2)$. We leave the formulas for the other matrices $\boldsymbol{\Gamma}_k$ as a problem for the interested reader. A simpler case, when the polynomial has real coefficients, will be detailed in Sect. 2.8.3.

Problem (*Min_poly_value*) The parameterization (2.79) can be used to solve optimization problems in the same way as the trace parameterization. For example, the minimum value of a given polynomial $R(\omega)$ can be computed by solving

$$\begin{aligned} \mu^\star = \max_{\mu} \ & \mu \\ \text{s.t.} \ & R(\omega) - \mu = \operatorname{tr}[\boldsymbol{\Gamma}_k \tilde{\boldsymbol{Q}}],\ k = 0:n \\ & \tilde{\boldsymbol{Q}} \succeq 0 \end{aligned} \tag{2.80}$$

This is obviously an SDP problem. The main difference with respect to (2.18) is that the Gram matrix is now real, even though $R(z)$ has complex coefficients. The size of the matrix is the same in both problems. The number of equality constraints in (2.80) is $2n + 1$, i.e., the number of real coefficients in (1.8); that is why, for example, the matrix (2.76) depends only on a single variable, q_{00}; the 6 distinct elements of a 3×3 symmetric matrix must satisfy 5 linear equalities. In (2.18), the number of equality constraints is only $n + 1$, but these are complex equalities. Generally, we expect that (2.80) is solved faster than (2.18). We note also that the problem (2.80) is equivalent to finding the most positive matrix $\boldsymbol{Q}$ for which (2.73) holds; see **P** 2.10. ■

Example 2.28 (continued) The most positive matrix (2.76) is obtained for $q_{00} = 2.55$. Its smallest eigenvalue is 0.2612. This leads to a minimum value of $R(\omega)$ equal to 0.5224. (See again problem **P** 2.10.) ■

Let us now look at the case where the degree of the polynomial is odd, $n = 2\tilde{n} + 1$. We have now $R(\omega) = |\tilde{H}(\omega)|^2$, with

$$\tilde{H}(z) = z^{\tilde{n}+\frac{1}{2}} H(z) = \sum_{k=-\tilde{n}}^{\tilde{n}+1} h_{k+\tilde{n}} z^{-k+\frac{1}{2}}. \tag{2.81}$$

With the basis vector

$$\tilde{\boldsymbol{\chi}}(\omega) = [\cos\tfrac{\omega}{2}\ \sin\tfrac{\omega}{2}\ \ldots\ \cos(\tilde{n}+\tfrac{1}{2})\omega\ \sin(\tilde{n}+\tfrac{1}{2})\omega]^T \tag{2.82}$$

of length $n+1 = 2(\tilde{n}+1)$, it results that (2.70) holds with $A(\omega) = \boldsymbol{a}^T\tilde{\boldsymbol{\chi}}(\omega)$, $B(\omega) = \boldsymbol{b}^T\tilde{\boldsymbol{\chi}}(\omega)$. Hence, Theorem 2.25 holds also for odd-order polynomials if we replace $\boldsymbol{\chi}(\omega)$ with $\tilde{\boldsymbol{\chi}}(\omega)$ in (2.73). However, in this case, although the polynomial $R(\omega)$ is expressed as a sum-of-squares, the terms of the sum-of-squares are trigonometric polynomials in $\omega/2$. This aspect has no consequence on optimization applications that can be carried on as for even-order polynomials. A parameterization like (2.79) also holds, but with different constant matrices $\boldsymbol{\Gamma}_k$.

2.8.2 Transformation to Real Polynomials

We consider now trigonometric polynomials with *real* coefficients, having thus the form (1.4). (Polynomials in which all terms are sine functions can be treated similarly.) As already written in (1.6), using the simple substitution $t = \cos\omega$, a polynomial $R \in \mathbb{R}_n[z]$ can be expressed on the unit circle as

$$R(\omega) \equiv P(t) = \sum_{k=0}^{n} p_k t^k, \quad t \in [-1, 1]. \tag{2.83}$$

We can parameterize nonnegative trigonometric polynomials with real coefficients by using results valid for real polynomials nonnegative on an interval, specifically Theorem 1.11. For simplicity, we consider only the case $n = 2\tilde{n}$. According to (1.30), a polynomial (2.83) which is nonnegative for $t \in [-1, 1]$, can always be written as

$$P(t) = F(t)^2 + (1 - t^2)G(t)^2, \tag{2.84}$$

where $F(t)$ and $G(t)$ are polynomials of degree $\tilde{n}$ and $\tilde{n} - 1$, respectively. Since $F(t)^2$ and $G(t)^2$ are globally nonnegative polynomials, they can be characterized via (2.56), using positive semidefinite Gram matrices $\boldsymbol{Q}_1$ and $\boldsymbol{Q}_2$, as follows:

$$\begin{aligned} F(t)^2 &= \boldsymbol{\psi}_{\tilde{n}}^T(t)\boldsymbol{Q}_1\boldsymbol{\psi}_{\tilde{n}}(t), \\ G(t)^2 &= \boldsymbol{\psi}_{\tilde{n}-1}^T(t)\boldsymbol{Q}_2\boldsymbol{\psi}_{\tilde{n}-1}(t). \end{aligned} \tag{2.85}$$

Replacing these equalities with their counterparts similar to (2.60), the coefficients of the polynomial $P(t)$, as resulting from the identity (2.84), are given by

$$p_k = \begin{cases} \mathrm{tr}[\boldsymbol{\Upsilon}_k \boldsymbol{Q}_1] + \mathrm{tr}[\boldsymbol{\Upsilon}_k \boldsymbol{Q}_2] - \mathrm{tr}[\boldsymbol{\Upsilon}_{k-2} \boldsymbol{Q}_2], & \text{if } k \ge 2, \\ \mathrm{tr}[\boldsymbol{\Upsilon}_k \boldsymbol{Q}_1] + \mathrm{tr}[\boldsymbol{\Upsilon}_k \boldsymbol{Q}_2], & \text{if } k < 2. \end{cases} \tag{2.86}$$

Hence, we can parameterize a nonnegative trigonometric polynomials with two positive semidefinite matrices, of sizes $(\tilde{n} + 1) \times (\tilde{n} + 1)$ and $\tilde{n} \times \tilde{n}$ ($\boldsymbol{Q}_1$ and $\boldsymbol{Q}_2$, respectively). (Note the ambiguity of notation in (2.86), where the size of a matrix $\boldsymbol{\Upsilon}_k$ is dictated by the size of the matrix multiplying it.) This is in contrast with the trace parameterization (2.6), where a single matrix of size $(n + 1) \times (n + 1)$ appears. In terms of complexity, the parameterization (2.86) seems more convenient, as the size of the matrices is twice smaller; we can hope that, at least asymptotically, the problems using (2.86) may be solved faster than when using (2.6). In practice, the speedup is not visible, since typically the former problems need significantly more iterations. Moreover, the use of (2.86) is hampered by numerical stability considerations. The transformation from $R(\omega)$ to $P(t)$, using a Chebyshev basis (see Sect. 1.5.1), is made using coefficients that have a broad range of values (practically, from 1 to 2^n); also, the Chebyshev transformation matrix from (1.44) has a large condition number. For these reason, the use of (2.86) is limited to, say, $n \leq 30$; even so, the solutions obtained using the trace parameterization (2.6) are more accurate.

We conclude that the transformation (2.83) is a bad idea, although it may seem attractive from a complexity viewpoint. However, since the Chebyshev transformation is the main troublemaker, we can try to use bases of trigonometric functions, as shown in the sequel.

2.8.3 Gram-Pair Matrix Parameterization

We consider again a nonnegative trigonometric polynomial $R(z)$ with *real* coefficients, whose degree is $n = 2\tilde{n}$. The polynomial (2.69) has the form (2.70), where

$$\begin{aligned} A(\omega) &= h_{\tilde{n}} + \textstyle\sum_{k=1}^{\tilde{n}} (h_{\tilde{n}-k} + h_{\tilde{n}+k}) \cos k\omega, \\ B(\omega) &= \textstyle\sum_{k=1}^{\tilde{n}} (h_{\tilde{n}-k} - h_{\tilde{n}+k}) \sin k\omega. \end{aligned} \tag{2.87}$$

As in (2.72), we obtain

$$R(\omega) = |\tilde{H}(\omega)|^2 = A(\omega)^2 + B(\omega)^2, \tag{2.88}$$

where now $A(\omega)$ is a polynomial with cosine terms, while $B(\omega)$ is a polynomial with sine terms. Let us denote the bases of such $\tilde{n}$th order polynomials with

$$\boldsymbol{\chi}_c(\omega) = [1 \;\; \cos\omega \;\; \ldots \;\; \cos\tilde{n}\omega]^T, \tag{2.89}$$

and

$$\boldsymbol{\chi}_s(\omega) = [\sin\omega \;\; \ldots \;\; \sin\tilde{n}\omega]^T. \tag{2.90}$$

With these bases, a Gram parameterization of $R(\omega)$ is possible, using two Gram matrices.

Theorem 2.29 *Let $R \in \mathbb{R}_n[z]$ be a trigonometric polynomial of order $n = 2\tilde{n}$. The polynomial is nonnegative if and only if there exist positive semidefinite matrices $\boldsymbol{Q} \in \mathbb{R}^{(\tilde{n}+1)\times(\tilde{n}+1)}$ and $\boldsymbol{S} \in \mathbb{R}^{\tilde{n}\times\tilde{n}}$ such that*

$$R(\omega) = \boldsymbol{\chi}_c^T(\omega)\boldsymbol{Q}\boldsymbol{\chi}_c(\omega) + \boldsymbol{\chi}_s^T(\omega)\boldsymbol{S}\boldsymbol{\chi}_s(\omega). \tag{2.91}$$

We name $(\boldsymbol{Q}, \boldsymbol{S})$ a Gram pair associated with $R(\omega)$.

Proof If there exist $\boldsymbol{Q} \succeq 0$, $\boldsymbol{S} \succeq 0$ such that (2.91) holds, it results that $R(\omega) \geq 0$. Reciprocally, if $R(\omega) \geq 0$, then the matrices $\boldsymbol{Q} \stackrel{\Delta}{=} \boldsymbol{a}\boldsymbol{a}^T \succeq 0$ and $\boldsymbol{S} \stackrel{\Delta}{=} \boldsymbol{b}\boldsymbol{b}^T \succeq 0$, where $\boldsymbol{a}$ and $\boldsymbol{b}$ are the vectors of coefficients of the polynomials $A(\omega)$ and $B(\omega)$ from (2.87), satisfy (2.91). ■

To be able to formulate optimization problems, we need the expressions of the coefficients of $R(\omega)$ that result from (2.91). We start by expanding the quadratic forms, thus obtaining

$$R(\omega) = \sum_{i,\ell=0}^{\tilde{n}} q_{i\ell} \cos i\omega \cos \ell\omega + \sum_{i,\ell=0}^{\tilde{n}-1} s_{i\ell} \sin(i+1)\omega \sin(\ell+1)\omega. \tag{2.92}$$

Using the trigonometric identities (2.77) in (2.92) and taking (1.4) into account, the coefficients of $R(\omega)$ are given by

$$\begin{aligned} r_0 &= q_{00} + \frac{1}{2}\sum_{i=1}^{\tilde{n}} q_{ii} + \frac{1}{2}\sum_{i=0}^{\tilde{n}-1} s_{ii}, \\ r_k &= \frac{1}{4}\left(\sum_{i+\ell=k} q_{i\ell} + \sum_{|i-\ell|=k} q_{i\ell} - \sum_{i+\ell+2=k} s_{i\ell} + \sum_{|i-\ell|=k} s_{i\ell}\right), \quad k \geq 1. \end{aligned} \tag{2.93}$$

Thus, we can formulate the following theorem, expressing the coefficients in the style of (2.6).

Theorem 2.30 *The relation (2.91) defining a Gram pair associated with the even order trigonometric polynomial $R(\omega)$ is equivalent to*

$$r_k = \mathrm{tr}[\boldsymbol{\Phi}_k \boldsymbol{Q}] + \mathrm{tr}[\boldsymbol{\Lambda}_k \boldsymbol{S}], \tag{2.94}$$

where the matrices $\boldsymbol{\Phi}_k \in \mathbb{R}^{(\tilde{n}+1)\times(\tilde{n}+1)}$ and $\boldsymbol{\Lambda}_k \in \mathbb{R}^{\tilde{n}\times\tilde{n}}$ are

$$\begin{aligned} \boldsymbol{\Phi}_0 &= \tfrac{1}{2}(\boldsymbol{\Upsilon}_0 + \boldsymbol{I}), \\ \boldsymbol{\Phi}_k &= \tfrac{1}{4}(\boldsymbol{\Upsilon}_k + \boldsymbol{\Theta}_k + \boldsymbol{\Theta}_{-k}), \quad k \geq 1, \end{aligned} \tag{2.95}$$

and, respectively

$$\begin{aligned} \boldsymbol{\Lambda}_0 &= \tfrac{1}{2}\boldsymbol{I}, \\ \boldsymbol{\Lambda}_k &= \tfrac{1}{4}(-\boldsymbol{\Upsilon}_{k-2} + \boldsymbol{\Theta}_k + \boldsymbol{\Theta}_{-k}), \quad k \geq 1. \end{aligned} \tag{2.96}$$

In the above relations, we assume that the matrices $\boldsymbol{\Upsilon}_k$ or $\boldsymbol{\Theta}_k$ are zero whenever k is out of range (i.e., negative or larger than the number of diagonals). (For example, in (2.96), $\boldsymbol{\Upsilon}_{k-2} = 0$ if $k = 1$ and $\boldsymbol{\Theta}_k = 0$ if $k \geq \tilde{n}$.)

We note that the matrices $\boldsymbol{\Phi}_k$ and $\boldsymbol{\Lambda}_k$ are symmetric. We can replace $\boldsymbol{\Theta}_k + \boldsymbol{\Theta}_{-k}$ with $2\boldsymbol{\Theta}_k$, in their expressions, and the parameterization (2.94) remains valid.

Example 2.31 If $n = 4$, and so $\tilde{n} = 2$, the first three pairs of constant matrices from (2.94) are

$$\boldsymbol{\Phi}_0 = \frac{1}{2}\left(\begin{bmatrix} 1&0&0\\0&0&0\\0&0&0 \end{bmatrix} + \begin{bmatrix} 1&0&0\\0&1&0\\0&0&1 \end{bmatrix}\right), \quad \boldsymbol{\Lambda}_0 = \frac{1}{2}\left(-\begin{bmatrix} 0&0\\0&0 \end{bmatrix} + \begin{bmatrix} 1&0\\0&1 \end{bmatrix}\right),$$

$$\boldsymbol{\Phi}_1 = \frac{1}{4}\left(\begin{bmatrix} 0&1&0\\1&0&0\\0&0&0 \end{bmatrix} + \begin{bmatrix} 0&1&0\\1&0&1\\0&1&0 \end{bmatrix}\right), \quad \boldsymbol{\Lambda}_1 = \frac{1}{4}\left(-\begin{bmatrix} 0&0\\0&0 \end{bmatrix} + \begin{bmatrix} 0&1\\1&0 \end{bmatrix}\right),$$

$$\boldsymbol{\Phi}_2 = \frac{1}{4}\left(\begin{bmatrix} 0&0&1\\0&1&0\\1&0&0 \end{bmatrix} + \begin{bmatrix} 0&0&1\\0&0&0\\1&0&0 \end{bmatrix}\right), \quad \boldsymbol{\Lambda}_2 = \frac{1}{4}\left(-\begin{bmatrix} 1&0\\0&0 \end{bmatrix} + \begin{bmatrix} 0&0\\0&0 \end{bmatrix}\right).$$

In general, for $k = 0 : \tilde{n}$, the matrices $\boldsymbol{\Phi}_k$ have a Toeplitz+Hankel structure, while $\boldsymbol{\Phi}_k = \boldsymbol{\Upsilon}_k$ for $k > \tilde{n}$. The matrices $\boldsymbol{\Lambda}_k$ are Toeplitz for $k = 0, 1$, Toeplitz+Hankel for $k = 2 : \tilde{n} - 1$ and Hankel for $k = \tilde{n} : n$. ■

Problem (*Min_poly_value*) Using the parameterization (2.91), the minimum value of a given polynomial $R(\omega)$ can be computed by solving

$$\begin{aligned} \mu^\star = \max_{\mu} \; & \mu \\ \text{s.t.} \; & R(\omega) - \mu = \boldsymbol{\chi}_c^T(\omega)\tilde{\boldsymbol{Q}}\boldsymbol{\chi}_c(\omega) + \boldsymbol{\chi}_s^T(\omega)\tilde{\boldsymbol{S}}\boldsymbol{\chi}_s(\omega) \\ & \tilde{\boldsymbol{Q}} \succeq 0, \quad \tilde{\boldsymbol{S}} \succeq 0 \end{aligned} \tag{2.97}$$

The form (2.94) confirms that this is an SDP problem. As in Sect. 2.8.2, the two matrices from (2.97) are twice smaller than the single matrix from the corresponding problem (2.18), where the trace parameterization (2.6) is used. As discussed in Remark 2.13, the complexity of such an SDP problem depends on the square of the size of the matrices (the number of equality constraints is the same in the two problems). So, we can expect that (2.97) is solved up to four times faster than its counterpart (2.18); however, since there are two matrices in (2.97), the speedup factor could be actually twice smaller. These are only qualitative considerations; the fact that the constant matrices from (2.94) and (2.6) are sparse makes the complexity analysis more difficult.

We also note that solving (2.97) is equivalent to finding the most positive matrices $\boldsymbol{Q}$, $\boldsymbol{S}$ for which (2.91) holds. By this, we understand that $\min(\lambda_{min}(\boldsymbol{Q}), \lambda_{min}(\boldsymbol{S}))$ is maximum. See problem **P** 2.11 for details. ■

Example 2.32 We give in Table 2.1 the times needed for finding the minimum value of a trigonometric polynomial with random coefficients by solving two SDP problems; the first is (2.18), based on the trace parameterization (2.6); the second is (2.97), based on the Gram-pair parameterization (2.91). The first two rows contain data from 2006, when the first edition of this book was written. The last two rows were obtained

Table 2.1 Times, in seconds, for finding the minimum value of a trigonometric polynomial via two SDP problems

Year	SDP problem	Parameterization	Order n				
			20	50	100	200	300
2006	(2.18)	Trace	0.26	1.00	7.0	120	800
	(2.97)	Gram pair	0.21	0.51	2.7	22	99
2016	(2.18)	Trace	0.10	0.50	1.2	5.5	18
	(2.97)	Gram pair	0.04	0.20	0.7	2.6	7.5

in 2016, when preparing the second edition. In both cases, the computers were rather average for the period. We see that the Gram-pair parameterization leads to a roughly twice faster solution for almost all sizes in 2016, but only for small sizes in 2006 (excepting very small problems, where overhead due to preparing data and other operations may be significant). The most likely reason for the much better behavior of the Gram-pair parameterization on the old computer is the lower memory requirement of this parameterization. We also note that, no matter the parameterization, we can solve in the same time problems that are twice larger than 10 years ago. We conclude that the Gram-pair parameterization is clearly faster and should be preferred to the trace parameterization. ■

If the order of the polynomial $R(\omega)$ is odd, $n = 2\tilde{n} + 1$, the pseudopolynomial (2.81) leads to an expression (2.88) where $A(\omega)$ and $B(\omega)$ depend linearly on the elements of the basis vectors

$$\tilde{\boldsymbol{\chi}}_c(\omega) = [\cos\tfrac{\omega}{2} \ \cos\tfrac{3\omega}{2} \ \ldots \ \cos(\tilde{n}+\tfrac{1}{2})\omega]^T \tag{2.98}$$

and

$$\tilde{\boldsymbol{\chi}}_s(\omega) = [\sin\tfrac{\omega}{2} \ \sin\tfrac{3\omega}{2} \ \ldots \ \sin(\tilde{n}+\tfrac{1}{2})\omega]^T \tag{2.99}$$

respectively. It is easy to see that Theorem 2.29 holds also for odd n, with $\tilde{\boldsymbol{\chi}}_c(\omega)$ and $\tilde{\boldsymbol{\chi}}_s(\omega)$ replacing $\boldsymbol{\chi}_c(\omega)$ and $\boldsymbol{\chi}_s(\omega)$, respectively, in (2.91). Also, we note that the matrices $\boldsymbol{Q}$ and $\boldsymbol{S}$ have the same size, namely $(\tilde{n}+1) \times (\tilde{n}+1)$, since the basis vectors (2.98) and (2.99) have the same length. The relations between the coefficients of $R(\omega)$ and the elements of the two Gram matrices are simpler than in the even case (2.93). They are

$$\begin{aligned} r_0 &= \frac{1}{2}\sum_{i=0}^{\tilde{n}}(q_{ii}+s_{ii}), \\ r_k &= \frac{1}{4}\left(\sum_{i+\ell+1=k}(q_{i\ell}-s_{i\ell}) + \sum_{|i-\ell|=k}(q_{i\ell}+s_{i\ell})\right), \quad k \geq 1. \end{aligned} \tag{2.100}$$

The proof of the above formulas, based on relations in the style of (2.77), is left to the reader. From (2.100) it results that, for odd order, the constant matrices $\boldsymbol{\Phi}_k$ and $\boldsymbol{\Lambda}_k$ appearing in (2.94) should be replaced with

$$\begin{aligned}\tilde{\boldsymbol{\Phi}}_k &= \tfrac{1}{4}(\boldsymbol{\Upsilon}_{k-1} + \boldsymbol{\Theta}_k + \boldsymbol{\Theta}_{-k}),\\ \tilde{\boldsymbol{\Lambda}}_k &= \tfrac{1}{4}(-\boldsymbol{\Upsilon}_{k-1} + \boldsymbol{\Theta}_k + \boldsymbol{\Theta}_{-k}),\end{aligned} \tag{2.101}$$

respectively, for $k = 0 : n$.

2.9 Interpolation Representations

Until now, in all parameterizations, we have defined the polynomials by their coefficients. Alternatively, we can use as parameters the values of the polynomial on a specified set of points. Let $\Omega = \{\omega_i\}_{i=1:2n+1} \subset (-\pi, \pi]$ be a set of $2n+1$ frequency points. We return to the general case of trigonometric polynomials with complex coefficients. If the values

$$\rho_i = R(\omega_i), \quad i = 1 : 2n+1, \tag{2.102}$$

of the nth order trigonometric polynomial $R(z)$ are known, then the polynomial is completely determined.

For simplicity, let us consider only the case where $n = 2\tilde{n}$. We work with the basis of trigonometric polynomials

$$\boldsymbol{\varphi}(\omega) = \boldsymbol{C}\boldsymbol{\chi}(\omega), \tag{2.103}$$

where $\boldsymbol{C}$ is a nonsingular matrix; so, we consider all bases that are similar to (2.71). We can now state a characterization of nonnegative polynomials defined in terms of the values (2.102).

Theorem 2.33 *The trigonometric polynomial $R \in \mathbb{C}_n[z]$ satisfying (2.102) is nonnegative if and only if there exists a positive semidefinite matrix $\widehat{\boldsymbol{Q}} \in \mathbb{R}^{(n+1)\times(n+1)}$ such that*

$$\rho_i = \boldsymbol{\varphi}^T(\omega_i) \cdot \widehat{\boldsymbol{Q}} \cdot \boldsymbol{\varphi}(\omega_i), \tag{2.104}$$

for all the $2n+1$ points $\omega_i \in \Omega$.

Proof The nth order polynomial $\boldsymbol{\varphi}^T(\omega)\widehat{\boldsymbol{Q}}\boldsymbol{\varphi}(\omega)$ has the values ρ_i for the $2n+1$ frequencies $\omega_i \in \Omega$ and so is identical to $R(\omega)$ (which satisfies the relations (2.102)). Due to (2.103), it results that $R(\omega) = \boldsymbol{\chi}^T(\omega)\boldsymbol{C}^T\widehat{\boldsymbol{Q}}\boldsymbol{C}\boldsymbol{\varphi}(\omega)$, which is nonnegative if and only if $\widehat{\boldsymbol{Q}} \succeq 0$. ■

In principle, any basis $\boldsymbol{\varphi}(\omega)$ and any set of points Ω may be used. However, some choices are more appealing by offering a simple interpretation of some elements of the matrix $\widehat{\boldsymbol{Q}}$. One interesting basis is given by the Dirichlet kernel

$$D_{\tilde{n}}(\omega) = \frac{1}{2\tilde{n}+1} \sum_{k=-\tilde{n}}^{\tilde{n}} \mathrm{e}^{-jk\omega} = \frac{1}{2\tilde{n}+1} \frac{\sin \frac{(2\tilde{n}+1)\omega}{2}}{\sin \frac{\omega}{2}}. \tag{2.105}$$

Denote $\tau = 2\pi/(2\tilde{n}+1)$. We note that

$$D_{\tilde{n}}(\ell\tau) = \begin{cases} 1, & \text{if } \ell = 0, \\ 0, & \text{if } \ell \in \mathbb{Z} \setminus \{0\}. \end{cases} \tag{2.106}$$

This means that the $2\tilde{n}+1$ polynomials from the vector

$$\boldsymbol{\varphi}(\omega) = [D_{\tilde{n}}(\omega + \tilde{n}\tau) \ \ldots \ D_{\tilde{n}}(\omega) \ \ldots \ D_{\tilde{n}}(\omega - \tilde{n}\tau)]^T \tag{2.107}$$

form of basis for the space of $\tilde{n}$th order trigonometric polynomials. Moreover, any $\tilde{n}$th order polynomial $S(\omega)$ can be expressed as

$$S(\omega) = \sum_{\ell=-\tilde{n}}^{\tilde{n}} S(\ell\tau) D_{\tilde{n}}(\omega - \ell\tau). \tag{2.108}$$

(It is clear that the above equality holds for the points $\ell\tau$; a dimensionality argument shows that it holds everywhere.)

We can use the basis (2.107) in the representation (2.104). The points $\ell\tau$, with $\ell = -\tilde{n} : \tilde{n}$, are very good candidates for the set Ω (note that other $2\tilde{n}$ points are needed to complete the set). Since $\boldsymbol{\varphi}(\ell\tau)$ is a unit vector, it results immediately that

$$R(\ell\tau) = \widehat{q}_{\ell+\tilde{n},\ell+\tilde{n}}, \quad \ell = -\tilde{n} : \tilde{n}, \tag{2.109}$$

i.e., the diagonal elements of the Gram matrix $\widehat{\boldsymbol{Q}}$ are equal to the values of the polynomial in the given points.

Example 2.34 Let us take $\tilde{n} = 1$. Since $\tau = 2\pi/3$, the vector (2.107) is

$$\boldsymbol{\varphi}(\omega) = [D_1(\omega + \tfrac{2\pi}{3}) \ D_1(\omega) \ D_1(\omega - \tfrac{2\pi}{3})]^T,$$

with

$$\begin{aligned} D_1(\omega) &= \tfrac{1}{3}(1 + 2\cos\omega), \\ D_1(\omega + \tfrac{2\pi}{3}) &= \tfrac{1}{3}(1 - \cos\omega - \sqrt{3}\sin\omega), \\ D_1(\omega - \tfrac{2\pi}{3}) &= \tfrac{1}{3}(1 - \cos\omega + \sqrt{3}\sin\omega). \end{aligned}$$

It results that the relation (2.103) becomes

$$\begin{bmatrix} D_1(\omega + \frac{2\pi}{3}) \\ D_1(\omega) \\ D_1(\omega - \frac{2\pi}{3}) \end{bmatrix} = \frac{1}{3} \begin{bmatrix} 1 & -1 & -\sqrt{3} \\ 1 & 2 & 0 \\ 1 & -1 & \sqrt{3} \end{bmatrix} \begin{bmatrix} 1 \\ \cos\omega \\ \sin\omega \end{bmatrix}.$$

We now represent the polynomial (2.75) using the basis (2.107), i.e., in the form $R(\omega) = \boldsymbol{\varphi}^T(\omega)\widehat{\boldsymbol{Q}}\boldsymbol{\varphi}(\omega)$. Using (2.76), the Gram matrices satisfying this relation have the form

$$\widehat{\boldsymbol{Q}} = \boldsymbol{C}^{-T}\boldsymbol{Q}\boldsymbol{C} = \begin{bmatrix} 4+2\sqrt{3} & -5-\frac{\sqrt{3}}{2}+\alpha & -\frac{7}{2}+\alpha \\ -5-\frac{\sqrt{3}}{2}+\alpha & 19 & -5+\frac{\sqrt{3}}{2}+\alpha \\ -\frac{7}{2}+\alpha & -5+\frac{\sqrt{3}}{2}+\alpha & 4-2\sqrt{3} \end{bmatrix},$$

where α is a free parameter (equal to $3q_{00}/2$, where q_{00} is the parameter from (2.76)). It is easy to see that the diagonal entries of the matrix $\widehat{\boldsymbol{Q}}$ are, in order, the values $R(-2\pi/3)$, $R(0)$ and $R(2\pi/3)$. ■

Problem (*Min_poly_value*) Using the parameterization (2.104), the minimum value of the polynomial $R(\omega)$ can be found by solving the SDP problem

$$\begin{aligned} \mu^\star = \max_{\mu}\ & \mu \\ \text{s.t.}\ & R(\omega_i) - \mu = \boldsymbol{\varphi}^T(\omega_i)\widehat{\boldsymbol{Q}}\boldsymbol{\varphi}(\omega_i),\ i = 1:2n+1 \\ & \widehat{\boldsymbol{Q}} \succeq 0 \end{aligned} \tag{2.110}$$

The difference with respect to (2.80) is that the equality constraints are defined using polynomial values (and not coefficients). We note that the constraints can be written in the equivalent form

$$R(\omega_i) - \mu = \text{tr}[\boldsymbol{A}_i\widehat{\boldsymbol{Q}}],\ \text{ with } \boldsymbol{A}_i = \boldsymbol{\varphi}(\omega_i)\boldsymbol{\varphi}^T(\omega_i). \tag{2.111}$$

The rank of the matrices $\boldsymbol{A}_i$ is 1. (Moreover, if the Dirichlet kernel is used for generating a basis (2.107) as discussed above, some of these matrices have only one diagonal element equal to 1, the others being zero.) ■

Remark 2.35 If the polynomial $R(z)$ has real coefficients, then $n+1$ values (2.102) are sufficient for describing it uniquely. Moreover, we can use the Gram-pair representation (2.91) to say that $R(z)$ is nonnegative if and only if there exist $\boldsymbol{Q} \succeq 0$ and $\boldsymbol{S} \succeq 0$ such that

$$\rho_i = \boldsymbol{\chi}_c^T(\omega_i)\boldsymbol{Q}\boldsymbol{\chi}_c(\omega_i) + \boldsymbol{\chi}_s^T(\omega_i)\boldsymbol{S}\boldsymbol{\chi}_s(\omega_i),\ i = 1:n+1. \tag{2.112}$$

Moreover, the bases $\boldsymbol{\chi}_c(\omega)$ and $\boldsymbol{\chi}_s(\omega)$ can be replaced by (different) linear combinations of themselves. ■

2.10 Mixed Representations

This section presents a new parameterization of the coefficients of a nonnegative polynomial, using ideas from interpolation representations to make a connection with discrete transforms. We start with a general presentation, going then to particular cases. Let $R \in \mathbb{C}_n[z]$ be a nonnegative trigonometric polynomial. Then, there exists $\boldsymbol{Q} \in \mathbb{C}^{(n+1)\times(n+1)}$, $\boldsymbol{Q} \succeq 0$, such that

$$R(\omega) = \boldsymbol{\varphi}_Q^H(\omega)\boldsymbol{Q}\boldsymbol{\varphi}_Q(\omega), \tag{2.113}$$

where $\boldsymbol{\varphi}_Q(\omega)$ is a basis vector, e.g., like in (2.73). Consider a set of N frequency points ω_i, with sufficiently large N. Let $\boldsymbol{x} \in \mathbb{C}^M$ be a vector representing the polynomial, such that $R(\omega_i) = \boldsymbol{\varphi}_R^T(\omega_i)\boldsymbol{x}$, where $\boldsymbol{\varphi}_R(\omega)$ is a vector of trigonometric functions (another basis vector). The equality

$$\boldsymbol{\varphi}_R^T(\omega_i)\boldsymbol{x} = \boldsymbol{\varphi}_Q^H(\omega_i)\boldsymbol{Q}\boldsymbol{\varphi}_Q(\omega_i),\ i = 1 : N, \tag{2.114}$$

can be written as

$$\boldsymbol{A}\boldsymbol{x} = \operatorname{diag}(\boldsymbol{B}^H\boldsymbol{Q}\boldsymbol{B}), \tag{2.115}$$

where $\boldsymbol{A} \in \mathbb{C}^{N\times M}$ and $\boldsymbol{B} \in \mathbb{C}^{(n+1)\times N}$ are given by

$$\boldsymbol{A} = \begin{bmatrix} \vdots \\ \boldsymbol{\varphi}_R^T(\omega_i) \\ \vdots \end{bmatrix}, \quad \boldsymbol{B} = [\ldots\ \boldsymbol{\varphi}_Q(\omega_i)\ \ldots]. \tag{2.116}$$

We assume that the matrix $\boldsymbol{A}$ has full column rank, which happens when N is large enough ($N \geq M$ anyway). Denoting $\boldsymbol{A}^{\#}$ the pseudoinverse of $\boldsymbol{A}$, the relation (2.115) becomes

$$\boldsymbol{x} = \boldsymbol{A}^{\#}\operatorname{diag}(\boldsymbol{B}^H\boldsymbol{Q}\boldsymbol{B}), \tag{2.117}$$

which is the desired parameterization. As the relation (2.117) is fairly abstract, we will see immediately its (probably) simplest particular case. In any case, it is apparent that this parameterization is useful if the vector $\boldsymbol{x}$ contains the coefficients of the polynomial and the matrices $\boldsymbol{A}^{\#}$, $\boldsymbol{B}$ have "nice" properties.

2.10.1 Complex Polynomials and the DFT

We take the points $\omega_i = 2\pi i/N$, $i = 0 : N-1$, with $N \geq 2n+1$. Let $M = N$ and the vector of parameters be

$$\boldsymbol{x} = [r_0\, r_1\, \ldots\, r_n\, 0\, \ldots\, 0\, r_{-n}\, \ldots\, r_{-1}]^T. \tag{2.118}$$

Since the polynomial is Hermitian, we are interested only in the first $n+1$ elements of $\boldsymbol{x}$, which form the vector $\boldsymbol{r}$. Taking

$$\boldsymbol{\varphi}_R^T(\omega) = [1\ \mathrm{e}^{-j\omega}\ \ldots\ \mathrm{e}^{-j(N-1)\omega}],$$

the matrix

$$\boldsymbol{A} = \left[\mathrm{e}^{-j\frac{2\pi\ell i}{N}}\right]_{\ell,i=0:N-1} \tag{2.119}$$

is the length-N DFT matrix. We split $\boldsymbol{A} = [\boldsymbol{W}\ \boldsymbol{A}_2]$, where $\boldsymbol{W}$ contains the first $n+1$ columns of $\boldsymbol{A}$. We take $\boldsymbol{\varphi}_Q(\omega) = \boldsymbol{\psi}_n(\mathrm{e}^{j\omega})$, i.e., the standard basis from (2.1). It can be seen immediately that $\boldsymbol{B} = \boldsymbol{W}^H$. Since $\boldsymbol{A}$ (and in particular $\boldsymbol{W}$) has orthogonal columns, it follows that

$$\boldsymbol{W}^H\boldsymbol{A} = \frac{1}{N}[\boldsymbol{I}\ \ \boldsymbol{0}].$$

By multiplying (2.115) with $\boldsymbol{W}^H$, we obtain the parameterization

$$\boldsymbol{r} = \frac{1}{N}\boldsymbol{W}^H\mathrm{diag}(\boldsymbol{W}\boldsymbol{Q}\boldsymbol{W}^H) \tag{2.120}$$

of the coefficients of a nonnegative polynomial in terms of a positive semidefinite matrix.

Remark 2.36 Since we have used the same basis for the Gram matrix expression of $R(\omega)$, i.e., $\boldsymbol{\varphi}_Q(\omega) = \boldsymbol{\psi}_n(\mathrm{e}^{j\omega})$, it results that (2.120) is identical with the trace parameterization (2.6). (The mapping between the elements of the Gram matrix and the coefficients of the polynomial is linear.) See **P** 2.14 for an explicit proof. However, the parameterization (2.120) can be used directly in fast algorithms, as discussed in Sect. 2.11. ■

2.10.2 Cosine Polynomials and the DCT

We consider now polynomials $R(z)$ with *real* coefficients; for simplicity, we look only at the even degree case, $n = 2\tilde{n}$. Using the Gram-pair parameterization (2.112), it results similarly to (2.115) that $R(z)$ is nonnegative if there exist $\boldsymbol{Q} \succeq 0$ and $\boldsymbol{S} \succeq 0$ such that

$$\boldsymbol{A}\boldsymbol{x} = \mathrm{diag}(\boldsymbol{B}_1^T\boldsymbol{Q}\boldsymbol{B}_1) + \mathrm{diag}(\boldsymbol{B}_2^T\boldsymbol{S}\boldsymbol{B}_2), \tag{2.121}$$

where $\boldsymbol{A}$ is defined as in (2.116) and

$$\boldsymbol{B}_1 = [\ldots\ \boldsymbol{\chi}_c(\omega_i)\ \ldots], \quad \boldsymbol{B}_2 = [\ldots\ \boldsymbol{\chi}_s(\omega_i)\ \ldots].$$

We take the points $\omega_i = \pi i/(N-1)$, $i = 0 : N-1$, with $N \geq n+1$. Let $M = N$ and the vector of parameters be

$$\boldsymbol{x} = [r_0 \; 2r_1 \; \ldots \; 2r_n \; 0 \; \ldots \; 0]^T \in \mathbb{R}^N. \tag{2.122}$$

With

$$\boldsymbol{\varphi}_R^T(\omega) = [1 \;\; \cos\omega \; \ldots \; \cos(N-1)\omega],$$

the matrix $\boldsymbol{A}$ is

$$\boldsymbol{A} = \left[\cos\frac{\pi \ell i}{N-1}\right]_{\ell,i=0:N-1}. \tag{2.123}$$

We remark that, denoting $\boldsymbol{D} = \mathrm{diag}(1/2, 1, \ldots, 1, 1/2)$, the matrix $\boldsymbol{AD}$ is the DCT-I transform. Moreover, the inverse of $\boldsymbol{A}$ is

$$\boldsymbol{A}^{-1} = \frac{2}{N-1}\boldsymbol{DAD}.$$

We denote $\boldsymbol{W} \in \mathbb{R}^{N\times(n+1)}$ the first $n+1$ columns of $\boldsymbol{A}^{-1}$ (which are also its first rows, as the matrix is symmetric), and so $\boldsymbol{W}^T\boldsymbol{A} = [\boldsymbol{I} \; \boldsymbol{0}]$. By the choice of frequency points, the other constant matrices from (2.121) are

$$\boldsymbol{B}_1 = \left[\cos\frac{\pi \ell i}{N-1}\right]_{\ell=0:n,i=0:N-1}, \quad \boldsymbol{B}_2 = \left[\sin\frac{\pi \ell (i+1)}{N-1}\right]_{\ell=0:n-1,i=0:N-1}.$$

With these notations, by multiplication with $\boldsymbol{W}^T$ in (2.121), we obtain the parameterization

$$\begin{bmatrix} r_0 \\ 2r_1 \\ \vdots \\ 2r_n \end{bmatrix} = \boldsymbol{W}^T \left(\mathrm{diag}(\boldsymbol{B}_1^T\boldsymbol{QB}_1) + \mathrm{diag}(\boldsymbol{B}_2^T\boldsymbol{SB}_2)\right). \tag{2.124}$$

Remark 2.37 For reasons similar to those exposed in Remark 2.36, i.e., identity of bases and linearity, the parameterization (2.124) is identical with the Gram-pair parameterization (2.94). ■

2.11 Fast Algorithms

In this book, the presentation is focused on parameterizations of positive polynomials suited to the use of off-the-shelf SDP libraries. This approach is not only very convenient, as the implementation effort is minimal, but also efficient for polynomials of low or medium order (going to more than 100). Alternatively, SDP algorithms can be tailored to the specific of positive polynomials, obtaining fast methods. This section aims to open the path for the reader interested in such methods.

As mentioned before, a typical SDP problem involving a nonnegative trigonometric polynomial of order n has an $O(n^4)$ complexity, if either the trace (2.6) or Gram-pair (2.94) parameterizations are used. However, these parameterizations are expressed with sparse matrices, which allows a complexity reduction by simply informing the SDP library of the sparseness (actually, the current version of SeDuMi assumes that all matrices are sparse). So, in this case, the constant hidden by the $O(\cdot)$ notation is relatively small.

Fast methods have an $O(n^3)$ complexity. The identity between the dual cone (2.32) and the space of positive semidefinite Toeplitz matrices, presented in Sect. 2.4, allows the fast computation of the Hessian and gradient of the barrier function, required by interior point methods for solving SDP problems. The method from [2] uses displacement rank techniques. The method from [3] uses the Levinson–Durbin algorithm and the DFT. Unfortunately, it appears that the numerical stability of these algorithms limits their use to polynomials of relatively small degrees (less than, e.g., 50, for some applications). So, they may have no significant practical advantage over the algorithms based on the trace or Gram-pair parameterization, which are robust and for which the only limitation on the degree of the polynomials appears to be due mostly to the time necessary to obtain the solution and possibly also to memory requirements.

Another fast method is based on the interpolation representation presented in the previous section. The fact that the constant matrices appearing in constraints such as (2.111) have rank equal to one can be used for building fast algorithms, using a dual solver [4]; again, the Hessian of the barrier function can be evaluated with low complexity, precisely $O(n^3)$ operations.

Finally, the method from [5] is based on representations such as (2.120) and (2.124), that can be exploited when solving the Newton equations appearing in interior point methods. Not only the special form of the representations helps in reducing the number of operations, but also the fact that matrix multiplication can be sped up via the FFT, due to the connection of the constant matrices from (2.120) and (2.124) with discrete transforms such as the DFT and the DCT. Moreover, it seems that this method does not suffer from numerical stability problems, like the others above.

Typically, for small orders, the fast $O(n^3)$ methods are not faster than the standard $O(n^4)$ methods (especially if sparseness is used). The order n_o for which the fast methods become indeed faster depends on the implementation, the SDP algorithm, the programming and running environments, and the problem solved. From the data available in the literature and the author's experiments, it seems that n_o may be around 100.

2.12 Details and Other Facts

2.12.1 Writing Programs with Positive Trigonometric Polynomials

Solving an optimization problem with positive trigonometric polynomials requires an SDP library. We give here three programs for finding the minimum value of a trigonometric polynomial by solving (2.18). The first uses directly the SDP library SeDuMi [6]. The second works at a higher level, calling the convex optimization software CVX [7], which includes SDP but also other types of convex optimization. CVX calls SeDuMi or SDPT3 [8], depending on user's choice. CVX has the great advantage of expressing the optimization problems in a form very close to the mathematical one. The third program uses Pos3Poly [9], which is a library dedicated to positive polynomials, covering all types and situations described in this book. Pos3Poly is built on top of CVX, taking advantage of the possibility to build convex sets offered by CVX. The parameterization is hidden for the user, who simply works directly with the coefficients of the polynomial.

Table 2.2 contains the SeDuMi program for solving (2.18). The problem needs to be expressed in a standard form, here the equality form, see Appendix A. The polynomial has the form (1.1) and is given through its vector of coefficients $\boldsymbol{r} = [r_0 \ \ldots \ r_n]^T$. The variable `K` contains a description of the optimization variables from (2.18): μ is a free scalar (it may have any real value), while $\tilde{\boldsymbol{Q}}$ is a matrix of size $n' \times n'$ (where $n' = n + 1$ is the number of distinct coefficients of the polynomial). Denoting the variables vector with $\boldsymbol{x} = [\mu \ \mathrm{vec}(\tilde{\boldsymbol{Q}})^T]^T$, the constraints of (2.18) are expressed in SeDuMi as a linear system `Ax=b`; so, the first column of A contains a single nonzero value, i.e., `A(1,1)`, which represent the coefficient of μ in the first constraint from (2.18). The rows of `A` contain, starting with the second column, the vectorized elementary Toeplitz matrices; this is due to the equality $\mathrm{tr}[\boldsymbol{\Theta}_k \tilde{\boldsymbol{Q}}] = \mathrm{vec}(\boldsymbol{\Theta}_k)^T \mathrm{vec}(\tilde{\boldsymbol{Q}})$. Finally, the vector `b` is the right hand side of the constraints of (2.18) and so it is equal to $\boldsymbol{r}$. In SeDuMi, the objective is to minimize $\mathtt{c}^T\mathtt{x}$ and so only the first component of `c` is nonzero and equal to -1, in order to maximize μ; the matrix variable $\tilde{\boldsymbol{Q}}$ does not appear in the objective. If the polynomial is complex, the variable `K.scomplex` specifies the positive semidefinite matrices that are complex, by their indices in the variable `K.s`; in our case, there is only one such variable, the Gram matrix. However, this is not enough; we should specify that the equality constraints should be regarded as equalities of complex numbers; this is done by specifying that the dual variables are complex, with `K.ycomplex`.

The program gives also the solution to problem (2.14), i.e., the most positive Gram matrix associated with the polynomial $R(z)$. The Gram matrix `Q` is computed using relation (2.19), from the solution of (2.18).

Running this program for `r = [6 -3 2]`, which represents the polynomial (2.4), gives the Gram matrix $\boldsymbol{Q}$ and the minimal value μ shown in Examples 2.11 and 2.12, respectively.

Table 2.2 SeDuMi program for solving the SDP problem (2.18)

```
function [m,Q] = minpol1(r)

n = length(r);  % length of the polynomial

K.f = 1;        % one free variable (m)
K.s = [n];      % one pos. semidef. matrix of size nxn (Q)
nrA = n;        % number of equality constraints
                % (one for each coefficient of r)
ncA = 1+n*n;    % number of scalar variables (in m and Q)

if ~isreal(r)   % specify complex data, if this is the case
  K.scomplex = 1;
  K.ycomplex = 1:n;
end

A = sparse(nrA,ncA);
b = r(:);       % right hand term of equality constraints
c = zeros(ncA,1);
c(1) = -1;      % the objective is to maximize m

e = ones(n,1);  % generate Toeplitz elementary matrices
for k = 1:n
  A(k,2:end) = vec( spdiags(e,k-1,n,n) )’;
end
A(1,1) = 1;     % coefficient of m in first constraint

[x,y,info] = sedumi(A,b,c,K); % call SeDuMi

m = x(1);      % recover the variables
Q = mat(x(2:end)) + (m/n)*eye(n);
```

The CVX program is shown in Table 2.3 and is practically self-explanatory. The optimization variables are declared explicitly and the constraints have a very natural expression. In general, it is not needed to put the problem in a standard form, which was very easy for (2.18), but sometimes may be cumbersome.

Finally, the Pos3Poly program is shown in Table 2.4. The function `sos_pol` can be used to declare all kinds of positive polynomials; it has two arguments: The first is a vector containing the degree of the polynomial and the size of the coefficients (here

Table 2.3 CVX program for solving (2.18)

```
function [m,Q] = minpol1_cvx(r)

n = length(r);       % length of the polynomial
cvx_begin
  variable m;
  if ~isreal(r)      % complex data
    variable Q(n,n) complex semidefinite;
  else               % real data
    variable Q(n,n) semidefinite;
  end
  maximize( m )      % variable for the minimum
  subject to         % equality constraints
    m + trace(Q) == r(1);
    e = ones(n,1);
    for k = 2:n
       vec( spdiags(e,k-1,n,n) )' * vec(Q) == r(k);
    end
cvx_end
Q = Q + (m/n)*eye(n);
```

they are scalars, but we will talk later about polynomials with matrix coefficients), and the second is a structure describing the type of the polynomial; in our case, it is useful only when declaring that the polynomial is complex. The output of `sos_pol` is a variable vector representing the positive trigonometric polynomial and containing its coefficients (like `r` contains the coefficients of $R(z)$). Hence, the constraint can be written as a single vector equality. Since Pos3Poly hides the parameterization, we do not have access to the Gram matrix.

It is clear that the programming effort decreases as we go from SeDuMi to CVX and then to Pos3Poly. If the reader intends to solve only a small number of simple problems involving positive polynomials, then CVX might be the easiest way. For those who need to solve many or more difficult problems, or for those who do not want to read this book in detail, Pos3Poly is probably the best choice. SeDuMi (note that Pos3Poly can also work directly on top of SeDuMi, without CVX; read the manual if interested) is hard to recommend for others than those already very experienced with it.

Table 2.4 Pos3Poly program for solving (2.18)

```
function m = minpol1_pos3poly(r)

n = length(r);  % length of the polynomial
r = r(:);       % force column vector
p = [n-1 1];    % degree and coefficients size (scalars)
ptype = [];     % for real data this variable is not necessary
if ~isreal(r)   % complex data
  ptype.complex_coef = 1;
end
cvx_begin
  variable m;   % variable for the minimum
  maximize( m )
  subject to    % equality constraints
    m*eye(n,1) + sos_pol(p, ptype) == r;
cvx_end
```

2.12.2 *Proof of Theorem 2.16*

Consider that the FIR filter $H(z)$ has a white noise $e(\ell)$ (of unit variance) at its input and the output is $y(\ell)$. The state-space model of $H(z)$ is

$$\begin{cases} \boldsymbol{\xi}(\ell+1) = \boldsymbol{\Theta}_1^T \boldsymbol{\xi}(\ell) + \tilde{\boldsymbol{h}} e(\ell), \\ \quad\;\; y(\ell) = \boldsymbol{c}^T \boldsymbol{\xi}(\ell) + h_0 e(\ell), \end{cases} \tag{2.125}$$

where $\boldsymbol{\xi} \in \mathbb{C}^n$ is the vector of states and $\tilde{\boldsymbol{h}}$ and $\boldsymbol{c}$ are defined in (2.47). Denote

$$\boldsymbol{\Xi} = E\{\boldsymbol{\xi}(\ell)\boldsymbol{\xi}^H(\ell)\}$$

the state autocorrelation matrix. Multiplying both sides of the first equation from (2.125) with their Hermitians and taking the average, it results that

$$\boldsymbol{\Xi} = \boldsymbol{\Theta}_1^T \boldsymbol{\Xi} \boldsymbol{\Theta}_1 + \tilde{\boldsymbol{h}}\tilde{\boldsymbol{h}}^H.$$

This is the last relation from (2.50).

Using now the second equation from (2.125), we obtain $r_0 = E\{y(\ell)y^*(\ell)\} = \boldsymbol{c}^T \boldsymbol{\Xi} \boldsymbol{c} + h_0^2$, which is the first relation from (2.50) (remind that we have assumed h_0 to be real).

Finally, combining both equations from (2.125), we get

$$E\{\boldsymbol{\xi}(\ell+1)y^*(\ell)\} = \boldsymbol{\Theta}_1^T \boldsymbol{\Xi} \boldsymbol{c} + h_0 \tilde{\boldsymbol{h}}. \tag{2.126}$$

Rewriting (2.125) for each scalar component of the state vector, we obtain

$$\begin{aligned} y(\ell) &= \boldsymbol{\xi}_{n-1}(\ell) + h_0 e(\ell), \\ \boldsymbol{\xi}_{n-1}(\ell+1) &= \boldsymbol{\xi}_{n-2}(\ell) + h_1 e(\ell), \\ &\vdots \\ \boldsymbol{\xi}_1(\ell+1) &= \boldsymbol{\xi}_0(\ell) + h_{n-1} e(\ell), \\ \boldsymbol{\xi}_0(\ell+1) &= h_n e(\ell) \end{aligned}$$

By substituting successively the expressions of states, these relations are equivalent to

$$\boldsymbol{\xi}_{n-k}(\ell+1) = y(\ell+k) - \sum_{i=0}^{k-1} h_i e(\ell+k-i), \quad k = 1:n.$$

With this, we obtain

$$E\{\boldsymbol{\xi}(\ell+1)y^*(\ell)\} = \tilde{\boldsymbol{r}},$$

which makes (2.126) identical with the second relation from (2.50). Thus, we have shown that all three relations from (2.50) hold. Since they are equivalent to (2.48), (2.49), the proof is ready.

2.12.3 *Proof of Theorem 2.19*

The relation (2.56) can be written as

$$P(t) = \operatorname{tr}[\boldsymbol{\psi}(t) \cdot \boldsymbol{\psi}^T(t) \cdot \boldsymbol{Q}] = \operatorname{tr}[\boldsymbol{\Psi}(t) \cdot \boldsymbol{Q}],$$

where

$$\boldsymbol{\Psi}(t) = \begin{bmatrix} 1 \\ t \\ \vdots \\ t^n \end{bmatrix} [1 \; t \; \dots \; t^n] = \begin{bmatrix} 1 & t & \dots & t^n \\ t & t^2 & \iddots & t^{n+1} \\ \vdots & \iddots & \iddots & \vdots \\ t^n & t^{n+1} & \dots & t^{2n} \end{bmatrix} = \sum_{k=0}^{2n} \boldsymbol{\Upsilon}_k t^k.$$

Combining the last two relations, we obtain

$$P(t) = \sum_{k=0}^{2n} \operatorname{tr}[\boldsymbol{\Upsilon}_k \boldsymbol{Q}] t^k,$$

which proves (2.60).

2.12.4 Proof of Theorem 2.21

If $\boldsymbol{Q} \succeq 0$ (or $\boldsymbol{Q} \succ 0$) exists such that (2.60) holds, then it results directly from (2.56) that $P(t) \geq 0$ (or $P(t) > 0$), $\forall t \in \mathbb{R}$.

Reciprocally, if $P(t) \geq 0$, then, as stated by Theorem 1.7, the polynomial can be written as

$$P(t) = F^2(t) + G^2(t) = \psi^T(t) \left(\boldsymbol{f}\boldsymbol{f}^T + \boldsymbol{g}\boldsymbol{g}^T \right) \psi(t),$$

which shows that $\boldsymbol{Q} = \boldsymbol{f}\boldsymbol{f}^T + \boldsymbol{g}\boldsymbol{g}^T$ is a (rank-2) Gram matrix associated with $P(t)$.

If $P(t) > 0$, then there exists $\varepsilon > 0$ such that $P_\varepsilon(t) = P(t) - \varepsilon(1 + t^2 + \ldots + t^{2n})$ is nonnegative. Let $\boldsymbol{Q}_\varepsilon \succeq 0$ be a Gram matrix associated with $P_\varepsilon(t)$ (obtained, e.g., as above). Since $\boldsymbol{I}$ is the Gram matrix of the polynomial $1 + t^2 + \ldots + t^{2n}$, it results that $\boldsymbol{Q} = \boldsymbol{Q}_\varepsilon + \varepsilon \boldsymbol{I} \succ 0$ is a Gram matrix associated with $P(t)$.

2.13 Bibliographical and Historical Notes

The trace parameterization of trigonometric polynomials (Theorem 2.5) and its use as optimization tool have been proposed independently by several researchers [2, 3, 10–12], starting from different applications. Many of these works were motivated by the high complexity of the KYP lemma parameterization (see Sect. 2.5) of positive polynomials proposed in [13] (for FIR filter design), [14] (for MA estimation), [15] (for compaction filter design), and [16] (for the design of orthogonal pulse shapes for communication). The jump from an $O(n^6)$ complexity to $O(n^4)$ allowed a much higher range of problems to be solved.

The Toeplitz quadratic optimization problem discussed in Sect. 2.3 has been analyzed in SDP terms in [17]. Some extensions can be found in [18]. The dual-cone formulation is a simple dualization exercise; the presentation from Sect. 2.4 is taken from [3]; an excellent lecture on convex optimization and, among many others, dual cones is [19]. For the discrete-time version of the KYP lemma and other results regarding positive real systems, we recommend [20, 21].

Spectral factorization using SDP has been proposed by several authors. The proof of Theorem 2.15 using the Schur complement was presented in [22]. For Robinson's energy delay property see [23], problem 5.66. The spectral factorization method based on the Riccati equation appeared in [24]; useful information can be found in [25] (including connections with Kalman filtering). Other spectral factorization algorithms are presented in Appendix B.

The Gram-pair factorization from Sect. 2.8.3 and the real Gram representation of polynomials with complex coefficients from Theorem 2.25 have appeared in [5], in their equivalent forms from Sect. 2.10. The explicit representations from Theorems 2.27 and 2.30 have been derived here in the style of other Gram parameterizations. The idea of using interpolation representations, as presented in Sect. 2.9, appeared first in [4].

Problems

P 2.1 Are there nonnegative polynomials $R(z)$ for which the set $\mathcal{G}(R)$ of associated Gram matrices contains a single positive semidefinite matrix?

P 2.2 (problems VI.50, VI.51 [26]) The polynomial $R \in \mathbb{C}_n[z]$ is nonnegative and has the free coefficient $r_0 = 1$.

(a) Show that $R(\omega) \le n + 1$.

(b) Show that $|r_n| \le 1/2$.

Hint: use the Gram matrix representation (2.6).

(a) A Gram matrix $\boldsymbol{Q} \succeq 0$ has nonnegative eigenvalues λ_i, $i = 1 : n + 1$. Since $\sum \lambda_i = \text{tr}\,\boldsymbol{Q} = r_0 = 1$, it results that $\max \lambda_i \le 1$. Note also that $\|\boldsymbol{\psi}(\omega)\|^2 = n + 1$. It results that $R(\omega) = \boldsymbol{\psi}^H(\omega)\,\boldsymbol{Q}\boldsymbol{\psi}(\omega) \le \|\boldsymbol{\psi}(\omega)\|^2 \max \lambda_i = n + 1$.

(b) The determinant of the 2×2 matrix containing the corner elements of $\boldsymbol{Q}$ is nonnegative. The sum of the diagonal elements of this 2×2 matrix is less than 1 and the other two elements are r_n and r_n^*.

P 2.3 (problems VI.57, VI.58 [26]) The polynomial $R \in \mathbb{C}_n[z]$ has the free coefficient $r_0 = 0$.

(a) Show that $R(\omega)$ cannot have the same sign for all values of ω, unless it is identically zero.

(b) Let $-m$ and M be the minimum and maximum, respectively, of the values $R(\omega)$ (note that $m \ge 0$, $M \ge 0$). Show that $M \le nm$, $m \le nM$.

Hint: any Gram matrix has $\text{tr}\,\boldsymbol{Q} = 0$ and hence both positive and negative eigenvalues, unless it is the null matrix.

P 2.4 Are there polynomials $P \in \mathbb{R}_{2n}[t]$ with a single associated Gram matrix? What is their degree?

P 2.5 Let $P \in \mathbb{R}_{2n}[t]$ be a positive polynomial. Let $\mu^\star$ be the minimum value of $P(t)$, i.e., the optimal value of the SDP problem (2.62). Let $\lambda^\star$ be the smallest eigenvalue of the most positive Gram matrix associated with $P(t)$, i.e., the optimum of (2.64). Show that $\mu^\star \ge \lambda^\star$.

Show that there exist polynomials for which $\mu^\star = \lambda^\star$. Hint: think at polynomials with nonzero coefficients only for even powers of t.

P 2.6 Let $R \in \mathbb{R}_n[z]$ be the polynomial whose coefficients are $r_k = n + 1 - k$. (This is the triangular, or Bartlett, window.) Show that $R(\omega) \ge 0$ by finding a positive semidefinite Gram matrix associated with $R(z)$.

P 2.7 Consider the quadratic optimization problem (2.24), in which the matrices $\boldsymbol{A}_\ell$, $\ell = 0 : L$, are Hankel (and not Toeplitz). Why cannot the problem (2.24) be solved from the solution of an SDP problem (2.31) with Hankel matrices (and so there is no analogous for real polynomials of the algorithm presented in Sect. 2.3 for trigonometric polynomials)?

P 2.8 (LMI form of Theorem 1.15) The polynomial $R \in \mathbb{C}_n[z]$ is nonnegative on the interval $[\alpha, \beta] \subset (-\pi, \pi)$ if and only if there exist positive semidefinite matrices $\boldsymbol{Q}_1 \in \mathbb{C}^{(n+1)\times(n+1)}$ and $\boldsymbol{Q}_2 \in \mathbb{C}^{n\times n}$ such that

$$r_k = \text{tr}[\boldsymbol{\Theta}_k \boldsymbol{Q}_1] + \text{tr}[(d_1 \boldsymbol{\Theta}_{k-1} + d_0 \boldsymbol{\Theta}_k + d_1^* \boldsymbol{\Theta}_{k+1}) \boldsymbol{Q}_2].$$

Here, d_0 and d_1 are the coefficients of the polynomial (1.34). Also, in the argument of the second trace operator, we use the notation convention that $\boldsymbol{\Theta}_k = 0$ if $k > n - 1$.

P 2.9 Let $R \in \mathbb{C}_{2\tilde{n}}[z]$ be a trigonometric polynomial. Show that the parameterization (2.73) is equivalent to

$$R(z) = \boldsymbol{\phi}^T(z) \cdot \boldsymbol{Q} \cdot \boldsymbol{\phi}(z),$$

with the basis vector

$$\boldsymbol{\phi}(z) = \begin{bmatrix} 1 \\ z + z^{-1} \\ \vdots \\ z^{\tilde{n}} + z^{-\tilde{n}} \\ j(z - z^{-1}) \\ \vdots \\ j(z^{\tilde{n}} - z^{-\tilde{n}}) \end{bmatrix}.$$

Notice the resemblance with the definition (2.56) of Gram matrices for real polynomials.

P 2.10 Show that the problem (2.80) (for computing the minimum value of a complex trigonometric polynomial) is equivalent to finding the most positive matrix $\boldsymbol{Q}$ for which (2.73) holds. Hint: notice that $\|\boldsymbol{\chi}(\omega)\|^2 = \tilde{n} + 1$.

P 2.11 The SDP problem (2.97) computes the minimum value of a trigonometric polynomial $R(z)$ with real coefficients. Denote $\lambda(\boldsymbol{Q})$, $\lambda(\boldsymbol{S})$ the sets of eigenvalues of the matrices $\boldsymbol{Q}$ and $\boldsymbol{S}$, respectively, from the Gram-pair parameterization (2.91) of $R(z)$. Show that (2.97) is equivalent to finding the matrices $\boldsymbol{Q}$ and $\boldsymbol{S}$ for which the smallest eigenvalue from $\lambda(\boldsymbol{Q}) \cup \lambda(\boldsymbol{S})$ is maximum.

Hint: notice that (2.91) is equivalent to

$$R(\omega) = [\boldsymbol{\chi}_c^T(\omega)\ \boldsymbol{\chi}_s^T(\omega)] \begin{bmatrix} \boldsymbol{Q} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{S} \end{bmatrix} \begin{bmatrix} \boldsymbol{\chi}_c(\omega) \\ \boldsymbol{\chi}_s(\omega) \end{bmatrix}$$

and that the vector $[\boldsymbol{\chi}_c^T(\omega)\ \boldsymbol{\chi}_s^T(\omega)]$ has constant norm.

P 2.12 (LMI form of Theorem 1.18) The polynomial $R \in \mathbb{R}_n[z]$, with $n = 2\tilde{n}$, is nonnegative on $[\alpha, \beta] \subset [0, \pi]$ if and only if there exist positive semidefinite matrices $\boldsymbol{Q}_1 \in \mathbb{R}^{(\tilde{n}+1)\times(\tilde{n}+1)}$ and $\boldsymbol{Q}_2 \in \mathbb{R}^{\tilde{n}\times\tilde{n}}$ such that (for brevity, we denote $\cos\alpha = a$, $\cos\beta = b$)

$$r_k = \mathrm{tr}[\boldsymbol{\Phi}_k \boldsymbol{Q}_1] + \mathrm{tr}\Big[\big((-ab - \tfrac{1}{2})\boldsymbol{\Phi}_k + \tfrac{a+b}{2}(\boldsymbol{\Phi}_{k-1} + \boldsymbol{\Phi}_{k+1}) \\ - \tfrac{1}{4}(\boldsymbol{\Phi}_{k-2} + \boldsymbol{\Phi}_{k+2})\big)\, \boldsymbol{Q}_2\Big],$$

where the matrices $\boldsymbol{\Phi}_k$ are defined in (2.95).

Derive a similar result for the case of odd degree n.

P 2.13 Let $R_1(z)$, $R_2(z)$ be two trigonometric polynomials of the same degree. Prove the following:

(a) $R_1(\omega) \geq R_2(\omega)$, $\forall \omega \in [-\pi, \pi]$, if and only if there exist Gram matrices $\boldsymbol{Q}_1$ and $\boldsymbol{Q}_2$, associated with $R_1(z)$ and $R_2(z)$, respectively (i.e., defined as in the trace parameterization (2.6)) such that $\boldsymbol{Q}_1 \succeq \boldsymbol{Q}_2$.

(b) $R_1(\omega) \geq R_2(\omega)$, $\forall \omega \in [-\pi, \pi]$, if and only if there exist Gram pairs $(\boldsymbol{Q}_1, \boldsymbol{S}_1)$ and $(\boldsymbol{Q}_2, \boldsymbol{S}_2)$, associated with $R_1(z)$ and $R_2(z)$, respectively (i.e., defined as in the Gram-pair parameterization (2.94)) such that $\boldsymbol{Q}_1 \succeq \boldsymbol{Q}_2$ and $\boldsymbol{S}_1 \succeq \boldsymbol{S}_2$.

Generalize this kind of results to polynomials that are positive on an interval.

P 2.14 Show that the trace parameterization (2.6) and the DFT parameterization (2.120) of a nonnegative polynomial are identical.

Hint [5]: Denote $\boldsymbol{w}_k$ the k-th column of the DFT matrix (2.119), which is also the k-th column of $\boldsymbol{W}$ from (2.120). It results from (2.120) that

$$\begin{aligned} r_k &= \tfrac{1}{N}\boldsymbol{w}_k^H \mathrm{diag}(\boldsymbol{W}\boldsymbol{Q}\boldsymbol{W}^H) = \tfrac{1}{N}\mathrm{tr}[\mathrm{diag}(\boldsymbol{w}_k^H)\boldsymbol{W}\boldsymbol{Q}\boldsymbol{W}^H] \\ &= \tfrac{1}{N}\mathrm{tr}[\boldsymbol{W}^H \mathrm{diag}(\boldsymbol{w}_k^H)\boldsymbol{W}\boldsymbol{Q}]. \end{aligned}$$

It remains to show that $\boldsymbol{W}^H \mathrm{diag}(\boldsymbol{w}_k^H)\boldsymbol{W} = N\boldsymbol{\Theta}_k$.

P 2.15 It is clear that if any polynomial $R \in \mathbb{C}[z]$ can be written as $R(\omega) = \mathrm{tr}[\boldsymbol{Q}\boldsymbol{P}(\omega)]$, with positive semidefinite matrices $\boldsymbol{Q}$ (which depends on R) and $\boldsymbol{P}(\omega)$ (the same for all polynomials), then it follows that $R(\omega) \geq 0$. Investigate what are the conditions for the reverse implication to hold. Describe the parameterizations from this chapter as particular cases of these conditions.

References

1. M.D. Choi, T.Y. Lam, B. Reznick, Sums of squares of real polynomials. Proc. Symp. Pure Math. **58**(2), 103–126 (1995)
2. Y. Genin, Y. Hachez, Y. Nesterov, P. Van Dooren, Optimization problems over positive pseudopolynomial matrices. SIAM J. Matrix Anal. Appl. **25**(1), 57–79 (2003)
3. B. Alkire, L. Vandenberghe, Convex optimization problems involving finite autocorrelation sequences. Math. Progr. Ser. A **93**(3), 331–359 (2002)
4. J. Löfberg, P.A. Parrilo. From coefficients to samples: a new approach to SOS optimization, in *43rd IEEE Conference on Decision and Control*, Bahamas (2004), pp. 3154–3159
5. T. Roh, L. Vandenberghe, Discrete transforms, semidefinite programming and sum-of-squares representations of nonnegative polynomials. SIAM J. Optim. **16**, 939–964 (2006)
6. J.F. Sturm. Using SeDuMi 1.02, a Matlab toolbox for optimization over symmetric cones. Optim. Methods Softw. **11**:625–653 (1999). http://sedumi.ie.lehigh.edu

7. M. Grant, S. Boyd, *CVX: Matlab Software for Disciplined Convex Programming*, version 2.1 (2014). http://cvxr.com/cvx
8. K.C. Toh, M.J. Todd, R.H. Tütüncü, SDPT3 – a Matlab software package for semidefinite programming. Optim. Meth. Software, **11**:545–581 (1999). http://www.math.nus.edu.sg/mattohkc/sdpt3.html
9. B.C. Şicleru, B. Dumitrescu. POS3POLY – a MATLAB preprocessor for optimization with positive polynomials. Optim. Eng. **14**(2):251–273 (2013). http://www.schur.pub.ro/pos3poly
10. Y. Nesterov, Squared functional systems and optimization problems, in *High Performance Optimiation*, ed. By J.G.B. Frenk, C. Roos, T. Terlaky, S. Zhang (Kluwer Academic, The Netherlands, 2000), pages 405–440
11. B. Dumitrescu, I. Tabuş, P. Stoica, On the parameterization of positive real sequences and MA parameter estimation. IEEE Trans. Signal Proc. **49**(11), 2630–2639 (2001)
12. T.N. Davidson, Z.Q. Luo, J.F. Sturm, Linear matrix inequality formulation of spectral mask constraints with applications to FIR filter design. IEEE Trans. Signal Proc. **50**(11), 2702–2715 (2002)
13. S.P. Wu, S. Boyd, L.Vandenberghe, FIR filter design via semidefinite programming and spectral factorization, in *Proceedings of 35th IEEE Conference on Decision Contr*, vol. 1 (Kobe, Japan, 1996), pp. 271–276
14. P. Stoica, T. McKelvey, J. Mari, MA estimation in polynomial time. IEEE Trans. Signal Process. **48**(7), 1999–2012 (2000)
15. J. Tuqan, P.P. Vaidyanathan, A state space approach to the design of globally optimal FIR energy compaction filters. IEEE Trans. Signal Process. **48**(10), 2822–2838 (2000)
16. T.N. Davidson, Z.Q. Luo, K.M. Wong, Design of orthogonal pulse shapes for communications via semidefinite programming. IEEE Trans. Signal Process. **48**(5), 1433–1445 (2000)
17. B. Dumitrescu, C. Popeea, Accurate computation of compaction filters with high regularity. IEEE Signal Proc. Lett. **9**(9), 278–281 (2002)
18. A. Konar, N.K. Sidiropoulos, Hidden convexity in QCQP with Toeplitz-Hermitian quadratics. IEEE Signal Proc. Lett. **22**(10), 1623–1627 (2015)
19. S. Boyd, L. Vandenberghe, *Convex Optimization* (Cambridge University Press, Cambridge, 2004)
20. B.D.O. Anderson, S. Vongpanitlerd, *Network Analysis and Synthesis* (Prentice Hall, Englewood Cliffs, NJ, 1973)
21. V.M. Popov, *Hyperstability of Control Systems* (Springer, New York, 1973) (Romanian edition 1966)
22. J.W. McLean, H.J. Woerdeman, Spectral factorizations and sums of squares representations via semidefinite programming. SIAM J. Matrix Anal. Appl. **23**(3), 646–655 (2002)
23. A.V. Oppenheim, R.W. Schafer, *Discrete-Time Signal Processing* (Prentice Hall, USA, 1999)
24. B.D.O. Anderson, K.L. Hitz, N.D. Diem, Recursive algorithm for spectral factorization. IEEE Trans. Circ. Syst. **21**(6), 742–750 (1974)
25. A.H. Sayed, T. Kailath, A survey of spectral factorization methods. Numer. Lin. Alg. Appl. **8**, 467–496 (2001)
26. G. Pólya, G. Szegö, *Problems and Theorems in Analysis II* (Springer, New York, 1976)

Chapter 3
Multivariate Polynomials

Abstract Are the notions and results presented in the previous two chapters valid in the multivariate case? The answer is mostly yes, but with some limitations. The notion of Gram matrix is related directly only to sum-of-squares polynomials. Unlike the univariate case, multivariate nonnegative polynomials are not necessarily sum-of-squares. However, *positive* trigonometric polynomials are sum-of-squares, but the degrees of the sum-of-squares factors may be arbitrarily high, at least theoretically. To benefit from the SDP computation machinery, we must relax the framework from nonnegative polynomials to sum-of-squares polynomials (whose factors have bounded degree). The principle of sum-of-squares relaxations, presented in Sect. 3.5, is central to the understanding of this chapter. It resides in the idea that (many interesting) optimization problems with nonnegative polynomials can be approximated with a sequence of problems with sum-of-squares, implemented via SDP. Larger the order of the sum-of-squares, better the approximation, but higher the complexity. This chapter is rather long, so here is an outline of its content. The first three sections present some important properties of nonnegative and sum-of-squares multivariate polynomials. The Gram matrix (or generalized trace) parameterization of sum-of-squares trigonometric polynomials is introduced in Sect. 3.4. After discussing sum-of-squares relaxations in Sect. 3.5, dealing with sparse polynomials is considered in Sect. 3.6. The similar notions for real polynomials are presented in Sect. 3.7. The connections between pairs of relaxations for trigonometric and real polynomials are investigated in Sect. 3.8. The Gram *pair* parameterization of sum-of-squares trigonometric polynomials is examined in Sect. 3.9; similarly to the univariate case, as discussed in Sect. 2.8.3, the Gram-pair matrices have half the size of the Gram matrix. Finally, in Sect. 3.10, the previous results are generalized for polynomials with matrix coefficients.

3.1 Multivariate Polynomials

We investigate multivariate trigonometric polynomials, in the indeterminate $z = (z_1, \dots, z_d) \in \mathbb{C}^d$. A monomial of degree $\boldsymbol{k} \in \mathbb{Z}^d$ is

$$\boldsymbol{z}^{\boldsymbol{k}} = z_1^{k_1} z_2^{k_2} \dots z_d^{k_d}.$$

B. Dumitrescu, *Positive Trigonometric Polynomials and Signal Processing Applications*, Signals and Communication Technology,
DOI 10.1007/978-3-319-53688-0_3

Its total degree is $\sum_{i=1}^{d} |k_i| = \|\boldsymbol{k}\|_1$. A (Hermitian) trigonometric polynomial of degree $\boldsymbol{n}$ is

$$R(\boldsymbol{z}) = \sum_{\boldsymbol{k}=-\boldsymbol{n}}^{\boldsymbol{n}} r_{\boldsymbol{k}} \boldsymbol{z}^{-\boldsymbol{k}}, \quad r_{-\boldsymbol{k}} = r_{\boldsymbol{k}}^*. \tag{3.1}$$

By the notation above, we understand that the sum is taken for all indices $\boldsymbol{k}$ such that $-\boldsymbol{n} \le \boldsymbol{k} \le \boldsymbol{n}$ (these inequalities are understood elementwise).

Example 3.1 The bivariate (2D) polynomial

$$R(z_1, z_2) = 4 + 3(z_1 + z_1^{-1}) + 2(z_1^{-1} z_2 + z_1 z_2^{-1}) + (z_1 z_2 + z_1^{-1} z_2^{-1}) \tag{3.2}$$

has degree (1, 1) and total degree 2. ■

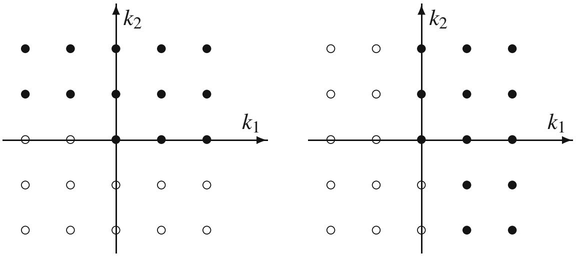
The sets of multivariate polynomials are denoted as in the univariate case, with appropriate bold letters; for example, the set of all polynomials (3.1) is $\mathbb{C}[\boldsymbol{z}]$; if the degree is (at most) a fixed $\boldsymbol{n}$, then the set is $\mathbb{C}_{\boldsymbol{n}}[\boldsymbol{z}]$; we omit from the notation the number of variables, denoted usually d or resulting from the context.

Unlike the univariate case, there is no unique definition of causality. The reason is that $\mathbb{Z}^d$ can be split in several ways into half-spaces. A half-space is a set $\mathcal{H} \subset \mathbb{Z}^d$ such that $\mathcal{H} \cap (-\mathcal{H}) = \{0\}$, $\mathcal{H} \cup (-\mathcal{H}) = \mathbb{Z}^d$, $\mathcal{H} + \mathcal{H} \subset \mathcal{H}$. A standard half-space can be defined recursively. In $\mathbb{Z}$, the standard half-space is $\mathcal{H}_1 = \mathbb{N} = \{0, 1, 2, \ldots\}$. In $\mathbb{Z}^d$, we say that $\boldsymbol{k} \in \mathcal{H}_d$ if $k_d > 0$ or $k_d = 0$ and $(k_1, \ldots, k_{d-1}) \in \mathcal{H}_{d-1}$. By permuting the order of dimensions in the definition above, we obtain $d!$ different half-spaces. We illustrate in Fig. 3.1 the two half-spaces thus obtained in 2D. The filled circles are in $\mathcal{H}$, the empty circles in $-\mathcal{H}$ (the origin is in both sets); the standard half-space is represented in the left side of the figure. There are other ways to build half-spaces, see, e.g., problem **P** 3.1. The causal part of the polynomial (3.1) is

$$R_+(\boldsymbol{z}) = \frac{r_{\boldsymbol{0}}}{2} + \sum_{\boldsymbol{k} \in \mathcal{H}, \boldsymbol{k} \neq \boldsymbol{0}} r_{\boldsymbol{k}} \boldsymbol{z}^{-\boldsymbol{k}}, \tag{3.3}$$

i.e., its support is in the standard half-space $\mathcal{H}$ (the free term is split between the causal and the "anticausal" parts; the latter is defined on $-\mathcal{H}$). From now on, when speaking about the support of a polynomial, by $\boldsymbol{k} \in \mathcal{H}$ we will understand also that

Fig. 3.1 Half-spaces in 2D



$|\boldsymbol{k}| \leq |\boldsymbol{n}|$ (or similar bounds). The number of coefficients of $R(\boldsymbol{z})$ belonging to a half-space is

$$M = \frac{1 + \prod_{i=1}^{d}(2n_i + 1)}{2} \tag{3.4}$$

and is actually the number of distinct coefficients of $R(\boldsymbol{z})$ (considering symmetry).

Example 3.2 The causal part of the polynomial (3.2) is $R_+(z_1, z_2) = 2 + 3z_1^{-1} + 2z_1 z_2^{-1} + z_1^{-1} z_2^{-1}$. ■

Positive orthant polynomials, which are a special class of causal polynomials, are defined by

$$H(\boldsymbol{z}) = \sum_{\boldsymbol{k}=\boldsymbol{0}}^{\boldsymbol{n}} h_{\boldsymbol{k}} \boldsymbol{z}^{-\boldsymbol{k}}. \tag{3.5}$$

On the unit d-circle $\mathbb{T}^d$, i.e., when

$$\boldsymbol{z} = \mathrm{e}^{j\boldsymbol{\omega}} = (\mathrm{e}^{j\omega_1}, \ldots, \mathrm{e}^{j\omega_d}), \quad \boldsymbol{\omega} \in [-\pi, \pi]^d, \tag{3.6}$$

the polynomial $R(\boldsymbol{z})$ has real values. If the coefficients are real, then

$$R(\boldsymbol{\omega}) \overset{\Delta}{=} R(\mathrm{e}^{j\boldsymbol{\omega}}) = 2\mathrm{Re}[R_+(\mathrm{e}^{j\boldsymbol{\omega}})] = r_{\boldsymbol{0}} + 2 \sum_{\boldsymbol{k} \in \mathcal{H}, \boldsymbol{k} \neq \boldsymbol{0}} r_{\boldsymbol{k}} \cos \boldsymbol{k}^T \boldsymbol{\omega}. \tag{3.7}$$

If $R(\boldsymbol{z})$ has complex coefficients, we can write

$$R(\boldsymbol{z}) = \sum_{\boldsymbol{k}=-\boldsymbol{n}}^{\boldsymbol{n}} (u_{\boldsymbol{k}} + j v_{\boldsymbol{k}}) \boldsymbol{z}^{-\boldsymbol{k}} = U(\boldsymbol{z}) + jV(\boldsymbol{z}), \tag{3.8}$$

where $U(\boldsymbol{z})$ is a symmetric polynomial, while $V(\boldsymbol{z})$ is antisymmetric, i.e., $v_{-\boldsymbol{k}} = -v_{\boldsymbol{k}}$ (in particular, $v_{\boldsymbol{0}} = 0$). On the unit d-circle, the polynomial (3.8) becomes

$$R(\boldsymbol{\omega}) = u_{\boldsymbol{0}} + 2 \sum_{\boldsymbol{k} \in \mathcal{H}, \boldsymbol{k} \neq \boldsymbol{0}} u_{\boldsymbol{k}} \cos \boldsymbol{k}^T \boldsymbol{\omega} + 2 \sum_{\boldsymbol{k} \in \mathcal{H}, \boldsymbol{k} \neq \boldsymbol{0}} v_{\boldsymbol{k}} \sin \boldsymbol{k}^T \boldsymbol{\omega}. \tag{3.9}$$

For $\boldsymbol{z} \in \mathbb{T}^d$, the transformation from (3.1) to (3.9) is also made by

$$\cos \boldsymbol{k}^T \boldsymbol{\omega} = \frac{\boldsymbol{z}^{\boldsymbol{k}} + \boldsymbol{z}^{-\boldsymbol{k}}}{2}, \qquad \sin \boldsymbol{k}^T \boldsymbol{\omega} = \frac{\boldsymbol{z}^{\boldsymbol{k}} - \boldsymbol{z}^{-\boldsymbol{k}}}{2j}. \tag{3.10}$$

Finally, let us define the set $\mathbb{R}[\boldsymbol{t}]$ of polynomials

$$P(\boldsymbol{t}) = \sum_{\boldsymbol{k}=\boldsymbol{0}}^{\boldsymbol{n}} p_{\boldsymbol{k}} \boldsymbol{t}^{\boldsymbol{k}} \tag{3.11}$$

of variable $\boldsymbol{t} \in \mathbb{R}^d$, with real coefficients. We will discuss later the connections between trigonometric polynomials and real polynomials. For the moment, let us note that, unlike the univariate case, there is no one-to-one correspondence between $\mathbb{R}_{\boldsymbol{n}}[\boldsymbol{z}]$ and $\mathbb{R}_{\boldsymbol{n}}[\boldsymbol{t}]$. A transformation such as (1.5), i.e., $t_i = \cos\omega_i$, is not enough, since the expression of $R(\boldsymbol{\omega})$ from (3.7) contains usually also sin terms (that appear when expressing $\cos \boldsymbol{k}^T\boldsymbol{\omega}$ with elementary cos and sin terms). The simplest example is $R(z_1, z_2) = 0.5(z_1^{-1}z_2^{-1} + z_1z_2)$, for which $R(\boldsymbol{\omega}) = \cos(\omega_1 + \omega_2) = \cos\omega_1 \cos\omega_2 - \sin\omega_1 \sin\omega_2$. By putting $t_i = \cos\omega_i$ and $t_{i+d} = \sin\omega_i$, we can express any d-variate trigonometric polynomial as a $2d$-variate real polynomial; however, many such real polynomials will correspond to the same trigonometric polynomial. Nevertheless, we will see that such a transformation is useful.

3.2 Sum-of-Squares Multivariate Polynomials

We are especially interested by trigonometric polynomials (3.1) that are nonnegative (i.e., $R(\boldsymbol{\omega}) \geq 0$) or positive ($R(\boldsymbol{\omega}) > 0$) on the unit d-circle. Factorable and sum-of-squares polynomials are defined as in the univariate case.

Definition 3.3 A trigonometric polynomial $R(\boldsymbol{z})$ defined as in (3.1) is *factorable* if it can be written as

$$R(\boldsymbol{z}) = H(\boldsymbol{z})H^*(\boldsymbol{z}^{-1}) \tag{3.12}$$

and *sum-of-squares* if it can be written in the form

$$R(\boldsymbol{z}) = \sum_{\ell=1}^{\nu} H_\ell(\boldsymbol{z})H_\ell^*(\boldsymbol{z}^{-1}), \tag{3.13}$$

where $H(\boldsymbol{z})$ and $H_\ell(\boldsymbol{z})$, $\ell = 1 : \nu$, are positive orthant polynomials and ν is a positive integer. ■

On the unit d-circle, a sum-of-squares polynomial has the expression

$$R(\boldsymbol{\omega}) = \sum_{\ell=1}^{\nu} |H_\ell(\boldsymbol{\omega})|^2 \tag{3.14}$$

and so is nonnegative. We have seen that the sets of nonnegative, factorable, and sum-of-squares *univariate* trigonometric polynomials are identical. For multivariate polynomials, this is no longer the case. The relations between the above-defined types of polynomials can be depicted by the following diagram, in which the inclusions are *strict*.

$$\begin{array}{c}\{\text{factorable}\}\\ \neq \\ \{\text{positive}\}\end{array} \subset \{\text{sum-of-squares}\} \subset \{\text{nonnegative}\} \tag{3.15}$$

Some of the relations are trivial. For instance, a factorable polynomial is a sum-of-squares with a single term. The following example shows that the inclusion is strict and suggests that the set of factorable polynomials is "small" in the set of nonnegative polynomials. So, in general, there is no spectral factorization of nonnegative multivariate polynomials and hence no correspondence for Theorem 1.1.

Example 3.4 Let us consider the polynomial

$$R(z) = 4 + (z_1 + z_1^{-1}) + (z_1 z_2 + z_1^{-1} z_2^{-1}). \tag{3.16}$$

Since $R(\boldsymbol{\omega}) = 4 + 2\cos\omega_1 + 2\cos(\omega_1 + \omega_2)$, it is clear that $R(z)$ is nonnegative. Moreover, the polynomial can be expressed as the sum-of-squares

$$R(z) = (1 + z_1)(1 + z_1^{-1}) + (1 + z_1 z_2)(1 + z_1^{-1} z_2^{-1}).$$

However, it can be easily proved that the polynomial is not factorable; see problem **P** 3.3. A free term larger than 4 in (3.16) makes the polynomial positive, but still not factorable. ■

To stress its importance, the most significant inclusion from the diagram (3.15) is stated formally here as a theorem.

Theorem 3.5 *Any polynomial (3.1) positive on the unit d-circle is sum-of-squares.*

The proof is delayed to Chap. 4, where it will turn out that the above result is a particular case of Theorem 4.11.

Remark 3.6 It is clear that sum-of-squares can be zero on the unit d-circle and so the inclusion $\{\text{positive}\} \subset \{\text{sum-of-squares}\}$ is strict.

Less trivial is an important aspect of Theorem 3.5, which makes an extra difference with respect to the univariate case. If $R(z)$ is a positive polynomial of degree $\boldsymbol{n}$ and so it can be written as in (3.13), then it is possible that the degrees of some factors $H_\ell(z)$ are *greater* than $\boldsymbol{n}$. By "greater", we mean that for some ℓ, we have the relation $\boldsymbol{m} = \deg H_\ell > \deg R = \boldsymbol{n}$. (Of course, it is possible that only the total degree of H_ℓ is greater than the total degree of R, or only $m_i > n_i$, for some $i \in 1 : d$.) Moreover, the degrees of the factors may be arbitrarily large. Here is an example of positive polynomial whose sum-of-squares expression has factors with degree larger than $\deg R$. ■

Example 3.7 ([1]) Consider the polynomial

$$\begin{aligned} R(z_1, z_2) = &\tfrac{7}{2} + (z_2 + z_2^{-1}) + \tfrac{1}{4}(z_2^2 + z_2^{-2}) \\ &+ (z_1 + z_1^{-1})[1 + (z_2 + z_2^{-1}) + \tfrac{1}{2}(z_2^2 + z_2^{-2})] \\ &+ (z_1^2 + z_1^{-2})[\tfrac{1}{4} + \tfrac{1}{2}(z_2 + z_2^{-1}) - \tfrac{1}{8}(z_2^2 + z_2^{-2})]. \end{aligned} \tag{3.17}$$

We notice that $\deg R = (2, 2)$. It can be shown that $R(\boldsymbol{\omega}) \geq 0$; the polynomial is obtained from a positive polynomial with real coefficients, by the transformation described in Sect. 3.11.1; the minimum value on the unit circle is actually zero and is obtained, for example, when $z_1 = z_2 = -1$. However, there is no sum-of-squares decomposition of $R(\boldsymbol{z})$ with factors with $\deg H_\ell \leq \deg R$. We will show later, in Example 3.19, that such a decomposition cannot be obtained even for $R(\boldsymbol{z}) + \alpha$, for any $0 < \alpha \leq 0.01$ (i.e., for a positive polynomial).

Nevertheless, we can express $R(\boldsymbol{z})$ as the sum-of-squares (3.13), with $\nu = 8$ and

$$\begin{aligned} H_1(z_1, z_2) &= (1 - z_1)^3(1 - z_2)^2(1 + z_2)/16, \\ H_2(z_1, z_2) &= (1 - z_1)^2(1 + z_1)(1 - z_2)^3/16, \\ H_3(z_1, z_2) &= (1 - z_1)^3(1 - z_2)(1 + z_2)^2/16, \\ H_4(z_1, z_2) &= (1 - z_1)(1 + z_1)^2(1 - z_2)^3/16, \\ H_5(z_1, z_2) &= (1 - z_1)^2(1 + z_1)(1 - z_2)^2(1 + z_2)/16, \\ H_6(z_1, z_2) &= (1 - z_1)(1 + z_1)^2(1 + z_2)^3/16, \\ H_7(z_1, z_2) &= (1 + z_1)^3(1 - z_2)(1 + z_2)^2/16, \\ H_8(z_1, z_2) &= (1 + z_1)^3(1 + z_2)^3/16. \end{aligned} \tag{3.18}$$

In this case, $\deg H_\ell = (3, 3) > \deg R$, for all $\ell = 1 : 8$. ■

Finally, we remind that any sum-of-squares polynomial is nonnegative. However, in at least three variables, there may be nonnegative trigonometric polynomials that are not sum-of-squares, no matter the degree of the factors. For two variables, however, {nonnegative} = {sum-of-squares}.

3.3 Sum-of-Squares of Real Polynomials

A polynomial $P \in \mathbb{R}_{2\boldsymbol{n}}[\boldsymbol{t}]$ is sum-of-squares if it can be written as

$$P(\boldsymbol{t}) = \sum_{\ell=1}^{\nu} F_\ell(\boldsymbol{t})^2, \tag{3.19}$$

where $F_\ell \in \mathbb{R}_{\boldsymbol{n}}[\boldsymbol{t}]$. We note that, unlike the trigonometric polynomials case, the degree of the factors F_ℓ is limited to $\boldsymbol{n}$.

Obviously, sum-of-squares polynomials are nonnegative over $\mathbb{R}^d$. The relations between different sets of nonnegative polynomials are depicted by the following diagram (note the differences with respect to the trigonometric polynomials diagram (3.15))

$$\begin{array}{ccc} \{\text{squares}\} \subset & \{\text{sum-of-squares}\} & \\ & \neq & \subset \{\text{nonnegative}\} \\ & \{\text{positive}\} & \end{array} \tag{3.20}$$

While the inclusions are clear, the relation between positive and sum-of-squares polynomials needs more explanations. Each of the two sets contains polynomials that do not belong to the other. Trivially, there are sum-of-squares that take the value zero, and so they are not strictly positive. Also, there are positive polynomials which are not sum-of-squares and, although their existence was proved by Hilbert in 1888, an example was given only in the 1960s. We give here the first example of Motzkin, adapted from [2].

Example 3.8 Consider the polynomial

$$P(t_1, t_2) = t_1^4 t_2^2 + t_1^2 t_2^4 - \alpha t_1^2 t_2^2 + 1, \tag{3.21}$$

where $0 < \alpha \leq 3$. (Motzkin's example was actually given with $\alpha = 3$.) Using the arithmetic-geometric means inequality

$$t_1^4 t_2^2 + t_1^2 t_2^4 + 1 \geq 3\sqrt[3]{t_1^4 t_2^2 \cdot t_1^2 t_2^4 \cdot 1} = 3 t_1^2 t_2^2 \geq \alpha t_1^2 t_2^2,$$

we see that the polynomial (3.21) is nonnegative for $\alpha = 3$ and positive for $0 < \alpha < 3$. (The first inequality is strict unless $t_1^4 t_2^2 = t_1^2 t_2^4 = 1$, i.e., $t_1^2 = t_2^2 = 1$; the second inequality is strict unless $t_1 = t_2 = 0$ or $\alpha = 3$.) If $P(t_1, t_2)$ were a sum-of-squares, it would have the form

$$P(t_1, t_2) = \sum_\ell (a_\ell t_1^2 + b_\ell t_1^2 t_2 + c_\ell t_1 + d_\ell t_1 t_2 + e_\ell t_1 t_2^2 + f_\ell + g_\ell t_2 + h_\ell t_2^2)^2.$$

By identification of the coefficients of t_1^4, we see that $\sum a_\ell^2 = 0$ and so $a_\ell = 0$; similarly, we obtain $h_\ell = 0$. Now, we look at the coefficients of t_1^2 and see that $\sum c_\ell^2 = 0$ and so $c_\ell = 0$; similarly, we obtain $g_\ell = 0$. Finally, the coefficients of $t_1^2 t_2^2$ say that $\sum d_\ell^2 = -\alpha$, which is impossible. So, $P(t_1, t_2)$ is not sum-of-squares. ■

Although not all nonnegative polynomials are sum-of-squares, there are important sum-of-squares characterizations. The first one is that any nonnegative polynomial is a sum-of-squares of *rational* functions.

Theorem 3.9 (Artin 1927) *For any nonnegative polynomial $P \in \mathbb{R}_{2\boldsymbol{n}}[\boldsymbol{t}]$, there exist polynomials $F_0, F_1, \ldots, F_\nu \in \mathbb{R}[\boldsymbol{t}]$ such that*

$$P(\boldsymbol{t}) \cdot F_0(\boldsymbol{t})^2 = \sum_{\ell=1}^{\nu} F_\ell(\boldsymbol{t})^2. \tag{3.22}$$

For a proof, see [3]. In this formulation, F_0 is the common denominator of the rational functions, i.e., it is possible that F_0 and some F_ℓ have some nontrivial divisor. Anyway, even after such simplifications, it is possible that the degrees of the polynomials are greater than $\boldsymbol{n}$.

The relation (3.22) is not appropriate to optimization due to the product $P(\boldsymbol{t})F_0(\boldsymbol{t})^2$ which involves two usually unknown polynomials. Much more interesting are sum-of-squares of rational functions in which the denominator F_0 is *fixed*. One result of this type is the following.

Theorem 3.10 (Reznick 1995) *For any* positive *polynomial $P \in \mathbb{R}_{2\boldsymbol{n}}[\boldsymbol{t}]$, there exist polynomials $F_1, \ldots, F_\nu \in \mathbb{R}[\boldsymbol{t}]$ and a positive integer κ such that*

$$P(\boldsymbol{t}) \cdot \left(1 + t_1^2 + \ldots + t_d^2\right)^\kappa = \sum_{\ell=1}^{\nu} F_\ell(\boldsymbol{t})^2. \tag{3.23}$$

We note that now the polynomial has to be strictly positive. Again, the degrees of the polynomials F_ℓ are higher than $\boldsymbol{n}$; they are actually at most $\boldsymbol{n} + \kappa$.

Example 3.8 (*continued*) Take $P(t_1, t_2) = t_1^4 t_2^2 + t_1^2 t_2^4 - t_1^2 t_2^2 + 1$, i.e., the polynomial from (3.21), with $\alpha = 1$. As shown before, it is not sum-of-squares. However, the polynomial

$$P(t_1, t_2)(1 + t_1^2 + t_2^2) = \tfrac{1}{2}(t_1^2 t_2^2 - 1)^2 + t_1^6 t_2^2 + t_1^2 t_2^6 + \tfrac{3}{2} t_1^4 t_2^4 + t_1^2 + t_2^2 + \tfrac{1}{2}$$

is clearly sum-of-squares. So, in this case, Theorem 3.10 holds for $\kappa = 1$. (Exercise: Show that the same is true for the polynomial (3.21) and any value $0 < \alpha \leq 3$.) ■

3.4 Gram Matrix Parameterization of Multivariate Trigonometric Polynomials

We now generalize the constructions from Sect. 2.1 for the case of multivariate polynomials. Using the notation (2.1), the canonical basis for d-variate polynomials of degree $\boldsymbol{n}$ is

$$\boldsymbol{\psi}_{\boldsymbol{n}}(\boldsymbol{z}) = \boldsymbol{\psi}_{n_d}(z_d) \otimes \ldots \otimes \boldsymbol{\psi}_{n_1}(z_1) = \bigotimes_{i=d}^{1} \boldsymbol{\psi}_{n_i}(z_i), \tag{3.24}$$

where $\otimes$ represents the Kronecker product. As before, we ignore the index $\boldsymbol{n}$ if it is obvious from the context. For instance, with $d = 2$, $n_1 = 2$, $n_2 = 1$, we have

$$\boldsymbol{\psi}(\boldsymbol{z}) = [1 \; z_2]^T \otimes [1 \; z_1 \; z_1^2]^T = [1 \; z_1 \; z_1^2 \; z_2 \; z_1 z_2 \; z_1^2 z_2]^T. \tag{3.25}$$

We note that the basis contains

$$N = \prod_{i=1}^{d} (n_i + 1) \tag{3.26}$$

monomials. A positive orthant trigonometric polynomial (3.5) can be written in the form

$$H(z) = \boldsymbol{\psi}^T(z^{-1})\boldsymbol{h}, \tag{3.27}$$

where $\boldsymbol{h} \in \mathbb{C}^N$ is a vector containing the coefficients of $H(z)$ ordered as corresponding to (3.24). For example, for $d = 2$, $n_1 = 2$, $n_2 = 1$, the vector is

$$\boldsymbol{h} = [h_{00}\ h_{10}\ h_{20}\ h_{01}\ h_{11}\ h_{21}]^T.$$

Definition 3.11 A Hermitian matrix $\boldsymbol{Q}$ is called a *Gram* matrix associated with the trigonometric polynomial (3.1) if

$$R(z) = \boldsymbol{\psi}^T(z^{-1}) \cdot \boldsymbol{Q} \cdot \boldsymbol{\psi}(z). \tag{3.28}$$

As in the univariate case, we denote $\mathcal{G}(R)$ the set of Gram matrices associated with $R(z)$. ■

Example 3.12 Let us take again $d = 2$, $n_1 = 2$, $n_2 = 1$ and consider polynomials with real coefficients. The equality (3.28) is equivalent to the following scalar equalities (we enumerate all coefficients in a half plane)

$$\begin{aligned}
r_{00} &= q_{00} + q_{11} + q_{22} + q_{33} + q_{44} + q_{55},\\
r_{10} &= q_{10} + q_{21} + q_{43} + q_{54},\\
r_{20} &= q_{20} + q_{53},\\
r_{-2,1} &= q_{32},\\
r_{-1,1} &= q_{31} + q_{42},\\
r_{01} &= q_{30} + q_{41} + q_{52},\\
r_{11} &= q_{40} + q_{51},\\
r_{21} &= q_{50}.
\end{aligned} \tag{3.29}$$

To understand easily the above equalities, we write, as in the proof of Theorem 2.3,

$$R(z) = \boldsymbol{\psi}^T(z^{-1}) \cdot \boldsymbol{Q} \cdot \boldsymbol{\psi}(z) = \mathrm{tr}[\boldsymbol{\psi}(z) \cdot \boldsymbol{\psi}^T(z^{-1}) \cdot \boldsymbol{Q}] = \mathrm{tr}[\boldsymbol{\Psi}(z) \cdot \boldsymbol{Q}]. \tag{3.30}$$

Taking (3.25) into account, the matrix $\boldsymbol{\Psi}(\boldsymbol{z})$ has the form

$$\boldsymbol{\Psi}(\boldsymbol{z}) = \begin{bmatrix} 1 & & & & & \\ z_1 & 1 & & & \mathrm{sym}^{-1} & \\ z_1^2 & z_1 & 1 & & & \\ z_2 & z_1^{-1}z_2 & z_1^{-2}z_2 & 1 & & \\ z_1z_2 & z_2 & z_1^{-1}z_2 & z_1 & 1 & \\ z_1^2z_2 & z_1z_2 & z_2 & z_1^2 & z_1 & 1 \end{bmatrix}. \tag{3.31}$$

Looking at the positions of some monomial in $\boldsymbol{\Psi}(\boldsymbol{z})$, we recover the expressions from (3.29). For example, the positions of z_1 (or the symmetric ones, for z_1^{-1}) give the indices of the elements of $\boldsymbol{Q}$ that appear in the expression of r_{10}. ■

The general relation between the coefficients of the polynomial $R(\boldsymbol{z})$ and an associated Gram matrix is given by the following theorem.

Theorem 3.13 *If $R \in \mathbb{C}_{\boldsymbol{n}}[\boldsymbol{z}]$ and $\boldsymbol{Q} \in \mathcal{G}(R)$, then the relation*

$$r_{\boldsymbol{k}} = tr[\boldsymbol{\Theta}_{\boldsymbol{k}} \cdot \boldsymbol{Q}] \tag{3.32}$$

holds, where

$$\boldsymbol{\Theta}_{\boldsymbol{k}} = \boldsymbol{\Theta}_{k_d} \otimes \ldots \otimes \boldsymbol{\Theta}_{k_1} \tag{3.33}$$

and the matrices $\boldsymbol{\Theta}_k$ are defined as in the body of Theorem2.3. We name (3.32) the generalized trace *parameterization of the trigonometric polynomial $R(\boldsymbol{z})$.*

Proof Using (3.24), the matrix $\boldsymbol{\Psi}(\boldsymbol{z})$ from (3.30) can be expressed as

$$\boldsymbol{\Psi}(\boldsymbol{z}) = \left[\bigotimes_{i=d}^{1} \boldsymbol{\psi}(z_i)\right] \cdot \left[\bigotimes_{i=d}^{1} \boldsymbol{\psi}^T(z_i^{-1})\right] = \bigotimes_{i=d}^{1} \left[\boldsymbol{\psi}(z_i) \cdot \boldsymbol{\psi}^T(z_i^{-1})\right]. \tag{3.34}$$

For the last equality above, we have used $d-1$ times the identity $(\boldsymbol{A} \otimes \boldsymbol{B})(\boldsymbol{C} \otimes \boldsymbol{D}) = (\boldsymbol{AC}) \otimes (\boldsymbol{BD})$, where $\boldsymbol{A}$, $\boldsymbol{B}$, $\boldsymbol{C}$, and $\boldsymbol{D}$ are matrices of appropriate sizes. Taking into account that each of the matrices $\boldsymbol{\psi}(z_i)\boldsymbol{\psi}^T(z_i^{-1})$ has the Toeplitz form shown in (2.7), we rewrite (3.34) as

$$\boldsymbol{\Psi}(\boldsymbol{z}) = \bigotimes_{i=d}^{1} \left[\sum_{k_i=-n_i}^{n_i} \boldsymbol{\Theta}_{k_i} z_i^{-k_i}\right] = \sum_{\boldsymbol{k}=-\boldsymbol{n}}^{\boldsymbol{n}} \boldsymbol{z}^{-\boldsymbol{k}} \left[\bigotimes_{i=d}^{1} \boldsymbol{\Theta}_{k_i}\right] = \sum_{\boldsymbol{k}=-\boldsymbol{n}}^{\boldsymbol{n}} \boldsymbol{\Theta}_{\boldsymbol{k}} \boldsymbol{z}^{-\boldsymbol{k}}. \tag{3.35}$$

Combining the last relation with (3.30), we obtain

$$R(\boldsymbol{z}) = \sum_{\boldsymbol{k}=-\boldsymbol{n}}^{\boldsymbol{n}} \mathrm{tr}[\boldsymbol{\Theta}_{\boldsymbol{k}} \boldsymbol{Q}] \boldsymbol{z}^{-\boldsymbol{k}},$$

which proves (3.32) after identification with (3.1).

Example 3.14 With $d = 2, n_1 = 2, n_2 = 1$, we look at (3.32) for $\boldsymbol{k} = (1, 0)$, i.e., at the expression of r_{10}. The matrix $\boldsymbol{\Theta}_{\boldsymbol{k}}$ has the form

$$\boldsymbol{\Theta}_{\boldsymbol{k}} = \boldsymbol{\Theta}_0 \otimes \boldsymbol{\Theta}_1 = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \otimes \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{bmatrix} = \left[\begin{array}{ccc|ccc} 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ \hline 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{array}\right].$$

Let us consider the symmetric polynomial

$$R(z) = \text{sym}^{-1} + 38 + 18z_1 + 4z_1^2 + z_1^{-2}z_2 + 2z_1^{-1}z_2 + z_2 - 8z_1z_2 - 5z_1^2z_2. \quad (3.36)$$

Then,

$$\boldsymbol{Q}_1 = \begin{bmatrix} 38 & & & & & \\ \underline{18} & 0 & & \text{sym} & & \\ 4 & \underline{0} & 0 & & & \\ 1 & 2 & 1 & 0 & & \\ -8 & 0 & 0 & \underline{0} & 0 & \\ -5 & 0 & 0 & 0 & \underline{0} & 0 \end{bmatrix}, \quad \boldsymbol{Q}_2 = \begin{bmatrix} 9 & & & & & \\ \underline{7} & 8 & & \text{sym} & & \\ 2 & \underline{2} & 2 & & & \\ 1 & 1 & 1 & 2 & & \\ -4 & -1 & 1 & \underline{2} & 8 & \\ -5 & -4 & 1 & 2 & \underline{7} & 9 \end{bmatrix} \quad (3.37)$$

are Gram matrices associated with $R(z)$. The sum of underlined elements is equal to the coefficient of z_1^{-1} (these elements are selected by the ones from $\boldsymbol{\Theta}_{\boldsymbol{k}}$ when computing $\text{tr}[\boldsymbol{\Theta}_{\boldsymbol{k}}\boldsymbol{Q}]$). The matrix $\boldsymbol{Q}_1$ is chosen such that each coefficient of $R(z)$ appears as one element. ■

The Gram matrix expression (3.32) of a polynomial allows a characterization of sum-of-squares polynomials by positive semidefinite matrices.

Theorem 3.15 *A polynomial $R \in \mathbb{C}_{\boldsymbol{n}}[z]$ is sum-of-squares, with* $\deg H_\ell \leq \boldsymbol{n}$ *for the factors from (3.13), if and only if there exists a positive semidefinite matrix $\boldsymbol{Q} \in \mathbb{C}^{N \times N}$ such that (3.32) holds.*

Proof The proof is similar to that from the univariate case, see Remark 2.9. The only change is to replace z with $\boldsymbol{z}$. The condition $\deg H_\ell \leq \boldsymbol{n}$ is necessary for obtaining Gram matrices of size $N \times N$. (We will discuss in the next section when larger Gram matrices are useful.) ■

Optimization problems with sum-of-squares polynomials can be formulated using the Gram matrix representation (3.32) and SDP, similarly to the univariate case. We discuss here one such problem.

Problem (*Most_positive_Gram_matrix*) Given the sum-of-squares polynomial $R \in \mathbb{C}_{\boldsymbol{n}}[z]$, we want to find the most positive Gram matrix (of size $N \times N$) associated with it. This is equivalent to finding the matrix from $\mathcal{G}(R)$ whose smallest eigenvalue is maximum. The corresponding optimization problem is

$$\begin{aligned} \lambda^\star = \max_{\lambda,\boldsymbol{Q}} &\ \lambda \\ \text{s.t.}\ &\ \mathrm{tr}[\boldsymbol{\Theta}_{\boldsymbol{k}} \boldsymbol{Q}] = r_{\boldsymbol{k}}, \quad \boldsymbol{k} \in \mathcal{H} \\ &\ \lambda \ge 0, \quad \boldsymbol{Q} \succeq \lambda \boldsymbol{I} \end{aligned} \tag{3.38}$$

where $\mathcal{H}$ is a half-space and is the multivariate version of (2.14). This SDP problem is transformed into standard equality form by denoting $\tilde{\boldsymbol{Q}} = \boldsymbol{Q} - \lambda \boldsymbol{I}$. Note that $\boldsymbol{\Theta}_{\boldsymbol{0}} = \boldsymbol{I}$, and so $\mathrm{tr}\boldsymbol{\Theta}_{\boldsymbol{0}} = N$, while for $\boldsymbol{k} \ne 0$ we have $\mathrm{tr}\boldsymbol{\Theta}_{\boldsymbol{k}} = 0$. We obtain

$$\begin{aligned} \lambda^\star = \max_{\lambda,\tilde{\boldsymbol{Q}}} &\ \lambda \\ \text{s.t.}\ &\ N\lambda + \mathrm{tr}\tilde{\boldsymbol{Q}} = r_{\boldsymbol{0}} \\ &\ \mathrm{tr}[\boldsymbol{\Theta}_{\boldsymbol{k}} \tilde{\boldsymbol{Q}}] = r_{\boldsymbol{k}}, \quad \boldsymbol{k} \in \mathcal{H} \setminus \{\boldsymbol{0}\} \\ &\ \lambda \ge 0, \quad \tilde{\boldsymbol{Q}} \succeq 0 \end{aligned} \tag{3.39}$$

The complexity of solving such a problem is $O(N^2M^2)$ (see Remark 2.13), where M is the number (3.4) of coefficients in a half-space. If $n_1 = \ldots = n_d = n$, then the complexity is $O(n^{4d})$. The complexity grows fast with the number of variables, but polynomially with respect to the degree of $R(\boldsymbol{z})$.

Example 3.16 The smallest eigenvalue of the matrix $\boldsymbol{Q}_2$ from (3.37) is 0.1853 and so the polynomial (3.36) is sum-of-squares (since $\boldsymbol{Q}_2 \succ 0$). We have actually taken $R(\boldsymbol{z}) = H(\boldsymbol{z})H(\boldsymbol{z}^{-1})$, with

$$H(\boldsymbol{z}) = 5 + 3z_1 + z_1^2 + z_2 - z_1z_2 - z_1^2z_2, \tag{3.40}$$

and so $R(\boldsymbol{z})$ is not only sum-of-squares, but also factorable. Solving the SDP problem (3.39), we obtain the Gram matrix

$$\boldsymbol{Q}_o = \begin{bmatrix} 8.9619 & & & & & \\ 6.8512 & 7.6099 & & & \text{sym} & \\ 1.9998 & 2.1482 & 2.4274 & & & \\ 1.3322 & 0.9998 & 1.0000 & 2.4275 & & \\ -4.0002 & -1.6646 & 1.0002 & 2.1485 & 7.6109 & \\ -5.0000 & -3.9998 & 1.3324 & 2.0002 & 6.8522 & 8.9625 \end{bmatrix},$$

whose smallest eigenvalue is $\lambda^\star = 0.3036$. ■

3.5 Sum-of-Squares Relaxations

As formulated in the previous section, the computation of the most positive Gram matrix is one of the few optimization problems genuinely related to sum-of-squares polynomials; we will see in this section that it can have another significance. More often, interesting problems can be formulated in terms of nonnegative polynomials.

However, although the set of multivariate nonnegative polynomials is convex, there is no known direct method for working in it. More important, many problems with nonnegative polynomials are NP-hard, i.e., in general there is no algorithm to solve them in polynomial time. Since {sum-of-squares} ⊂ {nonnegative} and optimization problems with sum-of-squares can be cast into SDP, a natural approach is to ignore that the above inclusion is strict and work with sum-of-squares instead of nonnegative polynomials. This approach, known as *sum-of-squares relaxation*, is discussed in this section.

3.5.1 Relaxation Principle

We have mentioned in Sect. 3.2 that some positive (or even nonnegative) polynomials $R(z)$ have only a sum-of-squares expression (3.13) in which the degrees of the factors $H_\ell(z)$ are larger than $\boldsymbol{n} = \deg R$. So, it makes sense to distinguish between sum-of-squares of same degree, but with factors of different degrees. For given degrees $\boldsymbol{n}$, $\boldsymbol{m}$, with $\boldsymbol{m} \geq \boldsymbol{n}$, we denote

$$\mathbb{RS}_{\boldsymbol{n}}^{\boldsymbol{m}}[z] = \{R \in \mathbb{R}_{\boldsymbol{n}}[z] \mid R(z) = \sum_{\ell=1}^{\nu} H_\ell(z) H_\ell^*(z^{-1}),\ \deg H_\ell \leq \boldsymbol{m}\}, \tag{3.41}$$

i.e., the set of sum-of-squares of degree $\boldsymbol{n}$, with factors of degree $\boldsymbol{m}$. It is clear that if $\boldsymbol{n} < \boldsymbol{m}$, then we have the following inclusions

$$\mathbb{RS}_{\boldsymbol{n}}^{\boldsymbol{n}}[z] \subset \mathbb{RS}_{\boldsymbol{n}}^{\boldsymbol{m}}[z] \subset \overline{\mathbb{RP}}_{\boldsymbol{n}}[z]. \tag{3.42}$$

(The inclusions are trivially justified by the fact that a factor H_ℓ can be considered as having a degree higher than its actual degree, by adding zero coefficients.) The inclusions may be strict or not, depending on the values of $\boldsymbol{n}$, $\boldsymbol{m}$, and d. In any case, a higher value of $\boldsymbol{m}$ gives a better approximation of $\overline{\mathbb{RP}}_{\boldsymbol{n}}[z]$ by $\mathbb{RS}_{\boldsymbol{n}}^{\boldsymbol{m}}[z]$.

Assume that we have a convex optimization problem whose variable is the nonnegative polynomial $R \in \overline{\mathbb{RP}}_{\boldsymbol{n}}[z]$. Since optimization with sum-of-squares polynomials can be expressed in terms of SDP (if the objective and the constraints are appropriate), as shown in the previous section, we can approximate the original problem with a new one, in which the variable is $R \in \mathbb{RS}_{\boldsymbol{n}}^{\boldsymbol{m}}[z]$. Thus, we replace a high complexity problem, for which there is no efficient and reliable algorithm, with a simpler problem having a convenient SDP formulation. This procedure is called *sum-of-squares relaxation*. If the solution $R^\star$ of the original problem is in $\mathbb{RS}_{\boldsymbol{n}}^{\boldsymbol{m}}[z]$, then it is also the solution of the relaxed problem. If $R^\star \in \overline{\mathbb{RP}}_{\boldsymbol{n}}[z] \setminus \mathbb{RS}_{\boldsymbol{n}}^{\boldsymbol{m}}[z]$, then the relaxed problem gives only an approximation of the original solution. This is the price of the relaxation. However, the approximation can be sufficiently good for practical purposes.

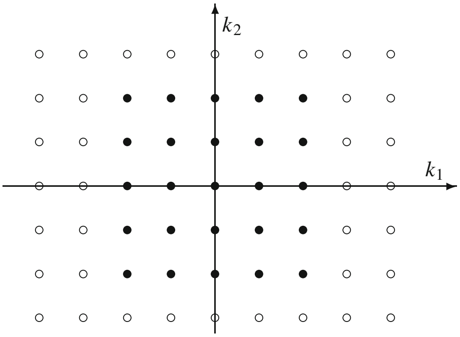


Fig. 3.2 Coefficients of a bivariate polynomial of degree (2, 2), padded with zeros up to degree (4, 3). *Filled circles* represent the original coefficients and *circles* the added zero coefficients

The Gram matrix parameterization of a sum-of-squares polynomial $R \in \mathbb{RS}_{\boldsymbol{n}}^{\boldsymbol{m}}[z]$ has the form (3.28), with a vector $\boldsymbol{\psi}(z)$ containing all the monomials of degree at most $\boldsymbol{m}$. Since the true degree of $R(z)$ is $\boldsymbol{n}$, we have to impose the condition

$$r_{\boldsymbol{k}} = 0, \quad \text{if } |k_i| > |n_i| \text{ for some } i \in 1 : d. \tag{3.43}$$

The zero padding is illustrated in Fig. 3.2, for $\boldsymbol{n} = (2, 2)$, $\boldsymbol{m} = (4, 3)$. Using the Gram matrix parameterization (3.32), the condition (3.43) is replaced by

$$\text{tr}[\boldsymbol{\Theta}_{\boldsymbol{k}} \cdot \boldsymbol{Q}] = 0, \quad \text{if } |k_i| > |n_i| \text{ for some } i \in 1 : d. \tag{3.44}$$

In the sequel, we will not write explicitly the condition (3.44) in optimization problems that are sum-of-squares relaxations, but we will assume that it is used. The size of the Gram matrix depends on the degree $\boldsymbol{m}$ of the relaxation. Specifically, the Gram matrix $\boldsymbol{Q}$ is $N \times N$, where N is no more given by (3.26), but by

$$N = \prod_{i=1}^{d} (m_i + 1), \tag{3.45}$$

in accordance with the number of monomials in $\boldsymbol{\psi}(z)$.

3.5.2 A Case Study

To illustrate the sum-of-squares relaxation idea, we consider a problem solved previously for univariate polynomials.

Problem (*Min_poly_value*) Let $R \in \mathbb{R}_{\boldsymbol{n}}[z]$ be a given polynomial. We want to find its minimum value on the unit d-circle

$$\mu^\star = \min_{\boldsymbol{\omega} \in [-\pi, \pi]^d} R(\boldsymbol{\omega}). \tag{3.46}$$

The problem is NP-hard. Traditionally, such a problem is solved by discretization or by using nonlinear optimization techniques; the former method leads to suboptimal solutions and has high complexity, and the latter may give a local optimum, since the problem is not convex. As in the univariate case, we transform (3.46) into a problem with nonnegative polynomials, namely into

$$\begin{aligned} \mu^\star = \max_{\mu} \ & \mu \\ \text{s.t.} \ & R(\boldsymbol{\omega}) - \mu \geq 0, \quad \forall \boldsymbol{\omega} \in [-\pi, \pi]^d \end{aligned} \tag{3.47}$$

However, in this formulation, we lack an appropriate description of the set of nonnegative polynomials. Thus, we can appeal to sum-of-squares relaxations and approximate (3.47) with

$$\begin{aligned} \mu^\star_{\boldsymbol{m}} = \max_{\mu} \ & \mu \\ \text{s.t.} \ & R(\boldsymbol{z}) - \mu \in \mathbb{RS}^{\boldsymbol{m}}_{\boldsymbol{n}}[\boldsymbol{z}] \end{aligned} \tag{3.48}$$

for some $\boldsymbol{m} \geq \boldsymbol{n}$. In view of (3.42), by solving (3.48) we obtain solutions that satisfy

$$\mu^\star_{\boldsymbol{n}} \leq \mu^\star_{\boldsymbol{m}} \leq \mu^\star. \tag{3.49}$$

These inequalities follow from the fact that for a given μ, the polynomial $R(\boldsymbol{z}) - \mu$ may be sum-of-squares with factors of degree $\boldsymbol{m}$, but not sum-of-squares with factors of degree $\boldsymbol{n}$. We have to note that the *nonnegative* polynomial $R(\boldsymbol{z}) - \mu^\star$ may be not sum-of-squares for any $\boldsymbol{m}$ (remind that Theorem 3.5 is valid only for strictly positive polynomials) and so it is theoretically possible that $\mu^\star_{\boldsymbol{m}} < \mu^\star$ for any $\boldsymbol{m}$.

Since $\mu^\star$ is not known (and often it cannot be known, see next subsection), the sum-of-squares relaxation process may be iterative:

1. Put $\boldsymbol{m} = \boldsymbol{n}$.
2. Solve (3.48).
3. If satisfied, stop. Otherwise, increase $\boldsymbol{m}$ and go back to 2.

Of course, the definition of the "satisfaction" from step 3 is loose. For example, if for some $\tilde{\boldsymbol{m}} > \boldsymbol{m}$ we obtain $\mu^\star_{\tilde{\boldsymbol{m}}} = \mu^\star_{\boldsymbol{m}}$, then we may assume that the true minimum is attained, although there are no guarantees that this happened indeed. Since the complexity of the problem grows with $\boldsymbol{m}$, we can solve (3.48) only once, for $\boldsymbol{m} = \boldsymbol{n}$ (or for a slightly larger value of $\boldsymbol{m}$). We will give immediately some examples.

Let us first note that using the generalized trace parameterization (3.32), the problem (3.48) can be cast into the following SDP form

$$\begin{aligned} \mu^\star_{\boldsymbol{m}} = \max_{\mu, \tilde{\boldsymbol{Q}}} \ & \mu \\ \text{s.t.} \ & \mu + \operatorname{tr} \tilde{\boldsymbol{Q}} = r_{\boldsymbol{0}} \\ & \operatorname{tr}[\boldsymbol{\Theta}_{\boldsymbol{k}} \tilde{\boldsymbol{Q}}] = r_{\boldsymbol{k}}, \quad \boldsymbol{k} \in \mathcal{H} \setminus \{\boldsymbol{0}\} \\ & \tilde{\boldsymbol{Q}} \succeq 0, \ \tilde{\boldsymbol{Q}} \in \mathbb{R}^{N \times N} \end{aligned} \tag{3.50}$$

Similarly to the univariate case, this problem is equivalent to the eigenvalue maximization (3.39) (from which the constraint $\lambda \geq 0$ has to be removed), with $\mu = N\lambda$ and N given by (3.45).

Example 3.17 Let us compute the minimum value on the unit bicircle of the polynomial

$$R(\boldsymbol{z}) = 5 + (z_1 + z_1^{-1}) + (z_1 z_2 + z_1^{-1} z_2^{-1})$$

of degree $\boldsymbol{n} = (1, 1)$. As shown in Example 3.4 (where the free term was 4 instead of 5), we have $R(\boldsymbol{\omega}) \geq 1$; moreover, for $z_1 = -1$, $z_2 = 1$, we have $R(\boldsymbol{z}) = 1$. So, the minimum value on the unit bicircle is 1. Moreover, for any $\mu \leq 1$, the polynomial

$$R(\boldsymbol{z}) - \mu = (\sqrt{1-\mu})^2 + (1 + z_1)(1 + z_1^{-1}) + (1 + z_1 z_2)(1 + z_1^{-1} z_2^{-1})$$

is sum-of-squares of degree $\boldsymbol{n}$. Due to the particular form of $R(\boldsymbol{z})$, it results that the sum-of-squares relaxation (3.48) with $\boldsymbol{m} = \boldsymbol{n}$ gives the exact solution. Therefore, solving the SDP problem (3.50) for this minimal value of $\boldsymbol{m}$ gives the exact solution. ■

Example 3.18 We solve the problem *Min_poly_value* for the polynomial (3.36), whose graph is displayed in Fig. 3.3. Its degree is $\boldsymbol{n} = (2, 1)$. Solving (3.50) for $\boldsymbol{m} = \boldsymbol{n}$, we obtain $\mu_{\boldsymbol{n}}^\star = 1.8214$. We note that $\mu_{\boldsymbol{n}}^\star = N\lambda^\star$, where the size of the Gram matrix is $N = 6$ and the most positive Gram matrix has the smallest eigenvalue $\lambda^\star = 0.3036$, see Example 3.16.

As we cannot check that we have indeed obtained the minimum value of $R(\boldsymbol{\omega})$, we solve (3.50) for greater values of $\boldsymbol{m}$. For all degrees of the relaxation less than or equal to (10, 9) (we have not tried larger values), we obtain the same minimum value (within a relative error of 10^{-10}, which is due to the numerical tolerance of the SeDuMi routine). So, it is safe to assume that we have indeed obtained the optimal value. ■

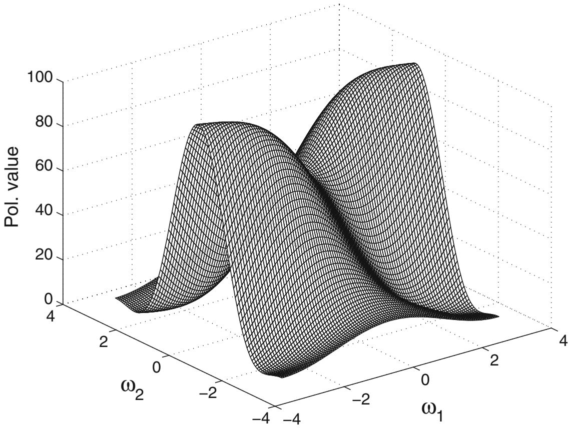


Fig. 3.3 Graph of the polynomial from Example 3.18

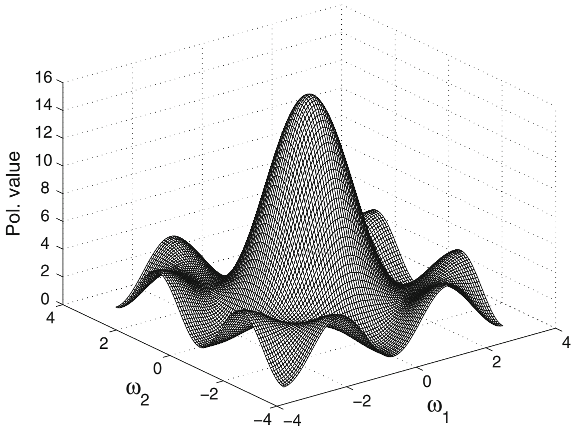


Fig. 3.4 Graph of the polynomial from Example 3.19

Example 3.19 Let us revisit Example 3.7. The polynomial (3.17) is nonnegative and its minimum value on the unit bicircle is zero. The graph of the polynomial is shown in Fig. 3.4. Its degree is $\boldsymbol{n} = (2, 2)$. Since the polynomial is not a sum-of-squares of degree $(2, 2)$, it follows that the sum-of-squares relaxation (3.50) with $\boldsymbol{m} = \boldsymbol{n}$ should have a negative solution. Indeed, we obtain $\mu_{\boldsymbol{n}}^\star = -0.01177$. (It follows that for any $0 > \mu > \mu_{\boldsymbol{n}}^\star$, the positive polynomial $R(z) - \mu$ is not sum-of-squares with factors of degree $(2, 2)$.)

For any $\boldsymbol{m} > \boldsymbol{n}$, including $\boldsymbol{m} = (3, 2)$ or $\boldsymbol{m} = (2, 3)$, i.e., the smallest values greater than $\boldsymbol{n}$, the solution of (3.50) is $\mu_{\boldsymbol{m}}^\star = 0$ (within the numerical accuracy). So, in this case, the first increase in the degree brings with it the optimal value. Even without knowing the solution, we may assume that it is $\mu^\star = 0$, since this is the value obtained for several relaxations with $\boldsymbol{m} > \boldsymbol{n}$. ■

The above examples suggest that in practical applications where the values of the degree $\boldsymbol{n}$ are moderate, sum-of-squares relaxations of degree only slightly larger than $\boldsymbol{n}$ are likely to give the optimal solution. This is indeed the case, at least as documented by the relatively scarce literature in this domain. By moderate degrees, we understand here the values of $\boldsymbol{m}$ for which the size (3.45) of the Gram matrix is at most a few hundreds maybe nearing one thousand. Otherwise, the time for obtaining the solution of the SDP problem becomes impractically large.

3.5.3 Optimality Certificate

As we have seen, the sum-of-squares relaxation cannot guarantee by itself the optimality of the obtained solution. So, we need to appeal to other means for certifying the optimality. Here, we present two such methods. Although they imply a computation effort that may be larger than the sum-of-squares relaxation and they may not succeed in providing the optimality certificate, these methods can be useful.

The first method attempts to bracket the optimal value. We describe it for the problem *Min_poly_value*, but it can be adapted to other problems. For solving (3.46), we simply use a nonlinear optimization algorithm and obtain a minimum value $\tilde{\mu}$. Since there is no certainty that we have not obtained a local optimum, it follows that $\tilde{\mu} \geq \mu^\star$. Solving the sum-of-squares relaxation (3.50) gives a solution $\mu_{\boldsymbol{m}}^\star$ satisfying (3.49). Thus, we bracket the true optimum by

$$\mu_{\boldsymbol{m}}^\star \leq \mu^\star \leq \tilde{\mu}. \tag{3.51}$$

If it happens that $\mu_{\boldsymbol{m}}^\star = \tilde{\mu}$, then we are sure to have obtained the true minimum. Otherwise, the distance between $\mu_{\boldsymbol{m}}^\star$ and $\tilde{\mu}$ indicates how near we are from the solution, in the worst case; if this distance is small, then we may be satisfied with the obtained results.

The second method is intimately related to sum-of-squares relaxations. As instantiated by (3.50), the relaxation provides only the minimum value of the objective, but not the values of the variable $\boldsymbol{\omega}$ for which this minimum is obtained. Were this $\boldsymbol{\omega} \overset{\Delta}{=} \boldsymbol{\omega}_{\boldsymbol{m}}^\star$ available, we could simply compute $\boldsymbol{R}(\boldsymbol{\omega}_{\boldsymbol{m}}^\star)$; if

$$\mu_{\boldsymbol{m}}^\star = \boldsymbol{R}(\boldsymbol{\omega}_{\boldsymbol{m}}^\star), \tag{3.52}$$

then we have indeed obtained the true minimum of (3.46); if $\mu_{\boldsymbol{m}}^\star < \boldsymbol{R}(\boldsymbol{\omega}_{\boldsymbol{m}}^\star)$, then at least we have found an interval where the minimum lies, i.e., a relation similar to (3.51).

What the solution of (3.50) provides is the optimal (and positive semidefinite) Gram matrix $\tilde{\boldsymbol{Q}}_{\boldsymbol{m}}^\star$, associated with the polynomial $\boldsymbol{R}(z) - \mu_{\boldsymbol{m}}^\star$. We consider the eigendecomposition

$$\tilde{\boldsymbol{Q}}_{\boldsymbol{m}}^\star = \sum_{\ell=1}^{\nu} \lambda_\ell \boldsymbol{x}_\ell \boldsymbol{x}_\ell^H, \tag{3.53}$$

where only the positive eigenvalues are considered; usually, several eigenvalues of the optimal Gram matrix are zero, and so $\nu = \operatorname{rank} \tilde{\boldsymbol{Q}}_{\boldsymbol{m}}^\star < N$. As in (2.12), it results that

$$R(\boldsymbol{\omega}) - \mu_{\boldsymbol{m}}^\star = \sum_{\ell=1}^{\nu} \lambda_\ell |H_\ell(\boldsymbol{\omega})|^2, \tag{3.54}$$

where $H_\ell(z) = \boldsymbol{\psi}^T(z^{-1})\boldsymbol{x}_\ell$. If we can find a solution to the system

$$|H_\ell(\boldsymbol{\omega})|^2 = 0, \quad \ell = 1 : \nu, \tag{3.55}$$

then from (3.54) we obtain $R(\boldsymbol{\omega}) = \mu_{\boldsymbol{m}}^\star$. Hence, the minimum is attained and so $\mu_{\boldsymbol{m}}^\star = \mu^\star$.

However, solving (3.55) is a difficult optimization problem and numerical algorithms may not find a solution to it. Nevertheless, even if an approximate solution can be obtained, it is helpful in evaluating how far from optimality is the sum-of-squares relaxation. Note that solving (3.55) means finding a common root on the unit circle of the polynomials $H_\ell(z)$, which opens the way for specific methods, like those using Gröbner bases.

Example 3.20 We consider again Examples 3.17–3.19.

We have attempted to solve (3.55) using the MATLAB functions `fsolve` and `fminsearch`; the second function actually founds a minimum of

$$f(\boldsymbol{\omega}) = \sum_{\ell=1}^{\nu} |H_\ell(\boldsymbol{\omega})|^2. \tag{3.56}$$

For both functions, the results may depend heavily on the initialization of $\boldsymbol{\omega}$. Of course, since we can check whether (3.55) indeed holds for the reported solution, we can run the functions for several initializations.

For Example 3.17, the solution of (3.55) is $\boldsymbol{\omega} = (\pi, 0)$, i.e., the one we already knew from the form of $R(z)$.

For Example 3.18 and $\boldsymbol{m} = \boldsymbol{n} = (2, 1)$, i.e., the minimum value, the solution is $\boldsymbol{\omega}_{\boldsymbol{n}}^\star = (2.3003, 3.4092)$. Moreover, we have $R(\boldsymbol{\omega}_{\boldsymbol{n}}^\star) = 1.8214 = \mu_{\boldsymbol{n}}^\star$, and so we know that we have obtained the optimal solution. There is no need of solving higher degree relaxations.

Finally, for Example 3.19 and $\boldsymbol{m} = \boldsymbol{n} = (2, 2)$, the minimum of (3.56) is obtained for $\boldsymbol{\omega}_{\boldsymbol{n}}^\star = (\pi, 0)$; this is not a solution of (3.55). As it can be easily checked by looking at (3.17) or (3.18), it results that $R(\boldsymbol{\omega}_{\boldsymbol{n}}^\star) = 0$. Since $\mu_{\boldsymbol{n}}^\star = -0.01177$, we know that we have not obtained the true optimum. However, when solving the sum-of-squares relaxation with $\boldsymbol{m} > \boldsymbol{n}$, we obtain $\mu_{\boldsymbol{m}}^\star = 0$. This confirms that we have obtained the optimum since we already have the frequency point $\boldsymbol{\omega}_{\boldsymbol{n}}^\star$ for which $R(\boldsymbol{\omega}_{\boldsymbol{n}}^\star) = 0$; there is no need to solve again (3.55) for the new Gram matrix $\tilde{\boldsymbol{Q}}_{\boldsymbol{m}}^\star$. ■

3.6 Gram Matrices from Partial Bases

Until now, the Gram matrices have been built using the vector (3.24), which contains all the d-variate monomials of degree at most $\boldsymbol{n}$. For sparse polynomials, which have only few nonzero coefficients, it may be interesting to use only part of the vector (3.24). Thus, Gram matrices of smaller size are obtained and the complexity of optimization problems decreases. As we will see, the price to pay is a possible loss of optimality.

3.6.1 Sparse Polynomials and Gram Representation

Let $\mathcal{I}_c = \{\boldsymbol{k} \in \mathbb{N}^d \mid \boldsymbol{k} \leq \boldsymbol{n}\}$ be the complete set of monomial degrees appearing in the vector $\boldsymbol{\psi}_{\boldsymbol{n}}(z)$ from (3.24). Let $\mathcal{I} \subset \mathcal{I}_c$ be a set of degrees and

$$\boldsymbol{\psi}_{\mathcal{I}}(z) = [\ldots\ z^{\boldsymbol{k}}\ \ldots]^T, \quad \text{with } \boldsymbol{k} \in \mathcal{I}. \tag{3.57}$$

We also can write

$$\boldsymbol{\psi}_{\mathcal{I}}(z) = \boldsymbol{C} \cdot \boldsymbol{\psi}_{\boldsymbol{n}}(z), \tag{3.58}$$

where $\boldsymbol{C}$ is a selection matrix of size $|\mathcal{I}| \times N$, having a single value of 1 on each row (all the other elements being zero).

Example 3.21 Taking $d = 2$, $n_1 = 2$, $n_2 = 1$, and $\mathcal{I} = \{(0,0), (2,0), (1,1)\}$, the vector (3.57) is

$$\boldsymbol{\psi}_{\mathcal{I}}(z) = [1\ z_1^2\ z_1 z_2]^T. \tag{3.59}$$

The selection matrix $\boldsymbol{C}$ from (3.58) is

$$\boldsymbol{C} = \begin{bmatrix} 1\ 0\ 0\ 0\ 0\ 0 \\ 0\ 0\ 1\ 0\ 0\ 0 \\ 0\ 0\ 0\ 0\ 1\ 0 \end{bmatrix},$$

due to the form (3.25) of the vector $\boldsymbol{\psi}_{\boldsymbol{n}}(z)$. ■

Assume now that the trigonometric polynomial $R(z)$ is such that we can express it using the basis (3.58) and an appropriate Gram matrix, similarly to relation (3.28), by

$$R(z) = \boldsymbol{\psi}_{\mathcal{I}}^T(z^{-1}) \cdot \boldsymbol{Q}_{\mathcal{I}} \cdot \boldsymbol{\psi}_{\mathcal{I}}(z), \tag{3.60}$$

with $\boldsymbol{Q}_{\mathcal{I}}$ of size $|\mathcal{I}| \times |\mathcal{I}|$. As in (3.30), we can write

$$R(z) = \operatorname{tr}[\boldsymbol{\Psi}_{\mathcal{I}}(z) \cdot \boldsymbol{Q}_{\mathcal{I}}], \quad \text{with } \boldsymbol{\Psi}_{\mathcal{I}}(z) = \boldsymbol{\psi}_{\mathcal{I}}(z) \cdot \boldsymbol{\psi}_{\mathcal{I}}^T(z^{-1}). \tag{3.61}$$

Example 3.21 (*continued*) Consider the polynomial

$$R(z_1, z_2) = \text{sym}^{-1} + 38 + 4z_1^2 + 2z_1^{-1}z_2 - 8z_1z_2. \tag{3.62}$$

(Remark that this polynomial is obtained by removing some of the monomials of (3.36).) We can see easily that we can write it in the form (3.61), using the basis vector (3.59), with

$$\boldsymbol{\Psi}_{\mathcal{I}}(z) = \begin{bmatrix} 1 & & \text{sym}^{-1} \\ z_1^2 & 1 & \\ z_1z_2 & z_1^{-1}z_2 & 1 \end{bmatrix}, \quad \boldsymbol{Q}_{\mathcal{I}} = \begin{bmatrix} q_{00} & & \text{sym} \\ 4 & q_{11} & \\ -8 & 2 & 38 - q_{00} - q_{11} \end{bmatrix}. \tag{3.63}$$

Remark that $\boldsymbol{\Psi}_{\mathcal{I}}(z)$ is a submatrix of $\boldsymbol{\Psi}(z)$ from (3.31). We also see that it is impossible to represent $R(z)$ on a sparser basis. However, we can freely add monomials to the basis and obtain a larger size Gram matrix (with more degrees of freedom). ■

We describe now the relation between the coefficients of $R(z)$ and the elements of the Gram matrix $\boldsymbol{Q}_{\mathcal{I}}$.

Theorem 3.22 *If $R \in \mathbb{C}_{\boldsymbol{n}}[z]$ can be represented as in (3.60) for some set of degrees $\mathcal{I}$, then the relation*

$$r_{\boldsymbol{k}} = tr[\boldsymbol{\Theta}_{\boldsymbol{k}}(\mathcal{I}) \cdot \boldsymbol{Q}_{\mathcal{I}}] \tag{3.64}$$

holds, where

$$\boldsymbol{\Theta}_{\boldsymbol{k}}(\mathcal{I}) = \boldsymbol{C} \cdot \boldsymbol{\Theta}_{\boldsymbol{k}} \cdot \boldsymbol{C}^T. \tag{3.65}$$

The matrices $\boldsymbol{\Theta}_{\boldsymbol{k}}$ and $\boldsymbol{C}$ are defined in (3.33) and (3.58), respectively.

Proof Combining relations (3.61), (3.58), and (3.35), we obtain

$$R(z) = \mathrm{tr}[\boldsymbol{\Psi}_{\mathcal{I}}(z)\boldsymbol{Q}_{\mathcal{I}}] = \mathrm{tr}[\boldsymbol{C}\boldsymbol{\Psi}(z)\boldsymbol{C}^T\boldsymbol{Q}_{\mathcal{I}}] = \sum_{\boldsymbol{k}=-\boldsymbol{n}}^{\boldsymbol{n}} \mathrm{tr}[\boldsymbol{C}\boldsymbol{\Theta}_{\boldsymbol{k}}\boldsymbol{C}^T\boldsymbol{Q}_{\mathcal{I}}]z^{-\boldsymbol{k}}. \tag{3.66}$$

By identification with (3.1), the proof is ready.

Example 3.21 (*continued*) Let $\boldsymbol{k} = (2, 0)$. The matrix $\boldsymbol{\Theta}_{\boldsymbol{k}}$ is

$$\boldsymbol{\Theta}_{\boldsymbol{k}} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \otimes \begin{bmatrix} 0 & 0 & 1 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} = \left[\begin{array}{ccc|ccc} 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ \hline 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{array}\right].$$

Using the basis (3.59), the matrix (3.65) is the submatrix obtained by taking the first, third, and fifth rows and columns of $\boldsymbol{\Theta}_{\boldsymbol{k}}$, i.e.,

$$\boldsymbol{\Theta}_{\boldsymbol{k}}(\mathcal{I}) = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}.$$

So, the relation (3.64) says that $r_{\boldsymbol{k}} = (\boldsymbol{Q}_{\mathcal{I}})_{10}$, which is visible in (3.63). ■

Remark 3.23 It is clear that if $\boldsymbol{Q}_{\mathcal{I}} \succeq 0$ in (3.64), then the polynomial $R(z)$ is nonnegative on $\mathbb{T}^d$. The reverse implication is not true, even if the polynomial is positive, in the sense that not any positive polynomial that can be represented as (3.60), for a *fixed* set $\mathcal{I}$, admits a positive semidefinite $\boldsymbol{Q}_{\mathcal{I}}$. We will shortly give an example, but the reason is obvious: by fixing $\mathcal{I}$, we limit the factors $H_\ell(z)$ of the sum-of-squares decomposition (3.13) to have only monomials with degrees in $\mathcal{I}$ (remind that these factors have expressions similar to (2.13), i.e., $H_\ell(z) = \lambda_\ell \boldsymbol{\psi}^T(z^{-1})\boldsymbol{x}_\ell$, where λ_ℓ and $\boldsymbol{x}_\ell$ are eigenvalues and, respectively, eigenvectors of the Gram matrix $\boldsymbol{Q}_\mathcal{I}$); by doing so, we also bound the degrees of the sum-of-squares associated with a Gram matrix. ■

Remark 3.24 The relation (3.64) can be described in a different way. Due to (3.61), the only monomials that can appear in $R(z)$ are those from the matrix $\boldsymbol{\Psi}_\mathcal{I}(z)$. Looking now at the expression of $\boldsymbol{\Psi}_\mathcal{I}(z)$, we see that $R(z)$ can have only monomials $r_{\boldsymbol{k}} z^{-\boldsymbol{k}}$ with degree $\boldsymbol{k}$ such that there are $\boldsymbol{i}, \boldsymbol{l} \in \mathcal{I}$, with $\boldsymbol{k} = \boldsymbol{i} - \boldsymbol{l}$. We denote

$$\mathcal{I} - \mathcal{I} = \{\boldsymbol{k} \in \mathbb{Z}^d \mid \boldsymbol{k} = \boldsymbol{i} - \boldsymbol{l},\ \ \boldsymbol{i}, \boldsymbol{l} \in \mathcal{I}\}. \tag{3.67}$$

So, a polynomial $R(z)$ having monomials with degrees belonging to a set $\mathcal{J} \subset \mathbb{Z}^d$ can have a Gram representation (3.64) if and only if $\mathcal{J} \in \mathcal{I} - \mathcal{I}$. Moreover, if we index the rows and columns of the Gram matrix $\boldsymbol{Q}_\mathcal{I}$ with the elements of $\mathcal{I}$, in the same order as they appear in the basis vector (3.57), then the relation between the coefficients of $R(z)$ and the elements of $\boldsymbol{Q}_\mathcal{I}$ is

$$r_{\boldsymbol{k}} = \sum_{\boldsymbol{i}-\boldsymbol{l}=\boldsymbol{k}} (\boldsymbol{Q}_\mathcal{I})_{\boldsymbol{i},\boldsymbol{l}} \,. \tag{3.68}$$

This relation results directly from (3.61). ■

Example 3.21 (*continued*) For $\boldsymbol{k} = (2, 0)$, it results from (3.63) that $r_{\boldsymbol{k}} = (\boldsymbol{Q}_\mathcal{I})_{10}$, as $\boldsymbol{k} = \boldsymbol{i} - \boldsymbol{l}$, with $\boldsymbol{i} = (2, 0)$, $\boldsymbol{l} = (0, 0)$, which are in positions 1 and 0, respectively, in the set $\mathcal{I}$ ordered as in (3.59).

For $\boldsymbol{k} = (-1, 1)$, we have $\boldsymbol{k} = \boldsymbol{i} - \boldsymbol{\ell}$, with $\boldsymbol{i} = (1, 1)$, $\boldsymbol{l} = (2, 0)$, which are in positions 2 and 1, respectively, in the set $\mathcal{I}$; it results that $r_{\boldsymbol{k}} = (\boldsymbol{Q}_\mathcal{I})_{21}$. ■

3.6.2 Relaxations

For sparse polynomials, the Gram matrix representation (3.60) is tempting due to the lower complexity associated with smaller Gram matrices. In this case, the sum-of-squares relaxations are based on the same idea as in Sect. 3.5: We replace a sparse nonnegative polynomial with a sum-of-squares whose Gram matrices are defined by a chosen set of degrees $\mathcal{I}$; of course, the set $\mathcal{I}$ includes the minimal set necessary to obtain $R(z)$ in (3.60). We denote

$$\mathbb{RS}_\mathcal{I}[z] = \{R \in \mathbb{R}[z] \mid R(z) \text{ as in (3.60)},\ \ \boldsymbol{Q}_\mathcal{I} \succeq 0\}, \tag{3.69}$$

the set of such sparse sum-of-squares. For two sets of degrees $\mathcal{I} \subset \mathcal{J}$ (with degrees less or equal $\boldsymbol{n}$), we clearly have (compare with (3.42))

$$\mathbb{RS}_\mathcal{I}[z] \subset \mathbb{RS}_\mathcal{J}[z] \subset \mathbb{RS}_{\boldsymbol{n}}^{\boldsymbol{n}}[z] \subset \overline{\mathbb{RP}}_{\boldsymbol{n}}[z]. \tag{3.70}$$

In the sparse case, the sum-of-squares relaxation means replacing a nonnegative polynomial $R \in \overline{\mathbb{RP}}_{\boldsymbol{n}}[z]$ with the sum-of-squares $R \in \mathbb{RS}_{\mathcal{I}}[z]$, with the advantages and limitations discussed in Sect. 3.5.

Problem (*Min_poly_value*) We come back again to our workhorse. Let $R \in \mathbb{R}_{\boldsymbol{n}}[z]$ be a given sparse polynomial, whose minimum value on $\mathbb{T}^d$ is sought. Given a set of degrees $\mathcal{I}$, the sum-of-squares relaxation of the basic problem (3.46) has a form similar to (3.50), namely

$$\begin{aligned} \mu_{\mathcal{I}}^{\star} = \max_{\mu, \tilde{\boldsymbol{Q}}_{\mathcal{I}}} \ & \mu \\ \text{s.t.} \ & \mu + \operatorname{tr} \tilde{\boldsymbol{Q}}_{\mathcal{I}} = r_{\boldsymbol{0}} \\ & \operatorname{tr}[\boldsymbol{\Theta}_{\boldsymbol{k}}(\mathcal{I}) \tilde{\boldsymbol{Q}}_{\mathcal{I}}] = r_{\boldsymbol{k}}, \quad \boldsymbol{k} \in (\mathcal{I} - \mathcal{I}) \cap (\mathcal{H} \setminus \{\boldsymbol{0}\}) \\ & \tilde{\boldsymbol{Q}}_{\mathcal{I}} \succeq 0, \ \ \tilde{\boldsymbol{Q}}_{\mathcal{I}} \in \mathbb{R}^{|\mathcal{I}| \times |\mathcal{I}|} \end{aligned} \tag{3.71}$$

Remark that the equality constraints are imposed only for the distinct coefficients from $\mathcal{I} - \mathcal{I}$ (some of them may be zero). The other coefficients are zero by construction (i.e., by the choice of $\mathcal{I}$), and so there is no need of an explicit constraint for them. So, the complexity of solving (3.71) is lower than that of (3.50) not only due to the smaller size of the Gram matrix $\tilde{\boldsymbol{Q}}_{\mathcal{I}}$, but also due to the smaller number of equality constraints.

Obviously, by solving (3.71), we obtain $\mu_{\mathcal{I}}^{\star} \le \mu^{\star}$. A certificate of optimality can be obtained as discussed in Sect. 3.5, e.g., by using the optimal Gram matrix. ■

Example 3.25 We compute the minimum value of the polynomial (3.62) by solving (3.71) for the set $\mathcal{I}$ giving the basis (3.59). We obtain the value $\mu_{\mathcal{I}}^{\star} = 10$. The same value is obtained when solving the complete problem (3.50), with $\boldsymbol{m} = \boldsymbol{n} = (2, 1)$; it is also the true optimal value. So, in this case, the relaxation has given the correct result.

Consider now the polynomial (obtained from (3.36) by removing the last monomial)

$$R(z) = \text{sym}^{-1} + 38 + 18z_1 + 4z_1^2 + z_1^{-2}z_2 + 2z_1^{-1}z_2 + z_2 - 8z_1z_2. \tag{3.72}$$

The smallest set $\mathcal{I}$ that can be used for representing this polynomial is

$$\mathcal{I} = \{(0,0), \ (2,0), \ (0,1), \ (1,1)\}$$

and the corresponding basis of monomials is

$$\boldsymbol{\psi}_{\mathcal{I}}(z) = [1 \ z_1^2 \ z_2 \ z_1z_2]^T.$$

The optimum of (3.71) is $\mu_{\mathcal{I}}^{\star} = -26$. However, solving the complete problem (3.50), with $\boldsymbol{m} = \boldsymbol{n} = (2, 1)$, leads to $\mu_{\boldsymbol{n}}^{\star} = -6$, which is the true optimum. So, in this case, the relaxation failed to provide the correct optimal value. This is due to the fact that the polynomial $R(z) + \alpha$, with $-26 < \alpha < -6$, although positive, has no sum-of-squares representation with the sparse factors generated by $\mathcal{I}$.

For the same polynomial (3.72), we use now the set of degrees

$$\mathcal{J} = \{(0,0),\ (1,0),\ (2,0),\ (0,1),\ (1,1)\},$$

obtaining the basis of monomials

$$\boldsymbol{\psi}_{\mathcal{J}}(z) = [1\ z_1\ z_1^2\ z_2\ z_1 z_2]^T.$$

The optimum of (3.71) is $\mu_{\mathcal{J}}^{\star} = -7.5$. As expected, we obtain $\mu_{\mathcal{J}}^{\star} > \mu_{\mathcal{I}}^{\star}$, i.e., a better approximation of the true optimum (however, still not the exact value).

This example has illustrated both sides of the relaxation. For highly sparse polynomials, as (3.62) could be an example, the relaxations may be successful even with relatively small bases. On the other hand, it is always useful to check the result by using a larger basis. ■

3.7 Gram Matrices of Real Multivariate Polynomials

Here, we generalize the Gram matrix parameterization presented in Sect. 2.7. Since the material is similar in spirit to that for trigonometric polynomials, we will give only the most important facts.

3.7.1 Gram Parameterization

Let $P \in \mathbb{R}_{2\boldsymbol{n}}[\boldsymbol{t}]$ be a real polynomial. A symmetric matrix $\boldsymbol{Q} \in \mathbb{R}^{N \times N}$, where N is given by (3.26), is called a Gram matrix associated with $P(\boldsymbol{t})$ if

$$P(\boldsymbol{t}) = \boldsymbol{\psi}_{\boldsymbol{n}}^T(\boldsymbol{t}) \cdot \boldsymbol{Q} \cdot \boldsymbol{\psi}_{\boldsymbol{n}}(\boldsymbol{t}), \tag{3.73}$$

where

$$\boldsymbol{\psi}_{\boldsymbol{n}}(\boldsymbol{t}) = \boldsymbol{\psi}_{n_d}(t_d) \otimes \ldots \otimes \boldsymbol{\psi}_{n_1}(t_1) = \bigotimes_{i=d}^{1} \boldsymbol{\psi}_{n_i}(t_i) \tag{3.74}$$

is a vector containing the canonical basis for d-variate polynomials of degree $\boldsymbol{n}$.

Theorem 3.26 *If $P \in \mathbb{R}_{2\boldsymbol{n}}[\boldsymbol{t}]$ and $\boldsymbol{Q}$ is a Gram matrix satisfying (3.73), then the relation*

$$p_{\boldsymbol{k}} = tr[\boldsymbol{\Upsilon}_{\boldsymbol{k}} \cdot \boldsymbol{Q}] \tag{3.75}$$

holds, where

$$\boldsymbol{\Upsilon}_{\boldsymbol{k}} = \boldsymbol{\Upsilon}_{k_d} \otimes \ldots \otimes \boldsymbol{\Upsilon}_{k_1} \tag{3.76}$$

and the matrices $\boldsymbol{\Upsilon}_k$ *are defined as in the body of Theorem 2.19.*

Moreover, the polynomial $P(\boldsymbol{t})$ *is sum-of-squares if and only if there exists a positive semidefinite matrix* $\boldsymbol{Q} \in \mathbb{C}^{N\times N}$ *such that (3.75) holds.*

Proof The proof is similar to that of Theorems 3.13 and 3.15 and is left as an exercise.

Example 3.27 Let $d = 2, n_1 = 2, n_2 = 1$. The vector containing the basis monomials is

$$\boldsymbol{\psi}(\boldsymbol{t}) = [1\ t_2]^T \otimes [1\ t_1\ t_1^2]^T = [1\ t_1\ t_1^2\ t_2\ t_1t_2\ t_1^2t_2]^T.$$

For $\boldsymbol{k} = (3, 1)$, the matrix $\boldsymbol{\Upsilon}_{\boldsymbol{k}}$ from (3.76) is

$$\boldsymbol{\Upsilon}_{\boldsymbol{k}} = \boldsymbol{\Upsilon}_1 \otimes \boldsymbol{\Upsilon}_3 = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} \otimes \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{bmatrix} = \left[\begin{array}{ccc|ccc} 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ \hline 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \end{array}\right]$$

and so the relation (3.75) gives $p_{\boldsymbol{k}} = 2(q_{51} + q_{42})$. ■

3.7.2 Sum-of-Squares Relaxations

For real polynomials, the basic idea of relaxation is the same: replace nonnegative polynomials with sum-of-squares and optimize using the Gram matrix representation. For example, if we seek the minimum value $\mu^\star = \min_{\boldsymbol{t}\in\mathbb{R}^d} P(\boldsymbol{t})$ of the polynomial $P \in \mathbb{R}_{2\boldsymbol{n}}[\boldsymbol{t}]$ (we assume that the minimum exists), we can relax this NP-hard problem to that of finding the maximum $\mu \in \mathbb{R}$ for which $\tilde{P}(\boldsymbol{t}) = P(\boldsymbol{t}) - \mu$ is sum-of-squares. Unlike the case of trigonometric polynomials, this formulation poses a difficulty: there may be polynomials for which $P(\boldsymbol{t}) - \mu$ is not a sum-of-squares for any value of μ.

Example 3.28 Consider again the polynomial (3.21). In Example 3.8, it is shown that this polynomial is not sum-of-squares. Reading again the proof, we see that this fact does not depend on the value of the free term. So, the polynomial $P(t_1, t_2) - \mu$ is not sum-of-squares for any μ. Trying to find the minimum value of the polynomial by a sum-of-squares relaxation leads to $\mu = -\infty$. ■

To make the relaxation more flexible, we can appeal to Theorem 3.10 and, for some $\kappa \in \mathbb{N}$, solve

$$\begin{aligned} \mu_\kappa^\star = {} & \max_{\mu} \ \mu \\ & \text{s.t.} \ \ (P(\boldsymbol{t}) - \mu)(1 + t_1^2 + \ldots + t_d^2)^\kappa \in \sum \mathbb{R}[\boldsymbol{t}]^2 \end{aligned} \tag{3.77}$$

For $\kappa = 0$, we obtain the basic relaxation described initially. Since it is possible that $P(\boldsymbol{t})(1 + t_1^2 + \ldots + t_d^2)^\kappa$ is not sum-of-squares, but $P(\boldsymbol{t})(1 + t_1^2 + \ldots + t_d^2)^{\kappa+1}$ is (and not the other way around), it follows that

$$\mu_0 \le \mu_1 \le \ldots \le \mu^\star. \tag{3.78}$$

We note that the coefficients of the sum-of-squares polynomial from (3.77) depend *linearly* on the variable μ (and, generally, on the coefficients of $P(\boldsymbol{t})$, which has no consequence here but may be important in other problems). So, when using the Gram matrix parameterization (3.75), we obtain a linear dependence between the elements of the Gram matrix $\boldsymbol{Q}$ and the other variables (μ, in the current case) of the optimization problem. Thus, the problem (3.77) can be implemented as an SDP problem.

3.7.3 Sparseness Treatment

Let $\mathcal{I} \in \mathbb{N}^d$ be a set of degrees (all less or equal a given $\boldsymbol{n}$). As in Sect. 3.6, we consider the vector of monomials

$$\boldsymbol{\psi}_{\mathcal{I}}(\boldsymbol{t}) = [\ldots \ \boldsymbol{t}^{\boldsymbol{k}} \ \ldots]^T, \quad \text{with } \boldsymbol{k} \in \mathcal{I}, \tag{3.79}$$

and note that $\boldsymbol{\psi}_{\mathcal{I}}(\boldsymbol{t}) = \boldsymbol{C}\boldsymbol{\psi}_{\boldsymbol{n}}(\boldsymbol{t})$, where $\boldsymbol{C}$ is a selection matrix of size $|\mathcal{I}| \times N$.

Theorem 3.29 *Let $P \in \mathbb{R}_{2\boldsymbol{n}}[\boldsymbol{t}]$ be a real polynomial that can be expressed as*

$$P(\boldsymbol{t}) = \boldsymbol{\psi}_{\mathcal{I}}^T(\boldsymbol{t}) \cdot \boldsymbol{Q}_{\mathcal{I}} \cdot \boldsymbol{\psi}_{\mathcal{I}}(\boldsymbol{t}), \tag{3.80}$$

for a given set of degrees $\mathcal{I}$ and a Gram matrix $\boldsymbol{Q}_{\mathcal{I}}$ of size $|\mathcal{I}| \times |\mathcal{I}|$. Then, the relation

$$p_{\boldsymbol{k}} = tr[\boldsymbol{\Upsilon}_{\boldsymbol{k}}(\mathcal{I}) \cdot \boldsymbol{Q}_{\mathcal{I}}] \tag{3.81}$$

holds, where $\boldsymbol{\Upsilon}_{\boldsymbol{k}}(\mathcal{I}) = \boldsymbol{C} \cdot \boldsymbol{\Upsilon}_{\boldsymbol{k}} \cdot \boldsymbol{C}^T$.

Proof Similar to that of Theorem 3.22. ■

Remark 3.30 An alternative way to describe relation (3.81) can be deduced as in Remark 3.24. The only monomials that can appear in a polynomial $P(\boldsymbol{t})$ of the form (3.80) are those from the matrix

$$\boldsymbol{\Psi}_{\mathcal{I}}(\boldsymbol{t}) = \boldsymbol{\psi}_{\mathcal{I}}(\boldsymbol{t})\boldsymbol{\psi}_{\mathcal{I}}^T(\boldsymbol{t}).$$

Their degrees belong to

$$\mathcal{I}+\mathcal{I}=\{\boldsymbol{k}\in\mathbb{N}^d \mid \boldsymbol{k}=\boldsymbol{i}+\boldsymbol{l},\ \boldsymbol{i},\boldsymbol{l}\in\mathcal{I}\}. \tag{3.82}$$

Moreover, if we index the rows and columns of the Gram matrix $\boldsymbol{Q}_{\mathcal{I}}$ with the elements of $\mathcal{I}$, in the same order as they appear in the basis vector (3.79), then we obtain the relation

$$p_{\boldsymbol{k}}=\sum_{\boldsymbol{i}+\boldsymbol{l}=\boldsymbol{k}}\left(\boldsymbol{Q}_{\mathcal{I}}\right)_{\boldsymbol{i},\boldsymbol{l}} \tag{3.83}$$

between the coefficients of the polynomial $P(\boldsymbol{t})$ and the elements of the Gram matrix. ■

Practically, given a polynomial $P\in\mathbb{R}_{2\boldsymbol{n}}[\boldsymbol{t}]$ whose support is $\mathcal{J}$, we are interested in finding a minimal set $\mathcal{I}$ such that $\mathcal{J}\subset\mathcal{I}+\mathcal{I}$. The following result shows that for real sum-of-squares polynomials, the set $\mathcal{I}$ can be nicely confined. (Remark that a similar result does not exist for trigonometric sum-of-squares!)

Theorem 3.31 *Let $P\in\mathbb{R}_{2\boldsymbol{n}}[\boldsymbol{t}]$ be a real polynomial whose support is $\mathcal{S}(P)$. If $P(\boldsymbol{t})$ is sum-of-squares, i.e., it can be written as in (3.19), then the support of the factors $F_\ell(\boldsymbol{t})$, $\ell=1:\nu$, is such that*

$$\mathcal{S}(F_\ell)\subset\frac{1}{2}conv\left(\mathcal{S}(P)\right), \tag{3.84}$$

where conv($\mathcal{A}$) is the convex hull of the set $\mathcal{A}$.

The set conv$(\mathcal{S}(P))$ is called *Newton polytope* associated with the polynomial $P(\boldsymbol{t})$.

Example 3.32 Let us consider the polynomial ($d=2, n_1=2, n_2=1$)

$$P(t_1,t_2)=3+4t_1^3+3t_1^4-4t_1^2t_2+4t_1^3t_2+2t_1^2t_2^2. \tag{3.85}$$

Its support is shown in Fig. 3.5. The polynomial has 6 monomials, marked with filled circles. The border of the convex hull of its support (obtained by all convex

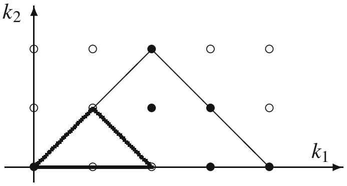


Fig. 3.5 *Filled circles* represent the support $\mathcal{S}(P)$ of the polynomial (3.85). The *thin line* is the border of conv$(\mathcal{S}(P))$. The *thick line* is the border of $\frac{1}{2}$conv$(\mathcal{S}(P))$, which is the support of the factors in the sum-of-squares

combinations between the points of $\mathcal{S}(P)$) is represented with a thin line. If this polynomial is a sum-of-squares (and it is!), then Theorem 3.31 says that the support of the factors $F_\ell(\boldsymbol{t})$ is inside the thick line. Indeed, one can represent the polynomial (3.85) as

$$P(t_1,t_2) = (1 - t_1 + t_1^2 + t_1t_2)^2 + (1 + t_1 - t_1^2 - t_1t_2)^2 + (1 - t_1^2)^2.$$

The support of the factors is $\{(0,0), (0,1), (0,2), (1,1)\}$, as seen in the figure.

Consider now the polynomial (3.21) from Example 3.8, which is nonnegative, but not sum-of-squares. Its support is $\{(0,0), (2,2), (4,2), (2,4)\}$. If it would be sum-of-squares, its factors would have the support limited to $\{(0,0), (1,1), (2,1), (1,2)\}$. Writing the expression (3.19) with such factors and looking at the coefficient of $t_1^2t_2^2$ results immediately in a contradiction. So, in this case, we can prove much faster than in Example 3.8 that the polynomial is not sum-of-squares. ■

Proof of Theorem 3.31. We only sketch a possible line of proof. We try to confine $\mathcal{S}(F_\ell)$ to a (convex) polyhedron $\mathcal{I}$. Initially, this polyhedron is the (hyper)-parallelepiped $[0,n_1] \times \ldots \times [0,n_d]$. We look at a vertex $\boldsymbol{v} \in \mathcal{I}$; the vertices are points in the polyhedron that cannot be expressed as a convex combination of other points of the polyhedron. Assume that $2\boldsymbol{v} \notin \mathcal{S}(P)$, i.e., $p_{2\boldsymbol{v}} = 0$. Looking at the terms $\boldsymbol{t}^{2\boldsymbol{v}}$ in the sum-of-squares expression (3.19), we see that $p_{2\boldsymbol{v}} = \sum_\ell (F_\ell)_{\boldsymbol{v}}^2$. (There are no other terms in the above equality since the monomials with degrees $\boldsymbol{i}, \boldsymbol{l} \in \mathcal{I}$ of a factor in the sum-of-squares produce a monomial with degree $\boldsymbol{k} = \boldsymbol{i} + \boldsymbol{l}$; if $\boldsymbol{k} = 2\boldsymbol{v}$ and $\boldsymbol{v}$ is a vertex of $\mathcal{I}$, then the only way of obtaining the monomial is with $\boldsymbol{k} = \boldsymbol{v}+\boldsymbol{v}$.) It follows that $(F_\ell)_{\boldsymbol{v}} = 0$, for any ℓ, and so $\boldsymbol{v} \notin \mathcal{I}$.

This trimming process continues until we have $2\boldsymbol{v} \in \mathcal{S}(P)$, for any vertex $\boldsymbol{v} \in \mathcal{I}$. It follows that $\mathcal{I} = \frac{1}{2}\mathrm{conv}(\mathcal{S}(P))$ and the theorem is proved. ■

The implication of Theorem 3.31 is clear. For sparse polynomials, we can safely reduce the degrees of the monomials in the factors of the sum-of-squares decomposition to the set $\mathcal{I} = \frac{1}{2}\mathrm{conv}(\mathcal{S}(P))$. Of course, this approach has to be combined with relaxations in the style of problem (3.77).

3.8 Pairs of Relaxations

Let us look again at the modality to obtain relaxations for trigonometric and real polynomials. We continue to use as a prototype the problem of finding the minimum value of a polynomial.

For a trigonometric polynomial $R \in \mathbb{R}_{\boldsymbol{n}}[z]$, the relaxations are obtained as in problem (3.48), by allowing the degrees of the sum-of-squares factors $H_\ell(\boldsymbol{z})$ from (3.13) to be larger than the degree of $R(\boldsymbol{z})$. The polynomial is not altered.

For a real polynomial $P \in \mathbb{R}_{2\boldsymbol{n}}[\boldsymbol{t}]$, the relaxation is obtained as in problem (3.77), by multiplying $P(\boldsymbol{t})$ with a fixed polynomial and expressing the result as a sum-of-squares. Naturally, the sum-of-squares factors $F_\ell(\boldsymbol{t})$ have degree larger than $\boldsymbol{n}$.

Although these two relaxations are different in form, we may ask whether they are related by the transformations described in Sect. 3.11.1. Precisely, the question is what we obtain if we transform the real polynomial $P(\boldsymbol{t})$ (appearing in some optimization problem) into an $R(z)$ via (3.121), relax the problem as for trigonometric polynomials, and go back via (3.120). The answer is already given by Remark 3.39, precisely by relation (3.123), showing how sum-of-squares are transformed. Using this approach, the relaxation has the form

$$\begin{aligned} \mu_{\boldsymbol{\kappa}}^\star = \ & \max_{\mu} \ \mu \\ & \text{s.t.} \ \ (P(\boldsymbol{t}) - \mu)(1+t_1^2)^{\kappa_1}\dots(1+t_d^2)^{\kappa_d} \in \sum \mathbb{R}[\boldsymbol{t}]^2 \end{aligned} \tag{3.86}$$

Remark the differences with respect to (3.77). The relaxation (3.86) may be more interesting because one can take, e.g., only one nonzero κ_i and thus obtain a sparser polynomial than in (3.77). However, the degree of the polynomial from (3.86) is potentially larger than in (3.77).

There is also a difference on the theoretical side (which probably has little importance in practice). We must be aware that (3.86) is not based on a result similar to Theorem 3.10. That is, it is not proved that for any positive polynomial $P(\boldsymbol{t})$, there exists $\boldsymbol{\kappa} \in \mathbb{N}^d$ such that $P(\boldsymbol{t})(1+t_1^2)^{\kappa_1}\dots(1+t_d^2)^{\kappa_d}$ is sum-of-squares. (This would be true if all nonnegative trigonometric polynomials would be sum-of-squares.)

We look now at the reverse transformation. Given a trigonometric polynomial $R(\boldsymbol{z})$, we transform it into a real polynomial $P(\boldsymbol{t})$ via (3.120), apply a relaxation as for real polynomials and go back via (3.121). Since the transformations apply separately to factors, this is equivalent to multiplying the polynomial $R(\boldsymbol{z})$ with the transform (3.121) of $(1+t_1^2+\dots+t_d^2)^\kappa$, i.e., with the polynomial

$$A_\kappa(\boldsymbol{z}) = \left(1 - \frac{(1-z_1)^2}{(1+z_1)^2} - \dots - \frac{(1-z_d)^2}{(1+z_d)^2}\right)^\kappa \prod_{i=1}^d \frac{(1+z_i)^2}{4z_i}. \tag{3.87}$$

So, we replace the relaxation (3.48) by

$$\begin{aligned} \mu_\kappa^\star = \ & \max_{\mu} \ \mu \\ & \text{s.t.} \ \ (R(\boldsymbol{z}) - \mu)A_\kappa(\boldsymbol{z}) \in \mathbb{RS}_{\boldsymbol{n}+\kappa}^{\boldsymbol{n}+\kappa}[\boldsymbol{z}] \end{aligned} \tag{3.88}$$

Remark that the degree of the sum-of-squares polynomial has increased to $\boldsymbol{n}+\kappa$, but the factors have the same degree. Therefore, the problem has approximately the same complexity as (3.48) for $\boldsymbol{m} = \boldsymbol{n}+\kappa$. However, it is simpler to write a program for (3.48), since there is no multiplication of $R(\boldsymbol{z}) - \mu$ with a fixed polynomial (depending, however, on the degree κ of relaxation).

We have thus presented two pairs of relaxations, (3.48)–(3.86) and (3.88)–(3.77). The first problem of the pair is for trigonometric polynomials, the second for real polynomials. The relaxations from the same pair are obtained by transformation from one to the other. It is difficult to give preference to a formulation. However, the basic problems (3.48) and (3.77) seem to have a slight advantage.

3.9 The Gram-Pair Parameterization

We investigate now the multivariate version of the Gram-pair parameterization presented in Sect. 2.8.3. Let $R(z)$ be a sum-of-squares d-variate trigonometric polynomial with *real* coefficients. Let the positive orthant polynomial $H(z)$ be a generic factor in a term of the sum-of-squares representation (3.13). We assume that $R \in \mathbb{RS}_{\boldsymbol{n}}^{\boldsymbol{n}}[z]$ and so the degree of $H(z)$ is $\boldsymbol{n}$. (It is the degree of the sum-of-squares factors that matters in all that follows.) We can write $\boldsymbol{n} = 2\tilde{\boldsymbol{n}} + \boldsymbol{\delta}$, where $\boldsymbol{\delta} = \boldsymbol{n} \bmod 2$; the elements of the vector $\boldsymbol{\delta}$ indicate the parity (0 means even and 1 means odd) of the elements of $\boldsymbol{n}$.

3.9.1 Basic Gram-Pair Parameterization

We define the (pseudo)-polynomial

$$\tilde{H}(z) = z^{\boldsymbol{n}/2} H(z) = \sum_{\boldsymbol{k}=\boldsymbol{0}}^{\boldsymbol{n}} h_{\boldsymbol{k}} z^{\boldsymbol{n}/2-\boldsymbol{k}}. \tag{3.89}$$

On the unit circle, we have

$$\tilde{H}(\boldsymbol{\omega}) = \sum_{\boldsymbol{k}=\boldsymbol{0}}^{\boldsymbol{n}} h_{\boldsymbol{k}} \left[\cos(\boldsymbol{k}-\boldsymbol{n}/2)^T\boldsymbol{\omega} - j\sin(\boldsymbol{k}-\boldsymbol{n}/2)^T\boldsymbol{\omega}\right] = A(\boldsymbol{\omega}) + jB(\boldsymbol{\omega}). \tag{3.90}$$

We perform now two simple operations in (3.90). First, since $\boldsymbol{k}-\boldsymbol{n}/2$ takes values that are symmetric with respect to the origin, we can group the terms in the sum such that $\boldsymbol{k}-\boldsymbol{n}/2$ is confined to a half-space $\mathcal{H}$; a similar operation has been done, e.g., in (2.87), in the univariate case. We then replace $\boldsymbol{k}$ with $\boldsymbol{k}-\tilde{\boldsymbol{n}}$. It results that

$$A(\boldsymbol{\omega}) = \boldsymbol{a}^T\boldsymbol{\chi}_c(\boldsymbol{\omega}), \quad B(\boldsymbol{\omega}) = \boldsymbol{b}^T\boldsymbol{\chi}_s(\boldsymbol{\omega}),$$

where

$$\begin{aligned} \boldsymbol{\chi}_c(\boldsymbol{\omega}) &= [\ \ldots\ \cos(\boldsymbol{k}-\boldsymbol{\delta}/2)^T\boldsymbol{\omega}\ \ldots\ ]^T, \\ \boldsymbol{\chi}_s(\boldsymbol{\omega}) &= [\ \ldots\ \sin(\boldsymbol{k}-\boldsymbol{\delta}/2)^T\boldsymbol{\omega}\ \ldots\ ]^T, \end{aligned} \quad -\tilde{\boldsymbol{n}} \le \boldsymbol{k} \le \tilde{\boldsymbol{n}}+\boldsymbol{\delta},\ \boldsymbol{k}-\boldsymbol{\delta}/2 \in \mathcal{H}, \tag{3.91}$$

are basis vectors of lengths N_c and N_s, respectively, and $\boldsymbol{a}$ and $\boldsymbol{b}$ are vectors of coefficients. The elements of $\boldsymbol{a}$ and $\boldsymbol{b}$ depend linearly on the coefficients of $H(z)$;

typically, they are a sum or difference of two coefficients, see e.g., (2.87) for the univariate case. Thus, there exist two matrices $\boldsymbol{C}_c \in \mathbb{R}^{N_c \times N}$ and $\boldsymbol{C}_s \in \mathbb{R}^{N_s \times N}$ such that

$$\begin{bmatrix} \boldsymbol{a} \\ \boldsymbol{b} \end{bmatrix} = \begin{bmatrix} \boldsymbol{C}_c \\ \boldsymbol{C}_s \end{bmatrix} \boldsymbol{h}, \tag{3.92}$$

where $\boldsymbol{h}$ is the vector (3.27) containing the coefficients of the polynomial. Moreover, the correspondence between a pair $(\boldsymbol{a}, \boldsymbol{b})$ and the vector $\boldsymbol{h}$ is one-to-one; so, the matrix from (3.92) is nonsingular and $N_c + N_s = N$ (as we will see again later).

From (3.90), we thus obtain

$$\begin{aligned} |H(\boldsymbol{\omega})|^2 = |\tilde{H}(\boldsymbol{\omega})|^2 &= A(\boldsymbol{\omega})^2 + B(\boldsymbol{\omega})^2 \\ &= \boldsymbol{\chi}_c^T(\boldsymbol{\omega})\boldsymbol{a}\boldsymbol{a}^T\boldsymbol{\chi}_c(\boldsymbol{\omega}) + \boldsymbol{\chi}_s^T(\boldsymbol{\omega})\boldsymbol{b}\boldsymbol{b}^T\boldsymbol{\chi}_s(\boldsymbol{\omega}). \end{aligned} \tag{3.93}$$

Using this relation, we can generalize Theorem 2.28 to the multivariate case.

Theorem 3.33 *Let $R \in \mathbb{R}_{\boldsymbol{n}}[z]$ be a trigonometric polynomial. The polynomial is sum-of-squares with factors of degree $\boldsymbol{n}$ (i.e., $R \in \mathbb{RS}_{\boldsymbol{n}}^{\boldsymbol{n}}[z]$) if and only if there exist positive semidefinite matrices $\boldsymbol{Q} \in \mathbb{R}^{N_c \times N_c}$ and $\boldsymbol{S} \in \mathbb{R}^{N_s \times N_s}$ such that*

$$R(\boldsymbol{\omega}) = \boldsymbol{\chi}_c^T(\boldsymbol{\omega})\boldsymbol{Q}\boldsymbol{\chi}_c(\boldsymbol{\omega}) + \boldsymbol{\chi}_s^T(\boldsymbol{\omega})\boldsymbol{S}\boldsymbol{\chi}_s(\boldsymbol{\omega}). \tag{3.94}$$

We name $(\boldsymbol{Q}, \boldsymbol{S})$ a Gram-pair associated with $R(\boldsymbol{\omega})$.

Proof The proof follows the already familiar pattern. If there exist $\boldsymbol{Q} \succeq 0$, $\boldsymbol{S} \succeq 0$ such that (3.94) holds, then by using the eigendecompositions of the matrices, it results that $R(z)$ is sum-of-squares as in Remark 2.9. Reciprocally, if $R(\boldsymbol{\omega})$ is sum-of-squares, then each term of the sum-of-squares can be expressed as in (3.93). It follows that the matrices $\boldsymbol{Q} \triangleq \sum \boldsymbol{a}\boldsymbol{a}^T$ and $\boldsymbol{S} \triangleq \sum \boldsymbol{b}\boldsymbol{b}^T$ (where the sums are taken for all the terms in the sum-of-squares decomposition), satisfy (3.94). ■

3.9.2 Parity Discussion

For univariate polynomials, we have seen in Sect. 2.8.3 that the parameterization (2.94) depended (through the constant matrices appearing there) on the parity of the degree. In the d-variate case, there are 2^d possible combinations of the parities of the degrees n_i, $i = 1 : d$. However, only $d + 1$ of them are essentially different, as we can reorder the variables such that the polynomial has, say, even order in the first variables and odd order in the others; so, the parity vector $\boldsymbol{\delta}$ is formed by a sequence of zeros followed by a sequence of ones. Here, we discuss two extreme parity cases.

(i) If all the degrees n_i are even and so $\boldsymbol{n} = 2\tilde{\boldsymbol{n}}$ (i.e., $\boldsymbol{\delta} = \boldsymbol{0}$), then the support of $H(z)$ contains its center of symmetry, as seen in the left of Fig. 3.6, for $d = 2$, $n_1 = 4, n_2 = 2$. After the translation of the support implied by (3.89), the support of

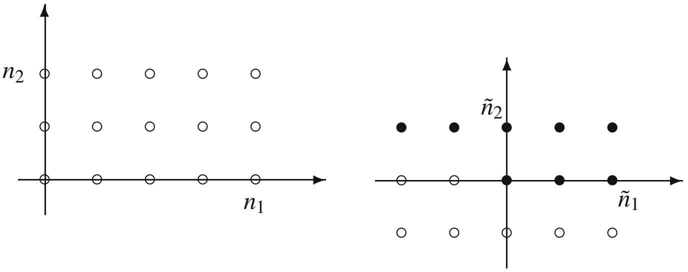


Fig. 3.6 Support of $H(z)$ (*left*) and $\tilde{H}(z)$ (*right*), for even degrees, in the 2D case. The degrees in the standard half plane are denoted with *bullets*

$\tilde{H}(z)$ is symmetric with respect to the origin and contains it. The restriction of this support to a half-space contains, as given by (3.4), a number of

$$N_c = \frac{1 + \prod_{i=1}^{d}(2\tilde{n}_i + 1)}{2} = \frac{1 + \prod_{i=1}^{d}(n_i + 1)}{2} = \frac{N+1}{2} \tag{3.95}$$

points, where N is defined in (3.26); N_c is the number of elements of the basis vector $\boldsymbol{\chi}_c(\boldsymbol{\omega})$ from (3.91). Since $\sin 0 = 0$, the vector $\boldsymbol{\chi}_s(\boldsymbol{\omega})$ has

$$N_s = \frac{N-1}{2} \tag{3.96}$$

useful elements. Note that $N_c + N_s = N$, as announced. The basis vectors (3.91) are

$$\begin{array}{ll} \boldsymbol{\chi}_c(\boldsymbol{\omega}) = [\ \dots\ \cos \boldsymbol{k}^T\boldsymbol{\omega}\ \dots\ ]^T, & \boldsymbol{k} \in \mathcal{H}, \\ \boldsymbol{\chi}_s(\boldsymbol{\omega}) = [\ \dots\ \sin \boldsymbol{k}^T\boldsymbol{\omega}\ \dots\ ]^T, & \boldsymbol{k} \in \mathcal{H} \setminus \{\boldsymbol{0}\}, \end{array} \quad -\tilde{\boldsymbol{n}} \le \boldsymbol{k} \le \tilde{\boldsymbol{n}}. \tag{3.97}$$

In our example, we have $N_c = 8$, $N_s = 7$ (and $N = 15$). Enumerating from left to right and upwards the points in the upper half plane, as in Fig. 3.9, the basis vectors (3.97) are

$$\boldsymbol{\chi}_c(\boldsymbol{\omega}) = \begin{bmatrix} 1 \\ \cos\omega_1 \\ \cos 2\omega_1 \\ \cos(-2\omega_1 + \omega_2) \\ \cos(-\omega_1 + \omega_2) \\ \cos\omega_2 \\ \cos(\omega_1 + \omega_2) \\ \cos(2\omega_1 + \omega_2) \end{bmatrix}, \quad \boldsymbol{\chi}_s(\boldsymbol{\omega}) = \begin{bmatrix} \sin\omega_1 \\ \sin 2\omega_1 \\ \sin(-2\omega_1 + \omega_2) \\ \sin(-\omega_1 + \omega_2) \\ \sin\omega_2 \\ \sin(\omega_1 + \omega_2) \\ \sin(2\omega_1 + \omega_2) \end{bmatrix}.$$

(ii) If at least one of the elements of $\boldsymbol{n}$ is odd, then the support of $H(z)$ no longer contains its center of symmetry, and thus the support of $\tilde{H}(z)$ does not contain the

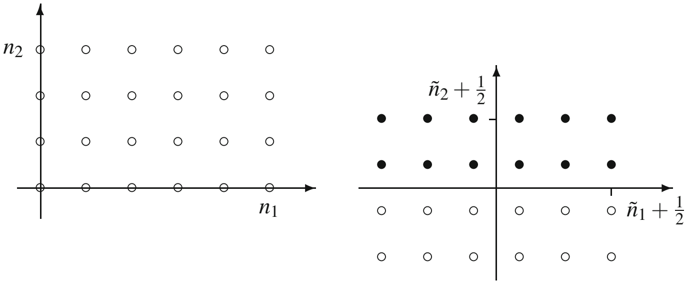


Fig. 3.7 Same as in Fig. 3.6, for odd degrees

origin. This situation is illustrated by Fig. 3.7, with $d = 2$, $n_1 = 5$, $n_2 = 3$. The number of points from the support of $\tilde{H}(z)$ that fall in the same half-space is exactly half the number of points in the support of $H(z)$ and so the basis vectors (3.91) have the same length

$$N_c = N_s = \frac{1}{2}\prod_{i=1}^{d}(n_i+1) = \frac{N}{2}.$$

If all the elements of $\boldsymbol{n}$ are odd (and so $\boldsymbol{\delta} = \mathbf{1}$), then the basis vectors are

$$\begin{array}{l} \boldsymbol{\chi}_c(\boldsymbol{\omega}) = [\ \dots\ \cos(\boldsymbol{k}-1/2)^T\boldsymbol{\omega}\ \dots\ ]^T, \\ \boldsymbol{\chi}_s(\boldsymbol{\omega}) = [\ \dots\ \sin(\boldsymbol{k}-1/2)^T\boldsymbol{\omega}\ \dots\ ]^T, \end{array} \quad -\tilde{\boldsymbol{n}} \le \boldsymbol{k} \le \tilde{\boldsymbol{n}}+1,\ \boldsymbol{k}-1/2 \in \mathcal{H}. \tag{3.98}$$

In our example, we have $N_c = N_s = 12$. The first basis vector from (3.98) has the form

$$\boldsymbol{\chi}_c(\boldsymbol{\omega}) = \begin{bmatrix} \cos(-5\omega_1+\omega_2)/2 \\ \cos(-3\omega_1+\omega_2)/2 \\ \cos(-\omega_1+\omega_2)/2 \\ \vdots \\ \cos(3\omega_1+3\omega_2)/2 \\ \cos(5\omega_1+3\omega_2)/2 \end{bmatrix}.$$

The basis vector $\boldsymbol{\chi}_s(\boldsymbol{\omega})$ is obtained by simply replacing cos with sin in the above formula.

3.9.3 LMI Form

Based on (3.94), we can state the multivariate version of Theorem 2.29, which is formally similar to the univariate version.

Theorem 3.34 *The trigonometric polynomial $R(z)$ is sum-of-squares with factors of degree $\boldsymbol{n}$ (i.e., $R \in \mathbb{RS}_{\boldsymbol{n}}^{\boldsymbol{n}}[z]$) if and only if there exist positive semidefinite matrices $\boldsymbol{Q} \in \mathbb{R}^{N_c \times N_c}$ and $\boldsymbol{S} \in \mathbb{R}^{N_s \times N_s}$ such that*

$$r_{\boldsymbol{k}} = tr[\boldsymbol{\Phi}_{\boldsymbol{k}} \boldsymbol{Q}] + tr[\boldsymbol{\Lambda}_{\boldsymbol{k}} \boldsymbol{S}], \tag{3.99}$$

where $\boldsymbol{\Phi}_{\boldsymbol{k}} \in \mathbb{R}^{N_c \times N_c}$ and $\boldsymbol{\Lambda}_{\boldsymbol{k}} \in \mathbb{R}^{N_s \times N_s}$ are constant matrices.

Proof The relations (3.94) and (3.99) express linear relations between the coefficients of $R(z)$ and the elements of the matrices $\boldsymbol{Q}$ and $\boldsymbol{S}$. The constant matrices $\boldsymbol{\Phi}_{\boldsymbol{k}}$ and $\boldsymbol{\Lambda}_{\boldsymbol{k}}$ result by simple identification. ■

Certainly, for implementing (3.99) we need the precise values of the matrices $\boldsymbol{\Phi}_{\boldsymbol{k}}$ and $\boldsymbol{\Lambda}_{\boldsymbol{k}}$. Although they result after detailing (3.94) and using trigonometric identities in the style of (2.77), an explicit formula for the matrices is not necessary for implementation. Instead, we can use relations similar to (2.93) to build the matrices. We assume that the rows and columns of the matrices $\boldsymbol{Q}$ and $\boldsymbol{S}$ are numbered via $\boldsymbol{i}, \boldsymbol{\ell} \in \mathbb{Z}^d$; the mapping between these d-dimensional numbers and the usual index range ($0 : N_c - 1$, for example), which is not unique, is not discussed; note, however, that the mostly used mappings are linear. (An example of program will be given later in Sect. 3.11.3.)

Let us consider only the case where all elements of the degree $\boldsymbol{n}$ are even. Using the basis vectors (3.97) and the identities (2.77), the relation (3.94) is equivalent to

$$\begin{aligned} R(\boldsymbol{\omega}) = \frac{1}{2} \sum_{\boldsymbol{i},\boldsymbol{\ell} \in \mathcal{H}} q_{\boldsymbol{i}\boldsymbol{\ell}} [\cos(\boldsymbol{i}+\boldsymbol{\ell})^T \boldsymbol{\omega} + \cos(\boldsymbol{i}-\boldsymbol{\ell})^T \boldsymbol{\omega}] \\ + \frac{1}{2} \sum_{\boldsymbol{i},\boldsymbol{\ell} \in \mathcal{H}^*} s_{\boldsymbol{i}\boldsymbol{\ell}} [-\cos(\boldsymbol{i}+\boldsymbol{\ell})^T \boldsymbol{\omega} + \cos(\boldsymbol{i}-\boldsymbol{\ell})^T \boldsymbol{\omega}] \end{aligned} \tag{3.100}$$

Here, we have denoted $\mathcal{H}^* = \mathcal{H} \setminus \{\boldsymbol{0}\}$; by $\boldsymbol{i} \in \mathcal{H}$, we understand implicitly that $|\boldsymbol{i}| \le \tilde{\boldsymbol{n}}$, as in (3.97). The coefficients of the polynomial are thus given by

$$\begin{aligned} r_{\boldsymbol{0}} &= q_{\boldsymbol{0}\boldsymbol{0}} + \frac{1}{2} \sum_{\boldsymbol{i} \in \mathcal{H}^*} q_{\boldsymbol{i}\boldsymbol{i}} + \frac{1}{2} \sum_{\boldsymbol{i} \in \mathcal{H}^*} s_{\boldsymbol{i}\boldsymbol{i}}, \\ r_{\boldsymbol{k}} &= \frac{1}{4} \left(\sum_{\substack{\boldsymbol{i}+\boldsymbol{\ell}=\boldsymbol{k} \\ \boldsymbol{i},\boldsymbol{\ell} \in \mathcal{H}}} q_{\boldsymbol{i}\boldsymbol{\ell}} + \sum_{\substack{\boldsymbol{i}-\boldsymbol{\ell}=\pm\boldsymbol{k} \\ \boldsymbol{i},\boldsymbol{\ell} \in \mathcal{H}}} q_{\boldsymbol{i}\boldsymbol{\ell}} - \sum_{\substack{\boldsymbol{i}+\boldsymbol{\ell}=\boldsymbol{k} \\ \boldsymbol{i},\boldsymbol{\ell} \in \mathcal{H}^*}} s_{\boldsymbol{i}\boldsymbol{\ell}} + \sum_{\substack{\boldsymbol{i}-\boldsymbol{\ell}=\pm\boldsymbol{k} \\ \boldsymbol{i},\boldsymbol{\ell} \in \mathcal{H}^*}} s_{\boldsymbol{i}\boldsymbol{\ell}} \right), \quad \boldsymbol{k} \neq \boldsymbol{0}. \end{aligned} \tag{3.101}$$

Example 3.35 Let us consider the simplest example where relations (3.101) are applicable and nontrivial, namely $d = 2$, $n_1 = n_2 = 2$ (and so $\tilde{n}_1 = \tilde{n}_2 = 1$). The standard half plane support of $R(z)$ is shown in Fig. 3.8 (with ×); the figure shows

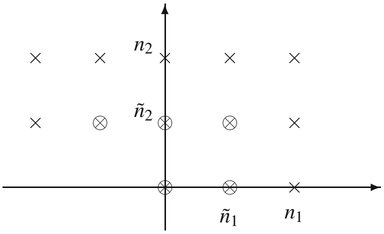


Fig. 3.8 With ×, support of $R(z)$. With *circles*, support of the basis vectors (3.97)

also the values $\boldsymbol{k}$ that appear in (3.97) (with circles). Accordingly, the sizes of the basis vectors (3.97) are $N_c = 5$, $N_s = 4$ (and $N = 9$).

A simple way to see the form of the matrices $\boldsymbol{\Phi}_{\boldsymbol{k}}$ and $\boldsymbol{\Lambda}_{\boldsymbol{k}}$ from (3.99) is to look at all the possible results $\boldsymbol{i} + \boldsymbol{\ell}$ and $\boldsymbol{i} - \boldsymbol{\ell}$, with $\boldsymbol{i}$ and $\boldsymbol{\ell}$ in the same half plane (i.e., the circles from Fig. 3.8). We enumerate the indices in a lexicographic order. For the sum, the table is

$\boldsymbol{i} \setminus \boldsymbol{\ell}$	$(0,0)$	$(1,0)$	$(-1,1)$	$(0,1)$	$(1,1)$
$(0,0)$	$(0,0)$	$(1,0)$	$(-1,1)$	$(0,1)$	$(1,1)$
$(1,0)$	$(1,0)$	$(2,0)$	$(0,1)$	$(1,1)$	$(2,1)$
$(-1,1)$	$(-1,1)$	$(0,1)$	$(-2,2)$	$(-1,2)$	$(0,2)$
$(0,1)$	$(0,1)$	$(1,1)$	$(-1,2)$	$(0,2)$	$(1,2)$
$(1,1)$	$(1,1)$	$(2,1)$	$(0,2)$	$(1,2)$	$(2,2)$

(3.102)

For the difference $\boldsymbol{i} - \boldsymbol{\ell}$, the table is

$\boldsymbol{i} \setminus \boldsymbol{\ell}$	$(0,0)$	$(1,0)$	$(-1,1)$	$(0,1)$	$(1,1)$
$(0,0)$	$(0,0)$	$(-1,0)$	$(1,-1)$	$(0,-1)$	$(-1,-1)$
$(1,0)$	$(1,0)$	$(0,0)$	$(2,-1)$	$(1,-1)$	$(0,-1)$
$(-1,1)$	$(-1,1)$	$(-2,1)$	$(0,0)$	$(-1,0)$	$(-2,0)$
$(0,1)$	$(0,1)$	$(-1,1)$	$(1,0)$	$(0,0)$	$(-1,0)$
$(1,1)$	$(1,1)$	$(0,1)$	$(2,0)$	$(1,0)$	$(0,0)$

(3.103)

For building $\boldsymbol{\Phi}_{\boldsymbol{k}}$ and $\boldsymbol{\Lambda}_{\boldsymbol{k}}$, we search the values $\boldsymbol{k}$ and $-\boldsymbol{k}$ in the tables (for $\boldsymbol{\Lambda}_{\boldsymbol{k}}$, we ignore the first rows and columns of the tables) and use the appropriate coefficients and signs as indicated by (3.101). Here are two pairs of matrices:

$$\boldsymbol{\Phi}_{(1,0)} = \frac{1}{4}\begin{bmatrix} 0 & 2 & 0 & 0 & 0 \\ 2 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 1 & 0 \end{bmatrix}, \quad \boldsymbol{\Lambda}_{(1,0)} = \frac{1}{4}\begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 0 \end{bmatrix},$$

Table 3.1 Times, in seconds, for finding the minimum value of 2D trigonometric polynomials using two parameterizations

Year	Parameterization	Order n						
		4	6	8	10	12	14	16
2006	Trace	0.36	1.1	5.4	18	65	270	540
	Gram pair	0.34	0.8	2.7	8.2	27	63	110
2016	Trace	0.08	0.15	0.5	1.1	3.0	7.0	18
	Gram pair	0.1	0.2	0.4	1.0	2.3	5.5	12

$$\boldsymbol{\Phi}_{(2,0)} = \frac{1}{4}\begin{bmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \end{bmatrix}, \quad \boldsymbol{\Lambda}_{(2,0)} = \frac{1}{4}\begin{bmatrix} -1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{bmatrix}.$$

A systematic way to build these matrices will be given in Sect. 3.11.3. ■

Remark 3.36 The Gram-pair parameterization (3.94) can be used instead of the generalized trace parameterization (3.32) without any restriction (for polynomials with real coefficients, obviously). For example, the multivariate version of problem (2.97) (which results formally by simply replacing ω with $\boldsymbol{\omega}$) can be used instead of (3.50) as a sum-of-squares relaxation for computing the minimum value of a trigonometric polynomial. As in the univariate case, we expect a speedup when using the Gram-pair parameterization, since the parameter matrices are twice smaller; remind that in (3.32), the Gram matrix is $N \times N$, while the sizes of the matrices from (3.94) are given by (3.95) and (3.96). Table 3.1 contains the times needed for solving the problem *Min_poly_value* using sum-of-squares relaxations (in $\mathbb{RS}_{\boldsymbol{n}}^{\boldsymbol{n}}[z]$), with the two parameterizations. The orders of the polynomials are $\boldsymbol{n} = (n, n)$. Like in Example 2.31, we repeat the experiments 10 years later. In the 2D case, in 2016, the advantage of the Gram-pair parameterization is smaller than in the univariate case (1.5 times faster vs twice faster). The largest size we tried is $\boldsymbol{n} = (20, 20)$ (not given in the table), in which case the size of the Gram matrix is $N = 441$, while the sizes of the Gram-pair matrices are $N_c = 221$, $N_s = 220$; the execution times are 75 and 48 s, respectively. So, again, on nowadays computers, limitations due to memory are apparently not visible at these degrees. We conclude that in the multivariate case, although the Gram-pair parameterization leads to faster solutions, it is less appealing, considering the implementation effort. ■

3.10 Polynomials with Matrix Coefficients

We extend the results presented until now to trigonometric polynomials with matrix coefficients, having the form

$$\boldsymbol{R}(z) = \sum_{\boldsymbol{k}=-\boldsymbol{n}}^{\boldsymbol{n}} \boldsymbol{R}_{\boldsymbol{k}} z^{-\boldsymbol{k}}, \quad \boldsymbol{R}_{-\boldsymbol{k}} = \boldsymbol{R}_{\boldsymbol{k}}^H. \tag{3.104}$$

The coefficients $\boldsymbol{R}_{\boldsymbol{k}} \in \mathbb{C}^{\kappa\times\kappa}$ are matrices, as well as the polynomial $\boldsymbol{R}(z)$ itself. The polynomial (3.104) is named positive if

$$\boldsymbol{R}(\boldsymbol{\omega}) \overset{\Delta}{=} \boldsymbol{R}(\mathrm{e}^{j\boldsymbol{\omega}}) \succ 0 \tag{3.105}$$

and nonnegative if the matrix $\boldsymbol{R}(\boldsymbol{\omega})$ is positive semidefinite. (Note that $\boldsymbol{R}(\boldsymbol{\omega})$ is Hermitian, i.e., $\boldsymbol{R}(\boldsymbol{\omega})^H = \boldsymbol{R}(\boldsymbol{\omega})$.)

A trigonometric polynomial is sum-of-squares if it can be written as

$$\boldsymbol{R}(z) = \sum_{\ell=1}^{\nu} \boldsymbol{H}_{\ell}(z)\boldsymbol{H}_{\ell}^H(z^{-1}), \tag{3.106}$$

where $\boldsymbol{H}_{\ell}(z)$ are positive orthant polynomials (defined as in (3.5), but with matrix coefficients). We note that Theorem 3.5 holds as stated for polynomials with matrix coefficients, i.e., any positive polynomial can be expressed as sum-of-squares, possibly with deg $\boldsymbol{H}_{\ell} > \boldsymbol{n}$.

Similarly to (2.1), we define

$$\boldsymbol{\psi}_n(z) = [\boldsymbol{I}_\kappa \; z\boldsymbol{I}_\kappa \; \ldots \; z^n\boldsymbol{I}_\kappa]^T \tag{3.107}$$

and interpret (3.24) as a canonical basis of monomials with matrix coefficients. A Hermitian matrix $\boldsymbol{Q} \in \mathbb{C}^{N\kappa\times N\kappa}$, where N is defined in (3.26), is called a Gram matrix associated with the polynomial (3.104) if

$$\boldsymbol{R}(z) = \boldsymbol{\psi}^T(z^{-1}) \cdot \boldsymbol{Q} \cdot \boldsymbol{\psi}(z). \tag{3.108}$$

In this context, it is natural to look at blocks of size $\kappa \times \kappa$. Let $\boldsymbol{A} \in \mathbb{C}^{p\kappa\times p\kappa}$ be a matrix split as

$$\boldsymbol{A} = [\boldsymbol{A}_{i\ell}]_{i,\ell=0:p-1}, \;\; \boldsymbol{A}_{i\ell} \in \mathbb{C}^{\kappa\times\kappa}. \tag{3.109}$$

We define a block trace operator by

$$\mathrm{TR}[\boldsymbol{A}] \overset{\Delta}{=} \sum_{i=0}^{p-1} \boldsymbol{A}_{ii}. \tag{3.110}$$

Now we can give the equivalent of Theorem 3.13 and Theorem 3.15. The proof is based on the same ideas and it will not be presented.

Theorem 3.37 *The relation between the coefficients of the polynomial (3.104) and the elements of the Gram matrix $\boldsymbol{Q} \in \mathbb{C}^{N\kappa\times N\kappa}$ from (3.108) is*

$$\boldsymbol{R_k} = TR[\boldsymbol{\Theta}_{\kappa,\boldsymbol{k}} \cdot \boldsymbol{Q}], \tag{3.111}$$

where

$$\boldsymbol{\Theta}_{\kappa,\boldsymbol{k}} = \boldsymbol{\Theta}_{\kappa_d} \otimes \ldots \otimes \boldsymbol{\Theta}_{\kappa_1} \otimes \boldsymbol{I}_\kappa = \boldsymbol{\Theta_k} \otimes \boldsymbol{I}_\kappa, \tag{3.112}$$

and $\boldsymbol{\Theta_k}$ is defined in (3.33). (The matrix $\boldsymbol{\Theta}_{\kappa,\boldsymbol{k}}$ is obtained from the $N \times N$ matrix $\boldsymbol{\Theta_k}$ by replacing the 1 values with $\kappa \times \kappa$ identity matrices and the 0 values with $\kappa \times \kappa$ zero matrices.)

Moreover, the polynomial (3.104) is sum-of-squares if and only if there exists a positive semidefinite matrix $\boldsymbol{Q} \in \mathbb{C}^{N\kappa \times N\kappa}$ such that (3.111) holds.

Example 3.38 With $d = 2$, $n_1 = n_2 = 1$, we consider the polynomial

$$\boldsymbol{R}(z) = \text{sym}^{-1} + \begin{bmatrix} 4 & 0 \\ 0 & 4 \end{bmatrix} + \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix} z_1 + \begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix} z_2 + \begin{bmatrix} 0 & -1 \\ 0 & 0 \end{bmatrix} z_1 z_2. \tag{3.113}$$

On the unit bicircle, the polynomial becomes

$$\boldsymbol{R}(\boldsymbol{\omega}) = \begin{bmatrix} 4 + 2\cos\omega_1 & -\cos(\omega_1 + \omega_2) \\ -\cos(\omega_1 + \omega_2) & 4 + 2\cos\omega_2 \end{bmatrix} + j \begin{bmatrix} 0 & -\sin(\omega_1 + \omega_2) \\ \sin(\omega_1 + \omega_2) & 0 \end{bmatrix}$$

and it is obvious that $\boldsymbol{R}(\boldsymbol{\omega})$ is a positive definite matrix for all $\boldsymbol{\omega}$. Let

$$\boldsymbol{Q} = \begin{bmatrix} \boldsymbol{Q}_{00} & \boldsymbol{Q}_{10}^T & \boldsymbol{Q}_{20}^T & \boldsymbol{Q}_{30}^T \\ \boldsymbol{Q}_{10} & \boldsymbol{Q}_{11} & \boldsymbol{Q}_{21}^T & \boldsymbol{Q}_{31}^T \\ \boldsymbol{Q}_{20} & \boldsymbol{Q}_{21} & \boldsymbol{Q}_{22} & \boldsymbol{Q}_{32}^T \\ \boldsymbol{Q}_{30} & \boldsymbol{Q}_{31} & \boldsymbol{Q}_{32} & \boldsymbol{Q}_{33} \end{bmatrix}$$

be a Gram matrix associated with $\boldsymbol{R}(z)$. The parameterization (3.111) is equivalent to the following equalities

$$\begin{aligned}
\boldsymbol{R}_{0,0} &= \begin{bmatrix} 4 & 0 \\ 0 & 4 \end{bmatrix} = \boldsymbol{Q}_{00} + \boldsymbol{Q}_{11} + \boldsymbol{Q}_{22} + \boldsymbol{Q}_{33}, \\
\boldsymbol{R}_{1,0} &= \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix} = \boldsymbol{Q}_{10} + \boldsymbol{Q}_{32}, \\
\boldsymbol{R}_{-1,1} &= \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix} = \boldsymbol{Q}_{21}, \\
\boldsymbol{R}_{0,1} &= \begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix} = \boldsymbol{Q}_{20} + \boldsymbol{Q}_{31}, \\
\boldsymbol{R}_{1,1} &= \begin{bmatrix} 0 & 0 \\ -1 & 0 \end{bmatrix} = \boldsymbol{Q}_{30}.
\end{aligned}$$

Remark that the above coefficients $\boldsymbol{R_k}$ are the transposed of those appearing in (3.113), due to the definition (3.104). An example of matrix (3.112) is

$$\boldsymbol{\Theta}_{2,(1,0)} = \left[\begin{array}{cc|cc|cc|cc} 0&0&1&0&0&0&0&0\\ 0&0&0&1&0&0&0&0\\ \hline 0&0&0&0&0&0&0&0\\ 0&0&0&0&0&0&0&0\\ \hline 0&0&0&0&0&0&1&0\\ 0&0&0&0&0&0&0&1\\ \hline 0&0&0&0&0&0&0&0\\ 0&0&0&0&0&0&0&0 \end{array}\right].$$

Two Gram matrices associated with $\boldsymbol{R}(z)$ are

$$\boldsymbol{Q}_1 = \left[\begin{array}{cc|cc|cc|cc} 4&0&&&\text{sym}&&&\\ 0&4&&&&&&\\ \cline{1-4} 1&0&0&0&&&&\\ 0&0&0&0&&&&\\ \cline{1-6} 0&0&0&0&0&0&&\\ 0&1&0&0&0&0&&\\ \hline 0&0&0&0&0&0&0&0\\ -1&0&0&0&0&0&0&0 \end{array}\right], \quad \boldsymbol{Q}_2 = \left[\begin{array}{cc|cc|cc|cc} 1.3&0&&&\text{sym}&&&\\ 0&0.7&&&&&&\\ \cline{1-4} 0.5&0&1&0&&&&\\ 0&0&0&1&&&&\\ \cline{1-6} 0&0&0&0&1&0&&\\ 0&0.5&0&0&0&1&&\\ \hline 0&0&0&0&0.5&0&0.7&0\\ -1&0&0&0.5&0&0&0&1.3 \end{array}\right]$$

(Check that $\boldsymbol{R}_{1,0} = \mathrm{TR}[\boldsymbol{\Theta}_{2,(1,0)} \cdot \boldsymbol{Q}_i]$, $i = 1, 2$.) The matrix $\boldsymbol{Q}_2$ is positive definite (its smallest eigenvalue is 0.0397), which confirms the positivity of $\boldsymbol{R}(z)$. (Again, the fact that $\boldsymbol{Q}_1$ is indefinite is of no consequence.) ■

Problem (*Most_positive_Gram_matrix*) The problem of finding the most positive Gram matrix (of size $N\kappa \times N\kappa$) associated with a polynomial with matrix coefficients is clearly well defined. Using the parameterization (3.111), we obtain an SDP problem similar to (3.38), namely

$$\begin{aligned} \lambda^\star = \max_{\lambda, \boldsymbol{Q}} \; & \lambda \\ \text{s.t.} \; & \mathrm{TR}[\boldsymbol{\Theta}_{\kappa,\boldsymbol{k}} \boldsymbol{Q}] = \boldsymbol{R}_{\boldsymbol{k}}, \quad \boldsymbol{k} \in \mathcal{H} \\ & \boldsymbol{Q} \succeq \lambda \boldsymbol{I} \end{aligned} \tag{3.114}$$

where $\mathcal{H}$ is a half-space. Since $\boldsymbol{\Theta}_{\kappa,\boldsymbol{0}} = \boldsymbol{I}$, it results that $\mathrm{TR}[\boldsymbol{\Theta}_0] = N\boldsymbol{I}_\kappa$; for $\boldsymbol{k} \neq 0$ we have $\mathrm{TR}[\boldsymbol{\Theta}_{\kappa,\boldsymbol{k}}] = 0$. Denoting $\tilde{\boldsymbol{Q}} = \boldsymbol{Q} - \lambda \boldsymbol{I}$, we transform (3.114) into the standard equality form SDP problem

$$\begin{aligned} \lambda^\star = \max_{\lambda, \tilde{\boldsymbol{Q}}} \; & \lambda \\ \text{s.t.} \; & N\lambda \boldsymbol{I}_\kappa + \mathrm{TR}[\tilde{\boldsymbol{Q}}] = \boldsymbol{R}_{\boldsymbol{0}} \\ & \mathrm{TR}[\boldsymbol{\Theta}_{\kappa,\boldsymbol{k}} \tilde{\boldsymbol{Q}}] = \boldsymbol{R}_{\boldsymbol{k}}, \quad \boldsymbol{k} \in \mathcal{H} \setminus \{\boldsymbol{0}\} \\ & \tilde{\boldsymbol{Q}} \succeq 0 \end{aligned} \tag{3.115}$$

Certainly, in an actual implementation, the $\kappa \times \kappa$ equalities from (3.115) are transformed into scalar equalities. The number of scalar equalities is $M\kappa^2 - \kappa(\kappa - 1)/2$, where M is the number of coefficients in a half-space, given by (3.4). (The first equality from (3.115) relates symmetric matrices, while the others involve general matrices.)

Since the size of the Gram matrix is $N\kappa \times N\kappa$, the complexity of solving (3.115) is $O(N^2 M^2 \kappa^6)$. ■

Example 3.38 (*continued*) Solving (3.115), we find that the most positive Gram matrix associated with the polynomial (3.113) is

$$\left[\begin{array}{rr|rr|rr|rr}
1.5483 & 0.0867 & & & & & & \\
0.0867 & 0.7784 & & & & \text{sym} & & \\
\cline{1-4}
0.4835 & -0.0377 & 0.6880 & -0.1721 & & & & \\
-0.2786 & 0.0986 & -0.1721 & 0.6880 & & & & \\
\cline{1-6}
-0.0986 & 0.0377 & 0 & 0 & 0.9854 & -0.0012 & & \\
0.2786 & 0.5165 & 0 & 0 & -0.0012 & 0.9854 & & \\
\cline{1-8}
0 & 0 & 0.0986 & -0.0377 & 0.5165 & 0.0377 & 0.7784 & 0.0867 \\
-1.0000 & 0 & -0.2786 & 0.4835 & 0.2786 & -0.0986 & 0.0867 & 1.5483
\end{array}\right].$$

Its smallest eigenvalue is $\lambda^\star = 0.25$. ■

As we have seen before, the problem of finding the most positive Gram matrix is related to that of computing the minimum value of a polynomial. In the case of polynomials with matrix coefficients, the correspondent of (2.17) is

$$\begin{aligned}
\mu^\star = \ & \max_{\mu} \ \mu \\
& \text{s.t.} \ \ \boldsymbol{R}(\boldsymbol{\omega}) - \mu \boldsymbol{I}_\kappa \succeq 0, \quad \forall \boldsymbol{\omega} \in [-\pi, \pi]^d
\end{aligned} \tag{3.116}$$

and it amounts to finding the minimal value of the *smallest eigenvalue* of the matrix $\boldsymbol{R}(\boldsymbol{\omega})$. We can solve a relaxed version of (3.116) by imposing the condition that $\boldsymbol{R}(z) - \mu \boldsymbol{I}_\kappa$ is sum-of-squares. The resulting problem is

$$\begin{aligned}
\mu^\star_{\boldsymbol{m}} = \ & \max_{\mu, \tilde{\boldsymbol{Q}}} \ \mu \\
& \text{s.t.} \ \ \mu \boldsymbol{I}_\kappa + \mathrm{TR}[\tilde{\boldsymbol{Q}}] = \boldsymbol{R}_{\boldsymbol{0}} \\
& \qquad \mathrm{TR}[\boldsymbol{\Theta}_{\kappa,\boldsymbol{k}} \tilde{\boldsymbol{Q}}] = \boldsymbol{R}_{\boldsymbol{k}}, \ \ \boldsymbol{k} \in \mathcal{H} \setminus \{\boldsymbol{0}\} \\
& \qquad \tilde{\boldsymbol{Q}} \succeq 0
\end{aligned} \tag{3.117}$$

The size of the Gram matrix $\tilde{\boldsymbol{Q}}$ is taken according to the degree $\boldsymbol{m}$ of the relaxation, as discussed in Sect. 3.5. We obtain $\mu^\star_{\boldsymbol{m}} \le \mu^\star$. We note that the solutions of problems (3.115) and (3.117) are related by $\mu^\star_{\boldsymbol{m}} = N\lambda^\star$ (we redefine N as in (3.45), such that the size of the Gram matrix is $N\kappa \times N\kappa$). We also remark that, in general, there is no $\boldsymbol{\omega}$ such that $\boldsymbol{R}(\boldsymbol{\omega}) = \mu^\star_{\boldsymbol{m}} \boldsymbol{I}_\kappa$. However, there exists an $\boldsymbol{\omega}$ such that the matrix $\boldsymbol{R}(\boldsymbol{\omega}) - \mu^\star \boldsymbol{I}_\kappa$ is singular (and, of course, positive semidefinite). So, if we find an $\boldsymbol{\omega}$ such that $\boldsymbol{R}(\boldsymbol{\omega}) - \mu^\star_{\boldsymbol{m}} \boldsymbol{I}_\kappa$ is singular, then we are sure that $\mu^\star_{\boldsymbol{m}} = \mu^\star$.

Example 3.38 (*continued*) The solution of (3.117) for the polynomial (3.113) is $\mu^\star = 1$; we use the smallest relaxation possible, with the degree of the sum-of-squares factors $\boldsymbol{m} = \boldsymbol{n} = (1, 1)$, and so $\mu^\star = 4\lambda^\star$. We conclude that $\boldsymbol{R}(\boldsymbol{\omega}) \succeq \boldsymbol{I}_\kappa$, for all $\boldsymbol{\omega}$. Since

$$\boldsymbol{R}(0,0) - \boldsymbol{I}_2 = \begin{bmatrix} 1 & -1 \\ -1 & 1 \end{bmatrix}$$

has the smallest eigenvalue equal to 0, we conclude that we have found the true optimum. ■

3.11 Details and Other Facts

3.11.1 Transformation Between Trigonometric and Real Nonnegative Polynomials

We are interested by a positivity-preserving transformation between trigonometric polynomials and real polynomials. The obvious candidate is the bilinear transform (1.58). For multivariate polynomials, we define it for each variable by

$$z_i = \frac{1 + jt_i}{1 - jt_i} = \frac{j - t_i}{j + t_i}, \quad i = 1 : d. \tag{3.118}$$

The inverse transform is

$$t_i = j\frac{1 - z_i}{1 + z_i}, \quad i = 1 : d. \tag{3.119}$$

We transform a trigonometric polynomial $R \in \mathbb{R}_{\boldsymbol{n}}[z]$ into a real polynomial $P \in \mathbb{R}_{2\boldsymbol{n}}[\boldsymbol{t}]$ by

$$P(\boldsymbol{t}) = R\left(\frac{j - t_1}{j + t_1}, \ldots, \frac{j - t_d}{j + t_d}\right) \prod_{i=1}^{d} (1 + t_i^2)^{n_i}. \tag{3.120}$$

So, we replace each complex variable by (3.118) and then multiply with the product $\prod_{i=1}^{d}(1 + t_i^2)^{n_i}$ in order to cancel the resulting denominator (see (1.62)). Since

$$1 + t_i^2 = \frac{4z_i}{(1 + z_i)^2},$$

the transformation inverse to (3.120) is

$$R(\boldsymbol{z}) = P\left(j\frac{1 - z_1}{1 + z_1}, \ldots, j\frac{1 - z_d}{1 + z_d}\right) \prod_{i=1}^{d} \frac{(1 + z_i)^{2n_i}}{(4z_i)^{n_i}}. \tag{3.121}$$

Since (3.118) maps the real axis to the unit circle (for each variable), a nonnegative polynomial $P(\boldsymbol{t})$ is mapped to a nonnegative polynomial $R(\boldsymbol{z})$ through (3.121) or, in the reverse sense, through (3.120). However, a positive $P(\boldsymbol{t})$ can be mapped to a nonnegative $R(\boldsymbol{z})$, i.e., strict positivity is not preserved by (3.121); this happens, for example, when the monomial $t_1^{2n_1} \cdots t_d^{2n_d}$ does not appear in $P(\boldsymbol{t})$; in this case, the relation (3.120) allows us to write

$$R(-1, \ldots, -1) = \lim_{t_i \to \infty,\ i=1:\ell} \frac{P(\boldsymbol{t})}{\prod_{i=1}^{d}(1+t_i^2)^{n_i}} = 0.$$

Remark 3.39 Moreover, the transformations (3.120) and (3.121) map sum-of-squares over $\mathbb{R}^d$ to sum-of-squares over $\mathbb{C}^d$. We prove this fact, giving also more details on the mapping.

Let $P \in \mathbb{R}_{2\boldsymbol{n}}[\boldsymbol{t}]$ be a sum-of-squares polynomial, having the form (3.19). Denoting

$$H_\ell(\boldsymbol{z}) = F_\ell\left(j\frac{1-z_1}{1+z_1}, \ldots, j\frac{1-z_d}{1+z_d}\right)\prod_{i=1}^{d}\frac{(1+z_i)^{n_i}}{2^{n_i}}$$

and noticing that for $\boldsymbol{z} \in \mathbb{T}^d$, we have

$$\begin{aligned} H_\ell^*(\boldsymbol{z}^{-1}) =& F_\ell\left(-j\frac{1-z_1^{-1}}{1+z_1^{-1}}, \ldots, -j\frac{1-z_d^{-1}}{1+z_d^{-1}}\right)\prod_{i=1}^{d}\frac{(1+z_i^{-1})^{n_i}}{2^{n_i}} \\ =& F_\ell\left(j\frac{1-z_1}{1+z_1}, \ldots, j\frac{1-z_d}{1+z_d}\right)^*\prod_{i=1}^{d}\frac{(1+z_i)^{n_i}}{2^{n_i}z_i^{n_i}}, \end{aligned}$$

it results that the trigonometric polynomial $R(\boldsymbol{z})$ defined by the transformation (3.121) is also sum-of-squares, with the form (3.13). Note that the degree of the factors $H_\ell(\boldsymbol{z})$ is at most $\boldsymbol{n}$.

In the opposite sense, let the trigonometric polynomial $R \in \mathbb{R}_{\boldsymbol{n}}[\boldsymbol{z}]$ be sum-of-squares, with factors of degree $\boldsymbol{n}$. We denote

$$F_\ell(\boldsymbol{t}) = H_\ell\left(\frac{j-t_1}{j+t_1}, \ldots, \frac{j-t_d}{j+t_d}\right)\prod_{i=1}^{d}(j+t_i)^{n_i}.$$

Since

$$\left(\frac{j-t_i}{j+t_i}\right)^* = \frac{j+t_i}{j-t_i},$$

it results that the transformation (3.120) gives

$$P(\boldsymbol{t}) = \sum_{\ell=1}^{\nu} F_\ell(\boldsymbol{t})F_\ell(\boldsymbol{t})^* = \sum_{\ell=1}^{\nu}(F_{\ell r}(\boldsymbol{t})^2 + F_{\ell i}(\boldsymbol{t})^2), \qquad (3.122)$$

where we have denoted $F_\ell(\boldsymbol{t}) = F_{\ell r}(\boldsymbol{t}) + jF_{\ell i}(\boldsymbol{t})$, with $F_{\ell r}$, $F_{\ell i}$ having real coefficients. So, the real polynomial $P(\boldsymbol{t})$ is sum-of-squares.

Now, let the trigonometric polynomial $R \in \mathbb{R}_{\boldsymbol{n}}[z]$ be sum-of-squares, with factors of degree $\boldsymbol{m} \geq \boldsymbol{n}$. Let $P(\boldsymbol{t})$ be the polynomial obtained by the transformation (3.120). As above, we denote

$$F_\ell(\boldsymbol{t}) = H_\ell\left(\frac{j-t_1}{j+t_1}, \ldots, \frac{j-t_d}{j+t_d}\right)\prod_{i=1}^{d}(j+t_i)^{m_i}.$$

The exponent of the rightmost factors is $\boldsymbol{m}$, otherwise $F_\ell(\boldsymbol{t})$ would not be a polynomial. It results that

$$\sum_{\ell=1}^{\nu} F_\ell(\boldsymbol{t})F_\ell(\boldsymbol{t})^* \stackrel{\Delta}{=} \tilde{P}(\boldsymbol{t}) = P(\boldsymbol{t}) \cdot \prod_{i=1}^{d}(1+t_i^2)^{m_i-n_i}. \tag{3.123}$$

So, if the degrees of the factors of the sum-of-squares trigonometric polynomial are larger than $\boldsymbol{n}$, then not $P(\boldsymbol{t})$ is sum-of-squares, but the polynomial $\tilde{P}(\boldsymbol{t})$ from (3.123). This is the way in which sum-of-squares are mapped by the transformations (3.120) and (3.121). ■

Example 3.40 We return to Example 3.7. The trigonometric polynomial (3.17) is obtained through the transformation (3.121) from the real polynomial

$$P(t_1, t_2) = t_1^4t_2^2 + t_1^2t_2^4 - t_1^2t_2^2 + 1. \tag{3.124}$$

This polynomial is positive, but not sum-of-squares, as shown in Example 3.8. So, the trigonometric polynomial (3.17) is not sum-of-squares with factors of degree $(2, 2)$. However, it is sum-of-squares with the factors of degree $(3, 3)$ shown in (3.18), which means that the real polynomial $\tilde{P}(t_1, t_2) = (1+t_1^2)(1+t_2^2)P(t_1, t_2)$ is sum-of-squares. We also notice that $R(-1, -1) = 0$, since there is no $t_1^4t_2^4$ term in $P(t_1, t_2)$. ■

3.11.2 Pos3Poly Program with Multivariate Polynomials

We give in Table 3.2 the multivariate version of the Pos3Poly [4] program from Sect. 2.12.1 for solving the problem *Min_poly_value*. The polynomial is given by the vector of its coefficients in a half-space, ordered as shown in Fig. 3.9 for the bivariate case; the order is lexicographic. For example, the polynomial (3.36) is represented by the vector `r = [38 18 4 1 2 1 -8 -5]`. The size of the vector is given by (3.4). Since this relation is not invertible, we need to feed the program with the degree of the polynomial. In our case, this is `n = [2 1]`. The equality constraint, which works with vectorized polynomials, models the fact the $R(z) - \mu$ is

Table 3.2 Pos3Poly program for the minimum of a multivariate trigonometric polynomial

```
function m = minpold_pos3poly(r)

r = r(:);         % force column vector
p = [n 1];        % degree and coefficients size (scalars)
ptype = [];
if ~isreal(r)     % complex data
  ptype.complex_coef = 1;
end
cvx_begin
  variable m;     % variable for the minimum
  maximize( m )
  subject to      % equality constraints
    m*eye( (prod(2*n+1)+1)/2, 1 ) + sos_pol(p, ptype) == r;
cvx_end
```

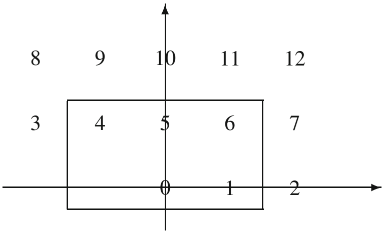


Fig. 3.9 Indices of the half plane coefficients of a symmetric polynomial with $d = 2, n_1 = n_2 = 2$

sum-of-squares; we could use `length(r)` instead of the explicit relation (3.4), which is there only for pedagogical purposes. Of course, running the program gives the value 1.8214 announced in Example 3.18. The sum-of-squares has the same degree as the polynomial; for using a higher degree sum-of-squares, the polynomial has to be padded with zeros and `n` has to be increased accordingly; the reader is invited to call the function in this manner.

Note that the differences with respect to the program from Table 2.4 are very small and that this program can also solve the univariate problem. This is the main incentive for using Pos3Poly.

3.11.3 A CVX Program Using the Gram-Pair Parameterization

We present here a CVX program for finding the minimum value of a bivariate trigonometric polynomial $R \in \mathbb{R}_n[z]$, using sum-of-squares relaxations in $\mathbb{RS}_n^n[z]$ (the adaptation of the program to a higher degree of the sum-of-squares factors is trivial). Note that Pos3Poly uses exclusively the trace parameterization. We discuss only the case where both elements of the order $\boldsymbol{n}$ are even. Using the Gram-pair parameterization (3.99), the SDP problem (in standard equality form) is

$$\begin{aligned} \mu_{\boldsymbol{n}}^{\star} = \max_{\mu, \tilde{\boldsymbol{Q}}, \tilde{\boldsymbol{S}}} \ & \mu \\ \text{s.t.} \ & \mu\delta_{\boldsymbol{k}} + \operatorname{tr}[\boldsymbol{\Phi}_{\boldsymbol{k}} \tilde{\boldsymbol{Q}}] + \operatorname{tr}[\boldsymbol{\Lambda}_{\boldsymbol{k}} \tilde{\boldsymbol{S}}] = r_{\boldsymbol{k}}, \quad \boldsymbol{k} \in \mathcal{H} \\ & \tilde{\boldsymbol{Q}} \succeq 0, \ \tilde{\boldsymbol{S}} \succeq 0 \end{aligned} \tag{3.125}$$

The program is presented in Table 3.3. The polynomial $R(z)$ is represented like in the previous subsection, by the vector `r` of coefficients in the standard half plane, in row major order. For the polynomial discussed in Example 3.35, the indices of the coefficients of $\operatorname{vec}(R)$ are as shown in Fig. 3.9.

The correspondence between the 2D index $\boldsymbol{k}$ and the index in `r` is given by the linear mapping

$$ind(\boldsymbol{k}) = (2n_1 + 1)k_2 + k_1. \tag{3.126}$$

Here, the indices start from 0, as in the whole book; in the MATLAB program, they start from 1. Note that if $\boldsymbol{k}$ belongs to the standard half plane, then $ind(\boldsymbol{k})$ is nonnegative.

For the Gram matrices, we also need the indices of the basis vectors (3.97). In Fig. 3.9, they are those inside the box. In the program, the vector `iv` contains the values $ind(\boldsymbol{k})$ for these indices. (To understand the second section of the program, note that each row of the standard half plane support of $R(z)$, excepting the lowest, contains $n_1 + 1$ elements; the value $ind(\boldsymbol{k})$ for the first element in the second lowest row is $2n_1 + 1 - \tilde{n}_1$.)

The fourth and fifth sections of the program build all the matrices $\boldsymbol{\Phi}_{\boldsymbol{k}}$ and $\boldsymbol{\Lambda}_{\boldsymbol{k}}$, initialized with zeros in the third section. The technique is to take all combinations of indices $\boldsymbol{i}$ and $\boldsymbol{\ell}$ of the basis vectors (3.97) and set, for each combination, the appropriate elements of the matrices $\boldsymbol{\Phi}_{\boldsymbol{k}}$ and $\boldsymbol{\Lambda}_{\boldsymbol{k}}$. Referring to Example 3.35, we build one by one the elements $\boldsymbol{k}$ of tables (3.102) and (3.103) and then set to $1/2$ (if $\boldsymbol{k} = \boldsymbol{0}$) or $\pm 1/4$ (otherwise) the element of $\boldsymbol{\Phi}_{\boldsymbol{k}}$ or $\boldsymbol{\Lambda}_{\boldsymbol{k}}$ in that position, as indicated by (3.101) (taking into account also the signs). This approach is possible due to the linearity of the index mapping (3.126), which means that

$$\boldsymbol{i} + \boldsymbol{\ell} = \boldsymbol{k} \iff ind(\boldsymbol{i}) + ind(\boldsymbol{\ell}) = ind(\boldsymbol{k}),$$

$$\boldsymbol{i} - \boldsymbol{\ell} = \pm\boldsymbol{k} \iff |ind(\boldsymbol{i}) - ind(\boldsymbol{\ell})| = ind(\boldsymbol{k}).$$

So, we do not have to work with 2D indices. Instead, we use the true indices `ii` and `ll` in the matrices $\boldsymbol{\Phi}_{\boldsymbol{k}}$ or $\boldsymbol{\Lambda}_{\boldsymbol{k}}$. The CVX part of the program is built directly from (3.125) with obvious correspondences.

Table 3.3 CVX program for solving the SDP problem (3.125)

```
function m = minpol2_g2_even_cvx(n, r)

nh = floor(n/2);          % half the order
Nc = (prod(n+1)+1)/2;     % number of coefs for cos base
Ns = Nc - 1;              % number of coefs for sin base

iv = (1:Nc)';             % halfplane indices
ii = nh(1)+1;
kk = 2*n(1) + 1 - nh(1);
for i2 = 1 : nh(2)
  iv( ii+1 : ii+n(1)+1 ) = kk+1 : kk+n(1)+1;
  ii = ii + n(1)+1;
  kk = kk + 2*n(1) + 1;
end

for k = 1 : length(r)     % initialize Phi and Lambda matrices
  Phi{k} = zeros(Nc);
  Lam{k} = zeros(Ns);
end

for ii = 1 : Nc           % compute Phi matrices
  for ll = 1 : Nc
    k = iv(ii) + iv(ll) - 1;      % coef affected by sum
    if k == 1, ad = 0.5; else, ad = 0.25; end
    Phi{k}(ii,ll) = Phi{k}(ii,ll) + ad;
    k = abs(iv(ii) - iv(ll)) + 1; % coef affected by difference
    if k == 1, ad = 0.5; else, ad = 0.25; end
    Phi{k}(ii,ll) = Phi{k}(ii,ll) + ad;
  end
end

iv = iv(2:end);           % no free term for sines
for ii = 1 : Ns           % compute Lambda matrices
  for ll = 1 : Ns
    k = iv(ii) + iv(ll) - 1;      % coef affected by sum
    if k == 1, ad = 0.5; else, ad = 0.25; end
    Lam{k}(ii,ll) = Lam{k}(ii,ll) - ad;
    k = abs(iv(ii) - iv(ll)) + 1; % coef affected by difference
    if k == 1, ad = 0.5; else, ad = 0.25; end
    Lam{k}(ii,ll) = Lam{k}(ii,ll) + ad;
  end
end

cvx_begin
  variable m;
  variable Q(Nc,Nc) semidefinite;
  variable S(Ns,Ns) semidefinite;
```

(continued)

Table 3.3 (continued)

```
 maximize m
 subject to
  m + vec(Phi{1})’ * vec(Q) + vec(Lam{1})’ * vec(S) == r(1);
  for k = 2 : length(r)
   vec(Phi{k})’ * vec(Q) + vec(Lam{k})’ * vec(S) == r(k);
  end
cvx_end
```

3.12 Bibliographical and Historical Notes

That any positive trigonometric polynomial is sum-of-squares is proved in different manners in several places [1, 5–7]. However, the problem did not attract special interest until the year 2000. The transformation described in Sect. 3.11.1 can be found in [1], but was used as early as in [8].

Nonnegative trigonometric polynomials in two variables are also sum-of-squares; this results from Theorem 3.3.1 from [9]. For more than two variables, one can build examples of nonnegative trigonometric polynomials that are not sum-of-squares; in a personal email, Markus Schweighofer described such a procedure that starts from a homogeneous nonnegative real polynomial that is not sum-of-squares.

The sum-of-squares problem for real polynomials is a classic theme in mathematics (Hilbert’s 17th problem) and good presentations can be found in [2, 3]. Theorem 3.10, the theoretical base for real sum-of-squares relaxations, appeared in [10].

Although not named so, the Gram matrix representation (for multivariate real polynomials) appeared for the first time in [11] as a tool to test the positivity of a polynomial; no optimization methods were suggested. The name *Gram* matrix was coined in [12], where the explicit relation to sum-of-squares was shown. An algorithm to find the sum-of-squares decomposition of a polynomial (or to infirm its existence) was proposed in [13], taking advantage of the convexity of the problem. The connection to SDP followed soon in [14–16] for real polynomials and in [17] for trigonometric polynomials, opening the way to solving a broad range of problems. The generalized trace parameterization as in Theorem 3.13 has been given in [18] and previous conference papers.

Sum-of-squares relaxations for trigonometric polynomials are discussed in [7, 18]. For real polynomials, the idea of sum-of-squares relaxations was originated by Shor [19]. Different SDP forms, based on different sum-of-squares decomposition are presented in [16, 20], with different approaches (to be discussed also in the next chapter). A good survey is [21]. The library SOSTOOLS [22] is an interface easing the implementation of such relaxations, based on the SDP library SeDuMi [23].

The Gram matrix representation can be immediately extended to hybrid polynomials, which are a mix of real and trigonometric polynomials. See **P** 3.10 and [24].

Sparseness was recognized from the beginning as an advantage, as the Gram matrices of sparse polynomials have smaller size. Theorem 3.22 (in the form given by Remark 3.24) was given in [17]. For real polynomials, the important Theorem 3.31 appeared already in [25]. More refined ways of identifying the monomials that appear in the terms of a sum-of-squares are presented in [26]. Sparse polynomials are dealt with in a very simple manner in SOSTOOLS.

The Gram-pair parameterization is a generalization of the univariate results from [27]; the simple proofs from Sect. 3.9 have appeared in [28]; more developments, including faster algorithms, are presented in [29].

Finally, positive polynomials with matrix coefficients were the natural presentation framework in [17, 30]. We have allocated Sect. 3.10 especially to them only for didactic purposes, since most of the scalar coefficient results generalize directly for matrix coefficients.

Problems

P 3.1 Show that the set $\mathcal{H} \in \mathbb{Z}^2$, defined by

$$\boldsymbol{k} \in \mathcal{H} \Leftrightarrow \begin{cases} k_1 \geq k_2, \text{ if } k_1 \geq 0, \\ k_1 > k_2, \text{ if } k_1 < 0, \end{cases}$$

is a half-space. Show how to build half-spaces in $\mathbb{Z}^2$ by using lines passing through the origin. Generalize to $\mathbb{Z}^d$.

P 3.2 Let $R, S \in \mathbb{C}[\boldsymbol{z}]$ be sum-of-squares polynomials. Show that $R + S$ and RS are also sum-of-squares. Same problem for $R, S \in \mathbb{R}[\boldsymbol{t}]$.

P 3.3 Show that the polynomial

$$R(z_1, z_2) = 4 + (z_1 + z_1^{-1}) + (z_1 z_2 + z_1^{-1} z_2^{-1})$$

is nonnegative, but cannot be factored as $R(\boldsymbol{z}) = H(\boldsymbol{z})H(\boldsymbol{z}^{-1})$. (Hint: take $H(\boldsymbol{z}) = a + bz_1 + cz_2 + dz_1z_2$, express the spectral factorization equation function of the coefficients a, b, c, d and show that there is no solution.)

P 3.4 The first example of polynomial $P \in \mathbb{R}[\boldsymbol{t}]$ that is positive but not sum-of-squares was given by Motzkin and is $P(t_1, t_2) = t_1^4 t_2^2 + t_1^2 t_2^4 - 3t_1^2 t_2^2 + 1$. Use the transformation (3.121) to obtain a nonnegative trigonometric polynomial that is not sum-of-squares with factors of degree (2, 2). Search a sum-of-squares decomposition with higher degree.

P 3.5 The general form of Motzkin's counterexample is $P(\boldsymbol{t}) = (t_1^2 + \ldots + t_d^2 - d - 1)t_1^2 \ldots t_d^2 + 1$. Show that this polynomial is nonnegative, but not sum-of-squares.

P 3.6 Show that the polynomial $(1 + t_1^2)(1 + t_2^2)P(t_1, t_2)$ is sum-of-squares, where $P(t_1, t_2)$ is given by (3.124). Show also that $(1 + t_1^2)P(t_1, t_2)$ is sum-of-squares.

P 3.7 (Generalized Toeplitz positivity conditions) Let $R \in \mathbb{C}_{\boldsymbol{n}}[\boldsymbol{z}]$ be a trigonometric polynomial. Remind that in the univariate case, Theorem 1.8 says that $R(z)$ is nonnegative if and only if the matrices

$$\boldsymbol{R}_m = \sum_{k=-n}^{n} r_k \boldsymbol{\Theta}_k$$

are positive semidefinite for any $m \geq n$, where the size of $\boldsymbol{\Theta}_k$ is $m \times m$.

Generalizing this result to the multivariate case, show that the polynomial $R(\boldsymbol{z})$ is sum-of-squares if and only if the matrices

$$\boldsymbol{R}_{\boldsymbol{m}} = \sum_{\boldsymbol{k}=-\boldsymbol{n}}^{\boldsymbol{n}} r_{\boldsymbol{k}} \boldsymbol{\Theta}_{\boldsymbol{k}}$$

are positive semidefinite for any $\boldsymbol{m} \geq \boldsymbol{n}$, where the matrices $\boldsymbol{\Theta}_{\boldsymbol{k}}$ are defined in (3.33).

Hint: Express the coefficients of $R(\boldsymbol{z})$ as a sum of MA autocorrelation sequences, generated by the processes $H_\ell(\boldsymbol{z})$ (i.e., use the same proof technique as in Sect. 1.3).

P 3.8 Let $R_1(\boldsymbol{z})$, $R_2(\boldsymbol{z})$ be two trigonometric polynomials. Prove the following:

(a) $R_1(\boldsymbol{z}) - R_2(\boldsymbol{z})$ is sum-of-squares if and only if there exist Gram matrices $\boldsymbol{Q}_1$ and $\boldsymbol{Q}_2$, associated with $R_1(\boldsymbol{z})$ and $R_2(\boldsymbol{z})$, respectively (i.e., defined as in the trace parameterization (3.32)) such that $\boldsymbol{Q}_1 \succeq \boldsymbol{Q}_2$.

(b) $R_1(\boldsymbol{z}) - R_2(\boldsymbol{z})$ is sum-of-squares if and only if there exist Gram pairs $(\boldsymbol{Q}_1, \boldsymbol{S}_1)$ and $(\boldsymbol{Q}_2, \boldsymbol{S}_2)$, associated with $R_1(\boldsymbol{z})$ and $R_2(\boldsymbol{z})$, respectively (i.e., defined as in the parameterization (3.99)) such that $\boldsymbol{Q}_1 \succeq \boldsymbol{Q}_2$ and $\boldsymbol{S}_1 \succeq \boldsymbol{S}_2$.

P 3.9 Prove the following multivariate version of Theorems 2.25 and 2.27. A trigonometric polynomial $R(\boldsymbol{z})$ with *complex* coefficients and order $\boldsymbol{n}$ is sum-of-squares with factors of degree $\boldsymbol{n}$ if and only if there exist a positive semidefinite matrix $\boldsymbol{Q} \in \mathbb{R}^{(N_c+N_s)\times(N_c+N_s)}$ such that

$$R(\boldsymbol{\omega}) = \boldsymbol{\chi}^T(\boldsymbol{\omega})\boldsymbol{Q}\boldsymbol{\chi}(\boldsymbol{\omega}), \tag{3.127}$$

where $\boldsymbol{\chi}(\boldsymbol{\omega}) = [\boldsymbol{\chi}_c^T(\boldsymbol{\omega})\ \boldsymbol{\chi}_s^T(\boldsymbol{\omega})]^T$ and all the other notations are as in Theorem 3.33.

Show that (3.127) is equivalent to

$$r_{\boldsymbol{k}} = \operatorname{tr}[\boldsymbol{\Gamma}_{\boldsymbol{k}}\boldsymbol{Q}] \tag{3.128}$$

and find the expressions of the constant matrices $\boldsymbol{\Gamma}_{\boldsymbol{k}}$.

P 3.10 A 2D *hybrid real-trigonometric* polynomial has the expression

$$R(t, z) = \sum_{k_1=0}^{n_1} \sum_{k_2=-n_2}^{n_2} r_{k_1,k_2} t^{k_1} z^{-k_2}, \tag{3.129}$$

with $t \in \mathbb{R}$, $z \in \mathbb{C}$, and satisfies the symmetry relation $r_{k_1,-k_2} = r^*_{k_1,k_2}$. The polynomial (3.129) is sum-of-squares if it can be written as

$$R(t,z) = \sum_{\ell=1}^{\nu} H_\ell(t,z) H^*_\ell(t, z^{-1}), \tag{3.130}$$

where $H(t,z)$ is causal in z. Show that (3.129) is sum-of-squares if and only if there exists $\boldsymbol{Q} \succeq 0$ such that

$$r_{k_1,k_2} = \mathrm{tr}[(\boldsymbol{\Theta}_{k_2} \otimes \boldsymbol{\Upsilon}_{k_1}) \cdot \boldsymbol{Q}]. \tag{3.131}$$

Extend the result to more than two variables.

P 3.11 When implementing an SDP problem using the block Gram parameterization (3.111), the equality must be expressed for each element of a block. Show that (3.111) is equivalent to

$$(\boldsymbol{R_k})_{i\ell} = \mathrm{tr}[((\boldsymbol{\Theta}_{k_d} \otimes \ldots \otimes \boldsymbol{\Theta}_{k_1}) \otimes \boldsymbol{E}_{i\ell}) \cdot \boldsymbol{Q}],$$

where $\boldsymbol{E}_{i\ell} \in \mathbb{R}^{\kappa \times \kappa}$ is the elementary matrix with 1 in position (i, ℓ) and zeros elsewhere.

References

1. M.A. Dritschel, On factorization of trigonometric polynomials. Integr. Equ. Oper. Theory **49**, 11–42 (2004)
2. B. Reznick, Some concrete aspects of Hilbert's 17th problem. Contemp. Math. **272**, 251–272 (2000). http://www.math.uiuc.edu/~reznick/hil17.pdf
3. A. Prestel, C.N. Delzell, *Positive Polynomials: From Hilbert's 17th Problem to Real Algebra*, Springer Monographs in Mathematics (Springer, Berlin, 2001)
4. B.C. Şicleru, B. Dumitrescu, POS3POLY – a MATLAB preprocessor for optimization with positive polynomials. Optim. Eng. **14**(2), 251–273 (2013). http://www.schur.pub.ro/pos3poly
5. W. Rudin, *Fourier Analysis on Groups* (Interscience Publishers, Berlin, 1962)
6. D.G. Quillen, On the representation of Hermitian forms as sums of squares. Invent. Math. **5**, 237–242 (1968)
7. A. Megretski, Positivity of trigonometric polynomials, in *Proceedings of the 42nd IEEE Conference on Decision Control (CDC)*, vol. 3 (Hawaii, USA, 2003), pp. 3814–3817
8. W. Rudin, The extension problem for positive definite functions. Ill. J. Math. **7**, 532–539 (1963)
9. C. Scheiderer, Positivity and sums of squares: a guide to recent results, in *Emerging Applications of Algebraic Geometry*, vol. 149, IMA Volumes in Mathematics and its Applications, ed. by M. Putinar, S. Sullivant (Springer, Berlin, 2009), pp. 271–324
10. B. Reznick, Uniform denominators in Hilbert's 17th problem. Math. Z. **220**, 75–98 (1995)
11. N.K. Bose, C.C. Li, A quadratic form representation of polynomials of several variables and its applications. IEEE Trans. Autom. Control **13**(4), 447–448 (1968)
12. M.D. Choi, T.Y. Lam, B. Reznick, Sums of squares of real polynomials. Proc. Symp. Pure Math. **58**(2), 103–126 (1995)
13. V. Powers, T. Wörmann, An algorithm for sum-of-squares of real polynomials. J. Pure Appl. Algebra **127**, 99–104 (1998)

14. P.A. Parrilo, Structured semidefinite programs and semialgebraic geometry methods in robustness and optimization. Ph.D. thesis, California Institute of Technology (2000)
15. Yu. Nesterov, Squared functional systems and optimization problems, in *High Performance Optimiation*, ed. by J.G.B. Frenk, C. Roos, T. Terlaky, S. Zhang (Kluwer Academic, New York, 2000), pp. 405–440
16. J.B. Lasserre, Global optimization with polynomials and the problem of moments. SIAM J. Optim. **11**(3), 796–814 (2001)
17. J.W. McLean, H.J. Woerdeman, Spectral factorizations and sums of squares representations via semidefinite programming. SIAM J. Matrix Anal. Appl. **23**(3), 646–655 (2002)
18. B. Dumitrescu, Multidimensional stability test using sum-of-squares decomposition. IEEE Trans. Circuit Syst. I **53**(4), 928–936 (2006)
19. N.Z. Shor, Class of global minimum bounds of polynomial functions. Cybernetics **23**(6), 731–734 (1987). (Russian orig.: Kibernetika, no. 6, pp. 9–11, 1987)
20. P.A. Parrilo, Semidefinite programming relaxations for semialgebraic problems. Math. Program. Ser. B **96**, 293–320 (2003)
21. M. Laurent, Sums of squares, moment matrices and optimization over polynomials, in *Emerging Applications of Algebraic Geometry*, vol. 149, IMA Volumes in Mathematics and its Applications, ed. by M. Putinar, S. Sullivant (Springer, Berlin, 2009), pp. 157–270
22. S. Prajna, A. Papachristodoulou, P.A. Parrilo, SOSTOOLS: sum of squares optimization toolbox for Matlab (2002). http://www.cds.caltech.edu/sostools
23. J.F. Sturm, Using SeDuMi 1.02, a Matlab toolbox for optimization over symmetric cones. Optim. Methods Softw. **11**, 625–653 (1999). http://sedumi.ie.lehigh.edu
24. B. Dumitrescu, B.C. Şicleru, R. Ştefan, Positive hybrid real-trigonometric polynomials and applications to adjustable filter design and absolute stability analysis. Circuit Syst. Signal Process. **29**(5), 881–899 (2010)
25. B. Reznick, Extremal PSD forms with few terms. Duke Math. J. **45**, 363–374 (1978)
26. M. Kojima, S. Kim, H. Waki, Sparsity in sums of squares of polynomials. Math. Program. **103**(1), 45–62 (2005)
27. T. Roh, L. Vandenberghe, Discrete transforms, semidefinite programming and sum-of-squares representations of nonnegative polynomials. SIAM J. Optim. **16**, 939–964 (2006)
28. B. Dumitrescu, Gram pair parameterization of multivariate sum-of-squares trigonometric polynomials, in *European Signal Processing Conference EUSIPCO* (Florence, Italy, 2006)
29. T. Roh, B. Dumitrescu, L. Vandenberghe, Multidimensional FIR filter design via trigonometric sum-of-squares optimization. IEEE J. Sel. Top. Signal Process. **1**(4), 641–650 (2007)
30. Y. Genin, Y. Hachez, Yu. Nesterov, P. Van Dooren, Optimization problems over positive pseudopolynomial matrices. SIAM J. Matrix Anal. Appl. **25**(1), 57–79 (2003)

Chapter 4
Polynomials Positive on Domains

Abstract In Sect. 1.4, we have presented the parameterizations of univariate polynomials that are positive on an interval. Here, we look at the generalization of this kind of results for multivariate polynomials that are positive on a domain $\mathcal{D}$, where $\mathcal{D} \subset \mathbb{R}^d$ for real polynomials and $\mathcal{D} \subset [-\pi, \pi]^d$ for trigonometric polynomials. The set $\mathcal{D}$ is characterized by the nonnegativity of some given polynomials. As in the case of globally positive polynomials, as discussed in the previous chapters, the parameterization can be expressed in terms of sum-of-squares polynomials and Gram matrices; again, SDP can be used as optimization tool for solving a relaxed version of the initial problem. We start by presenting, without proof, some recent results regarding real polynomials. More information can be found in [1, 2]. These results are used for deriving characterizations valid for trigonometric polynomials, in three standard frameworks: positivity on a frequency domain, Bounded Real Lemma, Positivstellensatz.

4.1 Real Polynomials Positive on Compact Domains

Let $g_\ell \in \mathbb{R}[\boldsymbol{t}]$, $\ell = 1 : L$, with $L \in \mathbb{N}$, be given d-variate polynomials. We define the set

$$\mathcal{D}(g) = \{\boldsymbol{t} \in \mathbb{R}^d \mid g_\ell(\boldsymbol{t}) \geq 0,\ \ell = 1 : L\}. \tag{4.1}$$

We assume that $\mathcal{D}(g)$ is not empty. Note that the set is defined by the nonnegativity of some given polynomials. This definition can accommodate equality constraints, since $g(\boldsymbol{t}) = 0$ is replaced by $g(\boldsymbol{t}) \geq 0$ and $-g(\boldsymbol{t}) \geq 0$. A set is called *semialgebraic* if it can be described as a Boolean combination (using intersection, union, complement) of sets of the form (4.1).

The problem we discuss in this section is how to describe polynomials that are positive on $\mathcal{D}(g)$. We present here several results of this type.

Let us assume first that the set $\mathcal{D}(g)$ is bounded (and thus compact). This is almost always true in practice, where finite solutions are interesting. With this assumption, the following remarkable result holds.

B. Dumitrescu, *Positive Trigonometric Polynomials and Signal Processing Applications*, Signals and Communication Technology,
DOI 10.1007/978-3-319-53688-0_4

Theorem 4.1 (Schmüdgen 1991) *If $\mathcal{D}(g)$ is bounded, then any polynomial $P \in \mathbb{R}[\boldsymbol{t}]$, with $P(\boldsymbol{t}) > 0$, for any $\boldsymbol{t} \in \mathcal{D}(g)$, can be written as*

$$P = \sum_{\boldsymbol{\alpha}\in\{0,1\}^L} g_1^{\alpha_1} \cdots g_L^{\alpha_L} s_{\boldsymbol{\alpha}}, \tag{4.2}$$

where $s_{\boldsymbol{\alpha}} \in \sum \mathbb{R}[\boldsymbol{t}]^2$.

This theorem says that a polynomial that is positive on $\mathcal{D}(g)$ can be parameterized as a function of 2^L sum-of-squares; the set of polynomials of the form (4.2) is called the preordering generated by the polynomials g_ℓ. Moreover, the dependence between the coefficients of $P(\boldsymbol{t})$ and those of the sum-of-squares from (4.2) is linear. However, Schmüdgen's Theorem is hardly practical, due to the large number of sum-of-squares, even for relatively small L. A stronger result gives a convenient alternative, at least in some cases. Let us arrange the products of g polynomials that appear in (4.2) in increasing order of the number of factors (i.e., in increasing order of the number of ones in the binary number α) as follows: 1, g_1, ..., g_L, g_1g_2, ..., $g_{L-1}\ g_L$, $g_1g_2g_3$, ..., $g_1 \dots g_L$. Denote the above polynomials as p_1, p_2, ..., p_{2^L}.

Theorem 4.2 (Jacobi-Prestel 2001) *If $\mathcal{D}(g)$ is bounded, then any polynomial $P \in \mathbb{R}[\boldsymbol{t}]$, with $P(\boldsymbol{t}) > 0$, for any $\boldsymbol{t} \in \mathcal{D}(g)$, can be written as*

$$P = \sum_{\ell=1}^{2^{L-1}+1} p_\ell s_\ell, \tag{4.3}$$

where $s_\ell \in \sum \mathbb{R}[\boldsymbol{t}]^2$.

Remark 4.3 Since only $2^{L-1} + 1$ terms are actually necessary in (4.2), it turns out that if $L = 2$, then $P = s_0 + g_1s_1 + g_2s_2$. If $L = 3$, then 5 sum-of-squares are enough in (4.2), instead of the 8 assessed by Theorem 4.2; in this case, we have $P = s_0 + g_1s_1 + g_2s_2 + g_3s_3 + g_1g_2s_4$; the sum includes only one product of g functions. ■

In this context, it is natural to ask when the polynomials that are positive on $\mathcal{D}(g)$ belong to the set

$$\mathcal{M}(g) = \{P \in \mathbb{R}[\boldsymbol{t}] \mid P = s_0 + \sum_{\ell=1}^{L} g_\ell s_\ell,\ s_\ell \in \sum \mathbb{R}[\boldsymbol{t}]^2\}, \tag{4.4}$$

i.e., have a form that is linear in the polynomials g_ℓ, as in the case $L = 2$ of Theorem 4.2. Of course, it is not this kind of linearity that is the most interesting (although it is mathematically beautiful), but the low number of sum-of-squares. The following theorems show what conditions can be added to the boundedness of $\mathcal{D}(g)$ in order to obtain such a form.

Theorem 4.4 (Putinar 1993) *If there exists a polynomial $p_0 \in \mathcal{M}(g)$ such that the set*

$$\mathcal{W}(p_0) = \{\boldsymbol{t} \in \mathbb{R}^d \mid p_0(\boldsymbol{t}) \geq 0\} \tag{4.5}$$

is bounded, then for any polynomial $P \in \mathbb{R}[\boldsymbol{t}]$, with $P(\boldsymbol{t}) > 0$, $\forall \boldsymbol{t} \in \mathcal{D}(g)$, it results that $P \in \mathcal{M}(g)$.

Theorem 4.5 (Jacobi 2001) *If there exists $N \in \mathbb{N}$ such that the polynomial*

$$p_0(\boldsymbol{t}) = N - \sum_{i=1}^{d} t_i^2 \tag{4.6}$$

belongs to $\mathcal{M}(g)$, then for any polynomial $P \in \mathbb{R}[\boldsymbol{t}]$, with $P(\boldsymbol{t}) > 0$, $\forall \boldsymbol{t} \in \mathcal{D}(g)$, it results that $P \in \mathcal{M}(g)$.

Remark 4.6 Theorem 4.5 is stronger than Theorem 4.4, since the set $\mathcal{W}(p_0)$ defined in (4.5) is clearly bounded for the polynomial p_0 from (4.6). However, both theorems give useful conditions to check whether the equivalence "P positive on $\mathcal{D}(g)$" $\Leftrightarrow$ "$P \in \mathcal{M}(g)$" holds.

The two theorems cover some particular forms of the set $\mathcal{D}(g)$. For example, the hypotheses of Theorems 4.4 and 4.5 hold if all the functions g_ℓ are linear (Handelman's theorem). Moreover, it is enough that only the first $L_0 < L$ functions are linear, provided the set $\mathcal{D}_0(g) = \{\boldsymbol{t} \in \mathbb{R}^d \mid g_\ell(\boldsymbol{t}) \geq 0,\ \ell = 1 : L_0\} \supset \mathcal{D}(g)$ is bounded. ■

Remark 4.7 In all the parameterizations given in this section, the degrees of the sum-of-squares may be greater than deg P (for s_0) or deg P − deg g_ℓ (for s_ℓ). The case of globally positive polynomials treated in Sect. 3.3 was different: The degrees of the sum-of-squares terms were less or equal deg P, but not all positive polynomials were expressible as sum-of-squares. ■

Remark 4.8 The reciprocals of all these theorems hold trivially. For instance, if $P \in \mathcal{M}(g)$, it follows that $P(\boldsymbol{t}) \geq 0$ for any $\boldsymbol{t} \in \mathcal{D}(g)$. So, some of the polynomials that are nonnegative on $\mathcal{D}(g)$ have sum-of-squares representations, but not all of them (a zero on $\mathcal{D}(g)$ may imply that such a representation is impossible, no matter the degrees of the sum-of-squares polynomials). ■

Remark 4.9 For univariate polynomials, similar but stronger results are given by Theorems 1.11 and 1.13. Note that there the degrees of the sum-of-squares (actually just squares) are minimal. Moreover, in Theorem 1.13, the domain $\mathcal{D}$ (actually an interval) is unbounded. ■

Remark 4.10 The algebraic structure at the basis of Theorem 4.5 is that of quadratic module. A set $\mathcal{M} \subset \mathbb{R}[\boldsymbol{t}]$ is a quadratic module if

$$\mathcal{M} + \mathcal{M} \subset \mathcal{M}, \quad \mathbb{R}[\boldsymbol{t}]^2 \cdot \mathcal{M} \subset \mathcal{M}, \quad 1 \in \mathcal{M}, \quad -1 \notin \mathcal{M}. \tag{4.7}$$

(Here, e.g., the second inclusion means that for any $p \in \mathbb{R}[\boldsymbol{t}]$, $q \in \mathcal{M}$, it results that $p^2 q \in \mathcal{M}$.) It is clear that $\mathcal{M}(g)$ is a quadratic module; the first three rules from (4.7) are satisfied from the mere definition (4.4), while the nonemptiness of $\mathcal{D}(g)$ implies $-1 \notin \mathcal{M}(g)$.

Moreover, the existence condition of the polynomial (4.6) is equivalent to $\mathcal{M}(g)$ to be Archimedean. A quadratic module $\mathcal{M}$ is Archimedean if for each $f \in \mathbb{R}[\boldsymbol{t}]$, there exists $N \in \mathbb{N}$ such that $N - f \in \mathcal{M}$.

So, Theorem 4.5 holds in the more general context of Archimedean quadratic modules of polynomials. However, the stated form of the theorem is enough for our (and most practical) purposes. ■

As a first application, we appeal again at the problem of finding the minimum value of a polynomial, this time with constraints.

Problem ***Constrained_min_poly_value*** Let $P \in \mathbb{R}_{2n}[\boldsymbol{t}]$ be a given polynomial; the set $\mathcal{D}(g)$ is defined as in (4.1), for L known polynomials. We want to find the minimum value of the polynomial on $\mathcal{D}(g)$, i.e., to solve the optimization problem

$$\mu^\star = \min_{\boldsymbol{t} \in \mathcal{D}(g)} P(\boldsymbol{t}). \tag{4.8}$$

Again, the problem is NP-hard. We note that generally neither the objective nor the domain $\mathcal{D}(g)$ are convex.

Since the problem (4.8) is equivalent to

$$\begin{aligned} \mu^\star = \max_{\mu}\ & \mu \\ \text{s.t.}\ & P(\boldsymbol{t}) - \mu \geq 0, \quad \forall \boldsymbol{t} \in \mathcal{D}(g) \end{aligned} \tag{4.9}$$

we approach it by solving

$$\begin{aligned} \mu^\star = \max_{\mu}\ & \mu \\ \text{s.t.}\ & P(\boldsymbol{t}) - \mu \in \mathcal{M}(g) \end{aligned} \tag{4.10}$$

Passing from (4.9) to (4.10) is possible, even though $\mathcal{D}(g)$ might not respect the conditions of Theorems 4.4 or 4.5. If $\mathcal{D}(g)$ is compact, then we add the polynomial

$$g_{L+1}(\boldsymbol{t}) = N - \sum_{i=1}^{d} t_i^2 \tag{4.11}$$

to the set of constraints, with N large enough for the hypersphere

$$\{\boldsymbol{t} \in \mathbb{R}[\boldsymbol{t}] \mid g_{L+1}(\boldsymbol{t}) \geq 0\}$$

to contain $\mathcal{D}(g)$. The constraint $g_{L+1}(\boldsymbol{t}) \geq 0$ is redundant but makes the new $\mathcal{M}(g)$ satisfy the conditions of Theorems 4.4 or 4.5.

If $\mathcal{D}(g)$ is not compact, we seek practically a finite minimum, if it exists, and thus proceed like above for a large N. If the solution appears to be too small (i.e., the original problem may have the solution $\mu^\star = -\infty$), then we solve it again for a larger N and are able to diagnose the case. So, from now on we assume that the polynomial (4.11) is part of the set of constraints defining $\mathcal{D}(g)$, if necessary (and redefine $\mathcal{D}(g)$ accordingly).

Although not all polynomials that are nonnegative on $\mathcal{D}(g)$ belong to $\mathcal{M}(g)$ (but surely the positive ones), passing from (4.8) to (4.10) is not restrictive, since $P(\boldsymbol{t}) - \mu \in \mathcal{M}(g)$ for any $\mu < \mu^\star$; for numerical computation, the distinction between positivity and nonnegativity is irrelevant. However, the problem (4.10) can be solved only in relaxed form, since we have to bound the degrees of the sum-of-squares polynomials. One possible relaxation is

$$\begin{aligned}
\mu_{\boldsymbol{\kappa}}^\star = \max_{\mu, s_0, \ldots, s_L} \quad & \mu \\
\text{s.t.} \quad & P(\boldsymbol{t}) - \mu = s_0(\boldsymbol{t}) + \textstyle\sum_{\ell=1}^{L} g_\ell(\boldsymbol{t}) s_\ell(\boldsymbol{t}) \\
& s_\ell \in \textstyle\sum \mathbb{R}[\boldsymbol{t}]^2, \ \ \ell = 0 : L \\
& \deg s_0 = 2(\boldsymbol{n} + \boldsymbol{\kappa}) \\
& \deg s_\ell = 2\lfloor \boldsymbol{n} + \boldsymbol{\kappa} - \tfrac{1}{2} \deg g_\ell \rfloor, \ \ \ell = 1 : L
\end{aligned} \tag{4.12}$$

The degrees of the sum-of-squares polynomials s_ℓ are chosen such that all the terms in the expression of $P(\boldsymbol{t}) - \mu$ have degrees as near as possible from $2(\boldsymbol{n} + \boldsymbol{\kappa})$; if the i-th component of $\deg g_\ell$ is odd, then the degree of t_i in $g_\ell s_\ell$ is odd and so equal to $2(n_i + \kappa_i) - 1$. Using the sum-of-squares parameterization (3.75) for the polynomials s_ℓ, $\ell = 0 : L$, the problem (4.12) becomes an SDP problem; it is essential that the relation between the coefficients of P and those of s_ℓ is linear. Since we will discuss in more detail a similar problem for trigonometric polynomials, the precise form of the SDP problem is left as an exercise. We note only that relaxations alternative to (4.12) can be obtained by bounding the *total* degree of the terms $g_\ell s_\ell$. ■

4.2 Trigonometric Polynomials Positive on Frequency Domains

We tackle now a problem similar to the basic one from the previous section, this time for trigonometric polynomials. Let

$$\mathcal{D} = \{\boldsymbol{\omega} \in [-\pi, \pi]^d \mid D_\ell(\boldsymbol{\omega}) \geq 0, \ \ell = 1 : L\} \tag{4.13}$$

be a nonempty frequency domain defined by the positivity of L given trigonometric polynomials $D_\ell(\boldsymbol{z})$, possibly with complex coefficients, as in (3.1). We want to characterize the set of trigonometric polynomials that are positive on $\mathcal{D}$. We note that $\mathcal{D}$ is bounded and so we expect parameterizations similar to those from the previous section. It turns out that without any supplementary condition, a "linear" sum-of-squares representation is valid, in the spirit of Theorems 4.4 and 4.5.

Theorem 4.11 *If a polynomial $R \in \mathbb{C}_{\boldsymbol{n}}[z]$ defined as in (3.1) is positive on $\mathcal{D}$ (i.e., $R(\boldsymbol{\omega}) > 0$, $\forall \boldsymbol{\omega} \in \mathcal{D}$), then there exist sum-of-squares polynomials $S_\ell(z)$, $\ell = 0 : L$, such that*

$$R(z) = S_0(z) + \sum_{\ell=1}^{L} D_\ell(z) \cdot S_\ell(z). \tag{4.14}$$

Moreover, if $R(z)$ and the polynomials $D_\ell(z)$ defining $\mathcal{D}$ have real coefficients, then the above sum-of-squares polynomials have also real coefficients.

The proof is given in Sect. 4.5. The idea is to transform trigonometric polynomials into real ones, apply Theorem 4.5 and go back.

Remark 4.12 The reciprocal of Theorem 4.11 holds in the sense that if the form (4.14) exists, then $R(\boldsymbol{\omega}) \geq 0, \forall \boldsymbol{\omega} \in \mathcal{D}$; the proof is trivial. So, some of the polynomials that are nonnegative on $\mathcal{D}$ have the form (4.14), but not all of them.

We note also that the degrees of the sum-of-squares polynomials from (4.14) may be arbitrarily high, at least theoretically. ■

Remark 4.13 What happens in the particular case where $\mathcal{D} = [-\pi, \pi]^d$, i.e., $L = 0$ in (4.13)? This is the case of globally positive trigonometric polynomials. From Theorem 4.11, we draw the simple conclusion that such a polynomial is sum-of-squares. This is actually Theorem 3.5! ■

Remark 4.14 Another natural question is: What do we obtain when $d = 1$, i.e., for univariate polynomials? In this case, the domain $\mathcal{D}$ can be only an interval $[\alpha, \beta]$ or an union of intervals.

In the case of a single interval, for complex coefficients, the simplest description (4.13) is via the polynomial (1.34–1.36) and so Theorems 4.11 becomes 1.15. In particular, relation (4.14) becomes (1.33); however, in the univariate case we know that the sum-of-squares polynomials (actually pure squares in the univariate case) have the minimum possible degree!

For real coefficients, we obtain Theorem 1.17, where (1.37) corresponds to the case where $\mathcal{D}$ is described by a single polynomial, namely $(\cos\omega - \cos\beta)(\cos\alpha - \cos\omega)$, while in (1.38) the "domain" $\mathcal{D}$ is described by two polynomials, $\cos\omega - \cos\beta$ and $\cos\alpha - \cos\omega$. Again, the degrees are minimal. Moreover, in (1.38) the term $S_0(z)$ from (4.14) is not present and so the sum-of-squares decomposition has two terms instead of three.

In the case of a union of intervals, which can be described by $D_1(z)$ only, relation (4.14) obviously holds with $L = 1$. Opposite to the case of a single interval, the degree of the sum-of-squares can be larger than the minimum. This is the trigonometric polynomial version of Remark 1.20. Exercise: Prove the real polynomials relation (1.41), using Theorem 4.4. Hint: Take $p_0(t) = 1 + g_1(t)$ and note that $\mathcal{W}(p_0)$ is bounded. ■

Remark 4.15 The matrix coefficients case has a very similar description. With the notation from Sect. 3.10, Theorem 4.11 stands with obvious modifications: If the polynomial (3.104) is positive definite on $\mathcal{D}$ (i.e., $\boldsymbol{R}(\boldsymbol{\omega}) \succ 0$, $\forall \boldsymbol{\omega} \in \mathcal{D}$), then there exist sum-of-squares polynomials $\boldsymbol{S}_\ell(z)$, $\ell = 0 : L$, such that

$$\boldsymbol{R}(z) = \boldsymbol{S}_0(z) + \sum_{\ell=1}^{L} D_\ell(z) \cdot \boldsymbol{S}_\ell(z). \tag{4.15}$$

All the above Remarks 4.12–4.14 apply as well, although there are no specific results for an interval $[\alpha, \beta]$. ■

4.2.1 Gram Set Parameterization

We now transform (4.14) by using the Gram matrix parameterization (3.32) of sum-of-squares polynomials.

Theorem 4.16 *If the trigonometric polynomial $R(z)$ is positive on the domain $\mathcal{D}$ defined as in (4.13), then there exist matrices $\boldsymbol{Q}_\ell \succeq 0$, $\ell = 0 : L$, such that*

$$r_{\boldsymbol{k}} = tr\left[\boldsymbol{\Theta}_{\boldsymbol{k}} \boldsymbol{Q}_0\right] + \sum_{\ell=1}^{L} tr\left[\boldsymbol{\Psi}_{\ell \boldsymbol{k}} \boldsymbol{Q}_\ell\right], \quad \boldsymbol{k} \in \mathcal{H}, \tag{4.16}$$

where $\mathcal{H}$ is a halfspace, the constant matrices $\boldsymbol{\Psi}_{\ell \boldsymbol{k}}$ are given by

$$\boldsymbol{\Psi}_{\ell \boldsymbol{k}} = \sum_{\boldsymbol{i}+\boldsymbol{l}=\boldsymbol{k}} (d_\ell)_{\boldsymbol{i}} \boldsymbol{\Theta}_{\boldsymbol{l}} \tag{4.17}$$

and the matrices $\boldsymbol{\Theta}_{\boldsymbol{k}}$ are defined by (3.33); by $(d_\ell)_{\boldsymbol{i}}$ we denote the coefficients of $D_\ell(z)$. If the polynomials $R(z)$ and $D_\ell(z)$ have real coefficients, then the matrices $\boldsymbol{Q}_\ell$ are real. Otherwise, the matrices $\boldsymbol{Q}_\ell$ are complex.

Proof The matrices $\boldsymbol{Q}_\ell$ are Gram matrices associated with the sum-of-squares $S_\ell(z)$ from (4.14) and so obey to relations similar to (3.32). The coefficient of $z^{-\boldsymbol{k}}$ of a product $D_\ell(z)S_\ell(z)$ from (4.14) is $\sum_{\boldsymbol{i}+\boldsymbol{l}=\boldsymbol{k}} (d_\ell)_{\boldsymbol{i}} (s_\ell)_{\boldsymbol{l}}$. Since the parameterization (3.32) says that $(s_\ell)_{\boldsymbol{l}} = \text{tr} \boldsymbol{\Theta}_{\boldsymbol{l}} \boldsymbol{Q}_\ell$, the relations (4.16) and (4.17) result immediately by identification in (4.14). ■

We name $\{\boldsymbol{Q}_\ell\}_{\ell=0:L}$ a *Gram set* associated with the polynomial $R(z)$ that is positive on $\mathcal{D}$. Generally, there are many Gram sets associated with a single polynomial.

Remark 4.17 The reciprocal of Theorem 4.16 holds in the sense that if the matrices $\boldsymbol{Q}_\ell \succeq 0$, $\ell = 0 : L$, exist such that (4.16) holds, then $R(\boldsymbol{\omega}) \geq 0$, $\forall \boldsymbol{\omega} \in \mathcal{D}$. (Note that similar to the reciprocal of Theorem 4.11, the strict positivity is replaced by nonnegativity.) The proof is immediate, since (4.16) is equivalent to (4.14). ■

Remark 4.18 The sizes of the Gram matrices $\boldsymbol{Q}_\ell$ from (4.16) depend on the degrees of the sum-of-squares polynomials from (4.14). Since these degrees must be bounded, we actually can implement only a sufficient positivity condition. Such a relaxation is similar to that discussed in Sect. 3.5. So, in a practical implementation, we use the degrees values

$$\begin{aligned} &\deg S_0 = \boldsymbol{m}, \\ &\deg S_\ell = \boldsymbol{m} - \deg D_\ell, \quad \ell = 1 : L, \end{aligned} \tag{4.18}$$

where $\boldsymbol{m} \geq \boldsymbol{n}$; the difference $\boldsymbol{m} - \boldsymbol{n}$ is usually small, preferably equal to zero. It is clear that a larger $\boldsymbol{m}$ allows a better approximation by (4.16) of the set of polynomials that are positive on $\mathcal{D}$, at the cost of higher complexity. ■

Problem ***Constrained_min_poly_value*** With trigonometric polynomials, the problem is

$$\begin{aligned} \mu^\star = \max_{\mu} \ & \mu \\ \text{s.t.} \ & R(\boldsymbol{\omega}) - \mu \geq 0, \quad \forall \boldsymbol{\omega} \in \mathcal{D} \end{aligned} \tag{4.19}$$

We appeal to Theorem 4.11, by using the expression (4.14) for the polynomial $R(z) - \mu$ that is positive on $\mathcal{D}$, and obtain the equivalent problem

$$\begin{aligned} \mu^\star = \max_{\mu, S_0, \ldots, S_L} \ & \mu \\ \text{s.t.} \quad & R(z) - \mu = S_0(z) + \textstyle\sum_{\ell=1}^L D_\ell(z) S_\ell(z) \\ & S_\ell \in \mathbb{RS}[z], \quad \ell = 0 : L \end{aligned} \tag{4.20}$$

We can solve only a relaxed version of this problem, by imposing the degree bounds (4.18) to obtain

$$\begin{aligned} \mu^\star_{\boldsymbol{m}} = \max_{\mu, \boldsymbol{Q}_0, \ldots, \boldsymbol{Q}_L} \ & \mu \\ \text{s.t.} \quad & r_{\boldsymbol{0}} - \mu = \operatorname{tr}[\boldsymbol{Q}_0] + \textstyle\sum_{\ell=1}^L \operatorname{tr}[\boldsymbol{\Psi}_{\ell \boldsymbol{0}} \boldsymbol{Q}_\ell] \\ & r_{\boldsymbol{k}} = \operatorname{tr}[\boldsymbol{\Theta}_{\boldsymbol{k}} \boldsymbol{Q}_0] + \textstyle\sum_{\ell=1}^L \operatorname{tr}[\boldsymbol{\Psi}_{\ell \boldsymbol{k}} \boldsymbol{Q}_\ell], \quad \boldsymbol{k} \in \mathcal{H} \setminus \{\boldsymbol{0}\} \\ & \boldsymbol{Q}_\ell \succeq 0, \ \ \boldsymbol{Q}_\ell \in \mathbb{R}^{N_\ell \times N_\ell}, \ \ \ell = 0 : L \end{aligned} \tag{4.21}$$

where N_ℓ is given by formulas similar to (3.45) for the degrees from (4.18). (In particular, N_0 has exactly the expression (3.45).) The solutions of (4.21) obey to the relation (3.49), i.e., a better approximation of the true solution is expected for larger $\boldsymbol{m}$. ■

Example 4.19 Let us consider the frequency domain

$$\mathcal{D} = \{\boldsymbol{\omega} \in [-\pi, \pi]^2 \mid \cos\omega_1 + \cos\omega_2 - 1 \geq 0\}. \tag{4.22}$$

It is one of the simplest possible domains, since it is defined by a single polynomial, i.e., $L = 1$ in (4.13), and the degree of this polynomial is (1, 1). The shape of the domain is illustrated in Fig. 4.1.

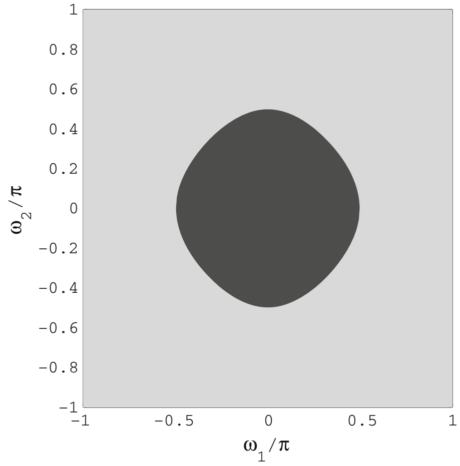


Fig. 4.1 Frequency domain (4.22), in *black*, and its complementary, in *gray*

We solve the problem (4.21) for the polynomial (3.36) and the domain (4.22). (See Example 3.18 for the global minimum of this polynomial and Fig. 3.3 for the graph of the polynomial.) We obtain $\mu^\star_{\boldsymbol{m}} = 26.7952$ for several values of $\boldsymbol{m} \geq \boldsymbol{n} = (2, 1)$, including $\boldsymbol{m} = \boldsymbol{n}$. So, in this case, we can assume that the smallest degree relaxation gives the true optimum.

We now solve the same problem, but for the complementary of the considered domain, i.e., $[-\pi, \pi]^2 \setminus \mathcal{D}$. The complementary is obtained by simply changing the sign of the polynomial from (4.22). With $\boldsymbol{m} = \boldsymbol{n}$, the solution of (4.21) is $\mu^\star_{\boldsymbol{n}} = 1.8214$, which coincides with the global optimum computed in Example 3.18. ■

Example 4.20 Let us now solve the problem (4.21) for the polynomial (3.17); the computation of its global minimum was discussed in Example 3.19; remind that the smallest degree sum-of-squares relaxation was unable to find the true minimum. The frequency domain on which the minimum is searched is $[-\pi, \pi]^2 \setminus \mathcal{D}$, with $\mathcal{D}$ defined by (4.22). We see from Figs. 4.1 and 3.4 that this domain contains the global minimum $\mu^\star = 0$. Solving the problem (4.21) with $\boldsymbol{m} = \boldsymbol{n} = (2, 2)$, we obtain the negative value $\mu^\star_{\boldsymbol{n}} = -0.01177$ (the same as for the global minimum, see Example 3.19; remark that this is actually the worst value we could obtain, since it corresponds to a decomposition $R + \mu^\star_{\boldsymbol{n}} = S_0 + D_1 S_1$ in which $S_1 = 0$; we could hope that S_1 would contribute to increasing the value of the minimum). Again, the minimum degree relaxation is not successful. However, for any $\boldsymbol{m} > \boldsymbol{n}$, the solution is $\mu^\star_{\boldsymbol{m}} = 0$, i.e., all other higher degree relaxations succeed in finding the true minimum. ■

Remark 4.21 Problems such as (4.19) can be solved with constraints that are an intersection, union, and/or complementary of sets defined as in (4.13). The intersection of several domains is intrinsic to the definition (4.13); the resulting domain is defined by all the polynomials defining the initial domains. The complementary of $\mathcal{D}$ from (4.13) is the union of domains characterized by the positivity of a single trigonometric polynomial; more exactly, the complementary is

$$\bigcup_{\ell=1}^{L} \{\boldsymbol{\omega} \in [-\pi, \pi]^d \mid -D_\ell(\boldsymbol{\omega}) \geq 0\}. \tag{4.23}$$

So, we have to see only how union can be treated. Consider two domains, $\mathcal{D}_1$ and $\mathcal{D}_2$, defined as in (4.13), with polynomials $D_{1\ell}(\boldsymbol{z})$, $\ell = 1 : L_1$, and, respectively, $D_{2l}(\boldsymbol{z})$, $l = 1 : L_2$. In order to solve the problem

$$\begin{array}{rl} \mu^\star = & \max\limits_{\mu} \ \mu \\ & \text{s.t.} \ \ R(\boldsymbol{\omega}) - \mu \geq 0, \quad \forall \boldsymbol{\omega} \in \mathcal{D}_1 \cup \mathcal{D}_2 \end{array} \tag{4.24}$$

we express the positive polynomial $R(\boldsymbol{z}) - \mu$ using (4.14) on each of the domains $\mathcal{D}_1$ and $\mathcal{D}_2$, obtaining

$$\begin{array}{rcl} \mu^\star = & \max\limits_{\mu, S_{10}, \ldots, S_{1L_1}, S_{20}, \ldots, S_{2L_2}} & \mu \\ & \text{s.t.} & R(\boldsymbol{z}) - \mu = S_{10}(\boldsymbol{z}) + \sum_{\ell=1}^{L_1} D_{1\ell}(\boldsymbol{z}) S_{1\ell}(\boldsymbol{z}) \\ & & R(\boldsymbol{z}) - \mu = S_{20}(\boldsymbol{z}) + \sum_{l=1}^{L_2} D_{2l}(\boldsymbol{z}) S_{2l}(\boldsymbol{z}) \\ & & S_{1\ell}, S_{2l} \in \mathbb{RS}[\boldsymbol{z}], \ \ \ell = 0 : L_1, \ l = 0 : L_2 \end{array} \tag{4.25}$$

This problem can be relaxed to an SDP form similar to (4.21); the number of positive semidefinite matrices is $L_1 + L_2 + 2$ (one for each sum-of-squares polynomial). We conclude that any intersection or union of domains (4.13) can be handled, the complexity of the SDP relaxation being roughly proportional to the total number of trigonometric polynomials defining the domains. An estimate of the complexity can be computed as for problem (3.39), depending on the degree of the relaxation.

Examples using unions of domains will be given later in the chapter dedicated to FIR filters design. ■

Remark 4.22 In view of Remark 4.15, the extension of Theorem 4.16 to matrix polynomials (3.104) is made by only changing (4.16) into

$$\boldsymbol{R}_{\boldsymbol{k}} = \text{TR}\left[(\boldsymbol{\Theta}_{\boldsymbol{k}} \otimes \boldsymbol{I}_\kappa)\boldsymbol{Q}_0\right] + \sum_{\ell=1}^{L} \text{TR}\left[(\boldsymbol{\Psi}_{\ell\boldsymbol{k}} \otimes \boldsymbol{I}_\kappa)\boldsymbol{Q}_\ell\right], \ \ \boldsymbol{k} \in \mathcal{H}. \tag{4.26}$$

The proof is the same, but Theorem 3.37 is used for the parameterization. ■

4.2.2 Gram-Pair Set Parameterization

For polynomials with real coefficients, an LMI form of Theorem 4.11 can be obtained via the Gram pair parameterization (3.99).

Theorem 4.23 *If the symmetric polynomial $R \in \mathbb{R}[z]$ is positive on the domain $\mathcal{D}$ defined as in (4.13), with $D_\ell \in \mathbb{R}[z]$, then there exist matrices $\boldsymbol{Q}_\ell \succeq 0$, $\boldsymbol{S}_\ell \succeq 0$, $\ell = 0 : L$, such that*

$$r_{\boldsymbol{k}} = tr\left[\boldsymbol{\Phi}_{\boldsymbol{k}} \boldsymbol{Q}_0\right] + \sum_{\ell=1}^{L} tr\left[\tilde{\boldsymbol{\Phi}}_{\ell \boldsymbol{k}} \boldsymbol{Q}_\ell\right] + tr\left[\boldsymbol{\Lambda}_{\boldsymbol{k}} \boldsymbol{S}_0\right] + \sum_{\ell=1}^{L} tr\left[\tilde{\boldsymbol{\Lambda}}_{\ell \boldsymbol{k}} \boldsymbol{S}_\ell\right], \quad \boldsymbol{k} \in \mathcal{H}, \tag{4.27}$$

where $\mathcal{H}$ is a halfspace, the coefficient matrices are defined by

$$\tilde{\boldsymbol{\Phi}}_{\ell \boldsymbol{k}} = \frac{1}{2} \sum_{\boldsymbol{i}+\boldsymbol{l}=\boldsymbol{k}} (d_\ell)_{\boldsymbol{i}} \boldsymbol{\Phi}_{\boldsymbol{l}} + \frac{1}{2} \sum_{\boldsymbol{i}-\boldsymbol{l}=\boldsymbol{k}} (d_\ell)_{\boldsymbol{i}} \boldsymbol{\Phi}_{\boldsymbol{l}}, \tag{4.28}$$

$$\tilde{\boldsymbol{\Lambda}}_{\ell \boldsymbol{k}} = \frac{1}{2} \sum_{\boldsymbol{i}+\boldsymbol{l}=\boldsymbol{k}} (d_\ell)_{\boldsymbol{i}} \boldsymbol{\Lambda}_{\boldsymbol{l}} + \frac{1}{2} \sum_{\boldsymbol{i}-\boldsymbol{l}=\boldsymbol{k}} (d_\ell)_{\boldsymbol{i}} \boldsymbol{\Lambda}_{\boldsymbol{l}}, \tag{4.29}$$

and the matrices $\boldsymbol{\Phi}_{\boldsymbol{k}}$, $\boldsymbol{\Lambda}_{\boldsymbol{k}}$ are those from (3.99).

Proof The matrices $(\boldsymbol{Q}_\ell, \boldsymbol{S}_\ell)$ are Gram pairs associated with the sum-of-squares $S_\ell(z)$ from (3.99) and so obey to relations similar to (3.99). Using the first trigonometric identity from (2.77), we can write

$$D_\ell(\boldsymbol{\omega}) S_\ell(\boldsymbol{\omega}) = \sum_{\boldsymbol{i}} \sum_{\boldsymbol{l}} (d_\ell)_{\boldsymbol{i}} (s_\ell)_{\boldsymbol{l}} \cdot \frac{1}{2} [\cos(\boldsymbol{i} + \boldsymbol{l})^T \boldsymbol{\omega} + \cos(\boldsymbol{i} - \boldsymbol{l})^T \boldsymbol{\omega}].$$

After inserting this relation into

$$R(\boldsymbol{\omega}) = S_0(\boldsymbol{\omega}) + \sum_{\ell=1}^{L} D_\ell(\boldsymbol{\omega}) \cdot S_\ell(\boldsymbol{\omega})$$

and identifying the coefficients of the “monomials” $\cos \boldsymbol{k}^T \boldsymbol{\omega}$, the theorem is proved. (Note that here the indices are not confined to a halfplane, as in (3.101).) ■

We name $\{(\boldsymbol{Q}_\ell, \boldsymbol{S}_\ell)\}_{\ell=0:L}$ a *Gram-pair set* associated with the polynomial $R(z)$. The remarks and the examples following Theorem 4.16 can be easily adapted to the parameterization (4.27).

4.3 Bounded Real Lemma

For univariate polynomials, the existence of the spectral factorization (1.11) allows the liberty of replacing, in many optimization problems, the causal polynomial $H(z)$ with a nonnegative $R(z)$, representing the squared magnitude $R(\omega) = |H(\omega)|^2$. The lack of a spectral factorization for multivariate polynomials makes impossible such an approach, narrowing the field of applications where positive polynomials could be used. However, another type of result can be derived, specifically a Bounded Real Lemma (BRL) for multivariate polynomials (or multidimensional FIR systems, if we adopt a signal processing terminology).

Let $H(\boldsymbol{z})$ be a d-variate positive orthant polynomial, defined as in (3.5). A BRL is a characterization of the inequality

$$|H(\boldsymbol{\omega})| \le \gamma, \quad \forall \boldsymbol{\omega} \in \mathcal{D}, \tag{4.30}$$

where the positive real number γ and the frequency domain $\mathcal{D}$ are given. Typically, the inequality (4.30) is desired globally, i.e., $\mathcal{D} = [-\pi, \pi]^d$. In this section, we will approximate (4.30) with two LMIs, for frequency domains defined as in (4.13). Practically, the approximations lead to relaxations similar to those already discussed.

4.3.1 Gram Set BRL

We start with two simple, but important results on the Gram set associated with a polynomial that is positive on $\mathcal{D}$.

Remark 4.24 Let $\boldsymbol{h}$ be a vector containing the coefficients of the positive orthant polynomial $H(\boldsymbol{z})$, as in (3.27). The polynomial

$$R_h(\boldsymbol{z}) \stackrel{\Delta}{=} H(\boldsymbol{z})H^*(\boldsymbol{z}^{-1}) = \boldsymbol{\psi}^T(\boldsymbol{z}^{-1})\boldsymbol{h}\boldsymbol{h}^H\boldsymbol{\psi}(\boldsymbol{z}) \tag{4.31}$$

is globally nonnegative. Comparing (4.31) with (3.28), we see that $\boldsymbol{Y}_0 = \boldsymbol{h}\boldsymbol{h}^H \succeq 0$ is a Gram matrix associated with $R_h(\boldsymbol{z})$. We can also interpret $R_h(\boldsymbol{z})$ as a polynomial nonnegative on $\mathcal{D}$, having trivially the form (4.14), with $S_0(\boldsymbol{z}) = R_h(\boldsymbol{z})$ and $S_\ell(\boldsymbol{z}) = 0$, $\ell = 1 : L$. So, a relation like (4.16) holds, with the Gram set

$$\boldsymbol{Y}_0 = \boldsymbol{h}\boldsymbol{h}^H, \quad \boldsymbol{Y}_\ell = \boldsymbol{0}, \; \ell = 1 : L. \tag{4.32}$$

We have thus associated a special Gram set with $R_h(\boldsymbol{z})$. ■

The second result relates the Gram sets of two polynomials that obey to a majorization relation on a frequency domain. The following result, which is a generalization of problem **P** 3.8a, holds even if we extend the notion of Gram set associated with a polynomial $R(z)$ to any matrices $\boldsymbol{Q}_\ell$ that respect (4.16), even if they are not positive semidefinite.

Lemma 4.25 *Let $R(z)$ and $\tilde{R}(z)$ be two trigonometric polynomials and let $\{\tilde{\boldsymbol{Q}}_\ell\}_{\ell=0:L}$ be a Gram set associated with $\tilde{R}(z)$. If $R(\omega) > \tilde{R}(\omega)$ on a frequency domain $\mathcal{D}$ defined as in (4.13), then there exists a Gram set $\{\boldsymbol{Q}_\ell\}_{\ell=0:L}$ associated with $R(z)$, such that $\boldsymbol{Q}_\ell \succeq \tilde{\boldsymbol{Q}}_\ell$, $\ell = 0 : L$.*

Reciprocally, if the Gram sets satisfy $\boldsymbol{Q}_\ell \succeq \tilde{\boldsymbol{Q}}_\ell$, $\ell = 0 : L$, then the associated polynomials respect the relation $R(\omega) \geq \tilde{R}(\omega)$, $\forall \omega \in \mathcal{D}$.

Proof The polynomial $X(z) = R(z) - \tilde{R}(z)$ is positive on $\mathcal{D}$ and so, according to Theorem 4.16, it has a Gram set $\{\boldsymbol{X}_\ell\}$, $\ell = 0 : L$, with $\boldsymbol{X}_\ell \succeq 0$. Due to the linearity of (4.16), it results that $\{\boldsymbol{Q}_\ell = \tilde{\boldsymbol{Q}}_\ell + \boldsymbol{X}_\ell\}$ is a Gram set associated with $R(z)$. It follows that $\boldsymbol{Q}_\ell \succeq \tilde{\boldsymbol{Q}}_\ell$.

The reciprocal results by going back in the above reasoning (and noticing that strict inequality may not be always obtained). ■

We can now state the BRL, for an inequality more general than (4.30).

Theorem 4.26 *Let $H(z)$ and $A(z)$ be positive orthant polynomials and $\mathcal{D}$ a frequency domain defined as in (4.13). Denote*

$$R(z) = A(z)A^*(z^{-1}). \tag{4.33}$$

If the inequality

$$|H(\omega)| < |A(\omega)|, \quad \forall \omega \in \mathcal{D}, \tag{4.34}$$

is satisfied, then there exist matrices $\boldsymbol{Q}_\ell \succeq 0$, $\ell = 0 : L$, such that the relations (4.16) and

$$\begin{bmatrix} \boldsymbol{Q}_0 & \boldsymbol{h} \\ \boldsymbol{h}^H & 1 \end{bmatrix} \succeq 0 \tag{4.35}$$

hold, where $\boldsymbol{h}$ is the vector of the coefficients of the filter $H(z)$ defined by (3.27).

Reciprocally, relations (4.16) and (4.35) imply $|H(\omega)| \leq |A(\omega)|$.

Proof The inequality (4.34) holds if and only if $|A(\omega)|^2 > |H(\omega)|^2$, which is the same with $R(\omega) > R_h(\omega)$, $\forall \omega \in \mathcal{D}$, where R_h is defined in (4.31). If the latter inequality holds, then, by applying Lemma 4.25 with $\tilde{R} = R_h$, there exists a Gram set $\{\boldsymbol{Q}_\ell\}_{\ell=0:L}$ associated with $R(z)$ such that $\boldsymbol{Q}_0 \succeq \boldsymbol{Y}_0$ and $\boldsymbol{Q}_\ell \succeq \boldsymbol{Y}_\ell$, $\ell = 1 : L$, where $\boldsymbol{Y}_\ell$ are the Gram matrices from (4.32), associated with $R_h(z)$. This means that (4.16) holds for $\boldsymbol{Q}_\ell \succeq 0$, $\ell = 0 : L$. Moreover, the inequality

$$\boldsymbol{Q}_0 \succeq \boldsymbol{h}\boldsymbol{h}^H \tag{4.36}$$

is equivalent to (4.35) via a Schur complement argument.

Following backwards the above reasoning and using the reciprocal of Lemma 4.25, if (4.16) and (4.35) hold for $\boldsymbol{Q}_i \succeq 0$, it results that $|H(\omega)| \leq |A(\omega)|$. ■

The traditional form of the BRL results by taking $A(z) = \gamma$ in Theorem 4.26.

Corollary 4.27 *If the inequality*

$$|H(\boldsymbol{\omega})| < \gamma, \ \forall \boldsymbol{\omega} \in \mathcal{D}, \tag{4.37}$$

holds, then there exist matrices $\boldsymbol{Q}_\ell \succeq 0$*,* $\ell = 0 : L$*, such that*

$$\gamma^2 \delta_{\boldsymbol{k}} = tr\left[\boldsymbol{\Theta}_{\boldsymbol{k}} \boldsymbol{Q}_0\right] + \sum_{\ell=1}^{L} tr\left[\boldsymbol{\Psi}_{\ell \boldsymbol{k}} \boldsymbol{Q}_\ell\right], \ \ \boldsymbol{k} \in \mathcal{H}, \tag{4.38}$$

where $\mathcal{H}$ *is a halfspace, and*

$$\begin{bmatrix} \boldsymbol{Q}_0 & \boldsymbol{h} \\ \boldsymbol{h}^H & 1 \end{bmatrix} \succeq 0,$$

where $\delta_{\boldsymbol{0}} = 1$ *and* $\delta_{\boldsymbol{k}} = 0$ *if* $\boldsymbol{k} \neq \boldsymbol{0}$.

Reciprocally, relations (4.38) and (4.35) imply (4.30).

Remark 4.28 Since the BRL is a consequence of Theorem 4.11, applied for the positive polynomial $|A(\boldsymbol{\omega})|^2 - |H(\boldsymbol{\omega})|^2$, we can implement only a sufficient bounded realness condition, by choosing a degree $\boldsymbol{m}$ of this polynomial. It is clear that we have to respect $\boldsymbol{m} \geq \max(\deg A, \deg H)$. For a given degree $\boldsymbol{m}$, the sizes of the Gram matrices result as discussed in Remark 4.18. Accordingly, zero coefficients are added to the polynomials (4.33) and $H(z)$ (for the latter, in the vector $\boldsymbol{h}$ from (3.27)) in order to formally raise their degree up to $\boldsymbol{m}$. ■

Remark 4.29 In the particular case of univariate polynomials ($d = 1$), there are only two Gram matrices, $\boldsymbol{Q}_0$ and $\boldsymbol{Q}_1$. If $\mathcal{D}$ is an interval, the matrices (4.17) are defined using the polynomial (1.34–1.36). For the reasons discussed in Remark 4.14, it is enough to take $\boldsymbol{m} = \max(\deg A, \deg H)$ for obtaining a condition equivalent to (4.30). ■

Remark 4.30 In the important particular case $\mathcal{D} = [-\pi, \pi]^d$, the standard BRL has the following form. If $|H(\boldsymbol{\omega})| < \gamma, \forall \boldsymbol{\omega} \in [-\pi, \pi]^d$, then there exists a matrix $\boldsymbol{Q} \succeq 0$ such that

$$\begin{gathered} \gamma^2 \delta_{\boldsymbol{k}} = \mathrm{tr}[\boldsymbol{\Theta}_{\boldsymbol{k}} \boldsymbol{Q}], \ \ \boldsymbol{k} \in \mathcal{H}, \\ \begin{bmatrix} \boldsymbol{Q} & \boldsymbol{h} \\ \boldsymbol{h}^H & 1 \end{bmatrix} \succeq 0. \end{gathered} \tag{4.39}$$

The comments from the previous two Remarks apply here as well. ■

Problem (H_∞*-norm*) A simple application of the BRL is the computation of the H_∞-norm of a FIR system described by the transfer function (3.5). The definition

$$\|H\|_\infty = \max_{\boldsymbol{\omega} \in [-\pi, \pi]^d} |H(\boldsymbol{\omega})| \tag{4.40}$$

can be written as

$$\begin{aligned} \|H\|_\infty = \min_{\gamma} \ & \gamma \\ \text{s.t.} \ & |H(\boldsymbol{\omega})| \le \gamma, \ \ \forall \boldsymbol{\omega} \in [-\pi, \pi]^d \end{aligned} \tag{4.41}$$

Using the BRL results presented above, we obtain the SDP relaxation

$$\begin{aligned} \|H\|_\infty = \min_{\gamma, \boldsymbol{Q}} \ & \gamma \\ \text{s.t.} \ & (4.39), \ \ \boldsymbol{Q} \succeq 0 \end{aligned} \tag{4.42}$$

We note that a practically equivalent (in nature, but not in form) solution could be obtained by computing the maximum value of the polynomial $|H(\omega)|^2$, using the approach from Sect. 3.5. Besides the numerical advantage of not having to square the polynomial, solving (4.42) is the only possible way when the coefficients of $H(z)$ *are not constant*, but depend linearly on some parameters; in this case, squaring destroys the convexity of the problem. An application of this type will be presented in Sect. 5.3, dedicated to deconvolution. ■

Example 4.31 We approximate the H_∞ norm of the 2-D FIR filter

$$H(z_1, z_2) = (z_1^{-1} + z_2^{-1})^3 + z_2^{-2} + z_2^{-1} \tag{4.43}$$

by solving (4.42), for the smallest degree of relaxation, i.e., $\boldsymbol{m} = \boldsymbol{n} = (3, 3)$ (the size of $\boldsymbol{Q}$ is thus 16×16). The result is 10 (within an error less than 10^{-8}), which is the true H_∞ norm of the system (4.43). Various other BRLs proposed in the literature for 2-D systems (not necessarily FIR) fail to give the correct value; for example, the BRL from [3] leads to a value of 10.2. The advantage of the BRL presented here comes from its adequacy to the FIR (i.e., polynomial) case. ■

4.3.2 BRL for Polynomials with Matrix Coefficients

Consider the positive orthant polynomial

$$\boldsymbol{H}(z) = \sum_{\boldsymbol{k}=\boldsymbol{0}}^{\boldsymbol{n}} \boldsymbol{H}_{\boldsymbol{k}} z^{-\boldsymbol{k}}, \tag{4.44}$$

with matrix coefficients $\boldsymbol{H}_{\boldsymbol{k}} \in \mathbb{R}^{p \times \kappa}$; note that they may be rectangular, unlike the square coefficients of a symmetric polynomial (3.104). The BRL inequality (4.30) is replaced by

$$\sigma_{\max}[\boldsymbol{H}(\boldsymbol{\omega})] \le \gamma, \ \ \forall \boldsymbol{\omega} \in \mathcal{D}, \tag{4.45}$$

where $\sigma_{max}[\cdot]$ is the maximum singular value of its matrix argument. This extension is obviously related to the H_∞ norm of MIMO systems, defined by

$$\|\boldsymbol{H}\|_\infty = \sup_{\boldsymbol{\omega}\in[-\pi,\pi]^d} \sigma_{\max}[\boldsymbol{H}(\boldsymbol{\omega})]. \tag{4.46}$$

Similarly to the scalar case (3.27), we stack the coefficients of (4.44) in a block column in lexicographic order of the indices

$$\bar{\boldsymbol{H}} = \begin{bmatrix} \boldsymbol{H_0} \\ \vdots \\ \boldsymbol{H_n} \end{bmatrix} \in \mathbb{R}^{Np\times\kappa}, \tag{4.47}$$

where N is defined in (3.26). The generalization of Corollary 4.27 to matrix coefficients is the following.

Theorem 4.32 *Let $\boldsymbol{H}(z)$ be the polynomial (4.44). If the inequality*

$$\sigma_{\max}[\boldsymbol{H}(\boldsymbol{\omega})] < \gamma,\ \forall\boldsymbol{\omega}\in\mathcal{D}, \tag{4.48}$$

holds, then there exist matrices $\boldsymbol{Q}_\ell \succeq 0$, $\ell = 0 : L$, such that

$$\gamma^2\delta_{\boldsymbol{k}}\boldsymbol{I}_p = \mathrm{TR}\left[(\boldsymbol{\Theta}_{\boldsymbol{k}}\otimes\boldsymbol{I}_p)\boldsymbol{Q}_0\right] + \sum_{\ell=1}^{L}\mathrm{TR}\left[(\boldsymbol{\Psi}_{\ell\boldsymbol{k}}\otimes\boldsymbol{I}_p)\boldsymbol{Q}_\ell\right],\quad \boldsymbol{k}\in\mathcal{H}, \tag{4.49}$$

where $\mathcal{H}$ is a halfspace, and

$$\begin{bmatrix} \boldsymbol{Q}_0 & \bar{\boldsymbol{H}} \\ \bar{\boldsymbol{H}}^T & \boldsymbol{I}_\kappa \end{bmatrix} \succeq 0. \tag{4.50}$$

Reciprocally, relations (4.49) and (4.50) imply $\sigma_{\max}[\boldsymbol{H}(\boldsymbol{\omega})] \le \gamma,\ \forall\boldsymbol{\omega}\in\mathcal{D}$.

Proof Similar to that of Corollary 4.27, using Remark 4.22. See problem **P** 4.6. ■

Remark 4.33 Since $\boldsymbol{H}(\boldsymbol{\omega})$ and $\boldsymbol{H}^T(\boldsymbol{\omega})$ have the same singular values, we can replace (4.49) and (4.50) by their correspondent for the transposed system, such that the size of the matrices $\boldsymbol{Q}_\ell$ is smaller. To gain efficiency, the transposed system should be used whenever $p > \kappa$. ■

Sections 5.3 and 9.2 will present applications of the BRL for polynomials with matrix coefficients.

4.3.3 Gram-Pair Set BRL

Returning to the scalar case, if the polynomial $H(z)$ has real coefficients, the Gram pair parameterization (3.99) can also be used for obtaining a BRL in the style of

Theorem 4.26. We follow the same preliminary steps. Firstly, we obtain an analogue of Remark 4.24. As in (3.93), we note that the polynomial

$$R_h(\boldsymbol{\omega}) \overset{\Delta}{=} |H(\boldsymbol{\omega})|^2 = \boldsymbol{\chi}_c^T(\boldsymbol{\omega})\boldsymbol{a}\boldsymbol{a}^T\boldsymbol{\chi}_c(\boldsymbol{\omega}) + \boldsymbol{\chi}_s^T(\boldsymbol{\omega})\boldsymbol{b}\boldsymbol{b}^T\boldsymbol{\chi}_s(\boldsymbol{\omega}), \tag{4.51}$$

is nonnegative (the coefficient vectors $\boldsymbol{a}$ and $\boldsymbol{b}$ are defined in (3.92) and the basis vectors $\boldsymbol{\chi}_c(\boldsymbol{\omega})$ and $\boldsymbol{\chi}_s(\boldsymbol{\omega})$ in (3.91)). Comparing with (3.94), we see that

$$\boldsymbol{Y}_0 = \boldsymbol{a}\boldsymbol{a}^T, \quad \boldsymbol{Z}_0 = \boldsymbol{b}\boldsymbol{b}^T \tag{4.52}$$

is a Gram pair associated with $R_h(\boldsymbol{\omega})$. Adding to it zero matrices $\boldsymbol{Y}_\ell = 0$, $\boldsymbol{Z}_\ell = 0$, $\ell = 1:L$, we obtain a Gram-pair set associated with $R_h(\boldsymbol{\omega})$, which is a polynomial (also) nonnegative on $\mathcal{D}$.

The following counterpart of Lemma 4.25 has a similar significance: that majorization of polynomials can be translated into element-wise majorization of Gram-pair sets. It is a generalization of problem **P** 3.8b.

Lemma 4.34 *Let $R(z)$ and $\tilde{R}(z)$ be two trigonometric polynomials with real coefficients and let $\{(\tilde{\boldsymbol{Q}}_\ell, \tilde{\boldsymbol{S}}_\ell)\}_{\ell=0:L}$ be a Gram-pair set associated with $\tilde{R}(z)$. If $R(\boldsymbol{\omega}) > \tilde{R}(\boldsymbol{\omega})$ on a frequency domain $\mathcal{D}$ defined as in (4.13), then there exists a Gram set $\{(\boldsymbol{Q}_\ell, \boldsymbol{S}_\ell)\}_{\ell=0:L}$ associated with $R(z)$, such that*

$$\boldsymbol{Q}_\ell \succeq \tilde{\boldsymbol{Q}}_\ell, \quad \boldsymbol{S}_\ell \succeq \tilde{\boldsymbol{S}}_\ell, \quad \ell = 0:L. \tag{4.53}$$

Reciprocally, if the Gram sets satisfy (4.53), then the associated polynomials respect the relation $R(\boldsymbol{\omega}) \geq \tilde{R}(\boldsymbol{\omega})$, $\forall \boldsymbol{\omega} \in \mathcal{D}$.

Proof Similar to that of Lemma 4.25 and based on the fact that the polynomial $R(z) - \tilde{R}(z)$ is positive on $\mathcal{D}$ and, due to the linearity of (4.27), has Gram-pair sets of the form $\{(\boldsymbol{Q}_\ell - \tilde{\boldsymbol{Q}}_\ell, \boldsymbol{S}_\ell - \tilde{\boldsymbol{S}}_\ell)\}_{\ell=0:L}$; at least one of these sets is made of positive semidefinite matrices. ■

We can formulate now the Gram-pair counterpart of Theorem 4.26.

Theorem 4.35 *Let $H(z)$ and $A(z)$ be positive orthant polynomials with real coefficients and $\mathcal{D}$ a frequency domain defined as in (4.13), with $D_\ell \in \mathbb{R}[z]$. Denote $R(z)$ as in (4.33). If the inequality (4.34) is satisfied, then there exist matrices $\boldsymbol{Q}_\ell \succeq 0$, $\boldsymbol{S}_\ell \succeq 0$, $\ell = 0:L$, such that the LMIs (4.27) and*

$$\begin{bmatrix} \boldsymbol{Q}_0 & \boldsymbol{C}_c\boldsymbol{h} \\ \boldsymbol{h}^T\boldsymbol{C}_c^T & 1 \end{bmatrix} \succeq 0, \qquad \begin{bmatrix} \boldsymbol{S}_0 & \boldsymbol{C}_s\boldsymbol{h} \\ \boldsymbol{h}^T\boldsymbol{C}_s^T & 1 \end{bmatrix} \succeq 0, \tag{4.54}$$

hold, where $\boldsymbol{h}$ is the vector of the coefficients of the filter $H(z)$ defined by (3.27) and the matrices $\boldsymbol{C}_c$, $\boldsymbol{C}_s$ are defined in (3.92).

Proof Similar to that of Theorem 4.26. The polynomial $|H(\boldsymbol{\omega})|^2$ has the remarkable Gram-pair set $(\boldsymbol{Y}_\ell, \boldsymbol{Z}_\ell)$, with $\boldsymbol{Y}_0$, $\boldsymbol{Z}_0$ defined in (4.52) and the other matrices equal to zero. From Lemma 4.34, it results that the polynomial $R(\boldsymbol{\omega}) = |A(\boldsymbol{\omega})|^2$ has a Gram-pair set $\{(\boldsymbol{Q}_\ell, \boldsymbol{S}_\ell)\}_{\ell=0:L}$ (respecting (4.27)) such that $\boldsymbol{Q}_\ell \succeq \boldsymbol{Y}_\ell$ and $\boldsymbol{S}_\ell \succeq \boldsymbol{Z}_\ell$. The inequalities $\boldsymbol{Q}_0 \succeq \boldsymbol{a}\boldsymbol{a}^T$ and $\boldsymbol{S}_0 \succeq \boldsymbol{b}\boldsymbol{b}^T$ are equivalent, via Schur complements and using (3.92), with (4.54). ■

The typical BRL form given by Corollary 4.27 and the remarks that follow Theorem 4.26 can be straightforwardly adapted to the Gram pair BRL from Theorem 4.35.

4.4 Positivstellensatz for Trigonometric Polynomials

The Positivstellensatz is a characterization, in terms of sum-of-squares polynomials, of the nonexistence of a solution to a system of polynomial inequalities and equalities. We present here a somewhat particular version that does not include inequations. Let $f_k \in \mathbb{R}[\boldsymbol{t}]$, $k = 1 : K$, $g_\ell \in \mathbb{R}[\boldsymbol{t}]$, $\ell = 1 : L$, be given d-variate polynomials. We define the set

$$\mathcal{D}(f, g) = \left\{ \boldsymbol{t} \in \mathbb{R}^d \;\middle|\; \begin{array}{l} f_k(\boldsymbol{t}) = 0,\ k = 1 : K \\ g_\ell(\boldsymbol{t}) \geq 0,\ \ell = 1 : L \end{array} \right\}. \tag{4.55}$$

Theorem 4.36 (Positivstellensatz, Stengle 1974) *The set (4.55) is empty if and only if*

$$1 + \sum_{k=1}^{K} f_k u_k + \sum_{\boldsymbol{\alpha} \in \{0,1\}^L} g_1^{\alpha_1} \cdots g_L^{\alpha_L} s_{\boldsymbol{\alpha}} = 0, \tag{4.56}$$

for some polynomials $u_k \in \mathbb{R}[\boldsymbol{t}]$ and sum-of-squares $s_{\boldsymbol{\alpha}} \in \sum \mathbb{R}[\boldsymbol{t}]^2$.

We note that sufficiency is obvious. It there were a $\boldsymbol{t} \in \mathcal{D}(f, g)$, it would follow that $f_k(\boldsymbol{t}) = 0$, $g_\ell(\boldsymbol{t}) \geq 0$ and so the left-hand term of (4.56) would be strictly positive, which would be a contradiction.

Similarly to other results of this type, the degrees of the polynomials u_k and $s_{\boldsymbol{\alpha}}$ can be arbitrarily high. A bounded degree relaxation of (4.56) is an SDP problem, due to the sum-of-squares polynomials. Such a relaxation provides only a sufficient condition that the set (4.55) is empty.

Remark 4.37 The Positivstellensatz may be used for characterizing polynomials that are positive on a set $\mathcal{D}(g)$ as in (4.1). If $P(\boldsymbol{t}) > 0$, $\forall \boldsymbol{t} \in \mathcal{D}(g)$, then the set $\mathcal{D}(g) \cap \{\boldsymbol{t} \in \mathbb{R}^d \mid -P(\boldsymbol{t}) \geq 0\}$ is empty and has the form (4.55), without equality constraints. Applying Theorem 4.36, it results that $aP = 1 + b$, for some polynomials a, b belonging to the preordering generated by g_ℓ, $\ell = 1 : L$ (i.e., a, b have the form (4.2)). Comparing this result with Schmüdgen's Theorem 4.1, we remark that its use for optimization is limited, due to the multiplication of P with a variable polynomial; it can be used only if the coefficients of P are fixed, which is a rare case. However, the Positivstellensatz has other applications. ■

The Positivstellensatz can take a simpler form under conditions similar to those that allow passing from Theorems 4.1 or 4.2 to Theorem 4.5.

Theorem 4.38 *Assume that the set of polynomials $\mathcal{M}(g)$ defined in (4.4) satisfies the hypothesis of Theorem 4.5 (i.e., $\mathcal{M}(g)$ is an Archimedean quadratic module). Then, the set (4.55) is empty if and only if*

$$1 + \sum_{k=1}^{K} f_k u_k + s_0 + \sum_{\ell=1}^{L} g_\ell s_\ell = 0, \tag{4.57}$$

for some polynomials $u_k \in \mathbb{R}[\boldsymbol{t}]$ and sum-of-squares $s_\ell \in \sum \mathbb{R}[\boldsymbol{t}]^2$.

Proof The sufficiency is obvious, as noted after Theorem 4.36. We prove now the necessity. The set $\mathcal{D}(g)$ defined in (4.1) is compact, otherwise the polynomial (4.6) could not belong to $\mathcal{M}(g)$. Then, there exist polynomials u_k, $k = 1 : K$, such that

$$v(\boldsymbol{t}) \overset{\Delta}{=} 1 + \sum_{k=1}^{K} f_k(\boldsymbol{t}) u_k(\boldsymbol{t}) < 0, \ \forall \boldsymbol{t} \in \mathcal{D}(g).$$

Indeed, we can take $u_k = -\alpha f_k$ (where $\alpha > 0$ has to be determined), which leads to $v(\boldsymbol{t}) = 1 - \alpha \sum_{k=1}^{K} f_k(\boldsymbol{t})^2$. Since $\mathcal{D}(g)$ is compact and $f_k(\boldsymbol{t}) \neq 0$, $\forall \boldsymbol{t} \in \mathcal{D}(g)$ (otherwise the set (4.55) would not be empty), it follows that there exists $\beta > 0$ such that $f_k(\boldsymbol{t})^2 \geq \beta$, $\forall \boldsymbol{t} \in \mathcal{D}(g)$. Taking $\alpha > 1/K\beta$ ensures the negativity of $v(\boldsymbol{t})$.

Since the conditions of Theorem 4.5 are satisfied, it results that $-v \in \mathcal{M}(g)$, i.e., there exist $s_\ell \in \sum \mathbb{R}[\boldsymbol{t}]^2$ such that

$$-v = s_0 + \sum_{\ell=1}^{L} g_\ell s_\ell,$$

which is exactly (4.57). ■

We note that Theorem 4.38 also holds if the hypothesis of Theorem 4.5 is replaced with that of Theorem 4.4.

We can now state a Positivstellensatz for trigonometric polynomials. Consider the set

$$\mathcal{D}_E = \left\{ \boldsymbol{\omega} \in [-\pi, \pi]^d \;\middle|\; \begin{array}{l} E_k(\boldsymbol{\omega}) = 0, \ k = 1 : K \\ D_\ell(\boldsymbol{\omega}) \geq 0, \ \ell = 1 : L \end{array} \right\}, \tag{4.58}$$

where $D_\ell, E_k \in \mathbb{C}[z]$, $k = 1 : K$, $\ell = 1 : L$, are given trigonometric polynomials.

Theorem 4.39 *The set (4.58) is empty if and only if*

$$1 + \sum_{k=1}^{K} E_k(z) U_k(z) + S_0(z) + \sum_{\ell=1}^{L} D_\ell(z) S_\ell(z) = 0, \tag{4.59}$$

for some polynomials $U_k(z)$ and sum-of-squares polynomials $S_\ell(z)$.

Proof There are (at least) two possible ways of proving the theorem. The first is similar to the proof of Theorem 4.11: the transformations from Sect. 3.11.1 are employed to obtain a similar problem with real polynomials, for which Theorem 4.38 holds.

The second proof is similar to that of Theorem 4.38. Since trigonometric polynomials are continuous functions and the domain (4.13) is compact, there exist polynomials $U_k(z)$, $k = 1 : K$, such that

$$V(\boldsymbol{\omega}) \stackrel{\Delta}{=} 1 + \sum_{k=1}^{K} E_k(\boldsymbol{\omega}) U_k(\boldsymbol{\omega}) < 0, \ \forall \boldsymbol{\omega} \in \mathcal{D}.$$

Applying Theorem 4.11 for the polynomial $-V(z)$, we obtain (4.59). ■

We note that if the polynomials $D_\ell(z)$ and $E_k(z)$ have complex coefficients, then the sum-of-squares $S_\ell(z)$ and the polynomials $U_k(z)$ from (4.59) have also complex coefficients. However, if the polynomials defining (4.58) have real coefficients, then $S_\ell(z)$ and $U_k(z)$ have also real coefficients. The implementation of the Positivstellensatz from Theorem 4.39 and its applications to stability tests are discussed in Sect. 7.1.3.

4.5 Proof of Theorem 4.11

Complex coefficients. We assume for the beginning that the polynomial $R(z)$ has complex coefficients. Using standard trigonometric equalities, the polynomial (3.9) can be transformed into the form

$$R(\boldsymbol{\omega}) = \sum_{\boldsymbol{k}=\boldsymbol{0}}^{(\boldsymbol{n},\boldsymbol{n})} c_{\boldsymbol{k}} \prod_{i=1}^{d} (\cos \omega_i)^{k_i} (\sin \omega_i)^{k_{i+d}}, \tag{4.60}$$

where the relation between the coefficients $c_{\boldsymbol{k}} \in \mathbb{R}$ and the coefficients of $R(z)$ needs no explicit form; we need only to know that one can go from (3.9) to (4.60) and back. Defining $\boldsymbol{t} \in \mathbb{R}^{2d}$ by

$$t_i = \cos \omega_i, \quad t_{i+d} = \sin \omega_i, \quad i = 1 : d, \tag{4.61}$$

we can write

$$R(\boldsymbol{\omega}) \stackrel{\Delta}{=} P(\boldsymbol{t}) = \sum_{\boldsymbol{k}=\boldsymbol{0}}^{(\boldsymbol{n},\boldsymbol{n})} c_{\boldsymbol{k}} \boldsymbol{t}^{\boldsymbol{k}}. \tag{4.62}$$

The polynomial $P \in \mathbb{R}[\boldsymbol{t}]$ is defined on the set

$$\mathcal{T} = \{\boldsymbol{t} \in \mathbb{R}^{2d} \mid t_i^2 + t_{i+d}^2 = 1,\ i = 1 : d\}. \tag{4.63}$$

Similarly, the trigonometric polynomials $D_\ell(\boldsymbol{\omega})$ from (4.13) are transformed into the real polynomials $d_\ell(\boldsymbol{t})$. We denote

$$\mathcal{D}_r = \{\boldsymbol{t} \in \mathbb{R}^{2d} \mid d_\ell(\boldsymbol{t}) \geq 0,\ \ell = 1 : L\}. \tag{4.64}$$

By the above transformation, the domain $\mathcal{D}$ defined in (4.13) is transformed into

$$\mathcal{D}(g) = \mathcal{D}_r \cap \mathcal{T}. \tag{4.65}$$

We can write $\mathcal{D}(g)$ in the form (4.1), using the following $L + 2d$ polynomials:

$$g_\ell(\boldsymbol{t}) = \begin{cases} d_\ell(\boldsymbol{t}), & \ell = 1 : L, \\ -t_{\ell-L}^2 - t_{\ell+d-L}^2 + 1, & \ell = L+1 : L+d, \\ t_{\ell-d-L}^2 + t_{\ell-L}^2 - 1, & \ell = L+d+1 : L+2d. \end{cases} \tag{4.66}$$

Since $\mathcal{D}$ is not empty, it follows that $\mathcal{D}(g)$ from (4.65) is also not empty. Thus, $\mathcal{M}(g)$ defined as in (4.4) with the polynomials (4.66) is a quadratic module.

Moreover, it results from (4.66) that

$$p_0(\boldsymbol{t}) \overset{\Delta}{=} d - \sum_{i=1}^{2d} t_i^2 = \sum_{\ell=L+1}^{L+d} g_\ell(\boldsymbol{t}) \cdot 1 \in \mathcal{M}(g). \tag{4.67}$$

This shows that the quadratic module $\mathcal{M}(g)$ is Archimedean. We can now apply Theorem 4.5. Since by construction we have $P(\boldsymbol{t}) > 0$, $\forall \boldsymbol{t} \in \mathcal{D}(g)$, it results that $P \in \mathcal{M}(g)$, i.e.,

$$P(\boldsymbol{t}) = s_0(\boldsymbol{t}) + \sum_{\ell=1}^{L+2d} g_\ell(\boldsymbol{t}) s_\ell(\boldsymbol{t}), \tag{4.68}$$

with $s_\ell \in \sum \mathbb{R}[\boldsymbol{t}]^2$. We transform back to trigonometric polynomials by using (4.61). Since $g_\ell(\boldsymbol{t}) = 0$ for any $\boldsymbol{t} \in \mathcal{T}$, $\ell = L+1 : L+2d$, we obtain

$$R(\boldsymbol{\omega}) = S_0(\boldsymbol{\omega}) + \sum_{\ell=1}^{L} D_\ell(\boldsymbol{\omega}) \cdot S_\ell(\boldsymbol{\omega}), \tag{4.69}$$

where $S_\ell(\omega)$ are sum-of-squares. The extension from $\mathbb{T}^d$ to $\mathbb{C}^d$ is made via (3.10) and so we obtain (4.14).

Real coefficients. If $R(\boldsymbol{z})$ and $D_i(\boldsymbol{z})$ have real coefficients, then the equality (4.14) still stands, but now the polynomials $S_\ell(\boldsymbol{z})$ are the "real parts" (in the sense of retaining the real part of the coefficients) of some sum-of-squares. It remains to prove that these $S_\ell(\boldsymbol{z})$ are still sum-of-squares. Denote

$$H(z) = \sum_{k=0}^{n} (u_k + jv_k) z^{-k} = U(z) + jV(z) \tag{4.70}$$

a positive orthant polynomial, with $U, V \in \mathbb{R}[z]$, and let

$$E(z) = H(z)H^*(z^{-1}) \tag{4.71}$$

be a term of a sum-of-squares polynomial. Using (4.70), it results that

$$E(z) = U(z)U(z^{-1}) + V(z)V(z^{-1}) + j[U(z^{-1})V(z) - U(z)V(z^{-1})]. \tag{4.72}$$

Thus, the real part of $E(z)$ is a sum of two squares. We conclude that the real part of a sum-of-squares is sum-of-squares.

4.6 Bibliographical and Historical Notes

Theorems 4.1, 4.2, 4.4 and 4.5 are authored by, respectively, Schmüdgen [4], Jacobi and Prestel [5], Putinar [6], and Jacobi [7] (a simpler proof of the latter theorem appears in [8]); some particular cases have been previously investigated, e.g., when the constraints from (4.1) are linear [9]. The remarkable synchronization between these results on positive polynomials and the development of semidefinite programming seems to be mostly coincidental. However, the applicative side of the mentioned theorems has been grasped quickly by researchers in optimization. For example, the relaxations for finding the minimum value of a polynomial, subject to polynomial positivity constraints have been proposed by Lasserre [10]. Some results for matrix polynomials can be found in [11].

The results from Sects. 4.2 and 4.3 appear in [12], in their Gram set form; the Gram-pair set form is a direct application of the results from Sect. 3.9. The first BRL for univariate trigonometric polynomials, including the matrix case, was given in [13]. Bounded Real Lemmas for discrete-time systems (in state-space representation) have been proposed previously in [3, 14]; they are valid for recursive systems (although they may be implemented with SDP only for FIR systems), but work only globally, i.e., not on a frequency domain like (4.13).

The Positivstellensatz Theorem 4.36 is due to Stengle [15]. Its use for solving optimization problems with polynomials using SDP was initiated by Parrilo [16, 17]. Theorem 4.39 appeared in [18].

Problems

P 4.1 A relaxation method for finding the unconstrained minimum of a real polynomial $P(\boldsymbol{t})$ was presented in Sect. 3.7, based on Theorem 3.10. Assuming that the optimal $\boldsymbol{t}$ is finite (and in a region that can be roughly guessed), devise a different relaxation, based on solving a constrained problem like (4.8). (Hint: use the polynomial (4.11).)

P 4.2 An optimization problem with polynomials that are positive on a frequency domain (4.13) is solved in relaxed form using (i) the Gram set parameterization (4.16) and (ii) the Gram pair set parameterization (4.27). When it is certain that the results coincide?

P 4.3 Using the Bounded Real Lemma from Corollary 4.27, formulate an SDP problem to find the value of the real parameters a, b for which the H_∞ norm of the system

$$H(z_1, z_2) = (z_1^{-1} + z_2^{-1})^3 + az_2^{-2} + bz_2^{-1}$$

is minimum.

P 4.4 (*BRL with real matrices for complex polynomials*) Let $H(z)$ be a n-th order causal polynomial with complex coefficients. Prove that we have $|H(\omega)| \le \gamma$ if and only if there exists a positive semidefinite matrix $\boldsymbol{Q} \in \mathbb{R}^{(n+1)\times(n+1)}$ such that

$$\gamma^2\delta_k = \operatorname{tr}[\boldsymbol{\Gamma}_k \boldsymbol{Q}],$$
$$\begin{bmatrix} \boldsymbol{Q} & \boldsymbol{a} & \boldsymbol{b} \\ \boldsymbol{a}^T & 1 & 0 \\ \boldsymbol{b}^T & 0 & 1 \end{bmatrix} \succeq 0,$$

where $\boldsymbol{a}$ and $\boldsymbol{b}$ depend on the coefficients of $H(z)$ as in the relations before (2.72) and the matrices $\boldsymbol{\Gamma}_k$ appear in (2.79).

Generalize this result to multivariate polynomials (after reviewing **P** 3.9).

Finally, generalize the result for the case where the inequality $H(\boldsymbol{\omega}) < \gamma$ is not valid globally, but on a frequency domain (4.13).

P 4.5 Let $H(z)$ and $G(z)$ be positive orthant polynomials and $\mathcal{D}$ a frequency domain (4.13). Show that

$$|H(\boldsymbol{\omega})|^2 + |G(\boldsymbol{\omega})|^2 < \gamma^2, \quad \forall \boldsymbol{\omega} \in \mathcal{D},$$

if and only if there exist matrices $\boldsymbol{Q}_\ell \succeq 0$, $\ell = 0 : L$, such that the LMIs (4.38) and

$$\begin{bmatrix} \boldsymbol{Q}_0 & \boldsymbol{h} & \boldsymbol{g} \\ \boldsymbol{h}^T & 1 & 0 \\ \boldsymbol{g}^T & 0 & 1 \end{bmatrix} \succeq 0$$

hold, where $\boldsymbol{h}$ and $\boldsymbol{g}$ are the vectors of coefficients.

Formulate a similar result using the Gram-pair set parameterization.

P 4.6 Prove the matrix polynomial BRL Theorem 4.32, on the following steps.

1. $\sigma_{\max}[\boldsymbol{H}(\boldsymbol{\omega})] < \gamma$ on $\mathcal{D}$ if and only if $\boldsymbol{R}(z) = \gamma^2\boldsymbol{I}_p - \boldsymbol{H}(z)\boldsymbol{H}^T(z^{-1})$ is positive on $\mathcal{D}$.

2. $\bar{\boldsymbol{H}}\bar{\boldsymbol{H}}^T$ is a Gram matrix for $\boldsymbol{H}(\boldsymbol{z})\boldsymbol{H}^T(\boldsymbol{z}^{-1})$.

3. $\boldsymbol{R}(\boldsymbol{z})$ is positive on $\mathcal{D}$ if there exists a Gram set $\boldsymbol{Q}_\ell \geq 0$ associated with $\gamma^2\boldsymbol{I}_p$, which is equivalent to (4.49), such that $\boldsymbol{Q}_0 \succeq \bar{\boldsymbol{H}}\bar{\boldsymbol{H}}^T$, which is equivalent to (4.50).

Extend the result for characterizing the inequality $\sigma_{\max}[\boldsymbol{H}(\boldsymbol{\omega})] < |A(\boldsymbol{\omega})|, \forall\boldsymbol{\omega} \in \mathcal{D}$, instead of (4.48), where $A(\boldsymbol{z})$ is given.

P 4.7 Remind the definition of 2D hybrid polynomials from **P** 3.10 and the notations therein. Assume that the domain

$$\mathcal{D} = \{(t, z) \in \mathbb{R} \times \mathbb{T} \mid D_\ell(t, z) \geq 0,\ \ell = 1 : L\}, \tag{4.73}$$

where $D_\ell(t, z)$ are defined as in (3.129), is bounded and thus $\mathcal{D} \subset [a, b] \times \mathbb{T}$ for some constants a and b; we can assume with no loss of generality that $D_L(t, z) = (t - a)(b - t)$. Prove that if a polynomial (3.129) is positive on $\mathcal{D}$, then there exist sum-of-squares $S_\ell(t, z)$, $\ell = 0 : L$, such that

$$R(t, z) = S_0(t, z) + \sum_{\ell=1}^{L} D_\ell(t, z) \cdot S_\ell(t, z). \tag{4.74}$$

Equivalently, there exist positive semidefinite matrices $\boldsymbol{Q}_\ell$, $\ell = 0 : L$, such that

$$r_{k1,k2} = \operatorname{tr}[\boldsymbol{T}_{k_1,k_2}\boldsymbol{Q}_0] + \sum_{\ell=1}^{L} \operatorname{tr}[\boldsymbol{\Psi}_{\ell,k_1,k_2}\boldsymbol{Q}_\ell], \tag{4.75}$$

where $\boldsymbol{T}_{k_1,k_2} = \boldsymbol{\Theta}_{k_2} \otimes \boldsymbol{\Upsilon}_{k_1}$ and

$$\boldsymbol{\Psi}_{\ell,k_1,k_2} = \sum_{i_1+j_1=k_1}\sum_{i_2+j_2=k_2} (d_\ell)_{i_1,i_2}\boldsymbol{T}_{j_1,j_2}. \tag{4.76}$$

P 4.8 (*BRL for hybrid polynomials*) Let $H(t, z)$ be a hybrid polynomial that is causal in z and $\boldsymbol{h}$ the vector of its coefficients, arranged as usual. If the inequality $|H(t, z)| < \gamma, \forall(t, z) \in \mathcal{D}$, holds for $\mathcal{D}$ defined in (4.73), then there exist matrices $\boldsymbol{Q}_\ell \succeq 0$, $\ell = 0 : L$, such that

$$\gamma^2\delta_{k_1k_2} = \operatorname{tr}[\boldsymbol{T}_{k_1,k_2}\boldsymbol{Q}_0] + \sum_{\ell=1}^{L} \operatorname{tr}[\boldsymbol{\Psi}_{\ell,k_1,k_2}\boldsymbol{Q}_\ell], \tag{4.77}$$

where the notations are like in the previous problem, $\delta_{k_1k_2}$ is the Kronecker symbol, and

$$\begin{bmatrix} \boldsymbol{Q}_0 & \boldsymbol{h} \\ \boldsymbol{h}^H & 1 \end{bmatrix} \succeq 0. \tag{4.78}$$

Conversely, (4.77) and (4.78) imply $|H(t, z)| \leq \gamma, \forall(t, z) \in \mathcal{D}$.

P 4.9 Let $R \in \mathbb{C}[z]$ be a polynomial that is positive on the domain (4.13). Using the Positivstellensatz Theorem 4.39, show that there exist sum-of-squares polynomials $S(z)$ and $S_\ell(z)$, $\ell \in 0 : L$, such that

$$S(z)R(z) = 1 + S_0(z) + \sum_{\ell=1}^{L} D_\ell(z)S_\ell(z).$$

Compare this result with Theorem 4.11.

References

1. A. Prestel, C.N. Delzell, *Positive Polynomials: From Hilbert's 17th Problem to Real Algebra* (Springer Monographs in Mathematics, Berlin, 2001)
2. C. Scheiderer. Positivity and Sums of Squares: A Guide to Recent Results. In M. Putinar and S. Sullivant, editors, *Emerging Applications of Algebraic Geometry*, volume 149, pages 271–324. IMA Volumes in Mathematics and its Applications, 2009
3. L. Xie, C. Du, C. Zhang, Y.C. Soh, H_∞ Deconvolution Filtering of 2-D Digital Systems. IEEE Trans. Signal Proc. **50**(9), 2319–2332 (2002)
4. K. Schmüdgen, The K-moment problem for compact semi-algebraic sets. Math. Ann. **289**(2), 203–206 (1991)
5. T. Jacobi, A. Prestel, Distinguished representations of strictly positive polynomials. J. Reine Angew. Math. **532**, 223–235 (2001)
6. M. Putinar, Positive polynomials on compact semi-algebraic sets. Ind. Univ. Math. J. **42**(3), 969–984 (1993)
7. T. Jacobi, A representation theorem for certain partially ordered commutative rings. Math. Z **237**, 259–273 (2001)
8. M. Schweighofer, Optimization of polynomials on compact semialgebraic Sets. SIAM J. Opt. **15**(3), 805–825 (2005)
9. D. Handelman, Representing polynomials by positive linear functions on compact convex polyhedra. Pacific J. Math. **132**(1), 35–62 (1988)
10. J.B. Lasserre, Global optimization with polynomials and the problem of moments. SIAM J. Opt. **11**(3), 796–814 (2001)
11. C.W. Scherer, C.W.J. Hol, Matrix sum-of-squares relaxations for robust semi-definite programs. Math. Prog. Ser. B **107**, 189–211 (2006)
12. B. Dumitrescu, Trigonometric polynomials positive on frequency domains and applications to 2-D FIR filter design. IEEE Trans. Signal Proc. **54**(11), 4282–4292 (2006)
13. B. Dumitrescu, Bounded real lemma for FIR MIMO systems. IEEE Signal Proc. Lett. **12**(7), 496–499 (2005)
14. C. Du, L. Xie, Y.C. Soh, H_∞ Filtering of 2-D Discrete Systems. IEEE Trans. Signal Proc. **48**(6), 1760–1768 (2000)
15. G. Stengle, A nullstellensatz and a positivstellensatz in semialgebraic geometry. Math. Ann. **207**, 87–97 (1974)
16. Parrilo, P.A.: Structured Semidefinite Programs and Semialgebraic Geometry Methods in Robustness and Optimization. Ph.D. thesis, California Institute of Technology (2000)
17. P.A. Parrilo, Semidefinite Programming Relaxations for Semialgebraic Problems. Math. Prog. Ser. B **96**, 293–320 (2003)
18. B. Dumitrescu, Positivstellensatz for trigonometric polynomials and multidimensional stability tests. IEEE Trans. Circ. Syst. II **54**(4), 353–356 (2007)

Chapter 5
Design of FIR Filters

Abstract Filter design is one of the perennial topics in signal processing. FIR filters are often preferred for their simple implementation and robustness, so they are an appropriate subject for this first chapter devoted to applications. All the design methods presented here are based on positive trigonometric polynomials and the associated optimization tools; they are optimal for 1D filters and practically optimal for 2D filters. This is in contrast with many other methods that approximate the optimum, either from a desire to obtain rapidly the solution or from a lack of instruments that give optimality. We treat here three basic design problems: 1D filters, 2D filters, and deconvolution. For each problem, we consider several design specifications. In the 2D (and multidimensional) case, at the time of apparition, the methods were very different from those present in the literature.

5.1 Design of FIR Filters

In this section, we present three optimization methods for FIR filters, based on the LMIs described in the previous chapters. Although, for most applications, FIR filters can be satisfactorily designed using approximate constraints—and not exactly, as below—the presented methods deserve careful study, as they solve the simplest instances of more general problems. Let

$$H(z) = \sum_{k=0}^{n} h_k z^{-k} \tag{5.1}$$

be an FIR filter of order n, with real coefficients. For simplicity, we design only lowpass filters, the generalization to other types being straightforward. The passband is $[0, \omega_p]$; the stopband is $[\omega_s, \pi]$; and their edges ω_p and ω_s are given. A typical design of such a filter is based on the peak constrained least squares (PCLS) optimization, in which the stopband energy is minimized, while the maximum error (with respect to 1 in the passband and 0 in the stopband) is kept below some prescribed bounds. Since

B. Dumitrescu, *Positive Trigonometric Polynomials and Signal Processing Applications*, Signals and Communication Technology,
DOI 10.1007/978-3-319-53688-0_5

there will be differences between the design specifications for the three methods, we give here only the common information. The stopband energy of an FIR filter is

$$E_s = \frac{1}{\pi}\int_{\omega_s}^{\pi} |H(\omega)|^2 d\omega. \tag{5.2}$$

Denoting, as usual, the vector of filter coefficients by $\boldsymbol{h}$, the stopband energy is given by the quadratic form

$$E_s = \boldsymbol{h}^T \boldsymbol{C}\boldsymbol{h}, \tag{5.3}$$

where $\boldsymbol{C} = \text{Toep}(c_0, c_1, \ldots, c_n) \succeq 0$ and

$$c_k = \begin{cases} 1 - \omega_s/\pi, & \text{if } k = 0, \\ -\dfrac{\sin k\omega_s}{k\pi}, & \text{if } k > 0. \end{cases} \tag{5.4}$$

In terms of the squared magnitude $R(\omega) = |H(\omega)|^2$, the stopband energy is the linear function

$$E_s = c_0 r_0 + 2\sum_{k=1}^{n} c_k r_k. \tag{5.5}$$

Refer to Sect. 2.3 for details on the transformation from (5.3) to (5.5). Notice also that both functions (5.3) (since $\boldsymbol{C}$ is positive semidefinite) and (5.5) are convex in their variables (the coefficients of $H(z)$ and $R(z)$, respectively).

We remind that the coefficients of a polynomial $\tilde{R}(z)$ which is nonnegative on an interval $[\alpha, \beta]$ can be parameterized via an LMI. For example, using Theorem 1.17 and denoting $\cos\alpha = a$, $\cos\beta = b$, positivity on $[\alpha, \beta]$ is equivalent to the existence of positive semidefinite matrices $\boldsymbol{Q}_1 \in \mathbb{R}^{(n+1)\times(n+1)}$, $\boldsymbol{Q}_2 \in \mathbb{R}^{(n-1)\times(n-1)}$ such that

$$\begin{aligned} \tilde{r}_k &= \text{tr}[\boldsymbol{\Theta}_k \boldsymbol{Q}_1] + \text{tr}\big[\big(-(ab + \tfrac{1}{2})\boldsymbol{\Theta}_k + \tfrac{a+b}{2}(\boldsymbol{\Theta}_{k-1} + \boldsymbol{\Theta}_{k+1}) \\ &\qquad - \tfrac{1}{4}(\boldsymbol{\Theta}_{k-2} + \boldsymbol{\Theta}_{k+2})\big)\, \boldsymbol{Q}_2\big] \\ &\overset{\Delta}{=} \mathcal{L}_{k,\alpha,\beta}(\boldsymbol{Q}_1, \boldsymbol{Q}_2). \end{aligned} \tag{5.6}$$

A similar equality, but with smaller matrices, can be derived from Theorem 1.18, as in problem **P** 2.12. In the remainder of this section, we will use generically the notation (5.6), without detailing the exact form of the LMI. If the polynomial is globally nonnegative, its coefficients are parameterized as in (2.6) or (2.94) (the trace and Gram-pair parameterizations, respectively). We denote these equalities by $\tilde{r}_k = \mathcal{L}_k(\boldsymbol{Q})$, where $\boldsymbol{Q} \succeq 0$ (we hide the second matrix that appears in (2.94)).

5.1.1 Optimization of Linear-Phase FIR Filters

We consider the optimization of linear-phase symmetric FIR filters of even order $n = 2\tilde{n}$. Since we have to optimize only the magnitude of the filter, we can work with the zero-phase filter

$$\tilde{H}(z) = \sum_{k=-\tilde{n}}^{\tilde{n}} \tilde{h}_k z^{-k}, \ \ \tilde{h}_{-k} = \tilde{h}_k. \tag{5.7}$$

This is a symmetric trigonometric polynomial and so $\tilde{H}(\omega)$ is real, having the form (1.4). The standard PCLS problem can be formulated as

$$\begin{array}{ll} \min\limits_{\tilde{H}\in\mathbb{R}_{\tilde{n}}[z]} & E_s \\ \text{s.t.} & |\tilde{H}(\omega) - 1| \le \gamma_p, \ \forall\omega \in [0, \omega_p] \\ & |\tilde{H}(\omega)| \le \gamma_s, \ \forall\omega \in [\omega_s, \pi] \end{array} \tag{5.8}$$

where γ_p and γ_s are given error bounds. The magnitude constraints are formulated only in passband and stopband. Additionally, we can enforce an upper bound on the frequency response in the transition band, in order to prevent undesirable spikes there; the constraint is $\tilde{H}(\omega) \le 1 + \gamma_p, \ \forall\omega \in [\omega_p, \omega_s]$. The spectral mask that contains the frequency response is shown in Fig. 5.1. We can formulate the design problem (5.8) by means of polynomials that are nonnegative on given intervals, obtaining

$$\begin{array}{ll} \min\limits_{\tilde{H}\in\mathbb{R}_{\tilde{n}}[z]} & E_s \\ \text{s.t.} & 1 + \gamma_p - \tilde{H}(\omega) \ge 0, \ \forall\omega \in [0, \omega_s] \\ & \tilde{H}(\omega) - 1 + \gamma_p \ge 0, \ \forall\omega \in [0, \omega_p] \\ & \gamma_s - \tilde{H}(\omega) \ge 0, \ \forall\omega \in [\omega_s, \pi] \\ & \tilde{H}(\omega) - \gamma_s \ge 0, \ \forall\omega \in [\omega_s, \pi] \end{array} \tag{5.9}$$

Using the parameterization (5.6), we transform (5.9) into an SDP problem. Before doing so, we remark that the first inequality constraint can be extended to the whole domain $[0, \pi]$, since in the stopband it holds obviously; this is an advantage, as the

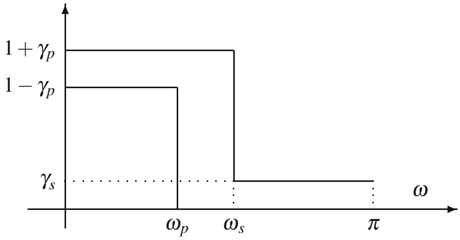


Fig. 5.1 Magnitude bounds for the frequency response of a lowpass filter

global nonnegativity LMI is simpler; using the trace parameterization (2.6), it can be expressed as a function of a single positive definite matrix, while two matrices are necessary for nonnegativity on an interval; with the Gram-pair parameterization (2.94), there are two matrices of the same size in both cases, so the advantage is marginal (simpler coefficient matrices). Also, we have to express the stopband energy function of the coefficients of $\tilde{H}(z)$. The vector $\tilde{\boldsymbol{h}} = [\tilde{h}_0\ \tilde{h}_1\ \dots\ \tilde{h}_{\tilde{n}}]^T$, containing the distinct coefficients of $\tilde{H}(z)$, generates $\boldsymbol{h}$ via

$$\boldsymbol{h} = \boldsymbol{P}\tilde{\boldsymbol{h}}, \quad \boldsymbol{P} = \begin{bmatrix} \boldsymbol{0} & \boldsymbol{J}_{\tilde{n}} \\ 1 & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{I}_{\tilde{n}} \end{bmatrix}, \tag{5.10}$$

where $\boldsymbol{J}_{\tilde{n}}$ is the counteridentity matrix of size $\tilde{n} \times \tilde{n}$. The stopband energy (5.3) can be expressed as

$$E_s = \tilde{\boldsymbol{h}}^T \tilde{\boldsymbol{C}} \tilde{\boldsymbol{h}}, \quad \text{with } \tilde{\boldsymbol{C}} = \boldsymbol{P}^T \boldsymbol{C} \boldsymbol{P} \succeq 0. \tag{5.11}$$

Finally, the SDP form of (5.9) is (actually, this is SQLP, due to the SOC constraint)

$$\begin{array}{ll} \min\limits_{\tilde{\boldsymbol{h}}, \boldsymbol{y}, \varepsilon, \boldsymbol{Q}_1, \dots, \boldsymbol{Q}_7} & \varepsilon \\ \text{s.t.} & \left.\begin{array}{l} (1+\gamma_p)\delta_k - \tilde{h}_k = \mathcal{L}_k(\boldsymbol{Q}_1) \\ \tilde{h}_k - (1-\gamma_p)\delta_k = \mathcal{L}_{k,0,\omega_p}(\boldsymbol{Q}_2, \boldsymbol{Q}_3) \\ \gamma_s\delta_k - \tilde{h}_k = \mathcal{L}_{k,\omega_s,\pi}(\boldsymbol{Q}_4, \boldsymbol{Q}_5) \\ \tilde{h}_k - \gamma_s\delta_k = \mathcal{L}_{k,\omega_s,\pi}(\boldsymbol{Q}_6, \boldsymbol{Q}_7) \end{array}\right\} \ k = 0 : \tilde{n} \\ & \boldsymbol{y} = \tilde{\boldsymbol{C}}^{1/2}\tilde{\boldsymbol{h}} \\ & \|\boldsymbol{y}\| \le \varepsilon, \ \boldsymbol{Q}_1 \succeq 0, \ \dots, \ \boldsymbol{Q}_7 \succeq 0 \end{array} \tag{5.12}$$

Example 5.1 We design a linear-phase filter with the specifications $n = 50$, $\omega_p = 0.2\pi$, $\omega_s = 0.25\pi$, $\gamma_p = 0.1$, and $\gamma_s = 0.0158$ (the last two values correspond to a passband ripple of 1.74 dB and a stopband attenuation of 36 dB, respectively). Solving (5.12), we obtain the filter whose frequency response is shown in Fig. 5.2. The stopband energy is $E_s = 4.36 \cdot 10^{-5}$. The frequency response has a typical shape; it is equiripple in the passband; in the stopband, the first ripples have height γ_s, while for higher frequencies, the attenuation is better. Varying γ_s, one can trade off stopband attenuation and energy. For $\gamma_s = -36.5$ dB, the frequency response is almost equiripple. Of course, further decreasing the value of γ_s leads to no solution to the problem; this is signaled when trying to solve (5.12). ■

5.1.2 Magnitude Optimization

We remove now the linear-phase constraint and assume no relations between the coefficients of the FIR filter (5.1). We optimize only the magnitude $R(\omega) = |H(\omega)|^2$,

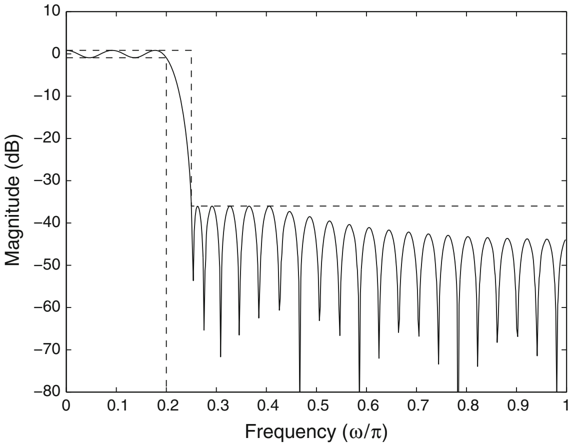


Fig. 5.2 Frequency response of the filter designed in Example 5.1

disregarding completely the phase information. With the spectral mask constraints from Fig. 5.1, the PCLS optimization can be formulated in terms of the magnitude $R(\omega)$, which is a nonnegative trigonometric polynomial. Using the same remarks that have led to (5.9) and the stopband energy formula (5.5), the design problem is formulated in terms of polynomials nonnegative on intervals as follows

$$\begin{array}{ll} \min\limits_{R\in\mathbb{R}_n[z]} & c_0 r_0 + 2\sum_{k=1}^n c_k r_k \\ \text{s.t.} & (1+\gamma_p)^2 - R(\omega) \geq 0,\ \forall\omega \\ & R(\omega) - (1-\gamma_p)^2 \geq 0,\ \forall\omega\in[0,\omega_p] \\ & \gamma_s^2 - R(\omega) \geq 0,\ \forall\omega\in[\omega_s,\pi] \\ & R(\omega) \geq 0,\ \forall\omega \end{array} \tag{5.13}$$

The equivalent SDP problem is

$$\begin{array}{ll} \min\limits_{\boldsymbol{Q}_1,\ldots,\boldsymbol{Q}_6} & c_0 r_0 + 2\sum_{k=1}^n c_k r_k \\ \text{s.t.} & \left.\begin{array}{l} (1+\gamma_p)^2\delta_k - r_k = \mathcal{L}_k(\boldsymbol{Q}_1) \\ r_k - (1-\gamma_p)^2\delta_k = \mathcal{L}_{k,0,\omega_p}(\boldsymbol{Q}_2,\ \boldsymbol{Q}_3) \\ \gamma_s^2\delta_k - r_k = \mathcal{L}_{k,\omega_s,\pi}(\boldsymbol{Q}_4,\ \boldsymbol{Q}_5) \\ r_k = \mathcal{L}_k(\boldsymbol{Q}_6) \end{array}\right\} \ k = 0:n \end{array} \tag{5.14}$$

After solving (5.14), the FIR filter $H(z)$ is recovered from $R(z)$ via spectral factorization.

Example 5.2 We use the same specifications as in Example 5.1, but with $\gamma_s = 0.01 = -40\,\text{dB}$. Solving (5.14), we obtain after spectral factorization the filter whose frequency response is shown in Fig. 5.3. Its stopband energy is $E_s = 3.29\cdot 10^{-6}$. We note that the performance is improved with respect to the linear-phase case, in terms of both stopband energy and attenuation. Decreasing the stopband attenuation bound

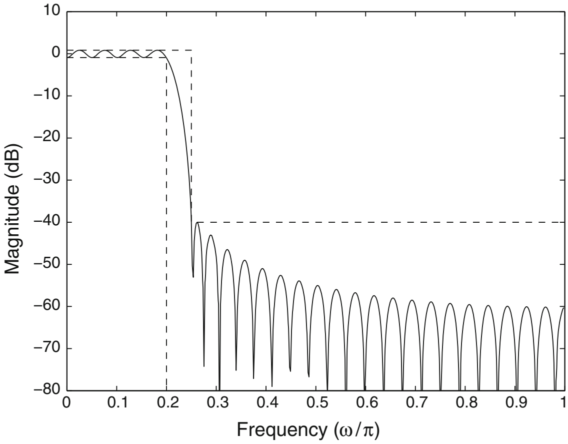


Fig. 5.3 Frequency response of the filter designed in Example 5.2, $\gamma_s = -40\,\text{dB}$

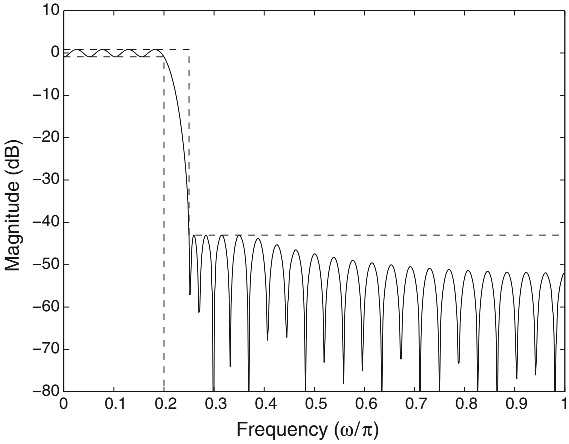


Fig. 5.4 Frequency response of the filter designed in Example 5.2, $\gamma_s = -43\,\text{dB}$

to $\gamma_s = -43\,\text{dB}$ leads to a stopband energy increased to $E_s = 7.19 \cdot 10^{-6}$ and the frequency response from Fig. 5.4. As for the linear-phase filters, the response is equiripple in the passband and in the initial part of the stopband. ■

5.1.3 Approximate Linear-Phase FIR Filters

The third design problem is a compromise between the previous two. The phase is not structurally constrained, but the optimization problem is formulated such that approximately linear phase is obtained. Given a desired group delay τ, we optimize the FIR filter as follows

$$\begin{array}{ll} \displaystyle\min_{H \in \mathbb{R}_{n+}[z]} & E_s \\ \text{s.t.} & |H(\omega) - \mathrm{e}^{-j\tau\omega}| \le \gamma_p, \ \forall\omega \in [0, \omega_p] \\ & |H(\omega)| \le \gamma_s, \ \forall\omega \in [\omega_s, \pi] \end{array} \tag{5.15}$$

In the passband, the error is considered with respect to an ideal *complex* response, that of a filter with group delay equal to τ. Typically, the interest is in low delay filters, with $\tau < n/2$. (If $\tau = n/2$, then the solution of (5.15) is a linear-phase filter identical to the solution of (5.8)). We note that the first constraint of (5.15) implies that $1 - \gamma_p \le |H(\omega)| \le 1 + \gamma_p$ and so γ_p serves also as an upper bound for the magnitude error in the passband. To express (5.15) as an SDP problem, let us remind the Bounded Real Lemma from Corollary 4.27. In the 1D case, it says that $|H(\omega)| \le \gamma$, $\forall\omega \in [\alpha, \beta]$, if and only if there exist positive semidefinite matrices $\boldsymbol{Q}_1$ and $\boldsymbol{Q}_2$ such that

$$\begin{array}{l} \gamma^2\delta_k = \mathcal{L}_{k,\alpha,\beta}(\boldsymbol{Q}_1, \boldsymbol{Q}_2), \quad k = 0 : n, \\ \begin{bmatrix} \boldsymbol{Q}_1 & \boldsymbol{h} \\ \boldsymbol{h}^T & 1 \end{bmatrix} \succeq 0. \end{array} \tag{5.16}$$

(A similar BRL can be derived from the Gram-pair form given by Theorem 4.35.) The stopband constraint from (5.15) has the standard BRL form, while the passband constraint has this form if the group delay τ is a nonnegative integer; if so, the BRL is formulated for the FIR filter $\hat{H}(z) = H(z) - z^{-\tau}$. Denoting $\boldsymbol{e}_\tau$ the unit vector with the value of 1 on the τ-th position, the SDP problem equivalent to (5.15) is

$$\begin{array}{ll} \displaystyle\min_{\boldsymbol{h}, \boldsymbol{y}, \varepsilon, \boldsymbol{Q}_1, \ldots, \boldsymbol{Q}_4} & \varepsilon \\ \text{s.t.} & \gamma_p^2\delta_k = \mathcal{L}_{k,0,\omega_p}(\boldsymbol{Q}_1, \boldsymbol{Q}_2), \quad k = 0 : n \\ & \begin{bmatrix} \boldsymbol{Q}_1 & \boldsymbol{h} - \boldsymbol{e}_\tau \\ \boldsymbol{h}^T - \boldsymbol{e}_\tau^T & 1 \end{bmatrix} \succeq 0 \\ & \gamma_s^2\delta_k = \mathcal{L}_{k,\omega_s,\pi}(\boldsymbol{Q}_3, \boldsymbol{Q}_4), \quad k = 0 : n \\ & \begin{bmatrix} \boldsymbol{Q}_3 & \boldsymbol{h} \\ \boldsymbol{h}^T & 1 \end{bmatrix} \succeq 0 \\ & \boldsymbol{y} = \boldsymbol{C}^{1/2}\boldsymbol{h} \\ & \|\boldsymbol{y}\| \le \varepsilon, \ \boldsymbol{Q}_2 \succeq 0, \boldsymbol{Q}_4 \succeq 0 \end{array} \tag{5.17}$$

Remark that there is no need to put explicitly the constraints $\boldsymbol{Q}_1 \succeq 0$, $\boldsymbol{Q}_3 \succeq 0$, since these matrices are principal blocks of larger positive semidefinite matrices. We note that this SDP problem has the smallest number of matrix variables, compared to (5.12) and (5.14), and hence the lower complexity. However, adding a magnitude constraint in the transition band (see problem **P** 5.3) increases the number of parameter matrices to six.

Example 5.3 We use the same specifications as in Example 5.1 and $\tau = 22$ (note that $\tau < n/2 = 25$). After solving (5.17), we obtain the filter whose frequency response is shown in Fig. 5.5 and a stopband energy $E_s = 1.92 \cdot 10^{-5}$. Comparing

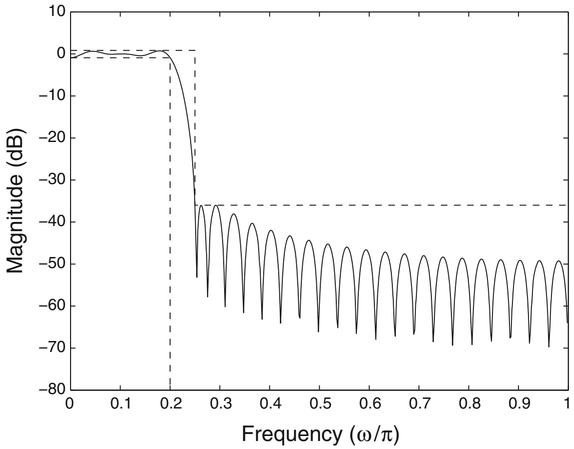


Fig. 5.5 Frequency response of the filter designed in Example 5.3

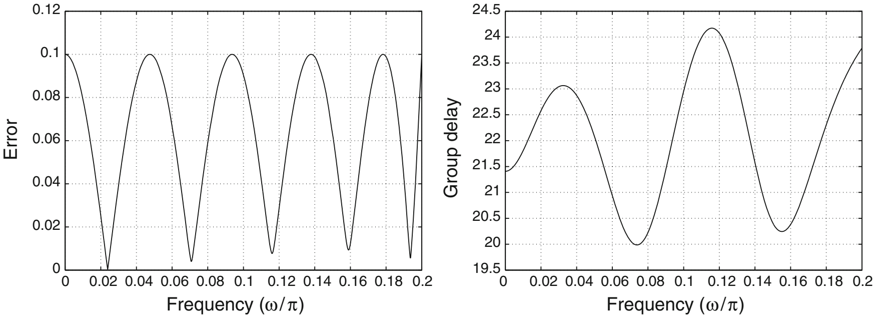


Fig. 5.6 Magnitude of complex passband error (*left*) and group delay (*right*) of the filter designed in Example 5.3

with Example 5.1, we see that relaxing the linear-phase constraint leads to a lower stopband energy, for the same error bounds. We note that the magnitude response is not equiripple in the passband. This is a normal behavior, since the passband error $|H(\omega) - \mathrm{e}^{-j\tau\omega}|$ is optimized; this error, shown in the left side of Fig. 5.6, is equiripple. In the right side of the figure, the group delay is shown in the passband. The group delay error with respect to τ is about 2.2. Of course, due to the expression of the error, we cannot optimize the magnitude and the group delay independently. However, the compromise is usually satisfactory. ■

5.2 Design of 2D FIR Filters

Only two of the three methods presented in the previous section can be generalized to multidimensional FIR filters. The method that obviously fails to generalize is that based on the optimization of the magnitude; we can optimize a positive trigonometric polynomial meant to signify $|H(\boldsymbol{\omega})|^2$, but the result cannot be spectrally factorized to recover $H(\boldsymbol{z})$. Still, we can use the properties of positive trigonometric polynomials described in Chap. 4, especially the parameterization of positive polynomials on frequency domains given by Theorem 4.11 and the LMI forms that result from it. We remind that if $R(\boldsymbol{z})$ is a multivariate trigonometric polynomial (3.1) that is positive on a domain $\mathcal{D}$ defined as in (4.13), then its coefficients can be parameterized via the LMIs (4.16) or (4.27). For brevity, we denote generically these identities by

$$r_{\boldsymbol{k}} = \mathcal{L}_{\boldsymbol{k},\mathcal{D}}(\boldsymbol{Q}_0, \ldots, \boldsymbol{Q}_L). \tag{5.18}$$

If the polynomial is globally positive, we denote $r_{\boldsymbol{k}} = \mathcal{L}_{\boldsymbol{k}}(\boldsymbol{Q})$.

Before going into details, let us point out the difficulties in generalizing the approach illustrated by e.g., the design problems (5.8), (5.12). Since this is the most practical case, many comments will be related to 2D filters.

- The frequency domains that represent the passband and the stopband may have various shapes. The immediate generalization of an interval $[\alpha, \beta]$ on the frequency axis is a Cartesian product $[\alpha_1, \beta_1] \times [\alpha_2, \beta_2]$. However, many other shapes (diamond, circle, fan, etc.) are interesting in the design of 2D FIR filters. We have to investigate if such shapes can be generated in the form required by Theorem 4.11.
- Typically, the LMIs that go together with multivariate positive polynomials implement relaxed versions of the optimization problems. It is possible that the solutions given by the relaxed problems are not optimal with respect to the original problem.

In the sequel, we will see that these difficulties can be overcome. The shapes of the frequency domains will be discussed in Sect. 5.2.1. The distance from optimality will be studied experimentally.

The design problem will be simpler than in the 1D case. Our study will be confined to minimax (or H_∞) optimization. Given a passband $\mathcal{D}_p$, a stopband $\mathcal{D}_s$, both being unions of frequency domains defined as in (4.13), and a maximum passband error γ_p, the maximum stopband error is minimized. This approach can be combined easily with stopband energy (least squares) minimization, as in the 1D case, but, besides its simpler form, the minimax optimization will allow us to evaluate accurately the effects of relaxation.

5.2.1 2D Frequency Domains

In classic design methods, the passbands and stopbands of 2D filters are delimited by simple curves (circle, ellipse, diamond), described by low-degree polynomials in $\boldsymbol{\omega}$.

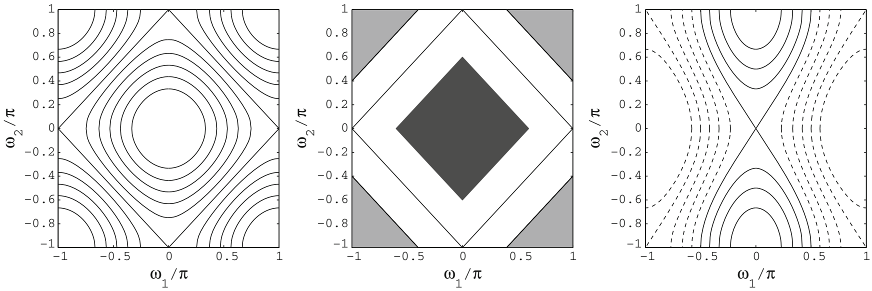


Fig. 5.7 *Left* borders of the domains defined by (5.20), for $c = -1.5 : 0.3 : 1.5$ (from exterior to interior). *Middle* diamond domains described by (5.21). *Right* borders of the domains defined by (5.22), for $c = -1.5 : 0.5 : 2.5$

Since we aim to obtain SDP methods based on Theorem 4.11, we consider domains $\mathcal{D}$ described by the positivity of some trigonometric polynomials, as in (4.13). On the one side, this approach reduces the number of possible shapes. On the other, it is more natural, since the frequency response of an FIR filter is also a trigonometric polynomial.

We give here few examples of frequency domains that can be obtained with simple trigonometric polynomials with real coefficients, in the 2D case; these examples suggest that other shapes can be obtained as well. We remind that intersection, union, and complementary of such domains can be used without restrictions, as discussed in Remark 4.21.

Rectangles. A rectangle in $[-\pi, \pi]^2$, whose sides are parallel to the axis, is defined by

$$\begin{aligned} D_1(\boldsymbol{\omega}) &= \cos\omega_1 - c_1 \geq 0, \\ D_2(\boldsymbol{\omega}) &= \cos\omega_2 - c_2 \geq 0. \end{aligned} \tag{5.19}$$

This rectangle is actually $[-\arccos c_1, \arccos c_1] \times [-\arccos c_2, \arccos c_2]$.

Low band. The simplest (in the sense that it is defined by a single polynomial) shape suited to describe low-frequency bands is defined by

$$D_1(\boldsymbol{\omega}) = \cos\omega_1 + \cos\omega_2 - c \geq 0. \tag{5.20}$$

The curves defined by $D_1(\boldsymbol{\omega}) = 0$, representing the borders of the domain defined by (5.20), are drawn on the left of Fig. 5.7, for several values of the parameter c. For values of c near 2, the shape is almost circular, while for c near 0, it is almost a diamond.

Diamond shapes of any size can be obtained with

$$\begin{aligned} D_1(\boldsymbol{\omega}) &= \cos(\omega_1 + \omega_2) - c \geq 0, \\ D_2(\boldsymbol{\omega}) &= \cos(\omega_1 - \omega_2) - c \geq 0, \\ D_3(\boldsymbol{\omega}) &= \cos\omega_1 + \cos\omega_2 \geq 0. \end{aligned} \tag{5.21}$$

In this case, the periodicity of trigonometric polynomials should be taken into account. Since $\omega_1 \pm \omega_2 \in [-2\pi, 2\pi]$, the first two polynomials from (5.21) define not only the desired central diamond shown in black in the middle of Fig. 5.7, but also the four gray triangles in the corners of $[-\pi, \pi]^2$. The third polynomial from (5.21) (which has the form (5.20)) has the purpose of removing these high-frequency areas; the line corresponding to $D_3(\boldsymbol{\omega}) = 0$ (which is the border of a diamond) separates the desired area from the undesired ones.

Fan. Shapes suited to fan filters are defined by e.g.,

$$D_1(\boldsymbol{\omega}) = 2\cos\omega_1 - \cos\omega_2 - c \ge 0 \tag{5.22}$$

and illustrated on the right of Fig. 5.7 where dashed lines correspond to $c < 1$ and solid lines to $c \ge 1$. It is clear that the coefficient of $\cos\omega_1$ affects the width of the fan on the ω_1 direction. Similar effects can be obtained in (5.20).

5.2.2 Linear-Phase Designs

We consider here symmetric FIR filters of odd degree, i.e., the simplest (and most common) case of linear-phase filters. With no loss of generality, we consider the zero-phase filter

$$H(\boldsymbol{z}) = \sum_{\boldsymbol{k}=-\boldsymbol{n}}^{\boldsymbol{n}} h_{\boldsymbol{k}} \boldsymbol{z}^{-\boldsymbol{k}}, \quad h_{-\boldsymbol{k}} = h_{\boldsymbol{k}}, \tag{5.23}$$

with real coefficients. This is a symmetric trigonometric polynomial and $H(\boldsymbol{\omega})$ is real. Denoting $\mathcal{D}_p$, $\mathcal{D}_s$, and $\mathcal{D}_t$ the passband, stopband, and transition bands, respectively, the minimax design problem is

$$\begin{array}{ll} \min\limits_{\gamma_s, H} & \gamma_s \\ \text{s.t.} & H(\boldsymbol{\omega}) - 1 \le \gamma_p, \ \forall \boldsymbol{\omega} \in \mathcal{D}_p \cup \mathcal{D}_t \\ & 1 - H(\boldsymbol{\omega}) \le \gamma_p, \ \forall \boldsymbol{\omega} \in \mathcal{D}_p \\ & |H(\boldsymbol{\omega})| \le \gamma_s, \ \forall \boldsymbol{\omega} \in \mathcal{D}_s \end{array} \tag{5.24}$$

where γ_p is the given passband error bound. Note that, as in the 1D case, we bound the frequency response in the transition band; similarly, we can replace the first constraint, posed on $\mathcal{D}_p \cup \mathcal{D}_t$, with a global one; in the 2D case, the saving in complexity is clearly greater, since global positivity is expressed by a single matrix (or two for the Gram-pair parameterization), while positivity on a domain by at least two matrices (or two pairs), often more. Using this remark and emphasizing the presence of nonnegative polynomials, the problem (5.24) becomes

$$
\begin{array}{ll}
\min\limits_{\gamma_s, H} & \gamma_s \\
\text{s.t.} & 1 + \gamma_p - H(\boldsymbol{\omega}) \geq 0, \ \forall \boldsymbol{\omega} \\
& H(\boldsymbol{\omega}) - 1 + \gamma_p \geq 0, \ \forall \boldsymbol{\omega} \in \mathcal{D}_p \\
& \gamma_s - H(\boldsymbol{\omega}) \geq 0, \ \forall \boldsymbol{\omega} \in \mathcal{D}_s \\
& H(\boldsymbol{\omega}) - \gamma_s \geq 0, \ \forall \boldsymbol{\omega} \in \mathcal{D}_s
\end{array}
\tag{5.25}
$$

We assume that the passband $\mathcal{D}_p$ and the stopband $\mathcal{D}_s$ are unions of domains defined by the positivity of some trigonometric polynomials. Thus, the passband has the form

$$
\mathcal{D}_p = \bigcup_{i=1}^{d_p} \mathcal{D}_{pi}, \tag{5.26}
$$

where $\mathcal{D}_{pi}$ are defined as in (4.13) by L_{pi} polynomials. The total number of polynomials necessary to define the passband is

$$
L_p = \sum_{i=1}^{d_p} L_{pi}. \tag{5.27}
$$

For the stopband, we use the same notations with the index s instead of p. For example, for the diamond passband defined by (5.21), there is a single domain in the union and so $d_p = 1$. The domain is defined by $L_p = L_{p1} = 3$ polynomials. A corresponding stopband is the complementary of a domain (5.21) and is a union having the form (4.23). It follows that $d_s = 3$, $L_{s1} = L_{s2} = L_{s3} = 1$, and $L_s = 3$.

Using Theorem 4.11 and its LMI equivalents from Sect. 4.2, and also the notations (5.18), (5.26), (5.27) the design problem (5.25) can be relaxed to the SDP form

$$
\begin{array}{ll}
\min\limits_{\substack{\gamma_s, \boldsymbol{h}, \boldsymbol{Q}, \\ \tilde{\boldsymbol{Q}}_{\ldots}, \hat{\boldsymbol{Q}}_{\ldots}, \check{\boldsymbol{Q}}_{\ldots}}} & \gamma_s \\
\text{s.t.} & \left.\begin{array}{l}
(1 + \gamma_p)\delta_{\boldsymbol{k}} - h_{\boldsymbol{k}} = \mathcal{L}_{\boldsymbol{k}}(\boldsymbol{Q}) \\
h_{\boldsymbol{k}} - (1 - \gamma_p)\delta_{\boldsymbol{k}} = \mathcal{L}_{\boldsymbol{k}, \mathcal{D}_{pi}}(\tilde{\boldsymbol{Q}}_{i,0}, \ldots, \tilde{\boldsymbol{Q}}_{i,L_{pi}}), \ i = 1 : d_p \\
\gamma_s \delta_{\boldsymbol{k}} - h_{\boldsymbol{k}} = \mathcal{L}_{\boldsymbol{k}, \mathcal{D}_{si}}(\hat{\boldsymbol{Q}}_{i,0}, \ldots, \hat{\boldsymbol{Q}}_{i,L_{si}}), \ i = 1 : d_s \\
h_{\boldsymbol{k}} - \gamma_s \delta_{\boldsymbol{k}} = \mathcal{L}_{\boldsymbol{k}, \mathcal{D}_{si}}(\check{\boldsymbol{Q}}_{i,0}, \ldots, \check{\boldsymbol{Q}}_{i,L_{si}}), \ i = 1 : d_s
\end{array}\right\} \boldsymbol{k} \in \mathcal{H} \\
& \boldsymbol{Q} \succeq 0, \ \tilde{\boldsymbol{Q}}_{\ldots} \succeq 0, \ \hat{\boldsymbol{Q}}_{\ldots} \succeq 0, \ \check{\boldsymbol{Q}}_{\ldots} \succeq 0
\end{array}
\tag{5.28}
$$

where $\mathcal{H}$ is a half plane and the notation e.g., $\tilde{\boldsymbol{Q}}_{\ldots}$ covers all possible indices (i, ℓ), with $i = 1 : d_p$ and $\ell = 1 : L_{pi}$. The number of positive semidefinite parameter matrices in (5.28) is $1 + L_p + d_p + 2(L_s + d_s)$ (this number must be doubled if the Gram-pair parameterization is used). Although the SDP problem (5.28) looks cumbersome, the program implementing it has (only) about 250 lines. The complexity of the problem depends on the degree of the relaxation. As discussed in Sect. 3.5, we can use a degree $\boldsymbol{m} \geq \boldsymbol{n}$ for the sum-of-squares that parameterize the positive polynomials from (5.25) and dictate the size of the Gram matrices from (5.28); for

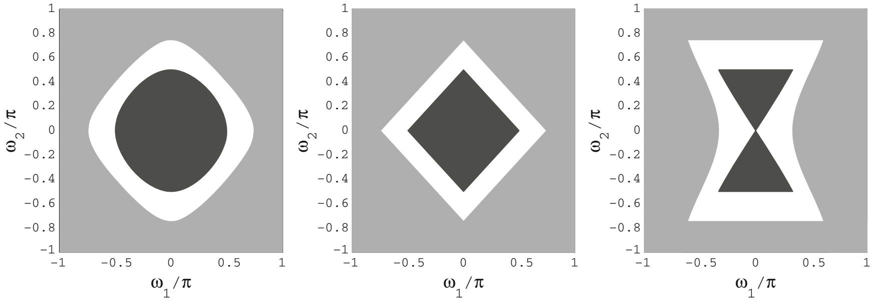


Fig. 5.8 Passbands (*black*) and stopbands (*gray*) for 2D filter design

the specific case of polynomials that are positive on domains, see relation (4.18). The typical question that we have to answer is: how far from optimality are the results of (5.28) if $\boldsymbol{m}$ is equal to $\boldsymbol{n}$ or only slightly larger? Since there is no useful theoretical answer, we search one through the design examples below.

In Fig. 5.8, we present the passband (in black) and stopband (in gray) frequency domains for three linear-phase FIR filters: a simple lowpass, a diamond lowpass, and a fan. The filters are designed by solving the SDP problem (5.28), with the specifications listed below. The passband error bound is $\gamma_p = 0.05$ (corresponding to a passband ripple of about 0.87 dB) for the first example and $\gamma_p = 0.1$ (1.74 dB ripple) for the other two. The size of the filters is 15 × 15, i.e., $\boldsymbol{n} = (7, 7)$ in (5.23). The degree of the relaxation is $\boldsymbol{m} = \boldsymbol{n}$ if not otherwise stated.

Example 5.4 The passband and the stopband are defined as in (5.20), by

$$\begin{aligned}\mathcal{D}_p &= \{\omega_1, \omega_2 \in [-\pi, \pi] \mid \cos\omega_1 + \cos\omega_2 - 1 \ge 0\},\\ \mathcal{D}_s &= \{\omega_1, \omega_2 \in [-\pi, \pi] \mid -\cos\omega_1 - \cos\omega_2 + 0.3 \ge 0\}.\end{aligned} \tag{5.29}$$

This is the simplest possible case, as each band is described by a single polynomial. The frequency response of the filter is shown in Fig. 5.9. The design time is about 40 s. The optimal value of the stopband error reported by the SDP program is $\gamma_s = 0.012496 = -38.06$ dB.

To evaluate the effect of the relaxation degree, we solve the SDP problem for values $\boldsymbol{m} > \boldsymbol{n}$. The optimal values of the stopband error γ_s are shown in Table 5.1. We see that taking $\boldsymbol{m} = \boldsymbol{n} + (1, 1) = (8, 8)$ improves the error to $\gamma_s = 0.012303 = -38.20$ dB, but further increase of the degree has almost no effect. The frequency response obtained with $\boldsymbol{m} = (8, 8)$ is given in Fig. 5.10; the equiripple character is more evident than in Fig. 5.9, where there are some small irregularities in the high-frequency area.

The largest filter we have designed has size 25 × 25; the design time was about 15 min and the optimal stopband error $\gamma_s = 1.624 \cdot 10^{-4} = -75.79$ dB. ■

Example 5.5 For the diamond lowpass filter, the frequency bands have the description

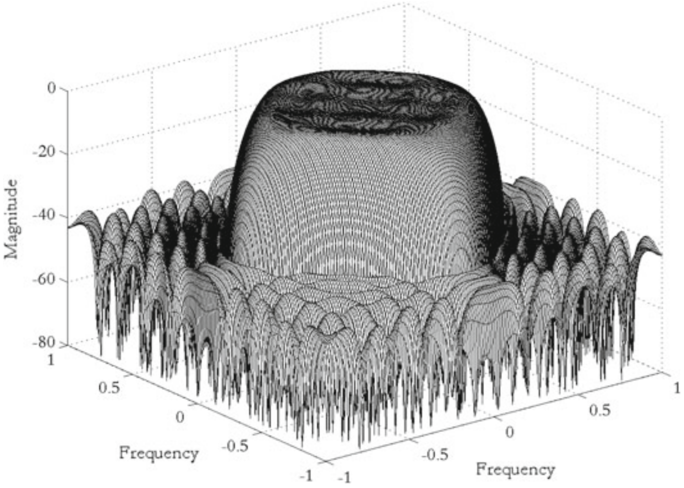


Fig. 5.9 Magnitude response of the filter from Example 5.4, $\boldsymbol{m} = \boldsymbol{n} = (7, 7)$

Table 5.1 Optimal values of γ_s in Examples 5.4–5.6

$\boldsymbol{m} - \boldsymbol{n}$	(0,0)	(1,1)	(2,2)	(3,3)
Example 5.4	0.012496	0.012303	0.012297	0.012297
Example 5.5	0.020174	0.020174	0.020174	-
Example 5.6	0.020837	0.020837	0.020837	-

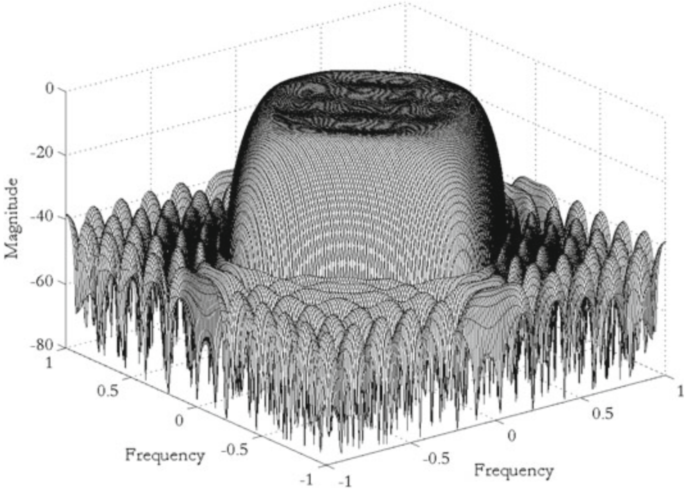


Fig. 5.10 Magnitude response of the filter from Example 5.4, $\boldsymbol{m} = \boldsymbol{n} + 1 = (8, 8)$

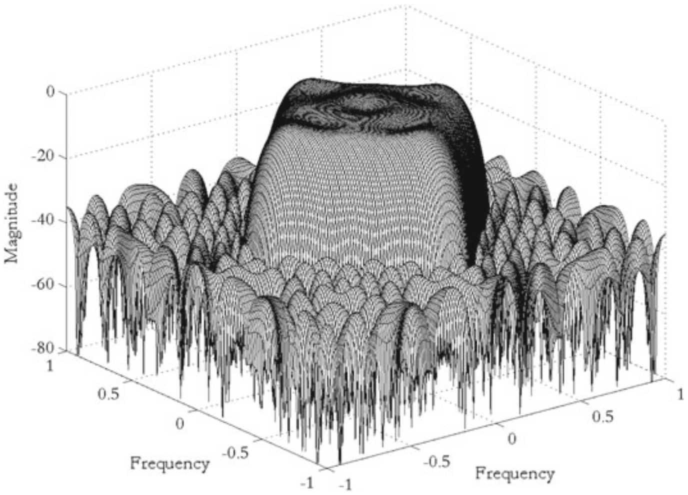


Fig. 5.11 Magnitude response of the filter from Example 5.5

$$
\begin{aligned}
\mathcal{D}_p &= \{\omega_{1,2} \mid \cos(\omega_1+\omega_2) - 0.1 \geq 0,\ \cos(\omega_1-\omega_2) - 0.1 \geq 0, \\
&\qquad\qquad \cos\omega_1 + \cos\omega_2 \geq 0\}, \\
\mathcal{D}_s &= \{\omega_{1,2} \mid -\cos(\omega_1+\omega_2) - 0.7 \geq 0\} \\
&\quad \cup \{\omega_{1,2} \mid -\cos(\omega_1-\omega_2) - 0.7 \geq 0\} \\
&\quad \cup \{\omega_{1,2} \mid -\cos\omega_1 - \cos\omega_2 \geq 0\}.
\end{aligned}
\tag{5.30}
$$

Now, the passband is defined by the positivity of three polynomials and the stopband by the union of three simple domains; this will increase the complexity of the SDP problem since there are more parameter matrices. Indeed, the design time is about 140 s, significantly greater than for Example 5.4. The frequency response of the optimal filter is shown in Fig. 5.11. The optimal stopband error is $\gamma_s = 0.020174 = -33.9\,\text{dB}$. As shown in Table 5.1, increasing the relaxation degree does not change the optimal stopband error; actually, there are changes, but only in the seventh significant digit and thus negligible. ■

Example 5.6 The passband and the stopband of the fan filter are defined by

$$
\begin{aligned}
\mathcal{D}_p &= \{\omega_{1,2} \mid 2\cos\omega_1 - \cos\omega_2 - 1 \geq 0,\ \cos\omega_2 \geq 0\}, \\
\mathcal{D}_s &= \{\omega_{1,2} \mid -2\cos\omega_1 + \cos\omega_2 \geq 0\} \cup \{\omega_{1,2} \mid -\cos\omega_2 - 0.7 \geq 0\}.
\end{aligned}
\tag{5.31}
$$

The design time is about 60 s, greater than for Example 5.4, but smaller than for Example 5.5; this would have been easy to forecast, due to the complexity of (5.31), compared with that of (5.29) or (5.30). The optimal stopband error is $\gamma_s = 0.020837 = -33.62\,\text{dB}$; again, this value does not change by increasing the degree of the relaxation. The frequency response of the filter is shown in Fig. 5.12. ■

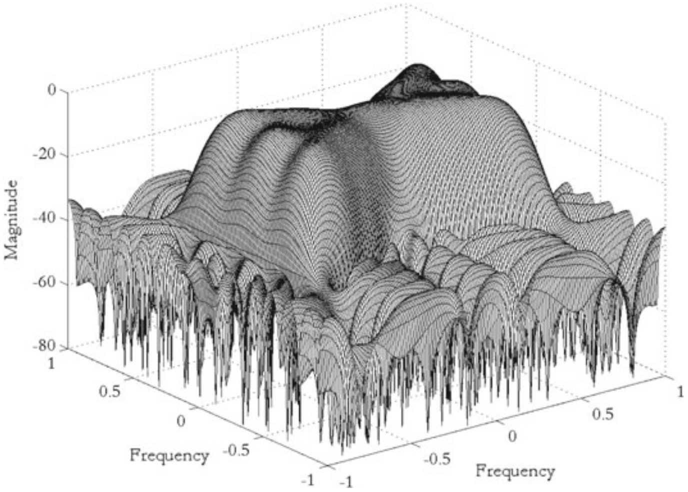


Fig. 5.12 Magnitude response of the filter from Example 5.6

Looking again at Table 5.1, we conclude that there is little departure from optimality even with $\boldsymbol{m} = \boldsymbol{n}$; we conjecture that for all practical purposes we can safely take $\boldsymbol{m} \leq \boldsymbol{n} + 1$ to obtain the optimal filter. Thus, we can extend to FIR filters and larger degrees, the practical remarks made in previous chapters for the simpler problem of computing the minimum value of a polynomial.

5.2.3 Approximate Linear-Phase Designs

We consider now positive orthant FIR filters (3.5). Given a desired group delay $\boldsymbol{\tau}$, we optimize the FIR filter as follows

$$\begin{aligned} \min_{\gamma_s, H} \ & \gamma_s \\ \text{s.t.} \ & |H(\boldsymbol{\omega}) - \mathrm{e}(-j\boldsymbol{\tau}^T\boldsymbol{\omega})| \leq \gamma_p, \ \forall \boldsymbol{\omega} \in \mathcal{D}_p \\ & |H(\boldsymbol{\omega})| \leq \gamma_s, \ \forall \boldsymbol{\omega} \in \mathcal{D}_s \end{aligned} \tag{5.32}$$

As in Sect. 5.1.3, we can use BRL results for transforming the above problem into an SDP one, this time obtaining only a relaxation. The group delay should have integer elements, i.e., $\boldsymbol{\tau} \in \mathbb{N}^d$. We denote $\boldsymbol{h}$ the vector containing the coefficients of $H(z)$ and $\boldsymbol{e}_{\boldsymbol{\tau}}$ the unit vector containing the coefficients of $z^{-\boldsymbol{\tau}}$; obviously, the same ordering of coefficients is adopted for both vectors. The SDP relaxation of (5.32) is

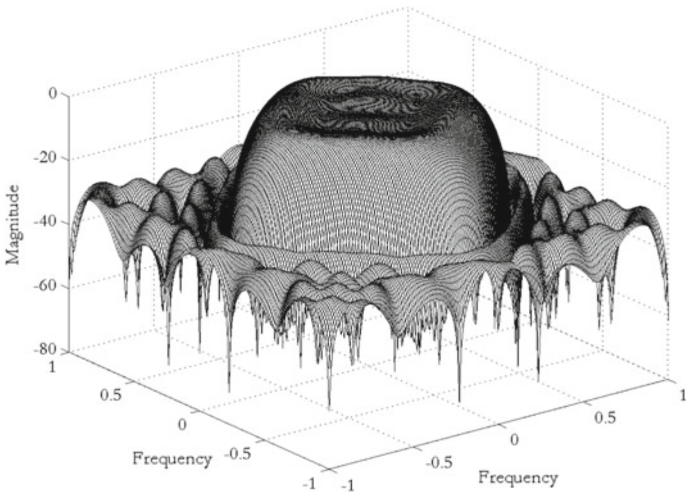


Fig. 5.13 Magnitude response of the filter from Example 5.7

$$
\begin{array}{ll}
\min\limits_{\boldsymbol{h},\gamma_s,\boldsymbol{Q}_{\dots},\tilde{\boldsymbol{Q}}_{\dots}} & \gamma_s \\
\text{s.t.} & \left.\begin{array}{l}
\left.\begin{array}{l}
\gamma_p^2 \delta_{\boldsymbol{k}} = \mathcal{L}_{\boldsymbol{k},\mathcal{D}_{pi}}(\boldsymbol{Q}_{i,0},\dots,\boldsymbol{Q}_{i,L_{pi}}) \\
\begin{bmatrix} \boldsymbol{Q}_{i,0} & \boldsymbol{h}-\boldsymbol{e}_{\boldsymbol{\tau}} \\ \boldsymbol{h}^T-\boldsymbol{e}_{\boldsymbol{\tau}}^T & 1 \end{bmatrix} \succeq 0
\end{array}\right\} i = 1:d_p \\
\left.\begin{array}{l}
\gamma_s^2 \delta_{\boldsymbol{k}} = \mathcal{L}_{\boldsymbol{k},\mathcal{D}_{si}}(\tilde{\boldsymbol{Q}}_{i,0},\dots,\tilde{\boldsymbol{Q}}_{i,L_{si}}) \\
\begin{bmatrix} \tilde{\boldsymbol{Q}}_{i,0} & \boldsymbol{h} \\ \boldsymbol{h}^T & 1 \end{bmatrix} \succeq 0
\end{array}\right\} i = 1:d_s
\end{array}\right\} \boldsymbol{k} \in \mathcal{H} \\
& \boldsymbol{Q}_{\dots} \succeq 0,\ \tilde{\boldsymbol{Q}}_{\dots} \succeq 0
\end{array}
\tag{5.33}
$$

Example 5.7 The stopband and passband are defined as in (5.29), i.e., identically to those in Example 5.4. We take $\boldsymbol{n} = (10, 10)$ (and so the filter size is 11×11), $\boldsymbol{\tau} = (4, 4)$ and $\gamma_p = 0.1$. The SDP problem (5.33) is solved in about 150 s, giving $\gamma_s = 0.0367562 = -28.69$ dB and the filter whose frequency response is shown in Fig. 5.13. Solving the problem with $\boldsymbol{m} = \boldsymbol{n} + 1$, we obtain $\gamma_s = 0.0367560$, i.e., virtually the same value; for other design specifications, we have noted a similar behavior, upholding the conclusion that lowest degree relaxations give practically optimal results. ■

5.3 FIR Deconvolution

Several signal processing problems, such as channel equalization, deconvolution, system inversion, can be modeled by the structure shown in Fig. 5.14. The signal $\boldsymbol{s}$

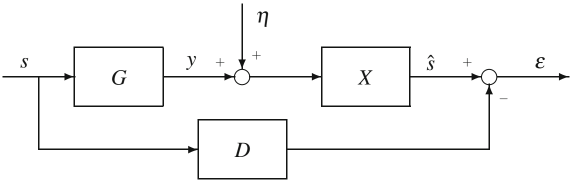


Fig. 5.14 Channel equalization scheme

passes through the channel $\boldsymbol{G}(z)$ and is contaminated by the noise $\boldsymbol{\eta}$. Our aim is to design a filter $\boldsymbol{X}(z)$ such that its output $\hat{\boldsymbol{s}}$ is an approximation of the ideal output $\boldsymbol{Ds}$. The filters $\boldsymbol{G}(z)$ and $\boldsymbol{D}(z)$ are given and very often $\boldsymbol{D}(z)$ is a simple delay; in this case, we actually compute an approximate inverse of $\boldsymbol{G}(z)$. For the generality of presentation, we assume that all filters have p inputs and p outputs, i.e., we work with (square) MIMO systems. We also assume that all systems are FIR; this is very often the case for the given systems (e.g., typical channels have a short impulse response); the FIR choice for $\boldsymbol{X}(z)$ is preferred for the ease and robustness of implementation; it also allows the safe computation of the optimal solution, as we will see.

The output error in Fig. 5.14 is

$$\boldsymbol{\varepsilon} = \hat{\boldsymbol{s}} - \boldsymbol{Ds} = \left(\boldsymbol{X}[\boldsymbol{G}\ \boldsymbol{I}_p] - [\boldsymbol{D}\ \boldsymbol{0}_p]\right)\begin{bmatrix}\boldsymbol{s}\\ \boldsymbol{\eta}\end{bmatrix} \overset{\Delta}{=} \boldsymbol{H}\begin{bmatrix}\boldsymbol{s}\\ \boldsymbol{\eta}\end{bmatrix}. \tag{5.34}$$

The error function $\boldsymbol{H}(z)$ has the general form

$$\boldsymbol{H}(z) = \boldsymbol{X}(z)\boldsymbol{A}(z) - \boldsymbol{B}(z), \tag{5.35}$$

where $\boldsymbol{A}(z)$, $\boldsymbol{B}(z)$ are given. We note that the coefficients of $\boldsymbol{H}(z)$ depend linearly on those of $\boldsymbol{X}(z)$. Without a priori information on the signal $\boldsymbol{s}$ and the noise $\boldsymbol{\eta}$, the error (5.34) is controlled by minimizing a norm of the error function $\boldsymbol{H}(z)$. The most used norms are the H_2 norm

$$\|\boldsymbol{H}(z)\|_2 = \left(\frac{1}{2\pi}\int_{-\pi}^{\pi} \mathrm{tr}[\boldsymbol{H}(\omega)^H \boldsymbol{H}(\omega)]d\omega\right)^{1/2} \tag{5.36}$$

and the H_∞ norm (4.46) which can also be interpreted as

$$\|\boldsymbol{H}(z)\|_\infty = \sup_{\|\boldsymbol{\xi}\|_2=1} \|\boldsymbol{H\xi}\|_2. \tag{5.37}$$

In (5.37), $\boldsymbol{\xi}$ is an input signal and $\boldsymbol{H\xi}$ is the corresponding output (error) signal. A minimum $\|\boldsymbol{H}(z)\|_2$ means minimum energy of the error signal, while a minimum $\|\boldsymbol{H}(z)\|_\infty$ means a smallest largest magnitude of the error over the whole frequency spectrum.

If $\boldsymbol{H}(z)$ is a FIR system (4.44) with m inputs and p outputs and coefficients $\boldsymbol{H}_k \in \mathbb{R}^{p\times m}$, then its H_2 norm is

$$\|\boldsymbol{H}(z)\|_2^2 = \sum_{k=0}^{n} \|\boldsymbol{H}_k\|_F^2 = \sum_{k=0}^{n}\sum_{i=0}^{p-1}\sum_{\ell=0}^{m-1}(\boldsymbol{H}_k)_{i\ell}^2 = \|\boldsymbol{h}\|_2^2, \tag{5.38}$$

where $\boldsymbol{h}$ is a vector of size $mp(n+1)$ containing the coefficients of $\boldsymbol{H}(z)$. If $\boldsymbol{H}(z)$ has the form (5.35) and $\boldsymbol{x}$ is a vector containing the coefficients of the FIR system $\boldsymbol{X}(z)$, then we have

$$\|\boldsymbol{H}(z)\|_2 = \|\boldsymbol{C}\boldsymbol{x} - \boldsymbol{f}\|, \tag{5.39}$$

where the constant matrix $\boldsymbol{C}$ and vector $\boldsymbol{f}$ depend (linearly) on the coefficients of $\boldsymbol{A}(z)$ and $\boldsymbol{B}(z)$. More precisely, using notations similar to (4.44) for $\boldsymbol{A}(z)$ and $\boldsymbol{B}(z)$, the equality (5.35) means that

$$\boldsymbol{H}_k = \sum_i \boldsymbol{X}_i \boldsymbol{A}_{k-i} - \boldsymbol{B}_k. \tag{5.40}$$

Using the matrix equality $\text{vec}(\boldsymbol{U}\boldsymbol{V}\boldsymbol{W}) = (\boldsymbol{W}^T \otimes \boldsymbol{U})\text{vec}\boldsymbol{V}$, the relation (5.40) becomes

$$\text{vec}(\boldsymbol{H}_k) = \sum_i (\boldsymbol{A}_{k-i}^T \otimes \boldsymbol{I}_p)\text{vec}(\boldsymbol{X}_i) - \text{vec}(\boldsymbol{B}_k), \tag{5.41}$$

which gives $\boldsymbol{h} = \boldsymbol{C}\boldsymbol{x} - \boldsymbol{f}$ as in (5.39).

The H_∞ norm can be characterized by the BRL Theorem 4.32, with coefficients stacked as in (4.47). Note that $\bar{\boldsymbol{H}} \in \mathbb{R}^{p(n+1)\times m}$.

5.3.1 Basic Optimization Problem

A general way of designing the system $\boldsymbol{X}(z)$ from Fig. 5.14 is based on the mixed H_2/H_∞ optimization

$$\begin{aligned} \min_{\boldsymbol{X}} \ & \|\boldsymbol{H}(z)\|_2 \\ \text{s.t.} \ & \|\boldsymbol{H}(z)\|_\infty \le \gamma \end{aligned} \tag{5.42}$$

where γ is a given error bound and $\boldsymbol{H}(z)$ is defined in (5.35). Denoting $\bar{\boldsymbol{H}} = \mathcal{L}(\boldsymbol{x})$ the linear dependence between the block vector (4.47) and the coefficients of $\boldsymbol{X}(z)$, and using the relations derived above, the problem (5.42) can be expressed in the SDP form

$$\begin{aligned} \min_{\varepsilon,\boldsymbol{x},\boldsymbol{Q}} \ & \varepsilon \\ \text{s.t.} \ & \gamma^2\delta_k \boldsymbol{I}_p = \text{TR}[\boldsymbol{\Theta}_{pk}\boldsymbol{Q}], \ k = 0:n \\ & \begin{bmatrix} \boldsymbol{Q} & \mathcal{L}(\boldsymbol{x}) \\ \mathcal{L}(\boldsymbol{x})^T & \boldsymbol{I}_m \end{bmatrix} \succeq 0 \\ & \|\boldsymbol{C}\boldsymbol{x} - \boldsymbol{f}\| \le \varepsilon \end{aligned} \tag{5.43}$$

Table 5.2 H_2 and H_∞ norms for the error system from Example 5.8

$\lVert \boldsymbol{H}(z) \rVert_2$	0.4066	0.3651	0.3505	0.3435	0.3348
$\lVert \boldsymbol{H}(z) \rVert_\infty$	0.4471	0.45	0.46	0.47	0.5134

(Compare with Theorem 4.32, with an inequality that holds globally, and note that $\boldsymbol{\Theta}_k \otimes \boldsymbol{I}_p = \boldsymbol{\Theta}_{pk}$.)

Example 5.8 The given channel model from Fig. 5.14 is

$$G(z) = 1 + 0.33562z^{-1} + 4.627z^{-2} - 0.14487z^{-3} + 1.6837z^{-4} \tag{5.44}$$

and the desired model is the delay $D(z) = z^{-2}$ (this is the first example from [1]). We take deg $X = 6$. In Table 5.2 we present the H_2 and H_∞ norms of the error system (of order 10), obtained by solving (5.43) for several values of the bound γ. The second column of the table corresponds to the H_∞ solution, obtained by a simple variation of (5.43); the last column gives the values for the H_2 solution (obtained via a simple quadratic optimization or by using a large value of γ in (5.43)). The behavior is typical and shows that the mixed H_2/H_∞ optimization offers compromise solutions that are not far from the minimum values obtained when a single error norm is optimized. ■

5.3.2 *Deconvolution of Periodic FIR Filters*

In the previous example, the systems $\boldsymbol{G}(z)$, $\boldsymbol{X}(z)$ were single-input single-output (SISO) and only the error system $\boldsymbol{H}(z)$ had size 1×2. Here, we give an example leading to MIMO systems. Let the channel model be the *periodic* SISO filter whose input–output behavior is given by

$$y(t) = \sum_{k=0}^{\nu} g_{k,t} s(t-k). \tag{5.45}$$

The period is p, which means that $g_{k,t} = g_{k,t+ip}$ for any integer i. We consider that in Fig. 5.14 the system $X(z)$ is also periodic. Formulated in terms of periodic filters, the problem (5.42) is nonlinear. To make it linear, but MIMO, the lifting technique can be used. All scalar signals are transformed in vector signals of size p, by grouping blocks of p successive samples, e.g.,

$$\boldsymbol{s}(t) = [s(tp) \;\; s(tp+1) \;\; \ldots \;\; s(tp+p-1)]^T.$$

Denoting

$$G_\ell(z) = \sum_{k=0}^{\nu} g_{k,\ell} z^{-k}, \ \ell = 0 : p-1,$$

and introducing the polyphase decomposition

$$G_\ell(z) = \sum_{i=0}^{p-1} z^{-i} G_{\ell,i}(z^p),$$

the input–output relation (5.45) can be modeled at block level as $\boldsymbol{y}(z) = \boldsymbol{G}(z)\boldsymbol{s}(z)$, with the transfer matrix

$$\boldsymbol{G}(z) = \begin{bmatrix} G_{0,0}(z) & G_{0,1}(z) & \ldots & G_{0,p-1}(z) \\ z^{-1}G_{1,p-1}(z) & G_{1,0}(z) & \ldots & G_{1,p-2}(z) \\ \vdots & \vdots & & \vdots \\ z^{-1}G_{p-1,1}(z) & z^{-1}G_{p-1,2}(z) & \ldots & G_{p-1,0}(z) \end{bmatrix} \overset{\Delta}{=} \sum_{n=0}^{N_g} \boldsymbol{G}_n z^{-n}. \quad (5.46)$$

In (5.46), the degree of the FIR MIMO system is $N_g = \lceil \nu/p \rceil$. The matrix $\boldsymbol{G}_0$ is upper triangular; also, some elements of $\boldsymbol{G}_{N_g}$ are zero and possibly some of $\boldsymbol{G}_{N_g-1}$. For example, if $\nu = 4$, $p = 3$, then we have

$$\boldsymbol{G}_0 = \begin{bmatrix} g_{0,0} & g_{1,0} & g_{2,0} \\ 0 & g_{0,1} & g_{1,1} \\ 0 & 0 & g_{0,2} \end{bmatrix}, \ \boldsymbol{G}_1 = \begin{bmatrix} g_{3,0} & g_{4,0} & 0 \\ g_{2,1} & g_{3,1} & g_{4,1} \\ g_{1,2} & g_{2,2} & g_{3,2} \end{bmatrix}, \ \boldsymbol{G}_2 = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ g_{4,2} & 0 & 0 \end{bmatrix}.$$

A similar lifting is valid for the FIR periodic inverse filter X, whose degree is μ, possibly different from ν. The degree of the MIMO system is $N_x = \lceil \mu/p \rceil$.

With vector signals, the deconvolution scheme is identical to that from Fig. 5.14. The relation between the norms of the periodic error system H and lifted error $\boldsymbol{H}(z)$ is

$$\|H\|_2^2 = \frac{1}{p}\|\boldsymbol{H}(z)\|_2^2, \quad \|H\|_\infty = \|\boldsymbol{H}(z)\|_\infty. \quad (5.47)$$

The first equality is the (natural) definition of the H_2 norm of a periodic FIR filter, while the second comes from (5.37) after remarking that the norms of the input and output signals are not affected by blocking. So, we can design a FIR periodic filter X by solving an optimization problem (5.43), subject to the additional constraints imposed by the structure of zeros in the coefficients of $\boldsymbol{X}(z)$, which can be cast more generally as a linear equality $\tilde{\boldsymbol{C}}\boldsymbol{x} = \tilde{\boldsymbol{f}}$ (in our case $\tilde{\boldsymbol{f}} = 0$). Adding a linear constraint to (5.43) does not change the SDP nature of the problem and has little influence on the complexity.

Example 5.9 Let us consider the toy example from [2], where $p = 2$, $\nu = 3$ and

$$H_0(z) = 5 + z^{-1} + 2z^{-2} - z^{-3},$$
$$H_1(z) = 3 + 2z^{-1} - 2z^{-2} + z^{-3}.$$

We seek a FIR periodic equalizer $X(z)$ of degree $\mu = 3$; the desired system is $D(z) = 1$, i.e., the delay is 0. For a bound $\gamma = 0.74$ of the H_∞ norm, the solution of (5.43) (with the extra linear equalities mentioned above) is

$$\begin{aligned} X_0(z) &= 0.1892 - 0.0571z^{-1} - 0.0464z^{-2} + 0.0396z^{-3}, \\ X_1(z) &= 0.2663 - 0.1111z^{-1} + 0.1439z^{-2} - 0.0594z^{-3}. \end{aligned}$$

The error H_2 norm is $\|E\|_2 = 0.3932$. Varying γ, a behavior similar to that described in Example 5.8 is obtained. ■

5.3.3 Robust H_∞ Deconvolution

Let us assume that the channel model $\boldsymbol{G}(z)$ from Fig. 5.14 is not known exactly. However, we know that its coefficients belong to a polytope with given vertices $\boldsymbol{G}_i(z)$, $i = 1 : I$, i.e.,

$$\boldsymbol{G}(z) = \sum_{i=1}^{I} \lambda_i \boldsymbol{G}_i(z), \qquad \sum_{i=1}^{I} \lambda_i = 1,\ \lambda_i \geq 0, \tag{5.48}$$

where the parameters λ_i of the convex combination are not known. We denote $\boldsymbol{A}_i(z)$ the systems that appear in (5.35) and correspond to the vertices $\boldsymbol{G}_i(z)$ and note that $\boldsymbol{A}(z)$ belongs to the polytope $\mathcal{A}$ with vertices $\boldsymbol{A}_i(z)$. Our aim is to design a filter $\boldsymbol{X}(z)$ that minimizes the H_∞ norm of the error in the worst case (over all the models in the polytope), i.e.,

$$\begin{aligned} &\min_{\boldsymbol{X}}\ \|\boldsymbol{X}(z)\boldsymbol{A}(z) - \boldsymbol{B}(z)\|_\infty \\ &\text{s.t.}\ \ \boldsymbol{A}(z) \in \mathcal{A} \end{aligned} \tag{5.49}$$

However, it is enough to minimize the H_∞ norm for the vertices, i.e., to solve

$$\begin{aligned} &\min_{\gamma,\boldsymbol{X}}\ \gamma \\ &\text{s.t.}\ \ \|\boldsymbol{X}(z)\boldsymbol{A}_i(z) - \boldsymbol{B}(z)\|_\infty \leq \gamma,\ \ i = 1 : I \end{aligned} \tag{5.50}$$

Indeed, any solution of (5.50) respects

$$\|\boldsymbol{X}(z)\sum_{i=1}^{I}\lambda_i \boldsymbol{A}_i(z) - \boldsymbol{B}(z)\|_\infty \leq \sum_{i=1}^{I}\lambda_i \|\boldsymbol{X}(z)\boldsymbol{A}_i(z) - \boldsymbol{B}(z)\|_\infty \leq \gamma$$

and so is a solution of (5.49). (The inverse implication is trivial.) Using Theorem 4.32, the problem (5.50) has the SDP form

$$\begin{array}{ll} \min\limits_{\gamma, \boldsymbol{x}, \boldsymbol{Q}_1, \ldots, \boldsymbol{Q}_I} & \gamma^2 \\ \text{s.t.} & \left. \begin{array}{l} \gamma^2 \delta_k \boldsymbol{I}_p = \operatorname{TR}[\boldsymbol{\Theta}_{pk} \boldsymbol{Q}_i], \ k = 0 : n \\ \begin{bmatrix} \boldsymbol{Q}_i & \mathcal{L}_i(\boldsymbol{x}) \\ \mathcal{L}_i(\boldsymbol{x})^T & \boldsymbol{I}_m \end{bmatrix} \succeq 0 \end{array} \right\} i = 1 : I \end{array} \tag{5.51}$$

where $\mathcal{L}_i(\boldsymbol{x})$ describes the linear function of the coefficients of $\boldsymbol{X}(z)$ that produces the block vector (4.47) corresponding to the "vertex" error system $\boldsymbol{X}(z)\boldsymbol{A}_i(z) - \boldsymbol{B}(z)$. Note that we use γ^2 as variable, in order to preserve linearity.

Example 5.10 Let us reconsider the model from Example 5.8 and assume that it is

$$G(g_1, g_2, z) = 1 + g_1 z^{-1} + g_2 z^{-2} - 0.14487 z^{-3} + 1.6837 z^{-4},$$

where $0.3 \leq g_1 \leq 0.4$ and $4.5 \leq g_2 \leq 4.8$. So, there are two coefficients taking unknown values inside a rectangle, which is our polytope (in a five-dimensional space); note that the model (5.44) is a point inside the polytope. The polynomials corresponding to the $I = 4$ corners (vertices) of the rectangle are $G_1(z) = G(0.3, 4.5, z)$, $G_2(z) = G(0.3, 4.8, z)$, $G_3(z) = G(0.4, 4.5, z)$, $G_4(z) = G(0.4, 4.8, z)$. Solving (5.51) with these data and $\deg X = 6$, we obtain $\gamma = 0.4722$ and

$$\begin{aligned} X(z) = 0.2273 &+ 0.0487 z^{-1} - 0.1079 z^{-2} - 0.0233 z^{-3} \\ &+ 0.0356 z^{-4} + 0.0246 z^{-5} - 0.0247 z^{-6}. \end{aligned}$$

If, instead of this filter, we use the H_∞ solution

$$\begin{aligned} X(z) = 0.2309 &+ 0.0452 z^{-1} - 0.1014 z^{-2} - 0.0273 z^{-3} \\ &+ 0.0320 z^{-4} + 0.0259 z^{-5} - 0.0216 z^{-6} \end{aligned}$$

from Example 5.8, which minimizes the H_∞ norm only for (5.44), not for the entire polytope, then the worst error norm in a corner of the rectangle is 0.4811, i.e., a larger value than for the robust filter. ■

5.3.4 2D H_∞ Deconvolution

We examine now the case where the systems $\boldsymbol{G}(\boldsymbol{z})$ and $\boldsymbol{X}(\boldsymbol{z})$ from Fig. 5.14 are 2D and FIR (possibly with matrix coefficients), applying the results from Sect. 4.3.2. The error system from the deconvolution problem (5.42) is

$$\boldsymbol{H}(\boldsymbol{z}) = \sum_{k_1=0}^{n_1} \sum_{k_2=0}^{n_2} \boldsymbol{H}_{k_1 k_2} z_1^{-k_1} z_2^{-k_2}, \quad \boldsymbol{H}_{k_1 k_2} \in \mathbb{R}^{p \times m}. \tag{5.52}$$

The BRL given by Theorem 4.32 can be directly applied, taking into account that the particular form of (4.47) for (5.52) is

$$\bar{\boldsymbol{H}} = \begin{bmatrix} \boldsymbol{H}_{00} \\ \vdots \\ \boldsymbol{H}_{n_1 0} \\ \boldsymbol{H}_{01} \\ \vdots \\ \boldsymbol{H}_{n_1 n_2} \end{bmatrix} \in \mathbb{R}^{p(n_1+1)(n_2+1)\times m}. \tag{5.53}$$

We study the H_∞ deconvolution problem

$$\begin{aligned} &\min_{\boldsymbol{X}} \ \gamma \\ &\text{s.t.} \ \ \|\boldsymbol{X}(z)\boldsymbol{A}(z) - \boldsymbol{B}(z)\|_\infty \leq \gamma \end{aligned} \tag{5.54}$$

with $\boldsymbol{A}(z)$ and $\boldsymbol{B}(z)$ depending on the given systems $\boldsymbol{G}$ and $\boldsymbol{D}$ from Fig. 5.14. We denote $\bar{\boldsymbol{H}} = \mathcal{L}(\boldsymbol{X})$ the linear dependence between the block vector (5.53) and the coefficients of $\boldsymbol{X}(z)$. Using Theorem 4.32, the problem (5.54) can be relaxed to the SDP form

$$\begin{aligned} &\min_{\gamma,\boldsymbol{X},\boldsymbol{Q}} \ \gamma^2 \\ &\text{s.t.} \ \ \gamma^2 \delta_{\boldsymbol{k}} \boldsymbol{I}_p = \operatorname{TR}\left[(\boldsymbol{\Theta}_{k_2} \otimes \boldsymbol{\Theta}_{k_1} \otimes \boldsymbol{I}_p)\boldsymbol{Q}\right], \ \boldsymbol{k} \in \mathcal{H} \\ &\qquad \begin{bmatrix} \boldsymbol{Q} & \mathcal{L}(\boldsymbol{X}) \\ \mathcal{L}(\boldsymbol{X})^T & \boldsymbol{I}_m \end{bmatrix} \succeq 0 \end{aligned} \tag{5.55}$$

Example 5.11 Let us consider the 2D channel model [3]

$$\begin{aligned} G(z_1, z_2) &= 0.1(z_1^{-1} + z_2^{-1})^3 + 0.1z_2^{-2} + 0.1z_2^{-1} + 8 \\ &= [1 \ z_1^{-1} \ z_1^{-2} \ z_1^{-3}] \begin{bmatrix} 8 & 0.1 & 0.1 & 0.1 \\ 0 & 0 & 0.3 & 0 \\ 0 & 0.3 & 0 & 0 \\ 0.1 & 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} 1 \\ z_2^{-1} \\ z_2^{-2} \\ z_2^{-3} \end{bmatrix}. \end{aligned} \tag{5.56}$$

We take $D(z) = 1$ and thus the solution of (5.54) is an approximate inverse of $G(z)$. Solving (5.55) with the smallest relaxation degree and several values of $\deg X = (\nu_1, \nu_2)$, we obtain the H_∞ norms shown in Table 5.3. We notice that the minimum error is already attained for a degree $\nu_1 = \nu_2 = 2$; the corresponding solution is

$$X(z_1, z_2) = [1 \ z_1^{-1} \ z_1^{-2}] \begin{bmatrix} 0.12402 & -0.00138 & -0.00150 \\ -0.00006 & -0.00026 & -0.00060 \\ -0.00006 & -0.00603 & 0.00012 \end{bmatrix} \begin{bmatrix} 1 \\ z_2^{-1} \\ z_2^{-2} \end{bmatrix}.$$

Table 5.3 H_∞ norms for the error system from Example 5.11

$\nu_1 \backslash \nu_2$	1	2	3	4
1	0.1604	0.1418	0.1397	0.1397
2	0.1434	0.1379	0.1379	0.1379
3	0.1404	0.1379	0.1379	0.1379
4	0.1404	0.1379	0.1379	0.1379

For comparison, in [3] where a sufficient BRL for 2D systems in the Fornasini–Marchesini model is employed, a value $\gamma = 0.15$ is reported for a degree equal to (3,3); the respective solution is not FIR.

Increasing the relaxation degree, which amounts to adding zero blocks in (5.53) in the appropriate positions, leads to solutions of (5.55) for which the error bound γ is improved only in the seventh or eighth significant digit with respect to the values from Table 5.3. ■

5.4 Bibliographical and Historical Notes

The first exact method for the magnitude design of FIR filters using polynomials positive on an interval and their parameterization was based on the KYP lemma [4, 5]. The method has been adapted to the trace parameterization, as in Sect. 5.1.2, in [6, 7]; in the latter paper, the design of linear-phase filters described in Sect. 5.1.1 is also proposed; the method is more flexible than the classic Remez algorithm. The design of approximately linear phase from Sect. 5.1.3 is a variation on the same theme. An overview of convex optimization methods used in FIR filter design can be found in [8].

The 2D FIR filter design methods described in Sect. 5.2 are taken from [9]. The only other method using positivity and leading to SDP is based on a 2D KYP lemma [10] and gives clearly suboptimal results; it also works only for rectangular frequency domains.

FIR H_∞ deconvolution using the KYP lemma and SDP was proposed in [1]. An algorithm for FIR H_2 equalization can be found in [11]. The design of periodic filters, as in Sect. 5.3.2, using the BRL for FIR MIMO systems, appeared in [12]. The H_2 optimal deconvolution of periodic filters using SDP is discussed in [2].

H_∞ deconvolution of 2D systems in Fornasini-Marchesini model, based on a sufficient condition in the form of an LMI, was proposed in [3]. The (practically) optimal FIR deconvolution from Sect. 5.3.4 has not been published, but is a consequence of the results from Sect. 4.3.

The 2D results from this chapter show that the design of FIR systems using positive trigonometric polynomials is almost always optimal and, in any case, clearly better than sufficient conditions for general systems, applied in particular form to FIR systems.

More applications of the BRL for matrix polynomials can be found in [13].

Hybrid polynomials results can be applied to the design of adjustable FIR filters, see [14].

Problems

P 5.1 A linear-phase symmetric FIR filter (5.1) of odd order $n = 2\tilde{n} + 1$ has a frequency response identical (modulo a phase shift) to

$$\tilde{H}(\omega) = \sum_{k=0}^{\tilde{n}} \tilde{h}_k \cos(k + \tfrac{1}{2})\omega.$$

Denoting $\theta = \omega/2$, it results that $\tilde{H}(\omega) = F(\theta) = \sum_{k=0}^{\tilde{n}} \tilde{h}_k \cos(2k+1)\theta$. Show that the PCLS design of a lowpass linear-phase symmetric filter of odd order is equivalent to

$$\begin{array}{ll} \min\limits_{F \in \mathbb{R}_n[z]} & E_s \\ \text{s.t.} & |F(\theta) - 1| \le \gamma_p, \ \forall\omega \in [0, \omega_p/2] \\ & |F(\theta)| \le \gamma_s, \ \forall\omega \in [\omega_s/2, \pi/2] \\ & f_{2k} = 0, \ \ k = 0 : \tilde{n} \end{array}$$

and express this as an SDP problem. (The notations are as in (5.8), the optimization problem for even order.)

P 5.2 We consider linear-phase *antisymmetric* FIR filters (5.1), of even order $n = 2\tilde{n}$.

a. Show that the magnitude of $H(z)$ is identical to that of the trigonometric polynomial

$$F(z) = \sum_{k=-\tilde{n}}^{\tilde{n}} f_k, \quad f_{-k} = f_k^*,$$

defined by $f_k = jh_{\tilde{n}-k}$, $f_0 = 0$. (Note that the polynomial $F(z)$ has imaginary coefficients, but its frequency response is real.)

b. Show that the PCLS design of a bandpass filter $H(z)$ is equivalent to

$$\begin{array}{ll} \min\limits_{F \in \mathbb{C}_{\tilde{n}}[z]} & E_s \\ \text{s.t.} & |F(\omega) - 1| \le \gamma_p, \ \forall\omega \in [\omega_{p1}, \omega_{p2}] \\ & |F(\omega)| \le \gamma_s, \ \forall\omega \in [0, \omega_{s1}] \cup [\omega_{s2}, \pi] \\ & \mathrm{Re} f_k = 0, \ \ k = 0 : \tilde{n} \end{array}$$

and express this as an SDP problem.

c. Using also the ideas from the previous problem, formulate an SDP problem for the design of linear-phase antisymmetric filters with *odd* order.

P 5.3 In the design problem (5.15), the frequency response is not constrained in the transition band. Show that adding an upper bound on the magnitude is possible and append the corresponding LMI to the SDP problem (5.17).

P 5.4 What are the changes in the three design problems discussed in Sect. 5.1 if the coefficients of the filter are *complex*?

P 5.5 Show that interpolation constraints $H(\omega_0) = b$, for a given ω_0 (e.g., $H(\omega_0) = 1$ for $\omega_0 = 0$), can be expressed as linear constraints for all the three design problems from Sect. 5.1. Append these constraints to the corresponding SDP problems.

P 5.6 (Frequency response fitting) We have the power spectrum measurements $|F(\omega_\ell)|^2 = R_\ell$, $\ell = 1 : L$, of a certain process F. We want to approximate it with an FIR process $H(z)$ of order n. Denoting $R(z) = H(z)H(z^{-1})$, we can find $H(z)$ by solving the minimax problem

$$\begin{array}{ll} \min\limits_{R\in\mathbb{R}_n[z]} & \max_{\ell=1:L} |R(\omega_\ell) - R_\ell| \\ \text{s.t.} & R(\omega) \geq 0, \;\; \forall\omega \in [-\pi, \pi] \end{array}$$

(followed by spectral factorization). Express this problem in SDP form.

Same requirement if the optimization objective is quadratic, i.e., equal to

$$\sum_{\ell=1}^{L} |R(\omega_\ell) - R_\ell|^2$$

P 5.7 In the linear-phase FIR filter design problem (5.8), the stopband constraint $|\tilde{H}(\omega)| \leq \gamma_s$, $\forall\omega \in [\omega_s, \pi]$, has been imposed with two positivity conditions, see (5.9). Show how the constraint (on the original filter) $|H(\omega)| \leq \gamma_s$, $\forall\omega \in [\omega_s, \pi]$, can be imposed using the LMI form (5.16) of the BRL. Compare the two approaches.

Extend the comparison to the 2D case.

P 5.8 (*Approximation of a fractional delay* [15]) Let $G(z)$ be a given FIR filter. Find the FIR filter $H(z)$ for which the approximation error $\|H(z) - z^{-1/2}G(z)\|_\infty$ is minimum. Thinking in the frequency domain and focusing on low frequencies, one can minimize $\|H(\omega) - \mathrm{e}^{-j\omega/2}G(\omega)\|$, over $\omega \in [0, \omega_0]$, where ω_0 is given. (Otherwise explained, find a sequence that best approximates another sequence shifted with 1/2.) Generalize to any fractional delay instead of 1/2.

Hint: note that the problem is equivalent to minimizing $\|H(2\omega) - \mathrm{e}^{-j\omega}G(2\omega)\|$, over $\omega \in [0, \omega_0/2]$ and express the problem using the BRL for trigonometric polynomials.

P 5.9 A 2D bandpass filter has the passband and the stopband as in the left of Fig. 5.15, bounded by rectangles of different sizes. Describe the passband and the stopband by unions of domains (4.13). Show that the parameters from (5.26), (5.27) have the following values: $d_p = 2$, $L_{p1} = L_{p2} = 3$ for the passband and $d_s = 3$, $L_{s1} = 2$, $L_{s2} = L_{s3} = 1$ for the stopband.

Same problem for the diamond bandpass filter from the right of Fig. 5.15.

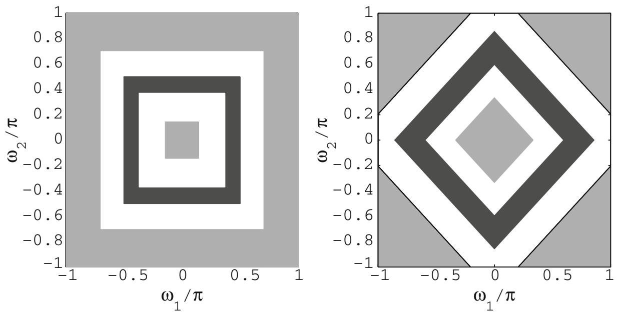


Fig. 5.15 Passbands (*black*) and stopbands (*gray*) for the filters from **P** 5.9

P 5.10 We assume that in the robust deconvolution problem (5.49) not only $\boldsymbol{A}(z)$ belongs to a polytope with L_a known vertices, but also $\boldsymbol{B}(z)$ belongs to another polytope, with L_b vertices. Write the optimization problems corresponding to (5.50). How many H_∞-norm inequality constraints (on vertices) are necessary? Write also the SDP equivalent problem.

Notice that the robustness deconvolution problem can be posed and solved similarly in the multidimensional case.

References

1. A.T. Erdogan, B. Hassibi, T. Kailath, FIR H_∞ equalization. Signal Process. **81**(5), 907–917 (2001)
2. H. Zhou, L. Xie, C. Zhang, A direct approach to H_2 optimal deconvolution of periodic digital channels. IEEE Trans. Signal Process. **50**(7), 1685–1698 (2002)
3. L. Xie, C. Du, C. Zhang, Y.C. Soh, H_∞ deconvolution filtering of 2-D digital systems. IEEE Trans. Signal Process. **50**(9), 2319–2332 (2002)
4. S.P. Wu, S. Boyd, L. Vandenberghe, FIR filter design via semidefinite programming and spectral factorization, in *Proceedings of 35th IEEE Conference on Decision and Control*, vol. 1 (Kobe, Japan, 1996), pp. 271–276
5. S.P. Wu, S. Boyd, L. Vandenberghe, FIR filter design via spectral factorization and convex optimization, in *Applied and Computational Control, Signals and Circuits*, ed. by B. Datta (Birkhauser, Basel, 1997), pp. 51–81
6. B. Alkire, L. Vandenberghe, Convex optimization problems involving finite autocorrelation sequences. Math. Program. Ser. A **93**(3), 331–359 (2002)
7. T.N. Davidson, Z.Q. Luo, J.F. Sturm, Linear matrix inequality formulation of spectral mask constraints with applications to FIR filter design. IEEE Trans. Signal Process. **50**(11), 2702–2715 (2002)
8. T.N. Davidson, Enriching the art of FIR filter design via convex optimization. IEEE Signal Process. Mag. **27**(3), 89–101 (2010)
9. B. Dumitrescu, Trigonometric polynomials positive on frequency domains and applications to 2-D FIR filter design. IEEE Trans. Signal Process. **54**(11), 4282–4292 (2006)
10. R. Yang, L. Xie, C. Zhang, Kalman–Yakubovich–Popov lemma for two-dimensional systems, in *IFAC World Congress* (Prague, Czech Republic, 2005)

11. N. Al-Dhahir, FIR channel-shortening equalizers for MIMO ISI channels. IEEE Trans. Commun. **49**(2), 213–218 (2001)
12. B. Dumitrescu, Bounded real lemma for FIR MIMO systems. IEEE Signal Process. Lett. **12**(7), 496–499 (2005)
13. B. Dumitrescu, Bounded real lemma for multivariate trigonometric matrix polynomials and FIR filter design applications, in *European Signal Processing Conference EUSIPCO* (Glasgow, UK, 2009)
14. B. Dumitrescu, B.C. Şicleru, R. Ştefan, Positive hybrid real-trigonometric polynomials and applications to adjustable filter design and absolute stability analysis. Circuits Syst. Signal Process. **29**(5), 881–899 (2010)
15. B. Dumitrescu, SDP approximation of a fractional delay and the design of dual-tree complex wavelet transform. IEEE Trans. Signal Process. **56**(9), 4255–4262 (2008)

Chapter 6
Orthogonal Filterbanks

Abstract In this chapter, we explore the use of positive polynomials in the design of FIR filterbanks (FB). The study is confined to a single class, that of orthogonal FBs. Two-channel FBs are discussed first, as the simplest instance of the problem; naturally related with it are the design of compaction filters or of signal-adapted wavelets. We go then to DFT-modulated FBs, with an arbitrary number of channels; similarly to the two-channel case, the free parameters of the whole FB are the coefficients of a single prototype filter. A typical requirement on FBs is that of perfect reconstruction (PR): the output signal is a delayed version of the input one. The connection between orthogonal FBs and positive polynomials is eased by the fact that PR amounts to simple (Nyquist) conditions on the squared magnitude of the prototype filter. Optimization problems that are nonconvex in the coefficients of the prototype filter become convex once expressed using its squared magnitude, which is a nonegative polynomial described by an appropriate Gram matrix parameterization. After solving the equivalent SDP problem, the prototype filter is recovered by spectral factorization.

6.1 Two-Channel Filterbanks

The scheme of a two-channel FB is presented in Fig. 6.1. In the first channel, the input signal x is filtered by the analysis filter $H_0(z)$, then decimated with a factor of two (every other sample is ignored); the resulting signal x_0 is interpolated with a factor of two (a zero sample is inserted between each two samples of x_0) and then filtered with the synthesis filter $F_0(z)$. Similar operations are performed in the second channel. The output y of the filter is the sum of the signals from the two channels.

The FB is composed of the analysis and synthesis banks. In applications, the subband signals x_0 and x_1 are processed in some manner (e.g., quantized or filtered), with the advantage of a lower sampling rate. However, when designing the FB, one often assumes that there is no subband processing. The main requirement on the FB is *perfect reconstruction* (PR): the output signal is a copy of the input one, with a delay Δ, i.e.,

$$Y(z) = z^{-\Delta} X(z). \tag{6.1}$$

B. Dumitrescu, *Positive Trigonometric Polynomials and Signal Processing Applications*, Signals and Communication Technology,
DOI 10.1007/978-3-319-53688-0_6

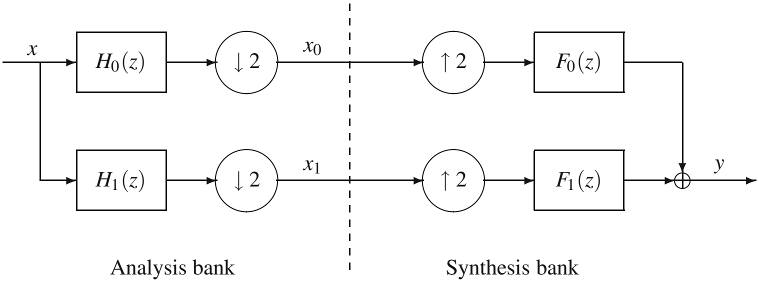


Fig. 6.1 Two-channel filterbank

This condition can be relaxed to near-PR; in this case, the output is only an approximate copy of the input; see Sect. 6.2 and problem **P** 6.6 for developments of this subject.

Let us first derive the input–output behavior of the FB. Due to decimation, the first subband signal is

$$X_0(z) = \frac{1}{2}\left[H_0(z^{1/2})X(z^{1/2}) + H_0(-z^{1/2})X(-z^{1/2})\right].$$

A similar expression holds for $X_1(z)$. Due to interpolation, the output is

$$Y(z) = F_0(z)X_0(z^2) + F_1(z)X_1(z^2).$$

We obtain the input–output relationship

$$Y(z) = T_d(z)X(z) + T_a(z)X(-z), \tag{6.2}$$

where

$$T_d(z) = \frac{1}{2}\left[F_0(z)H_0(z) + F_1(z)H_1(z)\right] \tag{6.3}$$

is the distortion transfer function (which accounts for the transformation of the input signal) and

$$T_a(z) = \frac{1}{2}\left[F_0(z)H_0(-z) + F_1(z)H_1(-z)\right] \tag{6.4}$$

is the aliasing transfer function (which describes the transformation of the aliased input $X(-z)$). From (6.2), it results that perfect reconstruction is equivalent to the conditions

$$T_d(z) = z^{-\Delta}, \quad T_a(z) = 0. \tag{6.5}$$

6.1.1 Orthogonal FB Design

The PR conditions (6.5) can be satisfied in different manners. The simplest choice, which leads to orthogonal (or conjugate quadrature) FBs, is to define all the filters as functions of a single one. We denote $H_0(z) = H(z)$ and assume that it is an FIR filter of *odd* order n, with complex coefficients; we will see later why the order must be odd; although we are interested mainly by FBs with real coefficients, we treat the complex case for the sake of generality. In an orthogonal FB, the other filters are defined by

$$\begin{aligned} H_1(z) &= -z^{-n} H^*(-z^{-1}), \\ F_0(z) &= H_1(-z) = z^{-n} H^*(z^{-1}), \\ F_1(z) &= -H_0(-z) = -H(-z). \end{aligned} \tag{6.6}$$

These conditions lead to simple relations between the frequency responses of the filters. The analysis and synthesis filters have the same magnitude, i.e., $|H_0(\omega)| = |F_0(\omega)|$, $|H_1(\omega)| = |F_1(\omega)|$. The magnitudes of filters from the same bank have a mirror property; they are symmetric about $\pi/2$, i.e.,

$$|H_1(\omega)| = |H_0(\pi - \omega)|. \tag{6.7}$$

Substituting (6.6) into (6.4), it results immediately that $T_a(z) = 0$, i.e., aliasing is perfectly canceled. On the unit circle, the trigonometric polynomial

$$R(z) = H(z)H^*(z^{-1}) \tag{6.8}$$

is the squared magnitude of the filter ($R(\omega) = |H(\omega)|^2$); in the FB context, $R(z)$ is often named *product filter*. With this notation, the distortion transfer function (6.3) becomes

$$T_d(z) = \tfrac{1}{2}\left[R(z) + R(-z)\right] z^{-n}. \tag{6.9}$$

Imposing a delay $\Delta = n$, the first condition from (6.5) is met if

$$R(z) + R(-z) = 2. \tag{6.10}$$

Since all the odd-indexed coefficients of $R(z) + R(-z)$ are zero, perfect reconstruction is achieved if and only if

$$r_{2k} = \delta_k, \tag{6.11}$$

i.e., the product filter has the Nyquist(2) property.

Reminding the relation (2.27) between the coefficients of $R(z)$ and $H(z)$, which is $r_k = \boldsymbol{h}^H \boldsymbol{\Theta}_k \boldsymbol{h}$ (where $\boldsymbol{h} \in \mathbb{R}^{n+1}$ contains the coefficients of $H(z)$), we note that the PR condition (6.11) is quadratic (with indefinite matrices) in the filter coefficients, but linear in those of the squared magnitude and thus more amenable to optimization.

Optimization objectives. There are several objectives for the optimization of an orthogonal FB. A first option may be the minimization of the energy in the stopband $[\omega_s, \pi]$, where ω_s is a given stopband edge. As we have seen in Sect. 5.1, the stopband energy (5.2) has a linear expression in the coefficients of $R(z)$, as shown by (5.5). Moreover, the PR condition (6.10) leads to

$$|H(\omega)|^2 + |H(\pi - \omega)|^2 = 2. \tag{6.12}$$

Due to the mirror property (6.7), this is equivalent to the power complementarity equality

$$|H_0(\omega)|^2 + |H_1(\omega)|^2 = 2. \tag{6.13}$$

The relation (6.12) tells that the minimization of the stopband energy (5.2) implies also an optimization of the passband error (more precisely, of the integral error between $|H(\omega)|^2$ and the ideal value 2), in the interval $[0, \pi - \omega_s]$. In general, for a Nyquist filter, it is enough to impose conditions on the stopband.

A second optimization objective arises in the design of *signal-adapted* FBs, strongly related to that of optimal *compaction* filters. Let us assume that the input signal x is a random process whose autocorrelations

$$\rho_k = E\{x(\ell)x^*(\ell - k)\}$$

are known. A compaction filter $H(z)$ maximizes the variance of its output $\xi(z) = H(z)x(z)$ and its squared magnitude (6.8) is a Nyquist filter. Equivalently, the first channel of the orthogonal FB with $H_0(z) = H(z)$ takes the most energy from the input signal; note that in an orthogonal FB, due to the power complementarity relation (6.13), the sum of the energies of the subband signals is equal to the energy of the input; maximizing the energy in the first channel is equivalent to minimizing the energy in the second channel. An orthogonal FB with this property is optimal for quantization purposes (the number of bits for each subband signal can be allocated optimally according to the variance of the signal).

The variance of the signal ξ is

$$\begin{aligned}
\sigma_\xi^2 \overset{\Delta}{=} E\{|\xi(\ell)|^2\} &= E\left\{\sum_{i=0}^{n} h_i x(\ell - i) \sum_{k=0}^{n} h_k^* x^*(\ell - k)\right\} \\
&= \sum_{i=0}^{n}\sum_{k=0}^{n} h_i h_k^* E\{x(\ell - i)x^*(\ell - k)\} = \sum_{i=0}^{n}\sum_{k=0}^{n} h_i h_k^* \rho_{k-i} \\
&= \sum_{k=-n}^{n} \rho_k \boldsymbol{h}^H \boldsymbol{\Theta}_k \boldsymbol{h} \overset{(2.27)}{=} \sum_{k=-n}^{n} \rho_k r_k.
\end{aligned} \tag{6.14}$$

Again, we have obtained a linear expression in the coefficients of the product filter.

Design problem. We conclude that the design of an orthogonal FB with real coefficients can be completed as follows. The degree n of the filters is an odd integer value; giving it an even value is useless, since from (6.11) it would result that $r_n = 0$ and so $h_n = 0$. We solve the problem

$$\begin{array}{ll} \min\limits_{R \in \mathbb{R}_n[z]} & \sum_{k=0}^{n} c_k r_k \\ \text{s.t.} & r_{2k} = \delta_k, \quad k = 0 : (n-1)/2 \\ & R(\omega) \geq 0, \quad \forall \omega \end{array} \tag{6.15}$$

where the coefficients of the linear objective come from one of the two objectives discussed above: stopband energy or variance of the first channel signal. The nonnegativity of the trigonometric polynomial $R(z)$ is expressed via the trace (2.6) or Gram pair (2.94) parameterization. After solving the SDP problem equivalent to (6.15) (see problem **P** 6.1), the filter $H(z)$ is found by spectral factorization and the filters of the orthogonal FB are derived from (6.6).

Example 6.1 We consider an input signal generated by an AR(1) process with the pole $\alpha = 0.9$. Its autocorrelations are $\rho_k = \alpha^k$. Such a process has a pronounced low-band spectral power density. We design a two-channel orthogonal FB maximizing the variance (6.14) of the first channel signal (or, in other words, we design a compaction filter adapted to the AR(1) process). After solving the SDP version of (6.15) for $n = 19$, we obtain the frequency responses shown in Fig. 6.2; the filter $H_0(z)$ is lowpass and $H_1(z)$ is highpass (we remind that the analysis and the synthesis filters have the same magnitude response). ■

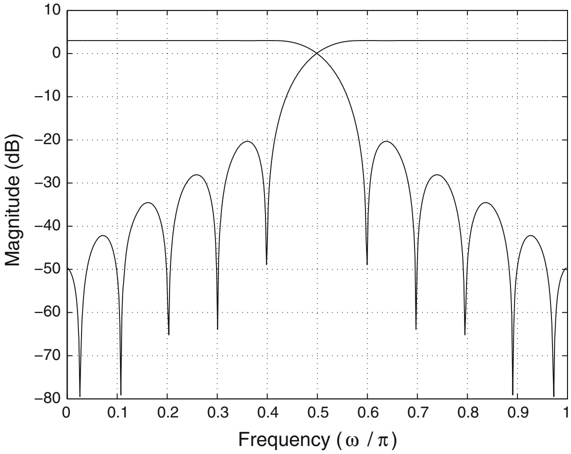


Fig. 6.2 Frequency response of the filterbank from Example 6.1

6.1.2 Towards Signal-Adapted Wavelets

Let us assume that $H(z)$ is the first analysis filter of an orthogonal FB, and thus, its product filter (6.8) satisfies the Nyquist (or orthogonality) condition (6.11). As before, $H(z)$ is an FIR filter of odd degree n. We also make a supplementary assumption, that the filter $H(z)$ has $n_r > 0$ degrees of regularity, i.e., it has n_r roots in $z = -1$. Consequently, it has the form

$$H(z) = (1 + z^{-1})^{n_r} U(z), \tag{6.16}$$

where $U(z)$ is an FIR filter of degree $n - n_r$.

We can associate with $H(z)$ the *dilation equation*

$$\phi(t) = \sqrt{2} \sum_{k=0}^{n} h_k \phi(2t - k), \tag{6.17}$$

whose solution, the function ϕ of real variable t, which exists and is unique, is named *scaling function*. The corresponding *wavelet function* is

$$\psi(t) = \sqrt{2} \sum_{k=0}^{n} (-1)^k h_{n-k} \phi(2t - k). \tag{6.18}$$

Under the above conditions on $H(z)$ (and some other mild conditions, e.g., an upper bound on $|U(\omega)|$), the set of functions (wavelets)

$$\left\{\psi_{i,\ell}(t) = 2^{i/2} \psi(2^i t - \ell)\right\}_{i,\ell \in \mathbb{Z}} \tag{6.19}$$

forms an orthogonal basis for $L_2(\mathbb{R})$; we remind that $L_2(\mathbb{R})$ is the set of real signals with finite energy; the scalar product of $f, g \in L_2(\mathbb{R})$ is $\int_{-\infty}^{\infty} f(t)g(t)dt$. The indices i and ℓ denote, respectively, the scale and the translation of the wavelet $\psi_{i,\ell}(t)$.

Given a function $f \in L_2(\mathbb{R})$ (viewed as the representative of a class of signals), we may want to approximate it with wavelets whose scale is less than or equal to an integer I, i.e., we ignore "details"; this is meaningful if the function has a finite support spectrum. Since the wavelet basis (6.19) is orthogonal, the approximation is made by projection onto the space generated by $\psi_{i,\ell}(t)$, $i \le I$. As the subspaces generated by wavelets with the same scale form a multiresolution analysis, the same approximation is obtained by projection onto the subspace generated by the scaling functions

$$\phi_{I,\ell}(t) = 2^{I/2} \phi(2^I t - \ell), \quad \ell \in \mathbb{Z} \tag{6.20}$$

of scale I. We thus approximate

$$f(t) \approx \sum_{\ell \in \mathbb{Z}} a_{I\ell} \phi_{I,\ell}(t), \tag{6.21}$$

where the coefficients are obtained by the orthogonal projection

$$a_{I\ell} = \int_{-\infty}^{\infty} f(t)\phi_{I,\ell}(t)dt.$$

We want to minimize the approximation error in (6.21). The natural optimization variables are the coefficients of the filter $H(z)$ that not only appear in (6.17) and (6.18), but also are effectively used in a practical approximation scheme, based on FB processing. Unfortunately, the approximation error is not convex in these variables. With arguments that are too long to present here, a simplified optimization objective is presented in [1], that is linear in the coefficients of the product filter (6.8), i.e., has the form (5.5) or (6.14). With this simplification, the design of signal-adapted wavelets is similar to the FB design problem (6.15), to which we have to add only the regularity constraints. This can be done in two ways, which we discuss in detail.

Explicit regularity constraints. We can force n_r roots of $H(z)$ to be equal to -1 by imposing conditions on the derivatives of $H(z)$, i.e., $H^{(\ell)}(-1) = 0$, $\ell = 0 : n_r - 1$, or, equivalently,

$$\sum_{k=0}^{n} (-1)^k k^\ell h_k = 0, \quad \ell = 0 : n_r - 1. \tag{6.22}$$

(When $k = \ell = 0$, we assume $k^\ell = 1$.) Since each root of $H(z)$ on the unit circle is a double root of the product filter $R(z)$, the constraint (6.22) can be written as

$$r_0 + 2\sum_{k=1}^{n} (-1)^k k^{2\ell} r_k = 0, \quad \ell = 0 : n_r - 1. \tag{6.23}$$

We have ignored half of the constraints due to the positivity of $R(\omega)$: the number of roots in -1 is always even. The constraint (6.23) is linear in the coefficients of the product filter. Adding this constraint to FB design problem (6.15), we obtain the optimization problem

$$\begin{array}{ll} \min\limits_{R\in\mathbb{R}_n[z]} & \sum_{k=0}^{n} c_k r_k \\ \text{s.t.} & r_{2k} = \delta_k, \quad k = 0 : (n-1)/2 \\ & r_0 + 2\sum_{k=1}^{n} (-1)^k k^{2\ell} r_k = 0, \quad \ell = 0 : n_r - 1 \\ & R(\omega) \geq 0, \quad \forall\omega \end{array} \tag{6.24}$$

This can be easily transformed into an SDP problem, in the style suggested by problem **P** 6.1. Although this approach is straightforward and easy to implement, it can cause numerical trouble for moderate values of n and n_r. The reason is the large range of values for the coefficients from (6.23). For instance, taking $n = 39$ and $n_r = 6$, the magnitude of the coefficients from (6.23) goes from 1 to $n^{2(n_r-1)} = 39^{10} \approx 10^{16}$; it is clear that the constraint can be imposed numerically with only low accuracy. Practical experiments have shown that even with values such as $n = 19$ and

$n_r = 4$, the roots of the designed filter $H(z)$ are spread around -1 (in the complex plane); the accuracy is worse as n and n_r grow.

Implicit regularity constraints. Another possibility is to work directly with the factorization (6.16), using $U(z)$ as variable. Denoting $V(z) = U(z)U^*(z^{-1})$ the product filter corresponding to $U(z)$, it results that

$$R(z) = (1 + z^{-1})^{n_r}(1 + z)^{n_r} V(z). \tag{6.25}$$

We denote

$$B(z) = (1 + z^{-1})^{n_r}(1 + z)^{n_r} = \sum_{i=0}^{n_r} \binom{n_r}{i} z^{-i} \sum_{\ell=0}^{n_r} \binom{n_r}{\ell} z^{\ell},$$

which is a symmetric polynomial whose coefficients are

$$b_k = \sum_{\ell-i=k} \binom{n_r}{i}\binom{n_r}{\ell} = \sum_{i=0}^{n_r} \binom{n_r}{i}\binom{n_r}{n_r-k-i} = \binom{2n_r}{n_r-k}. \tag{6.26}$$

We have used the fact that $\binom{n_r}{\ell} = \binom{n_r}{n_r-\ell}$; the last equality in (6.26) is the Vandermonde identity. Anyway, it is not the expression (6.26) that is important, but the linear dependence of the coefficients of $R(z)$ from those of $V(z)$, that results from (6.25) and is

$$r_k = \sum_{i=-n_r}^{n_r} b_i v_{k-i}. \tag{6.27}$$

Thus, adding regularity constraints to (6.15) gives

$$\begin{array}{ll} \min\limits_{R \in \mathbb{R}_{n-n_r}[z]} & \sum_{k=0}^{n} \sum_{i=-n_r}^{n_r} c_k b_i v_{k-i} \\ \text{s.t.} & \sum_{i=-n_r}^{n_r} b_i v_{2k-i} = \delta_k, \quad k = 0 : (n-1)/2 \\ & V(\omega) \ge 0, \quad \forall \omega \end{array} \tag{6.28}$$

The advantage over (6.24) is not only the smaller number of constraints (since there are no more regularity constraints on $V(z)$), but also the lower degree of the positive polynomial $V(z)$ (which is $n - n_r$). Hence, when expressing (6.28) as an SDP problem, the size of the Gram matrix (or matrices) will be smaller.

After solving (6.28), the FIR filter $U(z)$ is computed from $V(z)$ by spectral factorization and inserted in (6.16) to give the desired filter $H(z)$. Using this implicit approach, the roots in -1 can be possibly affected only by the computation involved by the convolution (6.16); typically, the actual roots are very near from -1.

Example 6.2 We consider the same design problem as in Example 6.1, but adding the regularity requirement of n_r roots in -1. For $n = 19$, $n_r = 4$, we solve the

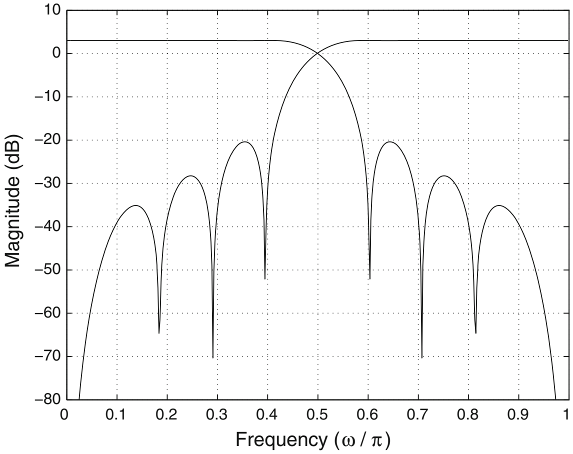


Fig. 6.3 Frequency response of the filterbank from Example 6.2

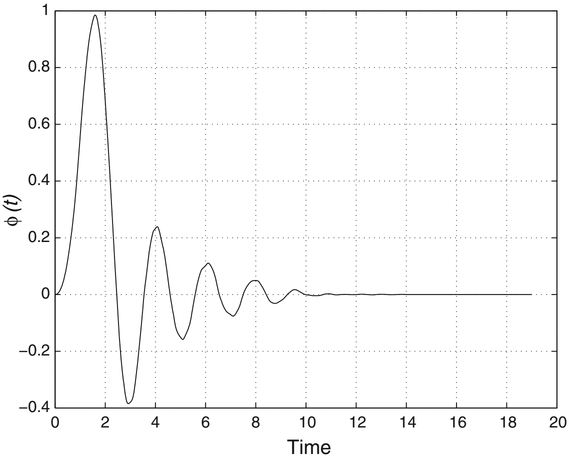


Fig. 6.4 Scaling function designed in Example 6.2

SDP version of (6.28) and obtain the FB whose frequency response is shown in Fig. 6.3. The effect of the regularity zeros is visible in the high frequency area. The corresponding scaling function $\phi(t)$, found by solving the dilation equation (6.17), is shown in Fig. 6.4; we remind that its support is the interval $[0, n]$. (This scaling function is quite similar to the one designed in [1] for the optimal approximation (6.21) of the sinc function $f(t) = \sin \pi t / \pi t$.) ■

6.1.3 Design of Symmetric Complex-Valued FBs

Until now, we have designed real-valued FBs, although the theoretical considerations were derived in the general case of complex-valued FBs. Here, we consider a special class of FBs that belong to the latter case.

FIR orthogonal FBs as described in Sect. 6.1.1 cannot have linear phase, in the sense that it is impossible to have $h_k = h^*_{n-k}$ or $h_k = -h^*_{n-k}$ (in this case, all filters (6.6) would have linear phase). However, it is possible to build such FBs that are *symmetric*. In this case, the coefficients of the first analysis filter obey to the condition

$$h_k = h_{n-k}. \tag{6.29}$$

In the real case, symmetry means linear phase, and thus, such symmetric FBs can only be complex-valued. The designation "symmetric FB" is actually misleading, since it results from (6.6) that the second analysis filter has *antisymmetric* coefficients. We note also that the scaling and wavelet functions given by (6.17) and (6.18) are symmetric and antisymmetric, respectively (and complex-valued).

To be able to design symmetric FBs using the method from Sect. 6.1.1, we must ensure that the product filter (6.8) is generated by a symmetric filter $H(z)$. Let us examine the properties of such a product filter.

Remark 6.3 The product filter has real coefficients. Indeed, since the symmetry condition (6.29) is equivalent to the equality

$$H(z^{-1}) = z^n H(z), \tag{6.30}$$

it results that the product filter satisfies the relation

$$R(z) \overset{\Delta}{=} H(z)H^*(z^{-1}) = z^n H(z) z^{-n} H^*(z^{-1}) = H(z^{-1})H^*(z) = R(z^{-1}).$$

As the definition (1.1) implies $R(z^{-1}) = R^*(z)$, it results that $R(z)$ has real coefficients. (A polynomial that is both Hermitian and symmetric has real coefficients.) ■

Remark 6.4 We examine now the zeros of the product filter. Since $H(z)$ is symmetric, if z is a zero of $H(z)$, then $1/z$ is also a zero. Therefore, $R(z)$ has (clusters of) four zeros: $z, 1/z$ (from $H(z)$), $z^*, 1/z^*$ (from $H^*(z^{-1})$). If z is complex, then this is the root configuration for *all* nonnegative trigonometric polynomials with real coefficients. What is specific to the case of symmetric spectral factors occurs when the zero z is real; the four zeros collapse to two *double* zeros: z and $1/z$; in particular, if 1 would be a zero of $R(z)$, then its multiplicity would be a multiple of 4 (typically there is no zero in 1, due to the usual lowpass nature of $H(z)$). There is a single exception to this rule, when the zero is in -1; since the degree of $H(z)$ is odd, it results from (6.30) that $H(-1) = -H(-1)$ and so -1 is always a root; its multiplicity is thus odd; so, -1 is a root of $R(z)$ of multiplicity $2n_r$, where n_r is odd.

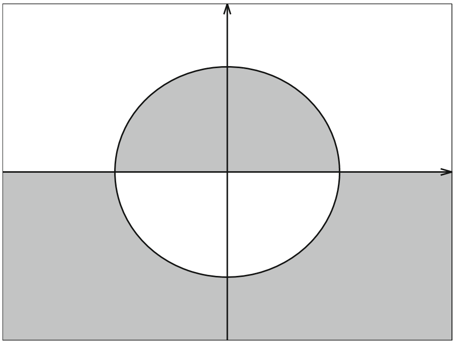


Fig. 6.5 Splitting of the complex plane: the *gray* area contains the roots of a symmetric spectral factor

The discussion above shows how to perform the special spectral factorization that, given $R(z)$, computes a symmetric spectral factor (assuming that it exists). The complex plane is split into two parts, such that if z belongs to a part, then $1/z$ belongs to the same part. An example of splitting is given in Fig. 6.5. At spectral factorization, we assign e.g., the roots from the gray region to $H(z)$, the other roots going to $H^*(z^{-1})$. The roots on the border between regions have even multiplicity (on the unit circle as a general rule for nonnegative trigonometric polynomials, and on the real axis as shown above for the case at hand) and they are split evenly between the two spectral factors. ■

Moreover, we can give the following Gram parameterization.

Theorem 6.5 *A nonnegative trigonometric polynomial $R \in \mathbb{R}[z]$ of order $n = 2\tilde{n} + 1$ has the form $R(z) = H(z)H^*(z^{-1})$, with $H(z)$ satisfying the symmetry condition (6.29) (in other words, $R(z)$ has a symmetric spectral factor, possibly with complex coefficients), if and only if there exists a positive semidefinite matrix $\boldsymbol{Q} \in \mathbb{R}^{(\tilde{n}+1)\times(\tilde{n}+1)}$ such that*

$$R(\omega) = \tilde{\boldsymbol{\chi}}_c^T(\omega)\boldsymbol{Q}\tilde{\boldsymbol{\chi}}_c(\omega), \tag{6.31}$$

where the vector $\tilde{\boldsymbol{\chi}}_c(\omega)$ is defined in (2.98). The coefficients of $R(z)$ have the form

$$r_k = \operatorname{tr}[\tilde{\boldsymbol{\Phi}}_k \boldsymbol{Q}], \quad k = 0 : n, \tag{6.32}$$

where the constant matrices $\tilde{\boldsymbol{\Phi}}_k$ are defined in (2.101).

Proof We note that (6.31) represents a "half" of the Gram pair parameterization presented in Sect. 2.8.3. The proof there is based on splitting the symmetric and antisymmetric parts of a spectral factor of $R(z)$. In our case, the antisymmetric part does not exist, since $H(z)$ is symmetric. Although this remark would seem almost sufficient for a proof, we will see immediately that the arguments are not at all immediate. It follows from (6.29) that

$$H(\omega) \overset{\Delta}{=} \sum_{k=0}^{n} h_k \mathrm{e}^{-jk\omega} = \mathrm{e}^{-jn\omega/2} \sum_{k=0}^{\tilde{n}} 2h_{\tilde{n}-k} \cos\left(k+\tfrac{1}{2}\right)\omega = \mathrm{e}^{-jn\omega/2} \tilde{\boldsymbol{\chi}}_c^T(\omega)\boldsymbol{a},$$

where

$$\boldsymbol{a} = 2[h_{\tilde{n}} \ \ldots \ h_1 \ h_0]^T = \boldsymbol{b} + j\boldsymbol{c} \tag{6.33}$$

with $\boldsymbol{b}, \boldsymbol{c} \in \mathbb{R}^{\tilde{n}+1}$. Hence, we obtain

$$|H(\omega)|^2 = \tilde{\boldsymbol{\chi}}_c^T(\omega)(\boldsymbol{b}\boldsymbol{b}^T + \boldsymbol{c}\boldsymbol{c}^T)\tilde{\boldsymbol{\chi}}_c(\omega). \tag{6.34}$$

So, if $R(\omega)$ is nonnegative and has a symmetric spectral factor, there exists $\boldsymbol{Q} = \boldsymbol{b}\boldsymbol{b}^T + \boldsymbol{c}\boldsymbol{c}^T$ such that (6.31) holds. This implication follows the pattern used in several proofs in this book. However, the reverse implication is not as trivial as usual.

If (6.31) holds, then $R(\omega) \geq 0, \forall\omega \in [-\pi, \pi]$. It remains to show that $R(z)$ has a symmetric spectral factor. We denote $t = \cos\omega/2$. Due to the recurrence (1.42), the Chebyshev polynomial $\cos(k+\frac{1}{2})\omega$, $k \in \mathbb{N}$, depends only on the odd powers of t. Hence, similarly to (1.44), we can write

$$\tilde{\boldsymbol{\chi}}_c(\omega) = \boldsymbol{A} \begin{bmatrix} t \\ t^3 \\ \vdots \\ t^{2\tilde{n}+1} \end{bmatrix} = \boldsymbol{A} t \boldsymbol{\psi}(t^2), \tag{6.35}$$

where $\boldsymbol{A}$ is a constant nonsingular matrix and $\boldsymbol{\psi}(t) = [1 \ t \ \ldots \ t^{\tilde{n}}]^T$. So, (6.31) becomes

$$R(\omega) = t^2 \boldsymbol{\psi}^T(t^2)\boldsymbol{A}^T \boldsymbol{Q}\boldsymbol{A}\boldsymbol{\psi}(t^2) = t^2 \tilde{R}(t^2). \tag{6.36}$$

The real polynomial $\tilde{R}(t) = \boldsymbol{\psi}^T(t)\boldsymbol{A}^T\boldsymbol{Q}\boldsymbol{A}\boldsymbol{\psi}(t)$ is nonnegative for any $t \in \mathbb{R}$, since $\boldsymbol{A}^T\boldsymbol{Q}\boldsymbol{A} \succeq 0$. Theorem 1.7 states that it can be expressed as the sum of two squares, i.e., there exist polynomials $\tilde{B}(t) = \boldsymbol{\psi}^T(t)\tilde{\boldsymbol{b}}$ and $\tilde{C}(t) = \boldsymbol{\psi}^T(t)\tilde{\boldsymbol{c}}$ such that

$$\tilde{R}(t) = \tilde{B}(t)^2 + \tilde{C}(t)^2 = \boldsymbol{\psi}^T(t)(\tilde{\boldsymbol{b}}\tilde{\boldsymbol{b}}^T + \tilde{\boldsymbol{c}}\tilde{\boldsymbol{c}}^T)\boldsymbol{\psi}(t).$$

Inserting this expression into (6.36) and taking (6.35) into account, we obtain

$$R(\omega) = \tilde{\boldsymbol{\chi}}_c^T(\omega)(\boldsymbol{b}\boldsymbol{b}^T + \boldsymbol{c}\boldsymbol{c}^T)\tilde{\boldsymbol{\chi}}_c(\omega),$$

with $\boldsymbol{b} = \boldsymbol{A}^{-T}\tilde{\boldsymbol{b}}$, $\boldsymbol{c} = \boldsymbol{A}^{-T}\tilde{\boldsymbol{c}}$. As in (6.34), the vectors $\boldsymbol{b}$ and $\boldsymbol{c}$ contain the real and imaginary parts, respectively, of the distinct coefficients of a symmetric spectral factor $H(z)$, as detailed by (6.33). We conclude that the trigonometric polynomial (6.31) has a symmetric spectral factor.

Finally, since (6.31) is equivalent to

$$R(\omega) = \frac{1}{2}\sum_{i=0}^{\tilde{n}}\sum_{\ell=0}^{\tilde{n}} q_{i\ell}[\cos(i+\ell+1)\omega + \cos(i-\ell)\omega],$$

the relation (6.32) follows immediately. (It is obtained by putting $s_{i\ell} = 0$ in (2.100).) ■

Remark 6.6 The existence of a Gram parameterization (6.31) ensures the existence of a symmetric spectral factor with *complex* coefficients. A parameterization of nonnegative polynomials with symmetric spectral factors with *real* coefficients is equivalent to a parameterization of real polynomials that are *squares*; hence, it seems not possible in the Gram formalism. For other parameterizations involving symmetry or antisymmetry, see problem **P** 6.5. ■

Design problem. A typical optimization of the FB consists of minimizing the stopband energy, to which we may add a bound on the stopband ripple, i.e., we impose $|H(z)| \le \gamma_s$, $\forall\omega \in [\omega_s, \pi]$, for a given bound γ_s. (Due to power complementarity, it is useless to impose a bound in the passband.) We have seen that the stopband energy is linear in the coefficients of the product filter $R(z)$, as in (6.15). We control stopband attenuation by requiring that the trigonometric polynomial $\gamma_s^2 - R(\omega)$ is nonnegative on $[\omega_s, \pi]$. Appealing to Theorem 1.18—the case of odd n, for which (1.40) holds—the stopband bound is satisfied if

$$\gamma_s^2 - R(\omega) = (\cos\omega + 1)\cdot\tilde{R}_1(\omega)^2 + (\cos\omega_s - \cos\omega)\cdot\tilde{R}_2(\omega)^2,$$

for some polynomials $R_1(\omega)$ and $R_2(\omega)$ of degree $\tilde{n}$. This is equivalent to the existence of positive semidefinite matrices $\boldsymbol{Q}_1, \boldsymbol{Q}_2 \in \mathbb{R}^{(\tilde{n}+1)\times(\tilde{n}+1)}$ such that

$$\gamma_s^2 - R(\omega) = (\cos\omega + 1)\boldsymbol{\chi}_c^T(\omega)\boldsymbol{Q}_1\boldsymbol{\chi}_c(\omega) + (\cos\omega_s - \cos\omega)\boldsymbol{\chi}_c^T(\omega)\boldsymbol{Q}_2\boldsymbol{\chi}_c(\omega),$$

where the vector $\boldsymbol{\chi}_c(\omega)$ is defined in (2.89). Using transformations similar to those from Sect. 2.8.3, it results that

$$\begin{aligned}\gamma_s^2\delta_k - r_k = \operatorname{tr}&\left[\left(\boldsymbol{\Phi}_k + \frac{1}{2}\boldsymbol{\Phi}_{k-1} + \frac{1}{2}\boldsymbol{\Phi}_{k+1}\right)\boldsymbol{Q}_1\right] \\ &+ \operatorname{tr}\left[\left(\cos\omega_s\boldsymbol{\Phi}_k - \frac{1}{2}\boldsymbol{\Phi}_{k-1} - \frac{1}{2}\boldsymbol{\Phi}_{k+1}\right)\boldsymbol{Q}_2\right],\end{aligned} \tag{6.37}$$

where the matrices $\boldsymbol{\Phi}_k$ are defined in (2.95). So, the symmetric FB design problem, expressed in terms of the product filter, is

$$\begin{array}{ll}\min\limits_{\boldsymbol{Q},\boldsymbol{Q}_1,\boldsymbol{Q}_2} & \sum_{k=0}^{n} c_k r_k \\ \text{s.t.} & r_k = \operatorname{tr}[\tilde{\boldsymbol{\Phi}}_k\boldsymbol{Q}], \quad k = 0:n \\ & r_{2k} = \delta_k, \quad k = 0:\tilde{n} \\ & (6.37), \quad k = 0:n \\ & \boldsymbol{Q} \succeq 0,\ \boldsymbol{Q}_1 \succeq 0,\ \boldsymbol{Q}_2 \succeq 0\end{array} \tag{6.38}$$

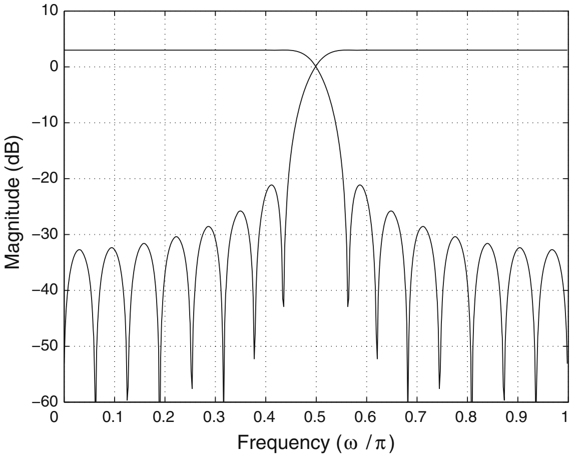


Fig. 6.6 Frequency response of symmetric FB from Example 6.7, with no stopband ripple bound

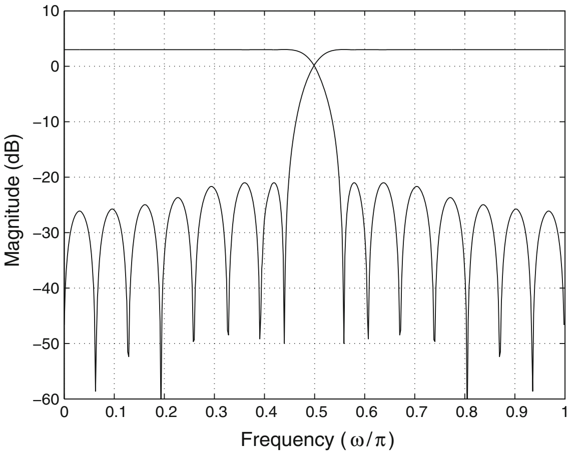


Fig. 6.7 Symmetric FB from Example 6.7, with $\gamma_s = -21$ dB

The constraints express, in order, the positivity of the product filter (ensuring also the existence of a symmetric spectral factor), the orthogonality of the FB and the stopband bound. We can add regularity constraints to this problem (not forgetting that $R(z)$ has always at least two zeros in -1), preferably in the implicit manner discussed in the previous section. After solving (6.38), the symmetric spectral factor $H(z)$ is found by selecting the roots of $R(z)$ as discussed in Remark 6.4.

Example 6.7 We take $n = 29$ and $\omega_s = 0.55\pi$. Firstly, we solve (6.38) for $\gamma_s =$ 1, i.e., for a large value that makes the constraint (6.37) inactive. The frequency response of this least squares optimized symmetric FB is shown in Fig. 6.6. The maximum stopband error is -17 dB. Then, we put $\gamma_s = -21$ dB and solve (6.38) again. The new frequency response is shown in Fig. 6.7. The stopband bound is visibly respected. ■

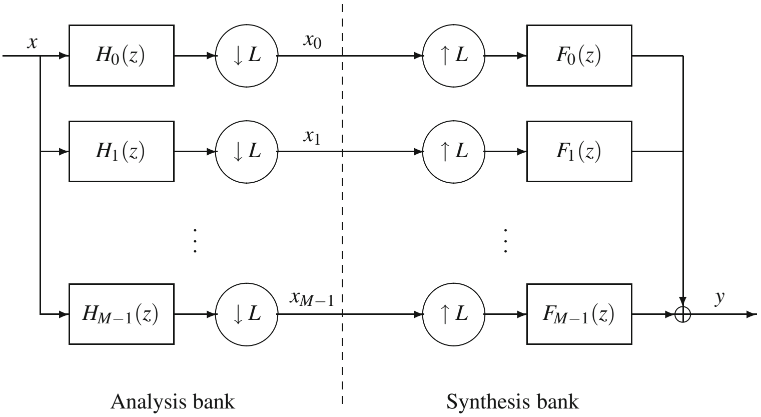


Fig. 6.8 M-channel filterbank

6.2 GDFT Modulated Filterbanks

We move now to filterbanks with more than two channels. An M-channel FB is shown in Fig. 6.8. Each channel has the same structure as in the two-channel case, but now the subband signals are obtained by downsampling with a factor of L. The same factor is used for upsampling. If $L = M$, then the FB is *critically* sampled. If $L < M$, the FB is *oversampled*. The FB has the perfect reconstruction property if (6.1) holds, where Δ is the delay.

6.2.1 GDFT Modulation: Definitions and Properties

Although in principle the analysis and synthesis filters can have independent coefficients, good performance and a simpler implementation can be obtained by building all the filters of a bank from a single prototype, by modulation. We discuss here a single type of modulation, that gives filters with complex coefficients; the main application of such FBs is in subband adaptive filtering. Let $H(z)$ and $F(z)$ be FIR prototype filters of degree n, with real coefficients (the degrees of the two filters can be different, but this does not change the developments below). The impulse responses of the filters from the FB are obtained by *generalized DFT* (GDFT) modulation, with

$$\begin{aligned} H_m(z) &= \textstyle\sum_{k=0}^{n} h_k \mathrm{e}^{j\pi(2m+1)(k-\Delta/2)/M} z^{-k}, \\ F_m(z) &= \textstyle\sum_{k=0}^{n} f_k \mathrm{e}^{j\pi(2m+1)(k-\Delta/2)/M} z^{-k}, \end{aligned} \tag{6.39}$$

where h_k, f_k, $k = 0 : n$, are the coefficients of the prototype filters.

Ideally, the analysis filters have the frequency responses from Fig. 6.9 (the synthesis filters have similar responses). The prototype filter $H(z)$ is lowpass, and since

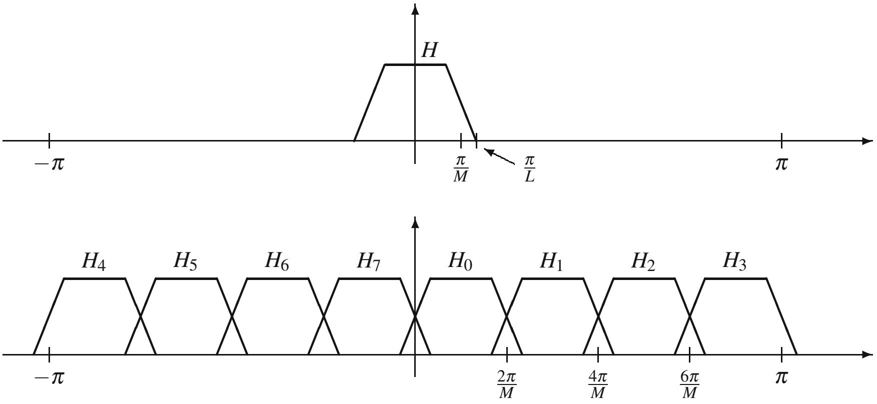


Fig. 6.9 Frequency responses (magnitude) of a prototype (*up*) and of the corresponding analysis filters (*down*), for $M = 8$

it has real coefficients, its magnitude response is symmetric with respect to $\omega = 0$. The responses of the analysis filters are obtained by shifting the response of the prototype and are asymmetric. The passband of a filter has (ideally) a width of $2\pi/M$, covering, for $H_m(z)$, the interval $[2m\pi/M, 2(m+1)\pi/M]$.

The advantage of oversampling is evident from Fig. 6.9. If the frequency response of the prototype filter is, like there, equal to zero outside the baseband $[-\pi/L, \pi/L]$, then the subband signals $x_m, m = 0 : M-1$, are not affected by aliasing. In this ideal situation, the processing of these signals (not shown in Fig. 6.8) uses only information that genuinely belongs to the respective frequency channels. In contrast, in critically sampled FBs, subband aliasing cannot be avoided and perfect reconstruction is realized by canceling the aliased components in the output. Subband processing and aliasing cancellation are usually independent processes and so PR is not a robust property of critically sampled FBs. Certainly, in practice one cannot have ideal frequency responses like those from Fig. 6.9, and so a certain amount of subband aliasing is unavoidable. However, oversampled FBs offer the potential of restricting the aliasing. Before examining the objectives for designing oversampled GDFT FBs, let us see first the relevant input–output relations.

The output of the FB from Fig. 6.8 is

$$Y(z) = T_0(z)X(z) + \sum_{\ell=1}^{L-1} T_\ell(z)X(ze^{-j2\pi\ell/L}), \tag{6.40}$$

where

$$T_0(z) = \frac{1}{L}\sum_{m=0}^{M-1} H_m(z)F_m(z) \tag{6.41}$$

is the distortion transfer function and

$$T_\ell(z) = \frac{1}{L} \sum_{m=0}^{M-1} H_m(z\mathrm{e}^{-j2\pi\ell/L}) F_m(z), \quad \ell = 1 : L-1, \tag{6.42}$$

are the aliasing transfer functions.

Proposition 6.8 *The distortion transfer function (6.41) of the FB generated by (6.39) has the expression*

$$T_0(z) = \frac{M}{L} \sum_{i,\ 0 \le \Delta + iM \le 2n} (-1)^i (\boldsymbol{h}^T \boldsymbol{\Upsilon}_{\Delta+iM} \boldsymbol{f}) z^{-(\Delta+iM)}, \tag{6.43}$$

where $\boldsymbol{h}$, $\boldsymbol{f}$ *are the vectors containing the coefficients of the analysis and synthesis prototypes, respectively (of length* $n+1$*).*

Proof Using the modulation expressions (6.39), it results that

$$H_m(z) = \boldsymbol{h}^T \begin{bmatrix} \vdots \\ \mathrm{e}^{j\pi(2m+1)(k-\Delta/2)/M} z^{-k} \\ \vdots \end{bmatrix}_{k=0:n}.$$

Hence, a term of the distortion transfer function (6.41) can be written as

$$H_m(z) F_m(z) = \boldsymbol{f}^T \boldsymbol{\Gamma}_m(z) \boldsymbol{h},$$

where $\boldsymbol{\Gamma}_m(z)$ is the Hankel matrix

$$\boldsymbol{\Gamma}_m(z) = \sum_{k=0}^{2n} \mathrm{e}^{j\pi(2m+1)(k-\Delta)/M} \boldsymbol{\Upsilon}_k z^{-k}. \tag{6.44}$$

Using the above relations, the distortion transfer function (6.41) becomes

$$\begin{aligned} T_0(z) &= \frac{1}{L} \boldsymbol{f}^T \left(\sum_{k=0}^{2n} \sum_{m=0}^{M-1} \mathrm{e}^{j\pi(2m+1)(k-\Delta)/M} \boldsymbol{\Upsilon}_k z^{-k} \right) \boldsymbol{h} \\ &= \frac{1}{L} \boldsymbol{f}^T \left(\sum_{k=0}^{2n} \mathrm{e}^{j\pi(k-\Delta)/M} \boldsymbol{\Upsilon}_k z^{-k} \sum_{m=0}^{M-1} \mathrm{e}^{j2\pi m(k-\Delta)/M} \right) \boldsymbol{h}. \end{aligned} \tag{6.45}$$

Since

$$\sum_{m=0}^{M-1} \mathrm{e}^{j2\pi m(k-\Delta)/M} = \begin{cases} M, & \text{if } (k-\Delta) \bmod M = 0, \\ 0, & \text{otherwise}, \end{cases}$$

and $\mathrm{e}^{j\pi(k-\Delta)/M} = (-1)^{(k-\Delta)/M}$ when $(k-\Delta) \bmod M = 0$, the expression (6.45) of the distortion transfer function is equivalent to (6.43). ■

It is clear from (6.40) that perfect reconstruction is achieved if $T_0(z) = z^{-\Delta}$ and $T_\ell(z) = 0$, $\ell = 1 : L - 1$. However, in the applications that use GDFT-modulated FBs, PR is not important and so we will consider FBs with near-PR property.

GDFT-modulated FBs as defined by (6.39) are described by the coefficients of *two* prototype filters; such FBs are named near-biorthogonal or two-prototype FBs. From now on, we consider GDFT-modulated FBs that are defined by a single prototype filter.

Definition 6.9 A GDFT-modulated FB is named *near-orthogonal* if the prototype filters appearing in (6.39) are related by

$$f_k = h_{n-k}. \tag{6.46}$$

(The word "near" is used in this definition because only PR FBs can be orthogonal.) ■

This class of FBs can be related to positive polynomials, as we will soon see.

Remark 6.10 For near-orthogonal FBs, it follows from (6.46) that

$$\boldsymbol{\Upsilon}_k \boldsymbol{f} = \boldsymbol{\Theta}_{n-k} \boldsymbol{h}.$$

Choosing a delay $\Delta = n$, the distortion transfer function (6.43) becomes

$$\begin{aligned} T_0(z) &= \frac{M}{L} z^{-n} \sum_{i,\ |i|M \le n} (-1)^i (\boldsymbol{h}^T \boldsymbol{\Theta}_{iM} \boldsymbol{h}) z^{-iM} \\ &= \frac{M}{L} z^{-n} \sum_{i,\ |i|M \le n} (-1)^i r_{iM} z^{-iM}, \end{aligned} \tag{6.47}$$

where $R(z)$ is the product filter (6.8); we have used (again!) the relation (2.27). We also remark that it is not enough to take $r_{iM} = (L/M)\delta_i$ to obtain perfect reconstruction, since aliasing is not necessarily canceled by this choice (like in the two-channel case). ■

6.2.2 Design of Near-Orthogonal GDFT-Modulated FBs

The optimization of near-orthogonal GDFT-modulated FBs is based on several objectives that can be expressed by conditions on the prototype filter.

- A good *frequency selectivity* of the filters is achieved by imposing conditions on the frequency response of the prototype filter only (since all responses are shifted versions of this response). Typically, we desire

$$\text{minimization of the stopband energy (5.2)} \tag{6.48}$$

and

$$|H(\omega)| \le \gamma_s, \quad \forall \omega \in [\omega_s, \pi], \tag{6.49}$$

where γ_s is a given bound and the stopband edge is usually

$$\omega_s = \pi / L, \tag{6.50}$$

as explained later.

- *Near perfect reconstruction* is obtained by putting conditions on the distortion transfer function (6.47) and also on the aliasing in the output.
- As explained in the beginning of the section, it is desirable to reduce the *aliasing in subbands*, such that subband processing uses only relevant information. However, since we aim to a general analysis, we will assume that there is no processing of subband signals.

The conditions (6.48) and (6.49) can be easily expressed using the product filter (6.8), as we have seen previously. We discuss now the other objectives. We assume (with no loss of generality) that the energy of the prototype filter is normalized to L/M and so we have

$$r_0 = \sum_{k=0}^{n} h_k^2 = \frac{L}{M}. \tag{6.51}$$

Distortion in the output. The deviation of the distortion transfer function (6.41) from the ideal value z^{-n} (remind that the delay of our FB is $\Delta = n$) can be measured by H_2 error norm

$$E_d = \frac{1}{\pi} \int_0^{\pi} |e^{-jn\omega} - T_0(\omega)|^2 d\omega.$$

Using (6.47), (6.51) and Parseval's theorem, we obtain

$$E_d = \frac{M}{L} \frac{1}{\pi} \int_0^{\pi} \left| \sum_{i \ne 0} (-1)^i r_{iM} \mathrm{e}^{-jiM\omega} \right|^2 d\omega = \frac{M}{L} \sum_{i \ne 0} r_{iM}^2 \tag{6.52}$$

This least squares error can be bounded by an imposed value.

Aliasing in the output. The frequency response of the aliasing transfer functions (6.42) is obtained from products between the frequency response of the prototype and a shifted version of itself (the shift being a multiple of $2\pi/L$). If the response of the prototype would be zero outside the baseband $[-\pi/L, \pi/L]$, then each such product and thus the aliasing transfer functions would be zero. Although this is not possible, it suggests that instead of seeking possible cancellations in the expressions of the aliasing transfer functions, we could reduce aliasing by imposing limitations on $H(\omega)$ for frequencies outside the baseband. That is, the natural conditions (6.48) and (6.49), with the choice (6.50), serve well to bound aliasing in the output.

Subband aliasing. The Fourier transform of the m-th subband signal is

$$X_m(\omega) = \frac{1}{L}\sum_{\ell=0}^{L-1} H_m\left(\frac{\omega - 2\pi\ell}{L}\right) X\left(\frac{\omega - 2\pi\ell}{L}\right),$$

i.e., it is a sum of L copies of $H_m(\omega)X(\omega)$, shifted with multiples of $2\pi/L$ and expanded L times. If $H_m(\omega)$ is zero outside an interval $\mathcal{I}$ of length $2\pi/L$ (centered in $(2m+1)\pi/M$, as in Fig. 6.9), then $X_m(\omega)$, on $[-\pi, \pi]$, contains essentially the same information as $X(\omega)$, on the interval $\mathcal{I}$. In this ideal case, there is no aliasing. To approximate the ideal, we can minimize the energy of $H_m(\omega)$ outside of $\mathcal{I}$. This is equivalent to minimizing the energy of the prototype $H(\omega)$ outside the baseband $[-\pi/L, \pi/L]$. We conclude that bounding subband aliasing can be realized by imposing the same conditions (6.48) and (6.49).

Optimization problem. We can now formulate the design problem. The design data are the number of channels M, the downsampling factor L, the degree n of the prototype (which is not a multiple of M), a distortion error bound γ_d and a stopband error bound γ_s. The optimization problem consists of the minimization of the stopband energy, subject to the constraints discussed above, i.e.,

$$\begin{array}{ll} \displaystyle\min_{R\in\mathbb{R}_n[z]} & \sum_{k=0}^{n} c_k r_k \\ \text{s.t.} & r_0 = L/M \\ & \sum_{i\neq 0} r_{iM}^2 \leq \gamma_d \\ & R(\omega) \leq \gamma_s^2, \quad \forall\omega \in [\pi/L, \pi] \\ & R(\omega) \geq 0, \quad \forall\omega \end{array} \tag{6.53}$$

This problem can be routinely transformed into an SQLP problem, as it is expressed in terms of nonnegative polynomials. The objective (stopband energy) and the first constraint (prototype energy normalization (6.51)) are linear. The second constraint (bound for the distortion error (6.52)) has a second order cone form. The third constraint is the stopband error bound (6.49) expressed in terms of the squared magnitude $R(\omega) = |H(\omega)|^2$. After solving the SQLP problem equivalent to (6.53), the prototype filter is found from the spectral factorization of the optimal product filter $R(z)$.

Example 6.11 We design a GDFT modulated filter bank with $M = 8$ channels and downsampling factor $L = 6$. The degree of the prototype filter is $n = 47$. The distortion error bound is $\gamma_d = 10^{-6}$ and the stopband error bound is given a large value (e.g., $\gamma_s = 1$) such that the third constraint from (6.53) is actually eliminated. After solving the SQLP version of (6.53), we obtain the prototype filter whose frequency response is shown in Fig. 6.10. The frequency response of the whole analysis bank (obtained by the modulation (6.39)) is shown in Fig. 6.11. We note that, for real signals, only the first $M/2$ filters have to be used (whose passbands cover $[0, \pi]$). Finally, the distortion error function $z^{-n} - T_0(z)$, where the distortion transfer function $T_0(z)$ is given by (6.47), has the frequency response shown in Fig. 6.12. The

Fig. 6.10 Frequency response of prototype filter from Example 6.11

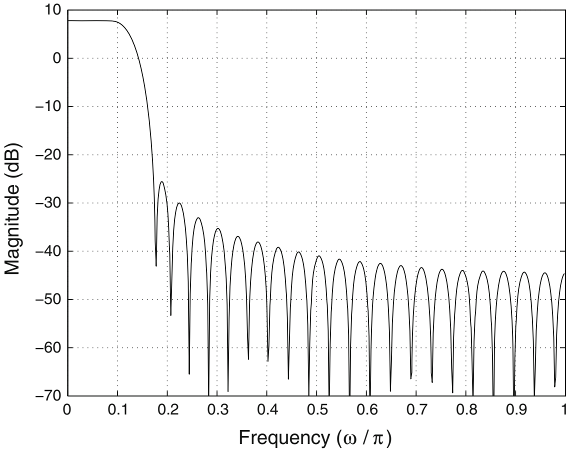
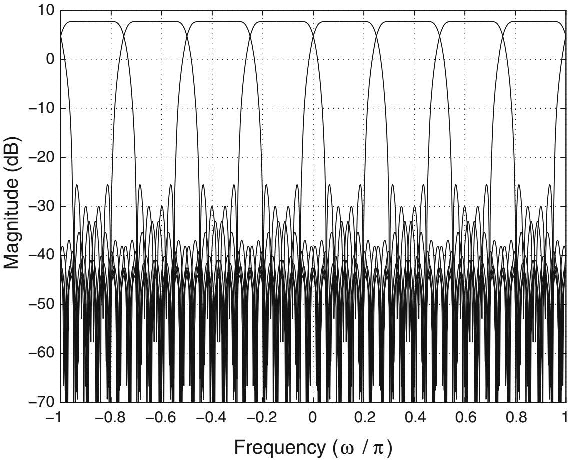


Fig. 6.11 Frequency responses of the analysis filters



maximum error is about −50dB. Alternatively, we can minimize an H_∞ error, as suggested in problem **P** 6.6. ■

6.3 Bibliographical and Historical Notes

The design of two-channel orthogonal filterbanks was the signal processing topic that led to the formulation and use of the trace parameterization. The first approaches for solving the compaction filter design problem (6.15) were based on semi-infinite optimization [2] or the Kalman–Yakubovich–Popov lemma [3]. The trace parameterization has been used in [4] (in a dual, not explicit, form) and [5] (for a related

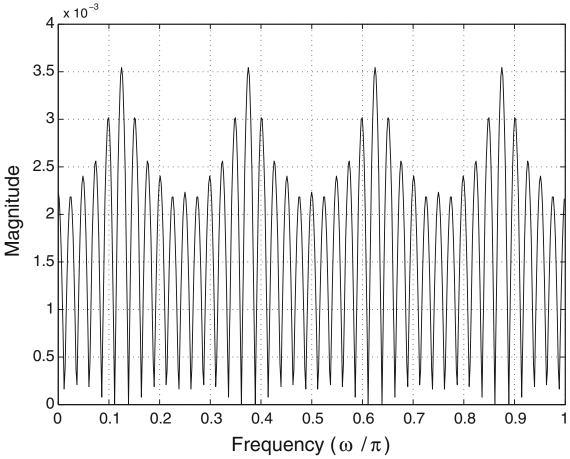


Fig. 6.12 Frequency response of the distortion error function

problem, that of pulse amplitude modulation; this problem belongs to a larger class, that of multiplexing, where the synthesis bank comes before the analysis one).

The use of positive polynomials for designing signal-adapted wavelets was proposed in [1]. The implicit expression of regularity constraints appeared in [1, 6]. A similar approach could be used for the optimization of other types of wavelets, as those from [7, 8].

The design of symmetric orthogonal complex-valued FBs and wavelets was discussed in many papers, among which [9, 10]. The use of positive polynomials that have a symmetric spectral factor was employed only more recently [11]; there, a parameterization equivalent to that from Theorem 6.5 was proposed, precisely the one found in problem **P** 6.5b.

The use of GDFT-modulated FBs for adaptive filters was advocated in [12–14], among others. The design of near-orthogonal GDFT-modulated FBs, as presented in Sect. 6.2, was proposed in [15]. Other papers [16, 17] deal with the design of biorthogonal GDFT-modulated FBs, preferred for their low-delay potential. In [18], biorthogonal FBs are designed with an iterative method, initialized with a near-orthogonal FB.

Problems

P 6.1 Show that the orthogonal two-channel FB design problem (6.15) can be written in the SDP form

$$\begin{array}{ll} \min\limits_{\boldsymbol{Q}} & \operatorname{tr}[\boldsymbol{C}\boldsymbol{Q}] \\ \text{s.t.} & \operatorname{tr}[\boldsymbol{\Theta}_{2k}\boldsymbol{Q}] = \delta_k, \quad k = 0 : (n-1)/2 \\ & \boldsymbol{Q} \succeq 0 \end{array} \tag{6.54}$$

where $\boldsymbol{C}$ is a Toeplitz matrix. (The coefficients of $R(z)$ from (6.8) are defined by the trace parameterization (2.6).)

Find a similar SDP problem, using the Gram pair parameterization (2.94).

P 6.2 Inspired by the previous problem, find the most economical SDP form of the two-channel FB design problem (6.28), which includes regularity constraints.

P 6.3 Consider a two-channel orthogonal FB with whose first analysis filter $H_0(z)$ is symmetric (and complex-valued). Show that $|H_0(1)| = \sqrt{2}$. (Hint: use power complementarity and the location of the roots of $H_0(z)$.)

P 6.4 Propose other partitions of the complex plane respecting the rule that generates Fig. 6.5. What can you say about the roots that belong to the border between the two regions? Have they always even multiplicity?

P 6.5 (Nonnegative trigonometric polynomials with symmetric or antisymmetric spectral factors—completion of Theorem 6.5)

a. We consider all the cases of nonnegative trigonometric polynomials $R \in \mathbb{R}[z]$ of order n with a symmetric or antisymmetric spectral factor $H(z)$ (with complex coefficients):
(i) n even ($n = 2\tilde{n}$), $H(z)$ symmetric ($h_k = h_{n-k}$),
(ii) n odd ($n = 2\tilde{n} + 1$), $H(z)$ symmetric,
(iii) n even, $H(z)$ antisymmetric ($h_k = -h_{n-k}$),
(iv) n odd, $H(z)$ antisymmetric.
Show that these polynomials can be parameterized by the Gram equality

$$R(\omega) = \boldsymbol{x}^T(\omega)\boldsymbol{Q}\boldsymbol{x}(\omega), \tag{6.55}$$

where the positive semidefinite matrix $\boldsymbol{Q}$ and the basis vector $\boldsymbol{x}(\omega)$ are:
(i) $\boldsymbol{Q} \in \mathbb{R}^{(\tilde{n}+1)\times(\tilde{n}+1)}$, $\boldsymbol{x}(\omega) = \boldsymbol{\chi}_c(\omega)$ (2.89),
(ii) $\boldsymbol{Q} \in \mathbb{R}^{(\tilde{n}+1)\times(\tilde{n}+1)}$, $\boldsymbol{x}(\omega) = \tilde{\boldsymbol{\chi}}_c(\omega)$ (2.98),
(iii) $\boldsymbol{Q} \in \mathbb{R}^{\tilde{n}\times\tilde{n}}$, $\boldsymbol{x}(\omega) = \boldsymbol{\chi}_s(\omega)$ (2.90),
(iv) $\boldsymbol{Q} \in \mathbb{R}^{(\tilde{n}+1)\times(\tilde{n}+1)}$, $\boldsymbol{x}(\omega) = \tilde{\boldsymbol{\chi}}_s(\omega)$ (2.99).

b. Remind that if $\tilde{H}(z)$ is an FIR filter of type ii–iv, then it can be written as, respectively, $(1 + z^{-1})H(z)$, $(1 + z^{-1})(1 - z^{-1})H(z)$, $(1 - z^{-1})H(z)$, where $H(z)$ is of type i). Extend this property to the corresponding product filter and show how to derive parameterizations for types ii–iv, starting from the relation (6.55) for type (i). For example, show that if $\tilde{R}(z)$ is of type ii), then $\tilde{R}(\omega) = (1 + \cos\omega)\boldsymbol{\chi}_c^T(\omega)\tilde{\boldsymbol{Q}}\boldsymbol{\chi}_c(\omega)$. Is this relation identical with (6.31)? (The answer is negative: the matrices $\boldsymbol{Q}$ from (6.31) and $\tilde{\boldsymbol{Q}}$ are not identical!)

P 6.6 (Near-PR orthogonal FBs)

a. Consider an orthogonal two-channel FB defined by the first analysis filter $H(z)$ of degree $n = 2\tilde{n} + 1$. Let $R(z)$ be the associated product filter (6.8), normalized such that $r_0 = 1$. Derive the PR error measure

$$E_d = \frac{1}{4\pi}\int_0^{\pi} |2 - R(\omega) - R(\pi - \omega)|^2 d\omega = 2\sum_{k=1}^{\tilde{n}} r_{2k}^2.$$

Given a PR error bound γ_d, append the near-PR condition $E_d \le \gamma_d$ to a FB design problem in SDP form, like (6.38) or (6.54).

b. Another near-PR condition can be $|e^{-jn\omega} - T_d(\omega)| \le \gamma_t$, for a given bound γ_t, where $T_d(z)$ is the distortion transfer function (6.9). Show how this constraint can be imposed using positive polynomials.

We denote $\tilde{R}(z) = \sum_{k=-\tilde{n}}^{\tilde{n}} r_{2k} z^{-k}$. Show that the above near-PR condition is equivalent to $|1 - \tilde{R}(\omega)| \le \gamma_t, \forall\omega \in [0, \pi/2]$. What is the advantage of this form of the condition?

Derive a similar result for GDFT-modulated FBs.

References

1. J.K. Zhang, T.N. Davidson, K.M. Wong, Efficient design of orthonormal wavelet bases for signal representation. IEEE Trans. Signal Process. **52**(7), 1983–1996 (2004)
2. P. Moulin, M. Aniţescu, K. Kortanek, F. Potra, The Role of Linear Semi-Infinite Programming in Signal-Adapted QMF Bank Design. IEEE Trans. Signal Process. **45**(9), 2160–2174 (1997)
3. J. Tuqan, P.P. Vaidyanathan, A state space approach to the design of globally optimal FIR energy compaction filters. IEEE Trans. Signal Process. **48**(10), 2822–2838 (2000)
4. C. Popeea, B. Dumitrescu, Optimal compaction gain by eigenvalue minimization. Signal Process. **81**(5), 1113–1116 (2001)
5. T.N. Davidson, Efficient design of waveforms for robust pulse amplitude modulation. IEEE Trans. Signal Process. **49**(12), 3098–3111 (2001)
6. B. Dumitrescu, C. Popeea, Accurate computation of compaction filters with high regularity. IEEE Signal Proc. Lett. **9**(9), 278–281 (2002)
7. I.W. Selesnick, The design of approximate hilbert transform pairs of wavelet bases. IEEE Trans. Signal Process. **50**(5), 1144–1152 (2002)
8. I.W. Selesnick, A higher density discrete wavelet transform. IEEE Trans. Signal Process. **54**(8), 3039–3048 (2006)
9. X.-P. Zhang, M.D. Desai, Y.-N. Peng, Orthogonal complex filter banks and wavelets: some properties and design. IEEE Trans. Signal Process. **47**(4), 1039–1048 (1999)
10. X.Q. Gao, T.Q. Nguyen, G. Strang, A study of two-channel complex-valued filterbanks and wavelets with orthogonality and symmetry properties. IEEE Trans. Signal Process. **50**(4), 824–833 (2002)
11. H.H. Kha, H.D. Tuan, B. Vo, T.Q. Nguyen, Symmetric Orthogonal Complex-Valued Filter Bank Design by Semidefinite Programming, *ICASSP*, volume 3 (Toulouse, France, 2006), pp. 221–224
12. B. Farhang-Boroujeny, Z. Wang, Adaptive filtering in subbands: design issues and experimental results for acoustic echo cancellation. Signal Process. **61**, 213–223 (1997)
13. M. Harteneck, S. Weiss, R.W. Stewart, Design of near perfect reconstruction oversampled filter banks for subband adaptive filters. IEEE Trans. Circ. Syst. II **46**(8), 1081–1085 (1999)
14. J.P. Reilly, M. Wilbur, M. Seibert, N. Ahmadvand, The complex subband decomposition and its applications to the decimation of large adaptive filtering problems. IEEE Trans. Signal Proc. **50**(11), 2730–2743 (2002)
15. M.R. Wilbur, T.N. Davidson, J.P. Reilly, Efficient design of oversampled NPR GDFT filterbanks. IEEE Trans. Signal Proc. **52**(7), 1947–1963 (2004)

16. K. Eneman, M. Moonen, DFT modulated filter bank design for oversampled subband systems. Signal Process. **81**, 1947–1973 (2001)
17. J.M. de Haan, N. Grbic, I. Claesson, S.E. Nordholm, Filter bank design for subband adaptive microphone array. IEEE Trans. Speech Audio Proc. **11**(1), 14–23 (2003)
18. Dumitrescu, B., Bregović, R., Saramäki, T.: Simplified design of low-delay oversampled NPR GDTF filterbanks. In: IEEE Conference Acoustics, Speech, Signal Processing ICASSP

Chapter 7
Stability

Abstract Stability is a basic property of dynamic systems. In this chapter, we explore several issues related to the stability of multidimensional discrete-time systems. First come stability tests: Given a system, we have to decide whether it is stable or not. Then, we discuss a robust stability problem, for the case where the coefficients of the system depend polynomially on some bounded parameters. Finally, we show how to build a convex stability domain around a given stable system. For all these problems, the solutions we present are based on the use of positive polynomials.

7.1 Multidimensional Stability Tests

The d-dimensional discrete-time system with transfer function

$$H(z) = \frac{B(z)}{A(z)}, \tag{7.1}$$

where the denominator is the (anticausal) positive orthant polynomial

$$A(z) = \sum_{\boldsymbol{k}=\boldsymbol{0}}^{\boldsymbol{n}} a_{\boldsymbol{k}} z^{\boldsymbol{k}} \tag{7.2}$$

is structurally stable if and only if

$$A(z) \neq 0, \quad \text{for } |z_1| \leq 1, \ldots, |z_d| \leq 1. \tag{7.3}$$

(Note that here, by exception to the rest of the book, we follow the traditional notation in stability and work with anticausal filters.) This definition of stability eliminates stable systems with nonessential singularities of the second kind, where the numerator and the denominator may become simultaneously zero on the unit d-circle; anyway, these systems have no practical importance since an infinitely small perturbation of the numerator may make them unstable.

B. Dumitrescu, *Positive Trigonometric Polynomials and Signal Processing Applications*, Signals and Communication Technology,
DOI 10.1007/978-3-319-53688-0_7

Multidimensional stability testing is an NP-hard problem. Moreover, the definition above does not look amenable to an implementation form. One of the simplifications of (7.3) is the DeCarlo–Strintzis test, saying that the system (7.1) is structurally stable if and only if the following univariate polynomials

$$\begin{array}{c} A(z_1, 1, \ldots, 1) \\ A(1, z_2, \ldots, 1) \\ \vdots \\ A(1, 1, \ldots, z_d) \end{array} \tag{7.4}$$

have no roots inside and on the unit circle and

$$A(z_1, \ldots, z_d) \neq 0, \quad \text{for } |z_1| = \ldots = |z_d| = 1. \tag{7.5}$$

Only condition (7.5) is difficult to implement, since it involves all d variables. In the rest of this section, we assume that the 1D conditions hold true and discuss only the multivariate condition (7.5), for whose implementation we propose two algorithms: The first is based on testing the positivity of a polynomial, while the second uses a special form of Positivstellensatz.

7.1.1 Stability Test via Positivity

Since we have to test if $A(\boldsymbol{\omega}) \neq 0$, $\forall \boldsymbol{\omega} \in [-\pi, \pi]^d$, a simple alternative is to test if $|A(\boldsymbol{\omega})| > 0$. We can transform this into a problem with positive trigonometric polynomials by defining

$$R(z) = A(z)A(z^{-1}) \tag{7.6}$$

and checking if $R(\boldsymbol{\omega}) > 0$. This can be done by computing the minimum value of $R(\boldsymbol{\omega})$ on the unit d-circle, i.e., by solving the problem (3.46). Of course, as discussed in Sect. 3.5, we can solve this problem only in relaxed form and find a minimum value $\mu_{\boldsymbol{m}}^\star$ (where $\boldsymbol{m}$ is the degree of the relaxation) that is smaller than the true minimum $\mu^\star$. If $\mu_{\boldsymbol{m}}^\star > 0$, then the system is stable. If $\mu_{\boldsymbol{m}}^\star = 0$, then we decide that the system is not stable, although it is possible that $\mu^\star > 0$. So, the test implements only a sufficient condition. In view of our experience with relaxations, we can safely presume that the test is practically accurate even with relaxations of smallest degree ($\boldsymbol{m} = \boldsymbol{n}$).

Another approach that leads to a feasibility SDP problem is to choose a constant $\varepsilon > 0$ and test whether $R(z) - \varepsilon$ is strictly positive; again, this can be implemented only in a relaxed version, by checking whether $R(z) - \varepsilon$ is sum-of-squares.

An important aspect of the decision is related to the numerical computation. The numerical value $\mu_{\boldsymbol{m}}^\star$ computed by an SDP solver will be never exactly zero, but very often positive. How small should $\mu_{\boldsymbol{m}}^{\star}$ be to say that it is actually zero? In the feasibility approach, how small should we choose ε?

It was found in [1] that the safest approach seems to be the following. Instead of finding the minimum value of $R(z)$, we solve the equivalent problem (3.38) of computing the most positive Gram matrix associated with $R(z)$; the minimum eigenvalue is $\lambda_{\boldsymbol{m}}^{\star}$, where again $\boldsymbol{m}$ is the degree of the relaxation. The SDP algorithms implemented in [2] and similar libraries solve the specified (primal) SDP problem *and its dual*; thus, they can return the gap δ between the values of the primal and dual optimal (numerical) values. Theoretically, the duality gap should be zero. Practically, the gap is nonzero and gives the order of magnitude of the accuracy of the computed optimal value of the problem. The polynomial $R(z)$ is deemed to be strictly positive if $\lambda_{\boldsymbol{m}}^{\star}$ is sufficiently large with respect to the gap. So, the stability test is as follows: if $\lambda_{\boldsymbol{m}}^{\star} \geq c\delta$, where, e.g., $c = 100$, then the system is stable. Otherwise, we decide that the system is unstable.

Example 7.1 A first example was already given in disguise. The system having as denominator the polynomial (3.40) is stable, since the minimum eigenvalue of the most positive Gram matrix associated with $R(z)$ is 0.3036, as seen in Example 3.16. We consider now polynomials with more than 2 variables:

$$A_1(z) = 5 + z_1^2 z_3^3 + z_3^3 z_4^2 + z_1^3 z_2 z_5 + z_1 z_2 z_3 z_4 z_5, \tag{7.7}$$

$$A_2(z) = 5.6 + 0.8 z_1 + 1.5 z_1^2 z_2 + 1.8 z_2^3 + 0.2 z_3 + 1.3 z_2 z_3^2. \tag{7.8}$$

The polynomial $A_1(z)$ is stable, since $\lambda_{\boldsymbol{n}}^{\star} = 0.0052$ and $\delta = 1.8 \cdot 10^{-13}$. The polynomial $A_2(z)$ was found to be unstable, with $\lambda_{\boldsymbol{n}}^{\star} = 1.1 \cdot 10^{-12}$ and $\delta = 6.8 \cdot 10^{-13}$; higher order relaxations give the same decision. However, a small perturbation of $A_2(z)$ may make it stable. For example, for the polynomial $A(0.9999z)$, we obtain $\lambda_{\boldsymbol{n}}^{\star} = 6.1 \cdot 10^{-8}$ and $\delta = 1.2 \cdot 10^{-12}$. This and other examples show that the test is indeed accurate. ■

However, the test may be costly. For the polynomial (7.7), the stability test took about 15 s (time measured in 2016), but the complexity grows quickly with the degree and number of variables. Since the polynomial has only few nonzero coefficients, a cure is to appeal to sparse bases for the Gram matrix, as in Sect. 3.6. The question is what set of indices $\mathcal{I}$ to use in (3.57). The first idea is to try a minimal (in the sense that its number of elements is minimum) set $\mathcal{I}_m$. For an arbitrary polynomial $R(z)$, finding $\mathcal{I}_m$ is a difficult task; however, in the case of (7.6), the minimal set is simply given by the degrees appearing in $A(z)$.

The minimal set $\mathcal{I}_m$ and the complete set $\mathcal{I}_c = \{\boldsymbol{k} \in \mathbb{N}^d \mid \boldsymbol{k} \leq \boldsymbol{n}\}$ are the extreme cases. Other sets could be used, based on heuristic choices, for example

$$\mathcal{I}_d = \{\boldsymbol{k} \in \mathbb{N}^d \mid \exists \boldsymbol{k}_m \in \mathcal{I}_m \text{ such that } \boldsymbol{k} \leq \boldsymbol{k}_m\}. \tag{7.9}$$

The monomials with degrees from $\mathcal{I}_d$ are divisors of the monomials that appear in $A(z)$. By construction, we have $\mathcal{I}_m \subset \mathcal{I}_d \subset \mathcal{I}_c$.

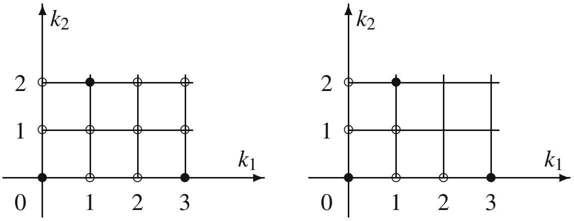


Fig. 7.1 Index sets for Example 7.2. *Left* complete set $\mathcal{I}_c$; bullets represent the basic set $\mathcal{I}_b$. *Right* divisors set $\mathcal{I}_d$

Example 7.2 Figure 7.1 presents the above sets of indices for the polynomial $A(z) = 2 + z_1^3 + z_1 z_2^2$. The basic set $\mathcal{I}_b$ is shown with bullets, and the complete set is on the left and the divisors set on the right. ■

Example 7.3 Using sparse bases, the polynomial (7.7) is still deemed stable. With the minimal set $\mathcal{I}_m$, the minimal eigenvalue is $\lambda_m^\star = 0.2$; with the divisors set $\mathcal{I}_d$, we obtain $\lambda_d^\star = 0.0189$. The execution time is 0.05 s for $\mathcal{I}_m$ and 1 s for $\mathcal{I}_d$, much smaller than for solving the complete problem. Of course, for (7.8) the decision is instability: If the computed minimal value of the polynomial is zero for the complete set, then it is zero for any index set included in it. ■

Other tests, performed on randomly generated polynomials and reported in [1], show that the test based on the minimal set often fails. On the contrary, the test based on the divisors set gave the correct decision in all experiments.

7.1.2 *Stability of Fornasini–Marchesini Model*

We assume now that the model of the multidimensional system is given not by the transfer function (7.1), but in a state-space form. We confine the discussion to the 2D case, although its generality will be clear. The Fornasini–Marchesini first model is

$$\begin{aligned}\boldsymbol{\xi}(\ell_1{+}1, \ell_2{+}1) &= \boldsymbol{A}_1\boldsymbol{\xi}(\ell_1{+}1, \ell_2) + \boldsymbol{A}_2\boldsymbol{\xi}(\ell_1, \ell_2{+}1) + \boldsymbol{A}_3\boldsymbol{\xi}(\ell_1, \ell_2) + \boldsymbol{B}\boldsymbol{v}(\ell_1, \ell_2)\\ \boldsymbol{\eta}(\ell_1, \ell_2) &= \boldsymbol{C}\boldsymbol{\xi}(\ell_1, \ell_2)\end{aligned} \tag{7.10}$$

This is the most general description of a linear 2D system; other models such as Roesser or Attasi can be brought to this form. The vectors $\boldsymbol{\xi} \in \mathbb{R}^n$, $\boldsymbol{v}$, and $\boldsymbol{\eta}$ represent the state, the input and the output of the system, respectively. We denote

$$\boldsymbol{F}(z_1, z_2) = \boldsymbol{I}_n - \boldsymbol{A}_1 z_1 - \boldsymbol{A}_2 z_2 - \boldsymbol{A}_3 z_1 z_2. \tag{7.11}$$

The system (7.10) is stable if $A(\boldsymbol{z}) = \det \boldsymbol{F}(\boldsymbol{z})$ respects the condition (7.3). So, the methods described in Sect. 7.1.1 are applicable. However, we can work directly with (7.11), which is a polynomial with matrix coefficients; the advantage is the low degree of this polynomial.

The system (7.10) is stable if $\boldsymbol{F}(\boldsymbol{z})$ has no zero eigenvalues for $\boldsymbol{z}$ inside or on the unit bicircle. The univariate DeCarlo–Strintzis conditions say that the polynomials

$$\begin{gathered}\det(\boldsymbol{I} - \boldsymbol{A}_2 - z(\boldsymbol{A}_1 + \boldsymbol{A}_3)),\\ \det(\boldsymbol{I} - \boldsymbol{A}_1 - z(\boldsymbol{A}_2 + \boldsymbol{A}_3))\end{gathered}$$

must be stable. These conditions can be expressed in several ways, for example by

$$\begin{gathered}\rho((\boldsymbol{I} - \boldsymbol{A}_2)^{-1}(\boldsymbol{A}_1 + \boldsymbol{A}_3)) < 1,\\ \rho((\boldsymbol{I} - \boldsymbol{A}_1)^{-1}(\boldsymbol{A}_2 + \boldsymbol{A}_3)) < 1,\end{gathered}$$

where $\rho(\cdot)$ is the spectral radius (the largest modulus of an eigenvalue). (To these conditions, we typically add the natural constraints $\rho(\boldsymbol{A}_1) < 1$, $\rho(\boldsymbol{A}_2) < 1$ that come from the requirement that $\boldsymbol{F}(z_1, 0)$ and $\boldsymbol{F}(0, z_2)$ are stable.) The multivariate condition (7.5) means that the eigenvalues of $\boldsymbol{F}(\boldsymbol{\omega})$ must be nonzero. Equivalently, the smallest eigenvalue of $\boldsymbol{R}(\boldsymbol{\omega})$, where

$$\boldsymbol{R}(\boldsymbol{z}) = \boldsymbol{F}(\boldsymbol{z})\boldsymbol{F}^T(\boldsymbol{z}^{-1}),$$

must be strictly positive. As discussed in Sect. 7.1.1 in the scalar case, a stability test consists of computing the most positive Gram matrix associated with $\boldsymbol{R}(\boldsymbol{z})$ an testing its strict positivity. We simply have to solve the SDP problem (3.115) and see whether the numerical value $\lambda^\star_{\boldsymbol{m}}$ (for a degree of relaxation $\boldsymbol{m}$) is significantly greater than the optimality gap.

Example 7.4 Let us consider the Fornasini–Marchesini model with

$$\boldsymbol{A}_1 = \begin{bmatrix} 0.5 & -0.1 & 0.2\\ 0.1 & -0.2 & 0.3\\ -0.2 & 0.1 & 0.3 \end{bmatrix},\quad \boldsymbol{A}_2 = \begin{bmatrix} 0.1 & -0.4 & 0.2\\ 0.1 & -0.4 & 0.3\\ -0.2 & 0.2 & 0.2 \end{bmatrix},$$
$$\boldsymbol{A}_3 = \begin{bmatrix} 0.2 & -0.3 & 0.2\\ 0.2 & -0.2 & 0.3\\ -0.1 & 0.4 & 0.3 \end{bmatrix}.$$

Solving (3.115) with the smallest relaxation degree $\boldsymbol{m} = (1, 1)$, we obtain $\lambda^\star_{\boldsymbol{m}} = 3.0 \cdot 10^{-5}$ and a gap $\delta = 1.8 \cdot 10^{-13}$ and so the system is stable. Applying the scalar stability test from Sect. 7.1.1 to the $\boldsymbol{n}$-th order polynomial $\det \boldsymbol{F}(\boldsymbol{z})$, we obtain $\lambda^\star_{\boldsymbol{n}} = 1.4 \cdot 10^{-5}$ and $\delta = 6.2 \cdot 10^{-13}$. These values suggest that the matrix polynomial test is potentially more accurate, as the smallest eigenvalue of $\boldsymbol{R}(\boldsymbol{\omega})$ is typically larger than the smallest value of the scalar polynomial. The execution times for the two tests are of about 0.1 s. ■

In general, the matrix polynomial test described here is faster than the scalar one from Sect. 7.1.1. For the smallest relaxation degree, the Gram matrix of $\boldsymbol{R}(\boldsymbol{z})$

has size $4n \times 4n$, since the degree of the polynomial (7.11) is (1, 1) and the size of its coefficients is $n \times n$. For the scalar algorithm, the polynomial det $\boldsymbol{F}(z)$ has (typically) degree (n, n), and so the Gram matrix has size $(n+1)^2 \times (n+1)^2$. Hence, in Example 7.4, the size of the Gram matrix is 12×12 for the matrix polynomial test and 16×16 for the scalar coefficients tests. For a larger n, we expect smaller execution times of the matrix polynomial test.

7.1.3 Positivstellensatz for Testing Stability

We return now to the transfer function model (7.1) and derive a stability test based on the Positivstellensatz for trigonometric polynomials described in Sect. 4.4. We assume that the polynomial $A(z)$ has real coefficients and its degree is even ($\boldsymbol{n} = 2\tilde{\boldsymbol{n}}$). We denote

$$\tilde{A}(z) = z^{\boldsymbol{n}/2} A(z) = A_s(z) + A_a(z), \tag{7.12}$$

where the polynomials

$$A_s(z) = \frac{\tilde{A}(z) + \tilde{A}(z^{-1})}{2}, \quad A_a(z) = \frac{\tilde{A}(z) - \tilde{A}(z^{-1})}{2} \tag{7.13}$$

have degree $\tilde{\boldsymbol{n}}$, $A_s(z)$ is symmetric, and $A_a(z)$ is antisymmetric, i.e., $A_a(z^{-1}) = -A_a(z)$. It follows that

$$\tilde{A}(\boldsymbol{\omega}) = A_s(\boldsymbol{\omega}) + A_a(\boldsymbol{\omega}),$$

where

$$A_s(\boldsymbol{\omega}) = \boldsymbol{a}_s^T \boldsymbol{\chi}_c(\boldsymbol{\omega}), \quad A_a(\boldsymbol{\omega}) = j\boldsymbol{a}_a^T \boldsymbol{\chi}_s(\boldsymbol{\omega}). \tag{7.14}$$

The vectors $\boldsymbol{a}_s$ and $\boldsymbol{a}_a$ are real and the basis vectors $\boldsymbol{\chi}_c(\boldsymbol{\omega})$ and $\boldsymbol{\chi}_s(\boldsymbol{\omega})$ are those defined in (3.97). Since $A_s(\boldsymbol{\omega})$ is real and $A_a(\boldsymbol{\omega})$ purely imaginary, the second DeCarlo–Strintzis condition (7.5) is equivalent to the requirement that the set

$$\mathcal{D}_A \overset{\Delta}{=} \{\boldsymbol{\omega} \in [-\pi, \pi]^d \mid A_s(\boldsymbol{\omega}) = 0 \text{ and } A_a(\boldsymbol{\omega}) = 0\} \tag{7.15}$$

is empty. This formulation hints immediately to the use of the Positivstellensatz given by Theorem 4.39. The only apparent impediment is that the theorem is formulated in terms of Hermitian polynomials, while in (7.15), we have also the antisymmetric polynomial $A_a(z)$. However, we can derive a result in the same style that takes into account the different symmetries.

Theorem 7.5 (Positivstellensatz stability test)

The even degree polynomial $A(z)$ has no roots on the unit d-circle if and only if there exist a symmetric polynomial $X(z)$, an antisymmetric polynomial $Y(z)$, and a sum-of-squares polynomial $R(z)$, all with real coefficients, such that

$$1 + X(z)A_s(z) + Y(z)A_a(z) + R(z) = 0, \tag{7.16}$$

where $A_s(z)$ and $A_a(z)$ are defined by (7.12), (7.13).

Proof The polynomials $E_1(z) = A_s(z)$ and $E_2(z) = jA_a(z)$ are symmetric and Hermitian, respectively (note that $E_2^*(z^{-1}) = -jA_a(z^{-1}) = jA_a(z) = E_2(z)$), and $E_2(z)$ has purely imaginary coefficients. Applying Theorem 4.39, the set (7.15) is empty if and only if there exist polynomials $U_1(z)$ and $U_2(z)$ and sum-of-squares $S_0(z)$, all with complex coefficients, such that

$$1 + U_1(z)E_1(z) + U_2(z)E_2(z) + S_0(z) = 0. \tag{7.17}$$

Denoting $U_1(z) = U_{1r}(z) + jU_{1i}(z)$ etc., where $U_{1r}(z)$ and $U_{1i}(z)$ have real coefficients, it results that (7.17) is equivalent to (we omit the argument for readability)

$$1 + U_{1r}A_s - U_{2i}A_a + S_{0r} + j(U_{1i}A_s + U_{2r}A_a + S_{0i}) = 0.$$

The above equality shows that we can take $U_{1i}(z) = 0$, $U_{2r}(z) = 0$, $S_{0i}(z) = 0$ and so we obtain (7.16) with $X(z) = U_{1r}(z)$, $Y(z) = -U_{2i}(z)$, and $R(z) = S_{0r}(z)$; the latter polynomial is sum-of-squares as the real part of a sum-of-squares, see the end of the proof of Theorem 4.11. ■

The equality (7.16) can be checked computationally using either the trace or the Gram-pair parameterizations. We detail here the latter variant. Since $X(z)$ and $Y(z)$ have real coefficients and are symmetric and, respectively, antisymmetric, we can write

$$X(\boldsymbol{\omega}) = \boldsymbol{x}^T \boldsymbol{\chi}_c(\boldsymbol{\omega}), \quad Y(\boldsymbol{\omega}) = j\boldsymbol{y}^T \boldsymbol{\chi}_s(\boldsymbol{\omega}),$$

with real vectors $\boldsymbol{x}$ and $\boldsymbol{y}$. Indexing these vectors and those from (7.14) with a single d-dimensional index (from a half-space $\mathcal{H}$), we have

$$\begin{aligned} X(\boldsymbol{\omega})A_s(\boldsymbol{\omega}) &= \textstyle\sum_{\boldsymbol{i},\boldsymbol{\ell}\in\mathcal{H}} x_{\boldsymbol{i}} a_{s\boldsymbol{\ell}} \cos \boldsymbol{i}^T\boldsymbol{\omega} \cos \boldsymbol{\ell}^T\boldsymbol{\omega}, \\ Y(\boldsymbol{\omega})A_a(\boldsymbol{\omega}) &= -\textstyle\sum_{\boldsymbol{i},\boldsymbol{\ell}\in\mathcal{H}} y_{\boldsymbol{i}} a_{a\boldsymbol{\ell}} \sin \boldsymbol{i}^T\boldsymbol{\omega} \sin \boldsymbol{\ell}^T\boldsymbol{\omega}. \end{aligned}$$

Using the Gram-pair parameterization (3.99) for the sum-of-squares $R(\boldsymbol{\omega})$ and the trigonometric identities (2.77) and (2.78), it results that the Positivstellensatz (7.16) is equivalent to

$$\delta_{\boldsymbol{k}} + \operatorname{tr}[\boldsymbol{\Phi}_{\boldsymbol{k}}\boldsymbol{Q}] + \operatorname{tr}[\boldsymbol{\Lambda}_{\boldsymbol{k}}\boldsymbol{S}] + \frac{1}{2}\sum_{\substack{\boldsymbol{i}+\boldsymbol{\ell}=\boldsymbol{k}\\ \boldsymbol{i},\boldsymbol{\ell}\in\mathcal{H}}}(x_{\boldsymbol{i}}a_{s\boldsymbol{\ell}} + y_{\boldsymbol{i}}a_{a\boldsymbol{\ell}}) + \frac{1}{2}\sum_{\substack{\boldsymbol{i}-\boldsymbol{\ell}=\pm\boldsymbol{k}\\ \boldsymbol{i},\boldsymbol{\ell}\in\mathcal{H}}}(x_{\boldsymbol{i}}a_{s\boldsymbol{\ell}} - y_{\boldsymbol{i}}a_{a\boldsymbol{\ell}}) = 0, \tag{7.18}$$

for any $\boldsymbol{k} \in \mathcal{H}$ and for some positive semidefinite matrices $\boldsymbol{Q}$ and $\boldsymbol{S}$. (To simplify the notation, we have introduced the coefficients $y_{\boldsymbol{0}} = a_{a\boldsymbol{0}} = 0$.) We have thus reduced the stability test to a feasibility SDP problem; again, we can solve the problem only in relaxed form, by choosing the degrees of the variable polynomials $X(z)$, $Y(z)$ and the sum-of-squares $R(z)$, and thus the sizes of the Gram-pair matrices. Regarding the accuracy, this approach has two slight advantages over the method from Sect. 7.1.1 (based on computing the minimum value of $|A(\boldsymbol{\omega})|^2$):

- The coefficients of $A(z)$ are combined only by addition, in (7.13); "squaring," as necessary in the computation of $|A(\boldsymbol{\omega})|^2$, is avoided;
- The numerical accuracy is that of the SDP algorithm; it is not necessary to interpret output data (more or less heuristically) when deciding stability.

The complexity of the feasibility SDP problem based on (7.16) is that typical to the Gram-pair parameterization. The natural choice of the degrees is $\deg X = \deg Y = \tilde{\boldsymbol{n}}$; it results that $\deg R = \boldsymbol{n} = 2\tilde{\boldsymbol{n}}$ and the Gram-pair matrices have sizes (3.95), (3.96). With these degrees, the Positivstellensatz stability test never failed in our experiments. (In principle, we could try to satisfy (7.16) with polynomials of smaller degree; however, in our experiments with such degrees, the equality (7.16) could not be satisfied for many stable polynomials.)

Example 7.6 We give here a very simple example, illustrative to the accuracy of the test. Let us consider the 2D system with denominator

$$A(z_1, z_2) = [1 \; z_1 \; z_1^2] \begin{bmatrix} 1 & -0.8 & 0.5 \\ -0.5 & 0.4 & -0.25 \\ 1-\varepsilon & -0.8 & 0.5 \end{bmatrix} \begin{bmatrix} 1 \\ z_2 \\ z_2^2 \end{bmatrix}.$$

The system is stable for small positive values of ε (e.g., $\varepsilon \le 0.4$), but is unstable for $\varepsilon = 0$; in this case, the polynomial has the separable form $A(\boldsymbol{z}) = (1 - 0.5z_1 + z_1^2)(1 - 0.8z_2 + 0.5z_2^2)$ and the roots of the first factor are on the unit circle. The test based on Theorem 7.5 and described in this section decides that the system is stable for ε as small as 10^{-8}. For comparison, the test from Sect. 7.1.1 gives an instability decision for values $\varepsilon = 10^{-4}$ and smaller and a stability decision only for $\varepsilon = 2 \cdot 10^{-4}$. In general, for other examples, the accuracy is always in favor of the Positivstellensatz test. ■

7.2 Robust Stability

We turn now to a robust stability problem that, although formulated for 1D systems, has a multivariate nature. Let

$$A(z, \boldsymbol{q}) = \sum_{k=0}^{n} a_k(\boldsymbol{q}) z^k \tag{7.19}$$

be the denominator of a discrete-time transfer function. The coefficients $a_k(\boldsymbol{q})$ depend on p parameters $\boldsymbol{q} \in \mathcal{Q}$. The robust stability problem consists of deciding whether the polynomial (7.19) is Schur, i.e., has no roots inside or on the unit circle, for any $\boldsymbol{q} \in \mathcal{Q}$. We study here the case where the coefficients $a_k(\boldsymbol{q})$ depend *polynomially* on the parameters $\boldsymbol{q}$ and each parameter q_ℓ, $\ell = 1 : p$, is bounded by some constants; without loss of generality, we can consider $|q_\ell| \le 1$.

If the parameters are complex, then $\mathcal{Q} = \mathbb{D}^p$, where $\mathbb{D} = \{z \in \mathbb{C} \mid |z| \le 1\}$ is the unit disk. Deciding whether the polynomial (7.19) is Schur is simply a multidimensional stability test on the $p+1$-variate polynomial $A(\boldsymbol{z})$ in the variable $\boldsymbol{z} = (z, q_1, \dots, q_p)$. The tests discussed in the previous section can be applied without any restriction.

In the remainder of this section, we treat the real parameters case, where $\mathcal{Q} = [-1, 1]^p$. We will transform the robust stability problem into two different Positivstellensatz: one with real polynomials and the other with trigonometric polynomials.

7.2.1 Real Polynomial Test

Since we aim to obtain real polynomials, we replace the complex variable z with two real variables. Denoting $z = \tau_1 + j\tau_2$, with $\tau_1, \tau_2 \in \mathbb{R}$, we can transform the polynomial (7.19) into

$$A(z, \boldsymbol{q}) = f_1(\tau_1, \tau_2, \boldsymbol{q}) + j f_2(\tau_1, \tau_2, \boldsymbol{q}), \tag{7.20}$$

where the polynomials $f_1(\cdot)$ and $f_2(\cdot)$ have real coefficients and depend on $d = p+2$ real variables. The degrees of the variables τ_1 and τ_2 in $f_1(\cdot)$ and $f_2(\cdot)$ are at most n. In the variable

$$\boldsymbol{t} = (\tau_1, \tau_2, \boldsymbol{q}) \in \mathbb{R}^d,$$

the polynomials $A(\boldsymbol{t})$, $f_1(\boldsymbol{t})$, and $f_2(\boldsymbol{t})$ have the same total degrees.

Since $f_1(\boldsymbol{t})$ and $f_2(\boldsymbol{t})$ have real values, the polynomial (7.20) is robustly Schur if and only if $f_1(\boldsymbol{t}) \neq 0$, $f_2(\boldsymbol{t}) \neq 0$ for all $\boldsymbol{t} \in \mathbb{R}^d$ such that $\tau_1^2 + \tau_2^2 \le 1$ and $\boldsymbol{q} \in \mathcal{Q}$. Denoting

$$g_\ell(\boldsymbol{t}) = \begin{cases} 1 - q_\ell^2, & \ell = 1 : p, \\ 1 - \tau_1^2 - \tau_2^2, & \ell = p+1, \end{cases} \tag{7.21}$$

it results that the polynomial is robustly Schur if and only if the set

$$\mathcal{D}(f, g) = \left\{ \boldsymbol{t} \in \mathbb{R}^d \;\middle|\; \begin{array}{l} f_1(\boldsymbol{t}) = 0, \;\; f_2(\boldsymbol{t}) = 0 \\ g_\ell(\boldsymbol{t}) \ge 0, \;\; \ell = 1 : p+1 \end{array} \right\} \tag{7.22}$$

is empty. This formulation hints immediately to a Positivstellensatz expression of the robust stability test.

Theorem 7.7 *The polynomial (7.19) is Schur for any* $\boldsymbol{q} \in \mathcal{Q} = [-1, 1]^p$ *if and only if there exist* $u_1, u_2 \in \mathbb{R}[t]$ *and* $s_\ell \in \sum \mathbb{R}[\boldsymbol{t}]^2$, $\ell = 0 : p+1$, *such that*

$$1 + f_1(\boldsymbol{t})u_1(\boldsymbol{t}) + f_2(\boldsymbol{t})u_2(\boldsymbol{t}) + s_0(\boldsymbol{t}) + \sum_{\ell=1}^{p+1} g_\ell(\boldsymbol{t})s_\ell(\boldsymbol{t}) = 0. \tag{7.23}$$

Proof The polynomials (7.21) have the property

$$p + 1 - \sum_{\ell=1}^{d} t_\ell^2 = \sum_{\ell=1}^{p+1} g_\ell(\boldsymbol{t}) \cdot 1 \in \mathcal{M}(G), \tag{7.24}$$

where $\mathcal{M}(G)$ is defined in (4.4). It results that the polynomial (4.6) belongs to $\mathcal{M}(g)$, with $N = p + 1$; as discussed in Remark 4.10, this means that $\mathcal{M}(g)$ is an Archimedean quadratic module. It follows that we can apply Theorem 4.38, of which the current theorem is a particular instance, as the set (7.22) is a particular case of (4.55). The relation (7.23) comes from (4.57), for $K = 2$, $L = p + 1$. ■

Remark 7.8 The test (7.23) can be implemented only in relaxed form, with bounded degrees of the polynomials $f_1(\boldsymbol{t})$, $f_2(\boldsymbol{t})$, $s_\ell(\boldsymbol{t})$. A sensible strategy is to take these polynomials such that the total degree m of each term appearing in (7.23) is the same. Since the sum-of-squares have even degree, it results that $m \geq m_0 = 2\lceil \text{tdeg} A/2 \rceil$. ■

Remark 7.9 The test proposed in [3] is based on the strict positivity of $|A(z, \boldsymbol{q})|^2 = f_1(\boldsymbol{t})^2 + f_2(\boldsymbol{t})^2$ (tested using Bernstein polynomials). This is equivalent to the test (7.23) in which $u_1(\boldsymbol{t})$, $u_2(\boldsymbol{t})$ are scalar multiples of $-f_1(\boldsymbol{t})$ and $-f_2(\boldsymbol{t})$, respectively. In this case, the degrees of the terms from (7.23) would be 2 tdeg$A$, i.e., about twice m_0. We will see that our test can be accurate for smaller values of the degree. ■

Remark 7.10 Since the roots of a polynomial are continuous functions of the coefficients and $\mathcal{Q}$ is connected, the stability test can be split in two parts:

(i) $A(z, \boldsymbol{q}_0)$ is Schur for some $\boldsymbol{q}_0 \in \mathcal{Q}$, e.g., $\boldsymbol{q}_0 = \boldsymbol{0}$;
(ii) $A(z, \boldsymbol{q})$ is Schur for any z with $|z| = 1$ (i.e., only on the unit circle) and any $\boldsymbol{q} \in \mathcal{Q}$.

Implementing (i) is trivial. As for (ii), we have to take into account that now $1 - \tau_1^2 - \tau_2^2 = 0$. So, in (7.22), the polynomial $g_{p+1}(\boldsymbol{t})$ is now involved in an equality constraint, instead of a positivity constraint. It results that in (7.23), $s_{p+1}(\boldsymbol{t})$ becomes a general polynomial instead of a sum-of-squares. Although in principle such modified test is less conservative, we have not noticed any practical difference between (7.23) and its modified version. ■

Example 7.11 Transforming the classic continuous-time example from [4] via the bilinear transformation, we obtain

$$\begin{aligned} A(z, q_1, q_2) = &\ (26.38 + \rho + 9.18 q_1 + 10.67 q_2 + 1.87 q_1 q_2) \\ &+ (49.64 + 3\rho + 22.44 q_1 + 25.41 q_2 + 5.61 q_1 q_2) z \\ &+ (45.14 + 3\rho + 20.74 q_1 + 23.21 q_2 + 5.61 q_1 q_2) z^2 \\ &+ (13.88 + \rho + 7.48 q_1 + 8.47 q_2 + 1.87 q_1 q_2) z^3. \end{aligned} \tag{7.25}$$

This polynomial is not robustly Schur for any $\rho \geq 0$, but Schur for small negative values of ρ. When $\rho = 0$, there is a single point in $\mathcal{Q}$ for which the polynomial is not Schur, and thus, gridding methods are prone to fail; these methods test stability only for a discrete set of points in $\mathcal{Q}$.

The total degree of the polynomial (7.25) is 5. We take the polynomials from (7.23) such that the total degree of the terms is $m_0 = 6$, the minimum possible; so, the total degrees of, e.g., $f_1(\boldsymbol{t})$, $s_0(\boldsymbol{t})$, and $s_1(\boldsymbol{t})$ are 1, 6, and 4, respectively. Solving the SDP relaxation of (7.23), the polynomial (7.25) is considered not robustly Schur for any $\rho > \rho_0 = -4 \cdot 10^{-6}$. The negative value of ρ_0, instead of exactly 0, exhibits the numerical accuracy of the test. Note that the value ρ_0 is very small with respect to the coefficients of the polynomial; anyway, the numerical inaccuracy may only make a stable system to be assessed as unstable. The polynomial (7.25) is considered robustly Schur for any $\rho \in [-2.799999, \rho_0]$. The execution time for a test is about 0.2 s.

Other few experiments suggest that the test is accurate with $m = 2(1 + \lfloor \text{tdeg} A/2 \rfloor)$; this is the minimum degree m_0 when tdeg$A$ is odd and $m_0 + 2$ when tdeg$A$ is even. ■

7.2.2 Trigonometric Polynomial Test

We give now an alternative Positivstellensatz, similar in spirit with that from Sect. 7.1.3. Since the parameters, q_ℓ, $\ell = 1 : p$, are real and subunitary, we transform (7.19) into a trigonometric polynomial by putting

$$q_\ell = \frac{\zeta_\ell + \zeta_\ell^{-1}}{2}. \tag{7.26}$$

On the unit circle, it results that $q_\ell = \cos\theta_\ell \in [-1, 1]$. We denote $A(z, \boldsymbol{\zeta})$ the polynomial (7.19) obtained after the substitution (7.26). We assume that $n = 2\tilde{n}$ in (7.19); if n is odd, similar developments are possible. We shift $A(z, \boldsymbol{\zeta})$ such that its support is symmetric, obtaining

$$\tilde{A}(z, \boldsymbol{\zeta}) = z^{n/2} A(z, \boldsymbol{\zeta}) = A_s(z, \boldsymbol{\zeta}) + A_a(z, \boldsymbol{\zeta}). \tag{7.27}$$

The polynomials

$$A_s(z, \boldsymbol{\zeta}) = \frac{\tilde{A}(z, \boldsymbol{\zeta}) + \tilde{A}(z^{-1}, \boldsymbol{\zeta})}{2}, \quad A_a(z, \boldsymbol{\zeta}) = \frac{\tilde{A}(z, \boldsymbol{\zeta}) - \tilde{A}(z^{-1}, \boldsymbol{\zeta})}{2} \tag{7.28}$$

are symmetric and antisymmetric, respectively, in their first variable, i.e.,

$$A_s(z^{-1}, \boldsymbol{\zeta}) = A_s(z, \boldsymbol{\zeta}), \quad A_a(z^{-1}, \boldsymbol{\zeta}) = -A_a(z, \boldsymbol{\zeta}).$$

Moreover, due to (7.26), the polynomial $\tilde{A}(z, \boldsymbol{\zeta})$ is symmetric in $\boldsymbol{\zeta}$. Denoting $d = p + 1$ and

$$\boldsymbol{z} = (z, \boldsymbol{\zeta}) \in \mathbb{C}^d,$$

it results that the polynomials (7.28) satisfy

$$A_s(\boldsymbol{z}^{-1}) = A_s(\boldsymbol{z}), \quad A_a(\boldsymbol{z}^{-1}) = -A_a(\boldsymbol{z}),$$

i.e., they are symmetric and antisymmetric, respectively. As noticed in Remark 7.10, the robust stability test can be split in two parts. The second (and more difficult) condition that $A(z, \boldsymbol{q})$ is Schur for any $\boldsymbol{q} \in \mathcal{Q}$ and for $z \in \mathbb{T}$ is equivalent to the requirement that the set (7.15) is empty, with $A_s(\boldsymbol{\omega})$ and $A_a(\boldsymbol{\omega})$ as defined by (7.28). Hence, we can apply Theorem 7.5 to obtain the following.

Corollary 7.12 *The polynomial (7.19) is Schur for any z on the unit circle and any $\boldsymbol{q} \in \mathcal{Q}$ if and only if there exist a symmetric polynomial $X(\boldsymbol{z})$, an antisymmetric polynomial $Y(\boldsymbol{z})$ and a sum-of-squares polynomial $R(\boldsymbol{z})$, all with real coefficients, such that*

$$1 + X(\boldsymbol{z})A_s(\boldsymbol{z}) + Y(\boldsymbol{z})A_a(\boldsymbol{z}) + R(\boldsymbol{z}) = 0.$$

Remark 7.13 Although $A_a(\boldsymbol{z})$ has also other symmetry properties which reduce the number of its distinct coefficients, it seems that they cannot be used to reduce the complexity of the problem, in the sense that $Y(\boldsymbol{z})$ has no particular structure apart from being antisymmetric. ■

Remark 7.14 As discussed in Sect. 7.1.3, the degrees of $X(\boldsymbol{z})$ and $Y(\boldsymbol{z})$ are practically taken equal to those of $A_s(\boldsymbol{z})$ and $A_a(\boldsymbol{z})$, respectively. With this choice, the degree of the sum-of-squares $R(\boldsymbol{z})$ is $n = 2\tilde{n}$ in the first variable and twice the degree of $A(z, \boldsymbol{\zeta})$ in the variables $\boldsymbol{\zeta}$. Comparing the robust stability tests from Corollary 7.12 and Theorem 7.7, the first test has the advantage of working with polynomials with $p+1$ variables, while in the second, there are $p+2$ variables; there is a single sum-of-squares polynomial in the first test and $p+2$ in the second. However, a drawback of the first test is the larger degree of the polynomials. Overall, we can appreciate that the trigonometric polynomial test given by Corollary 7.12 is better especially when the degree n of the polynomial (7.19) is relatively large with respect to the degrees of the parameters $\boldsymbol{q}$. ■

Example 7.15 We consider the polynomial [5]

$$A(z, q) = [a(q+1) - 6] + [a(1-q) + 6]z + b(1+q)z^3 + b(1-q)z^4, \quad (7.29)$$

where the parameter is $q \in [-1, 1]$ and a, b are real constants. By putting $q = (\zeta + \zeta^{-1})/2$ as in (7.26), we obtain

$$2\zeta A(z,\zeta) = [a(\zeta+1)^2 - 12\zeta] + [-a(\zeta-1)^2 + 12\zeta]z + b(\zeta+1)^2 z^3 - b(\zeta-1)^2 z^4. \tag{7.30}$$

For testing that the polynomial (7.29) is robustly Schur, we must check whether the polynomial (7.30) has no roots on the unit bicircle. (The multiplication with ζ in (7.30) does not affect the roots and allows us to obtain a quarter plane polynomial.) By using the Positivstellensatz from Theorem 7.5, we find that, with $a = -3$, this happens for $-2.99999999 \leq b \leq 2.99999998$. The test is accurate, as already suggested in Example 7.6. For $b = 3$, it results that $A(z,-1) = -6 + 6z^4$, whose roots are obviously on the unit circle; same remark is valid for $b = -3$, when $A(z,-1) = -6 - 6z^4$. ■

7.3 Convex Stability Domains

The set of Schur polynomials is not convex, and so, in optimization problems where a stable system is sought, it is customary to use diverse sufficient conditions. Some of these conditions amount to building a convex domain around a given Schur polynomial (7.2). In this section, we characterize such a domain, using a positive realness condition that takes the form of an LMI. We present the results directly for multivariate polynomials, although their illustration will be mostly in the univariate case.

7.3.1 Positive Realness Stability Domain

We aim to build the convex stability around a given Schur polynomial $A(\mathbf{z})$ (in optimization problems, this polynomial would be a nominal point of interest at a certain stage). Thus, we consider sets of polynomials $\tilde{A} = A + D$, for a variable $D(\mathbf{z})$; for simplicity, we consider the free terms of $\tilde{A}(\mathbf{z})$ and $A(\mathbf{z})$ both equal to 1, while the free term of $D(\mathbf{z})$ is zero ($a_{\mathbf{0}} = \tilde{a}_{\mathbf{0}} = 1$, $d_{\mathbf{0}} = 0$). The base for our construction is the following positive realness result.

Theorem 7.16 *Let $A(\mathbf{z})$, $\tilde{A}(\mathbf{z})$ be polynomials defined as in (7.2). If $A(\mathbf{z})$ is Schur (i.e., has no roots inside or on the unit circle) and*

$$Re\left[\frac{\tilde{A}(\boldsymbol{\omega})}{A(\boldsymbol{\omega})}\right] > 0, \quad \forall \boldsymbol{\omega} \in [-\pi, \pi]^d, \tag{7.31}$$

then $\tilde{A}(\mathbf{z})$ is also Schur.

The proof is presented in Sect. 7.3.3. The domain built using (7.31) has convenient properties.

Theorem 7.17 *Let $A(z)$, defined as in (7.2), be a Schur polynomial. The domain*

$$\mathcal{D}_A = \{\tilde{A}(z) \textit{ such that (7.31) holds}\} \tag{7.32}$$

is convex and $A(z)$ is an interior point of it.

Proof Let $\tilde{A}_1$, $\tilde{A}_2$ be arbitrary polynomials from $\mathcal{D}_A$. Denoting $\tilde{A}_\alpha = \alpha\tilde{A}_1 + (1-\alpha)\tilde{A}_2$, for $\alpha \in [0, 1]$, it results immediately that $\mathrm{Re}[\tilde{A}_\alpha(\boldsymbol{\omega})/A(\boldsymbol{\omega})] > 0$ and thus $\tilde{A}_\alpha \in \mathcal{D}_A$.

Now take an arbitrary polynomial $D(\boldsymbol{z})$, with zero free term. Since $D(\boldsymbol{\omega})$ has a finite maximum value, it follows that there is an $\varepsilon > 0$ such that $\mathrm{Re}[1 + \varepsilon D(\boldsymbol{\omega})/A(\boldsymbol{\omega})] > 0$, i.e., the distance from $A(\boldsymbol{z})$ to the border of $\mathcal{D}_A$ is nonzero in any direction $D(\boldsymbol{z})$. It is also clear from (7.31) that $A \in \mathcal{D}_A$. ■

We express now the positive realness condition (7.31) with the aid of positive polynomials. We denote

$$T_+(\boldsymbol{z}) = \frac{\tilde{A}(\boldsymbol{z})}{A(\boldsymbol{z})} = 1 + \frac{D(\boldsymbol{z})}{A(\boldsymbol{z})} \tag{7.33}$$

and

$$T(\boldsymbol{z}) = T_+(\boldsymbol{z}) + T_+(\boldsymbol{z}^{-1}) = \frac{R(\boldsymbol{z})}{A(\boldsymbol{z}) \cdot A(\boldsymbol{z}^{-1})}, \tag{7.34}$$

where

$$R(\boldsymbol{z}) = 2A(\boldsymbol{z})A(\boldsymbol{z}^{-1}) + A(\boldsymbol{z})D(\boldsymbol{z}^{-1}) + A(\boldsymbol{z}^{-1})D(\boldsymbol{z}) \tag{7.35}$$

is a trigonometric polynomial. Then, condition (7.31) is obviously equivalent with

$$R(\boldsymbol{\omega}) > 0, \quad \forall \boldsymbol{\omega} \in [-\pi, \pi]^d. \tag{7.36}$$

Since we aim to an implementable form of the condition (7.31), we relax (7.36) to the requirement that $R(\boldsymbol{z})$ is sum-of-squares.

Theorem 7.18 *Let $A(\boldsymbol{z})$ be a Schur polynomial defined as in (7.2), with $a_{\mathbf{0}} = 1$. Consider the domain*

$$\widehat{\mathcal{D}}_A = \{\tilde{A} = A + D \mid R \in \mathbb{RS}^{\boldsymbol{n}}_{\boldsymbol{n}}[\boldsymbol{z}], \quad R(\boldsymbol{\omega}) > 0, \ \forall \boldsymbol{\omega} \in [-\pi, \pi]^d\}, \tag{7.37}$$

where $R(\boldsymbol{z})$ is defined by (7.35) and D is a positive orthant polynomial of degree $\boldsymbol{n}$, with $d_{\mathbf{0}} = 0$. Then, the next affirmations are true:

(a) $\widehat{\mathcal{D}}_A \subset \mathcal{D}_A$ (and thus $A + D$ is Schur).

(b) The domain $\widehat{\mathcal{D}}_A$ is convex.

Proof (a) Obvious. (b) The relation (7.35) between the coefficients of the sum-of-squares $R(\boldsymbol{z})$ and of the polynomial $D(\boldsymbol{z})$ is linear. Since the set of positive sum-of-squares polynomials (with factors of a given degree) is convex, the set of $D(\boldsymbol{z})$ satisfying (7.35) is also convex, i.e., $\widehat{\mathcal{D}}_A$ is convex. ■

Remark 7.19 The advantage of the domain (7.37) lies on its description via an LMI. Since $R(z)$ is sum-of-squares, it can be expressed with the generalized trace (3.32) or Gram-pair (3.99) parameterizations. As $D(z)$ (and thus $\tilde{A}(z)$) is linearly related to $R(z)$, it results that the coefficients of $D(z)$ can be parameterized through an LMI. To stress the linearity of (7.35), we can write it in the equivalent form

$$\boldsymbol{r} = (\boldsymbol{F} + \boldsymbol{G})\boldsymbol{d} + 2\boldsymbol{G}\boldsymbol{a}, \tag{7.38}$$

where $\boldsymbol{F}$ and $\boldsymbol{G}$ are constant matrices (i.e., depending only on the coefficients of $A(z)$) and $\boldsymbol{r}$, $\boldsymbol{d}$, and $\boldsymbol{a}$ are vectors of the coefficients of the polynomials $R(z)$, $D(z)$, and $A(z)$, respectively (the vector $\boldsymbol{r}$ contains only the coefficients of $R(z)$ from a half-space). We define the vectors $\boldsymbol{d}$ and $\boldsymbol{a}$ as in (3.27). The relation (7.35) can be written as

$$R(z) = 2\boldsymbol{a}^T \boldsymbol{\Psi}(z^{-1})\boldsymbol{a} + \boldsymbol{a}^T \left[\boldsymbol{\Psi}(z^{-1}) + \boldsymbol{\Psi}(z)\right] \boldsymbol{d},$$

where $\boldsymbol{\Psi}(z)$ is defined in (3.34). By identification, we can derive the expressions of the matrices $\boldsymbol{F}$ and $\boldsymbol{G}$.

1-D case. For univariate polynomials (when $\boldsymbol{r} = [r_0 \; \dots \; r_n]^T$, etc.), the constant matrices from (7.38) are

$$\begin{aligned}
\boldsymbol{F} &= \overline{\mathrm{Toep}}(\boldsymbol{a}^R) \triangleq \begin{bmatrix} a_0 & a_1 & \dots & a_n \\ 0 & a_0 & \ddots & a_{n-1} \\ \vdots & \ddots & \ddots & \vdots \\ 0 & \dots & 0 & a_0 \end{bmatrix}, \\
\boldsymbol{G} &= \overline{\mathrm{Hank}}(\boldsymbol{a}) \triangleq \begin{bmatrix} a_0 & \dots & a_{n-1} & a_n \\ a_1 & \iddots & a_n & 0 \\ \vdots & \iddots & \iddots & \vdots \\ a_n & 0 & \dots & 0 \end{bmatrix}.
\end{aligned} \tag{7.39}$$

(The upper index in $\boldsymbol{a}^R$ denotes the reversal of the order of the elements of $\boldsymbol{a}$; the overline indicates a block of a larger matrix; the notation will become more relevant later.)

2-D case. If $d = 2$, we define $\boldsymbol{r}$ by concatenating the columns of the following table (giving the coefficients situated in a half plane)

$$\begin{array}{cccc}
 & r_{-n_1,1} & \dots & r_{-n_1,n_2} \\
 & \vdots & & \vdots \\
r_{0,0} & r_{0,1} & \dots & r_{0,n_2} \\
r_{1,0} & r_{1,1} & \dots & r_{1,n_2} \\
\vdots & \vdots & & \vdots \\
r_{n_1,0} & r_{n_1,1} & \dots & r_{n_1,n_2}
\end{array}$$

With this convention, the matrix $\boldsymbol{F}$ from (7.38) has the block Toeplitz structure

$$\boldsymbol{F} = \begin{bmatrix} \overline{\mathrm{Toep}}(\boldsymbol{a}_0^R) & \overline{\mathrm{Toep}}(\boldsymbol{a}_1^R) & \dots & \overline{\mathrm{Toep}}(\boldsymbol{a}_{n_2}^R) \\ 0 & \mathrm{Toep}(\boldsymbol{a}_0^R) & \ddots & \vdots \\ 0 & 0 & \ddots & \mathrm{Toep}(\boldsymbol{a}_1^R) \\ 0 & 0 & 0 & \mathrm{Toep}(\boldsymbol{a}_0^R) \end{bmatrix}, \tag{7.40}$$

where the blocks of size $(2n_1 + 1) \times (n_1 + 1)$ of $\boldsymbol{F}$ are Toeplitz matrices defined by

$$\mathrm{Toep}(\boldsymbol{a}_{k_2}^R) = \begin{bmatrix} a_{n_1,k_2} & 0 & 0 \\ \vdots & \ddots & 0 \\ a_{0,k_2} & \ddots & a_{n_1,k_2} \\ 0 & \ddots & \vdots \\ 0 & 0 & a_{0,k_2} \end{bmatrix}. \tag{7.41}$$

Also, the matrix $\boldsymbol{G}$ has the block Hankel structure

$$\boldsymbol{G} = \begin{bmatrix} \overline{\mathrm{Hank}}(\boldsymbol{a}_0) & \dots & \overline{\mathrm{Hank}}(\boldsymbol{a}_{n_2-1}) & \overline{\mathrm{Hank}}(\boldsymbol{a}_{n_2}) \\ \vdots & \iddots & \mathrm{Hank}(\boldsymbol{a}_{n_2}) & 0 \\ \mathrm{Hank}(\boldsymbol{a}_{n_2-1}) & \iddots & 0 & 0 \\ \mathrm{Hank}(\boldsymbol{a}_{n_2}) & 0 & 0 & 0 \end{bmatrix}, \tag{7.42}$$

where the blocks have size $(2n_1 + 1) \times (n_1 + 1)$ and the Hankel structure

$$\mathrm{Hank}(\boldsymbol{a}_{k_2}) = \begin{bmatrix} 0 & 0 & a_{0,k_2} \\ 0 & \iddots & \vdots \\ a_{0,k_2} & \iddots & a_{n_1,k_2} \\ \vdots & \iddots & 0 \\ a_{n_1,k_2} & 0 & 0 \end{bmatrix}. \tag{7.43}$$

Remark 7.20 Finally, we note that we could allow in (7.37) a relaxation degree higher than $\boldsymbol{n}$, obtaining a domain including $\widehat{\mathcal{D}}_A$. However, it seems that the implementation cost would not justify the (probably not significant) increase in the domain. ■

7.3.2 Comparisons and Examples

Another convex stability domain around a given Schur polynomial $A(z)$, used in several optimization methods for IIR filter design, is

$$\mathcal{D}_A^R = \{\tilde{A} = A + D \mid |D(\omega)| < |A(\omega)|, \ \omega \in [-\pi, \pi]^d\}. \tag{7.44}$$

The proof that $\mathcal{D}_A^R$ contains only Schur polynomials is based on Rouché's criterion and is suggested in problem **P** 7.3. More interestingly, for any $A(z)$, this domain is included in the positive realness domain (7.32); a proof is suggested in problem **P** 7.4.

We illustrate the shape and the size of the presented convex stability domains in the simplest case, that of univariate polynomials of degree $n = 2$. In this case, we have $A(z) = 1 + a_1z + a_2z^2$, $D(z) = d_1z + d_2z^2$ and $\tilde{a}_1 = a_1 + d_1$, $\tilde{a}_2 = a_2 + d_2$. The stability domain for polynomials of order two is the interior of a triangle in the parameter plane $(\tilde{a}_1, \tilde{a}_2)$, as shown in Figs. 7.2 and 7.3.

Example 7.21 For $A(z) = 1 - 0.5z + 0.6z^2$, Fig. 7.2 shows three convex stability domains. Besides $\mathcal{D}_A$ (dashed line) and $\mathcal{D}_A^R$ (dotted), we have also drawn, with solid line, the circle $\mathcal{D}_A^S$ with radius equal to the stability radius of $A(z)$ (the stability radius is the shortest Euclidean distance from $A(z)$ to an unstable polynomial, measured in the vector space of coefficients). It is clear that $\mathcal{D}_A^R \subset \mathcal{D}_A$, while $\mathcal{D}_A^S$ has points outside $\mathcal{D}_A$, although having a much smaller area. ■

Example 7.22 We take now $A(z) = 1 - 0.3z - 0.4z^2$ and obtain the domains shown in Fig. 7.3. It is interesting to remark that when $A(z)$ approaches the lower corner of the stability triangle, i.e., the polynomial $1 - z^2$, the positive realness domain $\mathcal{D}_A$ tends to be the whole triangle, while $\mathcal{D}_A^R$ and $\mathcal{D}_A^S$ tend to become empty (the stability radius domain tends to be empty whenever the distance to the border of the triangle becomes small). ■

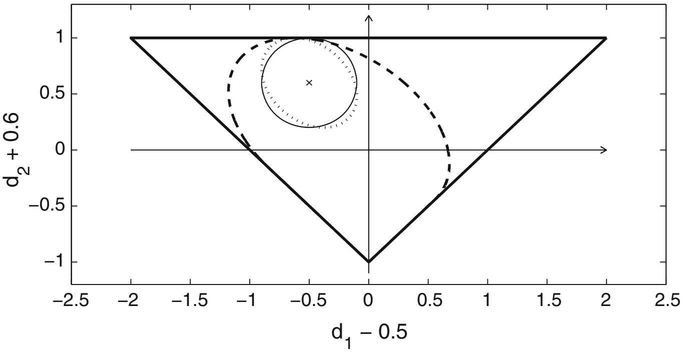


Fig. 7.2 Convex stability domains for Example 7.21. *Dashed line* positive realness. *Dotted* Rouché. *Solid* stability radius

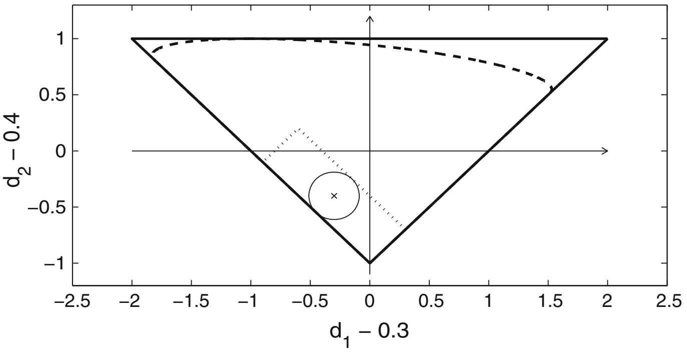


Fig. 7.3 Convex stability domains for Example 7.22. Same legend as in Fig. 7.2

7.3.3 *Proof of Theorem 7.16*

Proof for univariate polynomials. From (7.31), it results that $\tilde{A}(\omega) \neq 0$, $\forall \omega \in [-\pi, \pi]$, i.e., $\tilde{A}(z)$ has no roots on the unit circle. Let $z_0 = r\mathrm{e}^{j\theta}$ be an arbitrary point inside the unit circle ($0 \leq r < 1$). We have to show that $\tilde{A}(z_0) \neq 0$.

Since $A(z)$ and $\tilde{A}(z)$ are polynomials and $A(z)$ has no zeros on the unit disk, we can apply Cauchy's integral formula on the unit circle to obtain

$$\begin{aligned} \frac{\tilde{A}(z_0)}{A(z_0)} &= \frac{1}{2\pi j} \oint_{\mathbb{T}} \frac{\tilde{A}(z)}{(z - z_0)A(z)} dz \\ &\overset{z=\mathrm{e}^{j\omega}}{=} \frac{1}{2\pi j} \int_{-\pi}^{\pi} \frac{\tilde{A}(\omega)}{(\mathrm{e}^{j\omega} - r\mathrm{e}^{j\theta})A(\omega)} j\mathrm{e}^{j\omega} d\omega \\ &= \frac{1}{2\pi} \int_{-\pi}^{\pi} \frac{\tilde{A}(\omega)}{A(\omega)[1 - r\mathrm{e}^{j(\theta-\omega)}]} d\omega \\ &\overset{0\leq r<1}{=} \frac{1}{2\pi} \int_{-\pi}^{\pi} \frac{\tilde{A}(\omega)}{A(\omega)} \sum_{k=0}^{\infty} r^k \mathrm{e}^{jk(\theta-\omega)} d\omega. \end{aligned} \tag{7.45}$$

We can also apply Cauchy's integral theorem, again on the unit circle, to obtain, for any integer $\ell \geq 0$,

$$\oint_{\mathbb{T}} \frac{\tilde{A}(z)z^{\ell}}{A(z)} dz = 0$$

and thus

$$\int_{-\pi}^{\pi} \frac{\tilde{A}(\omega)\mathrm{e}^{j(\ell+1)\omega}}{A(\omega)} d\omega = 0.$$

It follows immediately that we have

$$\frac{1}{2\pi}\int_{-\pi}^{\pi}\frac{\tilde{A}(\omega)}{A(\omega)}\sum_{k=1}^{\infty}r^k \mathrm{e}^{jk(\omega-\theta)}d\omega = 0. \tag{7.46}$$

The sum

$$P(\omega) = \sum_{k=0}^{\infty}r^k \mathrm{e}^{-jk\omega} + \sum_{k=1}^{\infty}r^k \mathrm{e}^{jk\omega} \tag{7.47}$$

is the Poisson kernel. It is easy to see that $P(\omega)$ is real. Moreover, it can be proved that for $0 \le r < 1$, we have

$$P(\omega) = \frac{1-r^2}{1+r^2-2r\cos\omega} > 0, \quad \forall\omega \in [-\pi,\pi]. \tag{7.48}$$

Adding (7.45) and (7.46), we obtain

$$\frac{\tilde{A}(z_0)}{A(z_0)} = \frac{1}{2\pi}\int_{-\pi}^{\pi}\frac{\tilde{A}(\omega)}{A(\omega)}P(\omega-\theta)d\omega. \tag{7.49}$$

Due to (7.31) and to the positivity of the Poisson kernel, it results that the real part of the right-hand term of (7.49) is strictly positive, and thus, $\tilde{A}(z_0) \ne 0$, which is the desired conclusion.

The multivariate case is proved by means of the DeCarlo–Strintzis conditions. If (7.31) holds, then $\tilde{A}(\boldsymbol{\omega}) \ne 0$, $\forall\boldsymbol{\omega} \in [-\pi,\pi]^d$, and so the second condition (7.5) holds. We fix now $z_2 = \ldots = z_d = 1$. The univariate polynomial $A_1(z_1) = A(z_1, 1, \ldots, 1)$ is Schur. Denoting $\tilde{A}_1(z_1) = \tilde{A}(z_1, 1, \ldots, 1)$, it results from (7.31) that $\mathrm{Re}[\tilde{A}_1(\omega_1)/A_1(\omega_1)] > 0$, $\forall\omega_1 \in [-\pi,\pi]$. Applying the above proved univariate version of this theorem, we conclude that the univariate polynomial $\tilde{A}(z_1, 1, \ldots, 1)$ is Schur. Repeating this proof for z_2, ..., z_d, it results that all the univariate polynomials (7.4) are Schur and so the first DeCarlo–Strintzis conditions also hold.

7.4 Bibliographical and Historical Notes

To decide whether the stability condition (7.3) is true or not for a given system is an NP-hard problem [6] for $d \ge 2$. The DeCarlo–Strintzis test [7] was used, e.g., for a genetic algorithm approach [8]; its relation with positive polynomials and SDP was presented in [1]. Older algorithms for testing stability are based on solving systems of polynomial equations [9, 10]. The results from Sects. 7.1.2 and 7.1.3 have appeared in [11] and [12], respectively.

The robust stability problem from Sect. 7.2 has been solved with different types of gridding [4] (for which it is easy to generate counterexamples) or Bernstein polynomials expansions [3, 5]; the Positivstellensatz method from Sect. 7.2.1 has been presented in [13], while the trigonometric version from Sect. 7.2.2 appeared in [12].

The proof of Theorem 7.16 is inspired from the proof [14] of a simpler result, stating that if $\mathrm{Re}[\tilde{A}(\omega)] > 0, \forall\omega \in [-\pi, \pi]$, then $\tilde{A}(z)$ is Schur. (This is Theorem 7.16 for $A(z) = 1$.) The univariate version of the positive realness convex stability domain presented in Sect. 7.3.1 has been proposed in [15, 16], while the multivariate version appeared in [17]. The first use of the Rouché stability domain was in IIR filter design [18].

Problems

P 7.1 How many distinct nonzero coefficients may have a d-variate polynomial of complex variable $z = (z_1, \dots, z_d)$, that is, antisymmetric in z_1 and symmetric in z_2, …, z_d?

P 7.2 Generalize the robust stability tests from Sect. 7.2 to *multivariate* polynomials whose coefficients depend polynomially on some bounded parameters.

P 7.3 **a**. Show that the domain $\mathcal{D}_A^R$ defined by (7.44) is convex.

b. Show that $\mathcal{D}_A^R$ contains only Schur polynomials.

Hints. **a**. For each $\boldsymbol{\omega}$, the inequality $|D(\boldsymbol{\omega})| < |A(\boldsymbol{\omega})|$ defines a convex set. As their intersection, $\mathcal{D}_A^R$ is convex.

b. In the univariate case, this is a consequence of Rouché's criterion (if $|D(\omega)| < |A(\omega)|$, then $A(z)$ and $A(z) + D(z)$ have the same number of zeros inside the unit circle). For multivariate polynomials, use the DeCarlo–Strintzis conditions, as in the proof of Theorem 7.16.

P 7.4 Prove that $\mathcal{D}_A^R \subset \mathcal{D}_A$, for any Schur polynomial $A(z)$, where the domains $\mathcal{D}_A$ and $\mathcal{D}_A^R$ are defined in (7.32) and (7.44), respectively. (Hint: prove that $|D(\boldsymbol{\omega})|/|A(\boldsymbol{\omega})| < 1$ implies $1 + \mathrm{Re}[D(\boldsymbol{\omega})/A(\boldsymbol{\omega})] > 0$.)

References

1. B. Dumitrescu, Multidimensional stability test using sum-of-squares decomposition. IEEE Trans. Circ. Syst. I **53**(4), 928–936 (2006)
2. J.F. Sturm, Using SeDuMi 1.02, a matlab toolbox for optimization over symmetric cones. Optim. Methods Softw. **11**, 625–653 (1999). http://sedumi.ie.lehigh.edu
3. D.D. Siljak, D.M. Stipanovic, Robust D-stability via positivity. Automatica **35**(8), 1477–1484 (1999)
4. J. Ackermann, H.Z. Hu, D. Kaesbauer, Robustness analysis: a case study, in *Proceedings of the 27th Conference Decision Control* (Austin, Texas, 1988), pp. 86–91
5. J. Garloff, B. Graf, Robust schur stability of polynomials with polynomial parameter dependency. Multidim. Syst. Signal Proc. **10**(2), 189–199 (1999)
6. O. Toker, H. Özbay, On the complexity of pure μ computation and related problems in multidimensional systems. IEEE Trans. Auto. Control **43**(3), 409–414 (1998)
7. M.G. Strintzis, Tests of stability of multidimensional filters. IEEE Trans. Circ. Syst. CAS **24**(8), 432–437 (1977)

8. N.E. Mastorakis, I.F. Gonos, M.N.S. Swamy, Stability of multidimensional systems using genetic algorithms. IEEE Trans. Circ. Syst. I **50**(7), 962–965 (2003)
9. E. Zeheb, E. Wallach, Zero sets of multiparameter functions and stability of multidimensional systems. IEEE Trans. Acoust. Speech Signal Proc. ASSP **29**(2), 197–206 (1981)
10. E. Curtin, S. Saba, Stability and margin of stability tests for multidimensional filters. IEEE Trans. Circ. Syst. I **46**(7), 806–809 (1999)
11. B. Dumitrescu, LMI stability tests for the Fornasini-Marchesini model. IEEE Trans. Signal Proc. **56**(8), 4091–4095 (2008)
12. B. Dumitrescu, Positivstellensatz for trigonometric polynomials and multidimensional stability tests. IEEE Trans. Circ. Syst. II **54**(4), 353–356 (2007)
13. B. Dumitrescu, B.C. Chang, Robust schur stability with polynomial parameters. IEEE Trans. Circ. Syst. II **53**(7), 535–537 (2006)
14. A.T. Chottera, G.A. Jullien, A linear programming approach to recursive digital filter design with linear phase. IEEE Trans. Circ. Syst. CAS **29**(3), 139–149 (1982)
15. B. Dumitrescu, R. Niemistö, Multistage IIR filter design using convex stability domains defined by positive realness. IEEE Trans. Signal Proc. **52**(4), 962–974 (2004)
16. D. Henrion, M. Šebek, V. Kučera, Positive polynomials and robust stabilization with fixed-order controllers. IEEE Trans. Auto. Control **48**(7), 1178–1186 (2003)
17. B. Dumitrescu, Optimization of 2D IIR filters with nonseparable and separable denominator. IEEE Trans. Signal Proc. **53**(5), 1768–1777 (2005)
18. B. Dumitrescu, R. Bregović, T. Saramäki, Simplified design of low-delay oversampled NPR GDFT filterbanks. EURASIP J. Appl. Signal Process. 42961, 11 (2006)

Chapter 8
Design of IIR Filters

Abstract IIR filters can give the same magnitude performance with fewer parameters than FIR filters. However, they cannot have exact linear phase. Their design is more complicated due to the difficulty in ensuring stability and to the non-convexity of the optimization problems. In this short chapter, we give few guidelines for the optimization of IIR filters, insisting on algorithms that use positive polynomials. For 1D filters, we discuss two design problems, using magnitude and approximate linear phase as design criteria; in the latter case, stability domains based on positive realness are an important tool. The method for approximate linear phase is then extended to 2D, for the case when passband and stopband are described by the positivity of some polynomials.

8.1 Magnitude Design of IIR Filters

We consider IIR filters given by the transfer function

$$H(z) = \frac{B(z)}{A(z)} = \frac{\sum_{i=0}^{m} b_i z^{-i}}{\sum_{k=0}^{n} a_k z^{-k}}. \tag{8.1}$$

The orders of the numerator (m) and denominator (n) can be different. Since the multiplication with a constant of both the denominator and the numerator does not change the filter, a normalization constraint is imposed, usually on the denominator. In this section, the normalization constraint presets the energy of the denominator and is

$$\frac{1}{2\pi} \int_{-\pi}^{\pi} |A(\omega)|^2 d\omega = 1 \iff \sum_{k=0}^{n} a_k^2 = 1. \tag{8.2}$$

Similarly to the FIR case, presented in Chap. 5, we discuss here only the simplest design, that of lowpass filters, with given passband $[0, \omega_p]$ and stopband $[\omega_s, \pi]$. In this section, we confine our discussion to minimax optimization. The reason is that the stopband energy (5.2) is not a convex function of the coefficients of the denominator; although it is possible (and interesting) to deal with such an objective,

B. Dumitrescu, *Positive Trigonometric Polynomials and Signal Processing Applications*, Signals and Communication Technology,
DOI 10.1007/978-3-319-53688-0_8

we start by solving the easier problem. For given bounds γ_p and γ_s, we want to design a filter (8.1) that respects the magnitude constraints shown in Fig. 5.1; there are no constraints on the phase. That is, with given orders m and n, we want to solve the feasibility problem

$$\begin{array}{ll} \text{find} & a_k, b_i, \quad k = 0 : n, i = 0 : m \\ \text{s.t.} & |H(\omega)| \le 1 + \gamma_p, \ \forall \omega \\ & |H(\omega)| \ge 1 - \gamma_p, \ \forall \omega \in [0, \omega_p] \\ & |H(\omega)| \le \gamma_s, \ \forall \omega \in [\omega_s, \pi] \\ & \sum_{k=0}^{n} a_k^2 = 1 \\ & H(z) \text{ is stable} \end{array} \tag{8.3}$$

(Note that, similarly to the problems from Sect. 5.1, we have extended the first constraint from $[0, \omega_s]$ to the whole range of frequencies.) In this form, the problem is not convex. Since the problem is formulated only in terms of magnitude, it is natural to work with the squared magnitudes

$$R_a(z) = A(z)A(z^{-1}), \quad R_b(z) = B(z)B(z^{-1}), \tag{8.4}$$

as variables. We note that the normalization constraint (8.2) is equivalent to the simple condition $r_{a0} = 1$ on the free term of $R_a(z)$. In terms of polynomials that are nonnegative, globally or on some intervals, the problem (8.3) can be transformed into

$$\begin{array}{ll} \text{find} & R_a \in \mathbb{R}_n[z], \ R_b \in \mathbb{R}_m[z] \\ \text{s.t.} & (1 + \gamma_p)^2 R_a(\omega) - R_b(\omega) \ge 0, \ \forall \omega \\ & R_b(\omega) - (1 - \gamma_p)^2 R_a(\omega) \ge 0, \ \forall \omega \in [0, \omega_p] \\ & \gamma_s^2 R_a(\omega) - R_b(\omega) \ge 0, \ \forall \omega \in [\omega_s, \pi] \\ & R_a(\omega) \ge 0, \ R_b(\omega) \ge 0, \ \forall \omega \\ & r_{a0} = 1 \end{array} \tag{8.5}$$

This problem can be transformed easily into an SDP problem (similarly to the transformation from (5.13) to (5.14) from the FIR case). For example, using the notation (5.6), the second constraint from (8.5) can be written as the LMI

$$r_{bk} - (1 - \gamma_p)^2 r_{ak} = \mathcal{L}_{k,0,\omega_p}(\boldsymbol{Q}_1, \boldsymbol{Q}_2), \quad k = 0 : \max(m, n),$$

where $\boldsymbol{Q}_1$, $\boldsymbol{Q}_2$ are positive semidefinite parameter matrices. After solving the SDP form of (8.5), the numerator and denominator of $H(z)$ are recovered from $R_b(z)$ and $R_a(z)$, respectively, via spectral factorization.

Remark 8.1 There is no stability constraint present in (8.5). If, at the spectral factorization of $R_a(z)$, we compute a minimum phase denominator $A(z)$, then we are sure that $H(z)$ has no poles *outside* the unit circle. However, there may be poles on the unit circle or arbitrarily close to it. In this situation, there are two partial cures. The first is old in signal processing and consists of changing the design data! For example, a larger transition band may lead to poles farther from the unit circle. The second simple way to keep the poles away from the unit circle is to change the nonnegativity condition $R_a(\omega) \geq 0$ into $R_a(\omega) \geq \epsilon$, where ϵ is a given positive constant; thus, it will result that $R(z)$ is strictly positive and so it cannot have zeros on the unit circle. Of course, this method cannot control how far from the unit circle are actually the poles of $H(z)$; anyway, a larger ϵ implies a smaller pole radius. ■

Remark 8.2 Since (8.5) is a feasibility problem, the resulting filter is not optimal in a certain sense. However, it can be made so.

For example, it is possible to minimize the stopband error γ_s. For a given γ_s, the feasibility problem (8.5) may have or not a solution. If it has, we can decrease γ_s and try again. If it has not, we increase γ_s. Using a bisection procedure, the optimal γ_s is found. A similar approach can lead to a filter of minimum order that satisfies the magnitude constraints. Keeping, e.g., n fixed, we solve (8.5) for various values of m until the minimum one is found. (Of course, we could try to minimize n with fixed m, or even minimize the sum $n + m$, which gives the implementation complexity of the IIR filters, since $H(z)$ has $n + m + 1$ free coefficients.) ■

Example 8.3 As in Example 5.2, we take $\omega_p = 0.2\pi$, $\omega_s = 0.25\pi$, $\gamma_p = 0.1$, and $\gamma_s = 0.01$. We choose a denominator degree $n = 4$; for simple bandpass structures, as in our case, it is often indicated to choose $n < m$ and a small value of n. We solve (8.5) for several numerator orders, until finding the smallest one: $m = 7$. The magnitude response of the obtained IIR filter is shown in Fig. 8.1. The maximum magnitude of a pole is 0.9651, i.e., relatively near from the unit circle.

If we want poles with smaller radius, we have to increase the degree of the numerator (assuming that we keep $n = 4$). For example, putting $m = 20$, the filter resulting from (8.5) has poles with maximal radius equal to 0.9548. Imposing the positivity condition $R_a(\omega) \geq \epsilon$, with $\epsilon = 0.08$, leads to maximum pole magnitude of 0.9409. (Larger values, e.g., $\epsilon = 0.09$ lead to the infeasibility of (8.5).) ■

8.2 Approximate Linear-Phase Designs

Since IIR filters cannot have exact linear phase, we explore the possibilities of approximating the ideal frequency response (we stick to our basic lowpass filter)

$$H_{id}(\omega) = \begin{cases} e^{-j\tau\omega}, & \omega \in [0, \omega_p], \\ 0, & \omega \in [\omega_s, \pi], \end{cases} \tag{8.6}$$

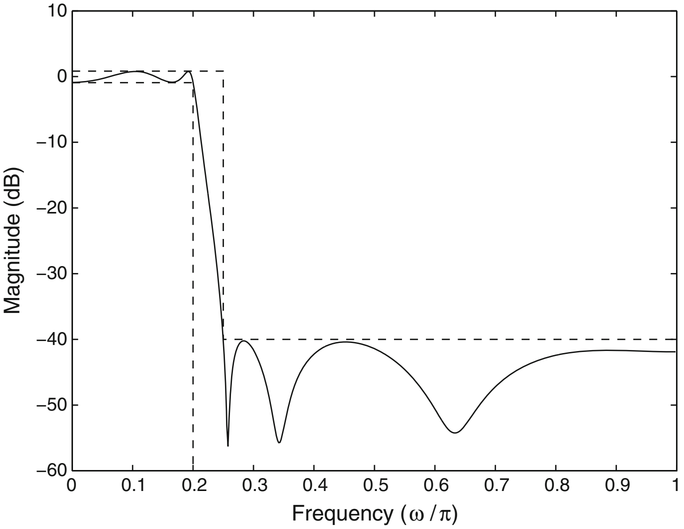


Fig. 8.1 Frequency response of the filter designed in Example 8.3

where the group delay $\tau \in \mathbb{R}$ is given. We have at least two options regarding the optimization objective. The first is based on a p-norm error measure. To simplify the numerical approach, we adopt a discretized form, on a set of frequencies ω_ℓ, $\ell = 1 : L$, that cover the passband and stopband. We denote $H_{id}(\omega_\ell) = H_\ell$, having in mind that ideal responses other than (8.6) could be used. The objective is

$$J_p(A, B) = \frac{1}{L} \sum_{\ell=1}^{L} \lambda_\ell \left| \frac{B(\omega_\ell)}{A(\omega_\ell)} - H_\ell \right|^p , \tag{8.7}$$

where $\lambda_\ell > 0$ are weights. We are interested especially by the least squares objective, when $p = 2$, and by the case of large p, when (8.7) may be used for approximating the Chebyshev (minimax) objective

$$J_\infty(A, B) = \max_{\ell=1:L} \lambda_{\ell,\infty} \left| \frac{B(\omega_\ell)}{A(\omega_\ell)} - H_\ell \right| . \tag{8.8}$$

If $\lambda_\ell = \lambda_{\ell,\infty}^p$, then (8.7) is a good approximation of (8.8) if the value of p is larger than, e.g., 50. Relatively low values of p, e.g., between 4 and 10, can serve to obtain an approximate PCLS design.

The second optimization option is the minimax problem

$$\begin{aligned} \min_{A\in\mathbb{R}_{n+}[z],\, B\in\mathbb{R}_{m+}[z],\, \gamma_s} \quad & \gamma_s \\ \text{s.t.} \quad & \left|\frac{B(\omega)}{A(\omega)} - \mathrm{e}^{-j\tau\omega}\right| \le \gamma_p,\ \forall\omega\in[0,\omega_p] \\ & \left|\frac{B(\omega)}{A(\omega)}\right| \le \gamma_s,\ \forall\omega\in[\omega_s,\pi] \end{aligned} \tag{8.9}$$

where γ_p is a given passband error bound. The solution of (8.9) is not necessarily a minimizer for $J_\infty(A, B)$ and vice versa. The coincidence (approximate only, due to the grid optimization inherent in (8.8)) appears if the weights in (8.8) are equal to, e.g., 1 in the passband and γ_p/γ_s^* in the stopband, where γ_s^* is the solution of (8.9).

The formulations above are non-convex. Obviously, we seek stable IIR filters, which is a supplementary complication. In this section, we will present several algorithms that give approximate solutions to the minimization of (8.7) or to (8.9). The solutions are typically only local minima, but obtained with a good trade-off between quality and complexity.

We do not seek PCLS solutions, inserting (8.7) into the objective of (8.9), since the resulting problem would be too complicated. We will focus on the simpler problems, trying to enhance the basic ideas in the treatment of non-convexity and of the stability constraint.

Since we will not appeal to squared magnitudes, the normalization constraint for (8.1) used in this section is the most natural one, i.e., $a_0 = 1$.

8.2.1 *Optimization with Fixed Denominator*

If the denominator of the IIR filter (8.1) is given, the optimization difficulties disappear almost completely. For the moment, we do not discuss how one could choose a denominator $A(z)$ and we just assume that it is known.

The p-norm error (8.7) becomes a convex function. In particular, for $p = 2$, the objective is

$$J_2(B) = \frac{1}{L}\sum_{\ell=1}^{L}\lambda_\ell\left|\sum_{k=0}^{m}\alpha_{k\ell}b_k - H_\ell\right|^2 = \boldsymbol{b}^T\boldsymbol{C}\boldsymbol{b} - 2\boldsymbol{f}^T\boldsymbol{b} + \text{ct}, \tag{8.10}$$

where $\alpha_{k\ell} = \mathrm{e}^{-jk\omega_\ell}/A(\omega_\ell)$ and the positive definite matrix $\boldsymbol{C}$ and the vector $\boldsymbol{f}$ can be computed easily. The minimization of (8.10) leads to the optimal vector of coefficients $\boldsymbol{b} = \boldsymbol{C}^{-1}\boldsymbol{f}$. We denote $B_{LS}(A, \lambda)$ this optimal numerator.

For a nonnegative integer τ, the minimax problem (8.9) becomes

$$\begin{array}{ll}\min\limits_{B\in\mathbb{R}_{m+}[z],\gamma_s} & \gamma_s \\ \text{s.t.} & |B(\omega)-\mathrm{e}^{-j\tau\omega}A(\omega)|\le\gamma_p|A(\omega)|,\ \forall\omega\in[0,\omega_p] \\ & |B(\omega)|\le\gamma_s|A(\omega)|,\ \forall\omega\in[\omega_s,\pi]\end{array}\tag{8.11}$$

and can be solved exactly by appealing to the Bounded Real Lemma from Theorem 4.26. Denoting again $R_a(z)=A(z)A(z^{-1})$ and also $m'=\max(m,n+\tau)$, the problem (8.11) can be written in the SDP form

$$\begin{array}{ll}\min\limits_{\boldsymbol{b},\gamma_s^2,\boldsymbol{Q}_1,\ldots,\boldsymbol{Q}_4} & \gamma_s^2 \\ & \gamma_p^2 r_{ak}=\mathcal{L}_{k,0,\omega_p}(\boldsymbol{Q}_1,\boldsymbol{Q}_2),\quad k=0:m' \\ & \begin{bmatrix}\boldsymbol{Q}_1 & \boldsymbol{g}\\ \boldsymbol{g}^T & 1\end{bmatrix}\succeq 0 \\ & g_k=b_k-a_{k+\tau},\quad k=0:m' \\ & \gamma_s^2 r_{ak}=\mathcal{L}_{k,\omega_s,\pi}(\boldsymbol{Q}_3,\boldsymbol{Q}_4),\quad k=0:\max(m,n) \\ & \begin{bmatrix}\boldsymbol{Q}_3 & \boldsymbol{b}\\ \boldsymbol{b}^T & 1\end{bmatrix}\succeq 0\end{array}\tag{8.12}$$

We have assumed that all coefficient vectors are padded with zeros whenever necessary.

Example 8.4 We use the same specifications as in Example 5.3, i.e., $n=50$, $\omega_p=0.2\pi$, $\omega_s=0.25\pi$ and $\tau=22$. We choose a denominator $A(z)$ of order two, with zeros in $0.8\mathrm{e}^{\pm j0.2\pi}$; this is by no means an optimal denominator, but a pole angle near the passband edge and a reasonably large pole radius are typical good choices. We design two numerators as described above. The first is obtained via the minimization of the least squares objective (8.10), on a grid of $L=200$ equidistant frequencies covering $[0,\pi]$; the weights λ_ℓ are equal to 1 in the passband, 100 in the stopband and zero in the transition band. The second numerator is obtained by solving the minimax problem (8.12) for $\gamma_p=0.1$. The frequency responses of the two filters are shown in Fig. 8.2. It is interesting to note that the SDP problem (8.12) is prone to numerical errors (at least with the algorithms used by SeDuMi); in our example, it is visible that the frequency response is not exactly equiripple in the stopband. The addition of only two poles gives a clearly better filter; the minimax FIR solution of order 50 has a stopband error of −38.19 dB, while in Fig. 8.2, the stopband error is −41.07 dB. (Compare also with the design from Example 5.3.) ■

If the group delay τ is not integer, then we cannot use the BRL and a formulation like (8.12) is impossible. However, there are other algorithms, many of them designed originally for FIR filters, that can be employed in the fixed denominator case.

The p-norm objective (8.7) can be minimized with any standard descent method (e.g., Newton or conjugate gradient). Also, iterative reweighted least squares (IRLS) [1] is a useful approach. This algorithm belongs to a family of methods in which the weights of (8.10) are changed iteratively until some values $\tilde{\lambda}_\ell$ are obtained, such

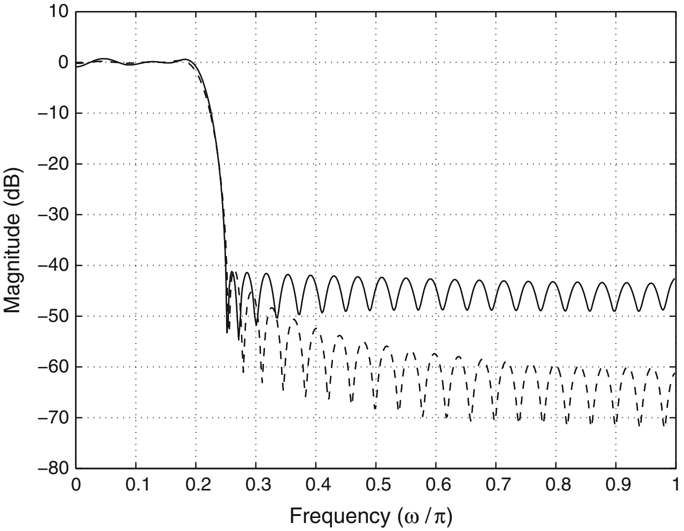


Fig. 8.2 Frequency responses of the filters designed in Example 8.4: minimax (*solid line*) and least squares (*dashed line*)

that $B_{LS}(A, \tilde{\lambda})$ is the minimizer of (8.7) or (8.8). These methods work well, even for large values of p, when the weights are uniform, i.e., $\lambda_\ell = 1$; otherwise, especially if the weights in the stopband are clearly larger than those in the passband, we have encountered numerical problems. However, the Chebyshev objective (8.8) can be minimized, apparently for arbitrary weight values, with another iterative reweighting algorithm, that from [2].

For solving (8.11), we can recur to discretization. For a given ω, a constraint of (8.11) has a SOC form, and so, on a discrete grid of frequencies, we end up with a SOCP problem. The obtained solution is, of course, only an approximation.

We conclude that there are many methods for optimizing IIR filters with fixed denominator. Barring the easy least squares case, none of the methods are perfect, but a good approximation of the solution can be expected.

8.2.2 IIR Filter Design Using Convex Stability Domains

We treat now the general case where both the numerator and the denominator of (8.1) are unknown. We study mainly the optimization of the least squares objective $J_2(A, B)$ given by (8.7) for $p = 2$. The methods described in this section have the general structure shown in Table 8.1. The initialization is usually trivial: if a good denominator is not known, we can take $A(z) = 1$. The best initial numerator is $B_{LS}(A, \lambda)$, i.e., the optimal numerator for the given denominator.

Table 8.1 Basic structure of IIR design method

Input: orders m, n of the filter (8.1); frequency points ω_ℓ and weights λ_ℓ of the least squares objective $J_2(A, B)$ from (8.7); a tolerance ε.

0. Initialize A and B.

1. Improve the objective: find D_A, D_B such that

$$J_2(A + D_A, B + D_B) < J_2(A, B), \quad A + D_A \in \mathcal{D}_A. \tag{8.13}$$

2. Put $\tilde{A} \leftarrow A + D_A$, $\tilde{B} \leftarrow B + D_B$.

3. If the relative improvement is small, i.e., the stop condition

$$\frac{J_2(A,B) - J_2(\tilde{A},\tilde{B})}{J_2(A,B)} < \varepsilon, \tag{8.14}$$

is satisfied, exit. Otherwise, put $A \leftarrow \tilde{A}$, $B \leftarrow \tilde{B}$ and go to 1.

The most important operation—common in nonlinear optimization—is to find a descent step (D_A, D_B) that improves the objective, compared to its value for the current IIR filter with numerator $B(z)$ and denominator $A(z)$. The distinguishing feature of (8.13) is that the new denominator $\tilde{A} = A + D_A$ must belong to a convex stability domain $\mathcal{D}_A$ containing $A(z)$. An obvious candidate for $\mathcal{D}_A$ is the positive realness domain described in Sect. 7.3. With a proper transformation of the least squares objective, we will be able to express each step of the iterative method from Table 8.1 as an SDP problem. We present here two such transformations.

The Steiglitz–McBride (SM) method is based on the approximation

$$\begin{aligned} J_2(\tilde{A}, \tilde{B}) &= \frac{1}{L} \sum_{\ell=1}^{L} \lambda_\ell \left| \frac{\tilde{B}(\omega_\ell)}{\tilde{A}(\omega_\ell)} - H_\ell \right|^2 \\ &\approx \frac{1}{L} \sum_{\ell=1}^{L} \frac{\lambda_\ell}{|A(\omega_\ell)|^2} \left| \tilde{B}(\omega_\ell) - H_\ell \tilde{A}(\omega_\ell) \right|^2 . \end{aligned} \tag{8.15}$$

The new objective is quadratic in the variables $\tilde{A}$ and $\tilde{B}$ (or D_A, D_B), with weights depending on the current denominator A.

The Gauss–Newton (GN) method is based on a first-order approximation of $H(\omega)$ as a function of its coefficients. We denote $\boldsymbol{d}_A, \boldsymbol{d}_B$ the vectors of coefficients of $D_A(z)$, $D_B(z)$, respectively, and $\boldsymbol{d} = [\boldsymbol{d}_A^T \ \boldsymbol{d}_B^T]^T$ the vector of optimization variables. Also, we denote $H(\omega_\ell, A, B)$ the value of (8.1) for $z = \mathrm{e}^{j\omega_\ell}$ and some denominator $A(z)$ and numerator $B(z)$. The Gauss–Newton approximation is

$$H(\omega_\ell, \tilde{A}, \tilde{B}) \approx H(\omega_\ell, A, B) + \nabla^T H(\omega_\ell, A, B) \cdot \boldsymbol{d}, \tag{8.16}$$

where $\nabla H(\omega_\ell, A, B)$ is the gradient of $H(\omega_\ell, A, B)$ with respect to the coefficients of the filter, evaluated in the current values A, B. Using (8.16), the optimization objective is approximated with

$$J_2(\tilde{A}, \tilde{B}) \approx \frac{1}{L} \sum_{\ell=1}^{L} \lambda_\ell \left| \nabla^T H(\omega_\ell, A, B) \cdot \boldsymbol{d} + \frac{B(\omega_\ell)}{A(\omega_\ell)} - H_\ell \right|^2 . \tag{8.17}$$

This is also a quadratic form in $\boldsymbol{d}$.

It results that both (8.15) and (8.17) have the form

$$J_2(\tilde{A}, \tilde{B}) \approx \boldsymbol{d}^T \boldsymbol{C} \boldsymbol{d} + \boldsymbol{f}^T \boldsymbol{d} + \text{ct}, \tag{8.18}$$

where $\boldsymbol{C}$ is a positive semidefinite matrix and $\boldsymbol{f}$ a vector, both known. Using the positive realness stability domain from Sect. 7.3, an iteration of the method from Table 8.1 consists of solving the problem

$$\begin{array}{cl} \min\limits_{\boldsymbol{d} \in \mathbb{R}^{m+n+1},\ R \in \mathbb{R}_n[z]} & \boldsymbol{d}^T \boldsymbol{C} \boldsymbol{d} + \boldsymbol{f}^T \boldsymbol{d} \\ \text{s.t.} & \boldsymbol{r} = (\boldsymbol{F} + \boldsymbol{G}) \boldsymbol{d}_A + 2\boldsymbol{G}\boldsymbol{a} \\ & R(\omega) \geq 0,\ \forall \omega \in [-\pi, \pi] \end{array} \tag{8.19}$$

where the matrices $\boldsymbol{F}$ and $\boldsymbol{G}$ are defined in (7.39) and the vector $\boldsymbol{r}$ contains the distinct coefficients of the nonnegative trigonometric polynomial $R(z)$. The first constraint from (8.19) is (7.38). Using the trace (2.6) or Gram-pair (2.94) parameterizations, we transform (8.19) into an SQLP problem.

The solutions D_A, D_B can be used as descent steps (as in Table 8.1, especially for the Steiglitz–McBride method) or as descent directions (especially for the Gauss–Newton method). In the latter case, we perform a unidimensional search to find the optimal value $\alpha \in [0, 1]$ for which $J_2(A + \alpha D_A, B + \alpha D_B)$ is minimum and then put $\tilde{A} = A + \alpha D_A$, $\tilde{B} = B + \alpha D_B$ in step 2 from Table 8.1.

Remark 8.5 (Robust stability) As shown in Examples 7.21 and 7.22, the border of the positive realness domain $\mathcal{D}_A$ may coincide with that of the set of stable polynomials. Consequently, the iterative algorithm described above may produce IIR filters with poles arbitrarily close to the unit circle. In applications, a certain robustness of stability is usually required. For example, we can impose the constraint that the poles lie inside a circle of radius $\rho < 1$, denoted $\mathcal{C}_\rho$. We assume that, in a certain iteration of the algorithm from Table 8.1, the zeros of the current denominator $A(z)$ are in $\mathcal{C}_\rho$. We define $A^\rho(z) = A(\rho z)$; then, the zeros of $A^\rho(z)$ are in $\mathcal{C}_1$, i.e., $A^\rho(z)$ is stable. Denoting $\boldsymbol{\Gamma} = \text{diag}(\rho^{-1}, \rho^{-2}, \ldots, \rho^{-n})$, we have $\boldsymbol{a}^\rho = \boldsymbol{\Gamma}\boldsymbol{a}$, where $\boldsymbol{a}^\rho$ is the vector of the coefficients of $A^\rho(z)$. We also denote $D_A^\rho(z) = D_A(\rho z)$ and $\tilde{A}^\rho(z) = \tilde{A}(\rho z) = A^\rho(z) + D_A^\rho(z)$. To have the zeros of $\tilde{A}(z)$ in $\mathcal{C}_\rho$, we impose the stability conditions on $\tilde{A}^\rho$ instead of $\tilde{A}$. The problem (8.19) becomes

$$\begin{array}{cl} \min\limits_{\boldsymbol{d}, R} & \boldsymbol{d}^T \boldsymbol{C} \boldsymbol{d} + \boldsymbol{f}^T \boldsymbol{d} \\ \text{s.t.} & \boldsymbol{r} = (\boldsymbol{F}_\rho + \boldsymbol{G}_\rho) \boldsymbol{\Gamma} \boldsymbol{d}_A + 2\boldsymbol{G}_\rho \boldsymbol{\Gamma} \boldsymbol{a} \\ & R(\omega) \geq 0,\ \forall \omega \in [-\pi, \pi] \end{array} \tag{8.20}$$

where $\boldsymbol{F}_\rho$ and $\boldsymbol{G}_\rho$ are the matrices from (7.39) with A^ρ replacing A. ■

Remark 8.6 An extensive study [3] showed that it is difficult to say that one of the SM or GN methods is better than the other. However, using first SM and then GN, initialized with the result of SM, gives significantly better results than a single method. Moreover, refining the solution with a simple descent method (see problem **P** 8.2) is useful in some cases. We name SMGNR this combined method. ■

Example 8.7 In Example 8.4, with $m = 50$ and $n = 2$, we have chosen the poles heuristically. We perform now the full optimization with SMGNR, initialized trivially with $A(z) = 1$. We set the maximal stability radius to $\rho = 0.8$. The objective $J_2(A, B)$ is improved with about 9% with respect to the fixed denominator case ($1.078 \cdot 10^{-5}$ with SMGNR versus $1.177 \cdot 10^{-5}$ in Example 8.4). The optimized poles are $0.8 \cdot \mathrm{e}^{\pm j0.2353\pi}$. We remark that the poles have maximal radius, which is the intuitively correct result. ■

Example 8.8 We keep the same design data as before, but increasing the denominator order to $n = 4$ and the maximum pole radius to $\rho = 0.9$. The filter designed with SMGNR has the frequency response shown in Fig. 8.3 with dashed line. The value of the least squares objective is $3.42 \cdot 10^{-6}$, i.e., about 3 times smaller than for $n = 2$, $\rho = 0.8$. The poles are $0.9 \cdot \mathrm{e}^{\pm j0.2187\pi}$ and $0.865 \cdot \mathrm{e}^{\pm j0.2256\pi}$.

A quick minimax design can be obtained by applying the iterative reweighting algorithm from [2], for the denominator designed with SMGNR. The equiripple response from Fig. 8.3 (solid line) was obtained by minimizing the objective (8.8) for weights equal to 1 in the passband and 10 in the stopband. (As a result, the maximal error is 0.0435 in the passband and 0.00433 = −47.26 dB in the stopband.) ■

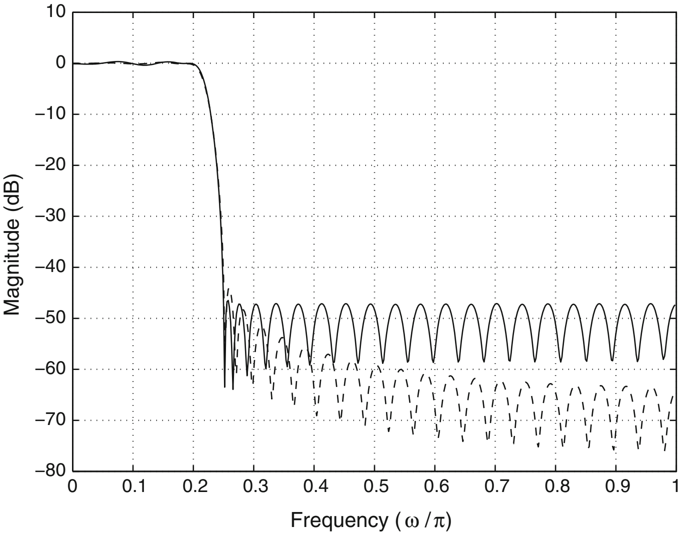


Fig. 8.3 Frequency responses of the filters designed in Example 8.8: Least squares (*dashed line*) and minimax (*solid line*)

8.3 2D IIR Filter Design

Many of the ideas for IIR filter design can be generalized to the 2D case, where the transfer function is (compare with (8.1))

$$H(z) = \frac{B(z)}{A(z)} = \frac{\sum_{i=0}^{m} b_i z^{-i}}{\sum_{k=0}^{n} a_k z^{-k}}, \tag{8.21}$$

where $\boldsymbol{n} = (n_1, n_2)$ etc. and $a_{\mathbf{0}} = 1$. There are two main difficulties: the increased complexity of the optimization problems and the more demanding treatment of stability. If not disregarded at all (by the hands-on approach: “design first, then check”), the stability issue can be alleviated in several ways, with a loss of optimality. One can express $A(z)$ as a product of factors of degree $(1, 1)$, for which the stability conditions are linear; besides the suboptimality, this gives an optimization objective with more local minima. Also, we can work with a separable denominator $A(z_1, z_2) = A_1(z_1)A_2(z_2)$, which is a product of univariate polynomials; in this case, the treatment of stability is identical to the 1D case. Although lacking generality, these factorization approaches have the advantage that the (hardware or software) implementation of the 2D IIR filter is more efficient. So, a trade-off optimality/complexity may become interesting.

In this section, we discuss only the general case of nonseparable denominator. Since, as in the FIR case, the magnitude design is impossible in 2D due to the lack of spectral factorization, we examine the approximate linear-phase design. Given two frequency domains, $\mathcal{D}_p$ for passband and $\mathcal{D}_s$ for stopband, the ideal frequency response is

$$H_{id}(\omega_1, \omega_2) = \begin{cases} e^{-j(\tau_1\omega_1 + \tau_2\omega_2)}, & \omega \in \mathcal{D}_p, \\ 0, & \omega \in \mathcal{D}_s, \end{cases} \tag{8.22}$$

where $\boldsymbol{\tau} = (\tau_1, \tau_2)$ is the group delay. The objectives (8.7) and (8.8) generalize immediately to 2D, using a set of frequencies $\boldsymbol{\omega}_\ell$, $\ell = 1 : L$; typically, the set is an $L_1 \times L_2$ grid.

If the denominator is fixed, the least squares reduce to a positive semidefinite quadratic expression, similar to (8.10), and so become an easy problem. The minimax optimization can be approached, as in the 1D case, via the Bounded Real Lemma; see problem **P** 8.3.

The algorithm structure from Table 8.1 is valid also for the complete optimization of a 2D IIR filter. The convex stability domain used in the 2D case is the sum-of-squares version of the positive realness domain, described by Theorem 7.18. However, unlike the 1D case, only the Gauss–Newton method has given good results in the author’s experiments [4] (while the greedy character of the Steiglitz–McBride method appears to prevent the approach of “good” local minima). Using the GN transformation (8.17), (8.18) and the usual notations (e.g., $\boldsymbol{d}_A$ for the vector of coefficients of $D_A(z)$), the descent direction in the 2D GN method is found by solving the SDP problem

$$\begin{aligned} \min_{\boldsymbol{d},\,R} \;& \boldsymbol{d}^T \boldsymbol{C} \boldsymbol{d} + \boldsymbol{f}^T \boldsymbol{d} \\ \text{s.t.}\;& \boldsymbol{r} = (\boldsymbol{F} + \boldsymbol{G}) \boldsymbol{d}_A + 2\boldsymbol{G}\boldsymbol{a} \\ & R \in \mathbb{RS}^{\boldsymbol{n}}_{\boldsymbol{n}}[z] \end{aligned} \tag{8.23}$$

where the matrices $\boldsymbol{F}$ and $\boldsymbol{G}$ are given by (7.40) and (7.42), respectively. The descent step is found by unidimensional search. The poles of the IIR filter can be forced to have radius less than a given ρ, as shown in Remark 8.5.

The initialization of the GN algorithm becomes more important than in the 1D case. Since the main danger seems to be a fast advance toward the border of the (robust) stability region, a cure is the following. The GN algorithm is run first with a reduced maximal pole radius, e.g., equal to 0.9ρ. The result is used as initialization for a new run, this time with the nominal value ρ.

Example 8.9 We consider the ideal response

$$|H_{id}(\omega_1, \omega_2)| = \begin{cases} 1, & if \sqrt{\omega_1^2 + \omega_2^2} \le \omega_p, \\ 0, & if \sqrt{\omega_1^2 + \omega_2^2} \ge \omega_s, \end{cases} \tag{8.24}$$

with $\omega_p = 0.5\pi$, $\omega_s = 0.7\pi$. The passband and stopband have circular border, as shown in Fig. 8.4. The filter orders are $\boldsymbol{m} = (12, 12)$ and $\boldsymbol{n} = (4, 4)$. The group delay is $\boldsymbol{\tau} = (7, 7)$. The frequency grid is uniform and has 80 × 80 points. (Actually only half of them are sufficient, those covering a half plane.) The weights are 1 in the passband and stopband and 0 in the transition band. The maximal pole radius is $\rho = 0.9$. The GN algorithm produces the filter whose frequency response is shown in Fig. 8.5. The value of the least squares objective is $J_2(A, B) = 1.63 \cdot 10^{-5}$. ■

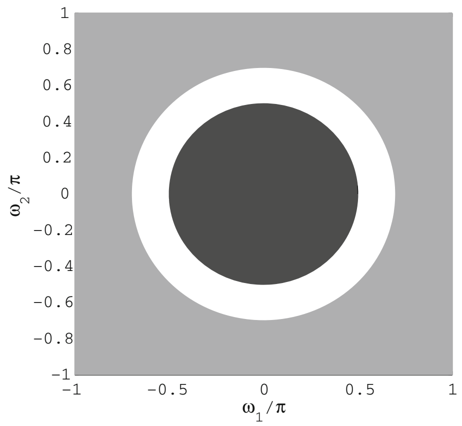


Fig. 8.4 Passband (*black*) and stopband (*gray*) for the filter designed in Example 8.9

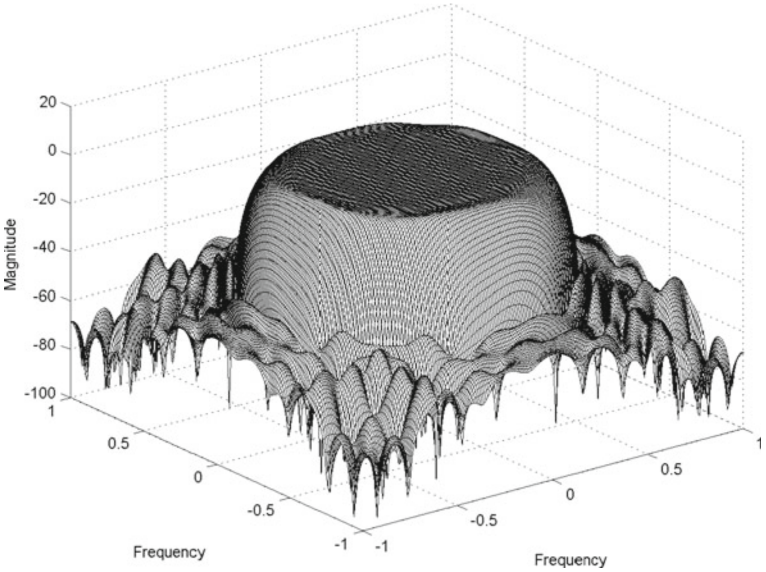


Fig. 8.5 Frequency response of the filter designed in Example 8.9

8.4 Bibliographical and Historical Notes

There is a huge literature on IIR filter design, and so we select only few references relevant to the contents of this chapter.

The magnitude optimization method based on positive polynomials, presented in Sect. 8.1, has been proposed in [5]. Previous methods for magnitude optimization were based mostly on modifications of the Remez exchange algorithm, with no guarantee of convergence.

Ensuring stability has been performed by various means. A simple idea, in the context of descent methods, is to reduce (e.g., to halve) the descent step if the new denominator is unstable. The fixed (and restrictive) stability domain $\mathrm{Re}[\tilde{A}(\omega)] > 0$ was used in [6, 7]. Stability domains built around the current denominator (in an iterative process) may be based on Rouché's criterion [8] or the positive realness condition from Theorem 7.16 [3]. Other means to obtain stability are a barrier term added to the objective [9] or a Lyapunov condition leading to an SDP formulation [10]. A stability constrain based on the argument principle was proposed in [11].

Optimization methods for IIR filters based on standard descent methods and used in the 1980s are usually not successful, unless a good initialization is available. The methods reported in Sect. 8.2 are Steiglitz–McBride [7, 12], Gauss–Newton [8] and their successive use [3]. The idea to relate the search direction with a stability domain appeared in [8], in an implicit form, and was conceptualized in [3]. Other optimization ideas include sequentially constrained least squares [13], SDP relaxation [14], and second-order cone programming [15].

Some methods for the optimization of IIR filters with fixed denominator are reviewed in Sect. 8.2.1. That a minimax IIR filter, obtained as in Example 8.8 by using the poles generated by the least squares optimization with SMGNR and the iterative reweighting method from [2], can be near-optimal (and easy to design) has been argued in [16].

There are relatively few ideas for treating stability in the design of 2D IIR filter with nonseparable denominator. The sufficient condition Re$[\tilde{A}(\boldsymbol{\omega}] > 0$ was used in [17]. A kind of barrier based on the distance to a stable spectral factor has been employed in [18]. The sum-of-squares stability domain from Theorem 7.18 has been used in [4].

It is hard to argue which one of the separable or nonseparable denominator IIR filters are better [4]. However, there are examples [19] where nonseparable denominators can have better performance even if the ideal response is quadrantally symmetric. (Filters with quadrantally symmetric frequency response have separable denominators [20].)

Besides those described in this chapter, there are other methods using convex programming in each iteration, e.g., linear programming [17], SOCP [21], SDP [7] or more general [8].

Problems

P 8.1 (Frequency response fitting with IIR model [5]) We have the power spectrum measurements $|F(\omega_\ell)|^2 = R_\ell$, $\ell = 1 : L$, of a certain process F. We want to approximate it with an IIR model (8.1). Using the notations (8.4), we can find $H(z)$ by solving the minimax problem

$$\begin{array}{cl} \min\limits_{R_a \in \mathbb{R}_n[z], R_b \in \mathbb{R}_m[z]} & \max_{\ell=1:L} \left| \frac{R_b(\omega_\ell)}{R_a(\omega_\ell)} - R_\ell \right| \\ \text{s.t.} & R_a(\omega) \geq 0, \;\; R_b(\omega) \geq 0, \;\; \forall \omega \in [-\pi, \pi] \\ & r_{a0} = 1 \end{array}$$

Express this problem in SDP form.

Can we obtain an SDP problem if the optimization objective is quadratic, i.e., $\sum_{\ell=1}^{L} \left| \frac{R_b(\omega_\ell)}{R_a(\omega_\ell)} - R_\ell \right|^2$? Compare with the FIR case from problem **P** 5.6.

P 8.2 In the basic algorithm from Table 8.1, a descent direction D_A, D_B can be found using standard nonlinear optimization algorithms (like Newton or conjugate gradient). Assuming that D_A is known, a maximum descent step α_m (in the sense that $A + \alpha_m D_A$ is on the border of $\mathcal{D}_A$) can be found by solving the problem

$$\begin{array}{rl} \alpha_m = & \max\limits_{\alpha} \; \alpha \\ & \text{s.t.} \;\; 1 + \alpha \text{Re} \frac{D_A(\omega)}{A(\omega)} > 0, \;\; \forall \omega \in [-\pi, \pi] \end{array}$$

Put this optimization problem in SDP form.

P 8.3 If the denominator of the 2D IIR filter (8.21) is known, then the minimax optimization of the filter can be formulated as

$$\begin{aligned} &\min_{B,\gamma_s} \ \gamma_s \\ &\text{s.t.} \ \ |B(\boldsymbol{\omega}) - \mathrm{e}^{-j(\tau_1\omega_1+\tau_2\omega_2)}A(\boldsymbol{\omega})| \le \gamma_p |A(\boldsymbol{\omega})|, \ \forall\omega \in \mathcal{D}_p \\ &\qquad |B(\boldsymbol{\omega})| \le \gamma_s |A(\boldsymbol{\omega})|, \ \forall\omega \in \mathcal{D}_s \end{aligned} \tag{8.25}$$

If the passband $\mathcal{D}_p$ and the stopband $\mathcal{D}_s$ are frequency domains defined as in (4.13), by the positivity of some trigonometric polynomials, and the group delays τ_1 and τ_2 are integers, relax the problem (8.25) to an SDP form, using the BRL Theorem 4.26. Compare this SDP problem with the 1D case (8.12).

References

1. C.S. Burrus, J.A. Barreto, I.W. Selesnick, Iterative reweighted least-squares design of FIR filters. IEEE Trans. Sign. Proc. **42**(11), 2926–2936 (1994)
2. Y.C. Lim, J.-H. Lee, C.K. Chen, R.H. Yang, A weighted least-squares algorithm for quasi-equiripple FIR and IIR filter design. IEEE Trans. Signal Process. **40**(3), 551–558 (1992)
3. B. Dumitrescu, R. Niemistö, Multistage IIR filter design using convex stability domains defined by positive realness. IEEE Trans. Signal Proc. **52**(4), 962–974 (2004)
4. B. Dumitrescu, Optimization of 2D IIR filters with nonseparable and separable denominator. IEEE Trans. Signal Proc. **53**(5), 1768–1777 (2005)
5. B. Alkire, L. Vandenberghe, Convex optimization problems involving finite autocorrelation sequences. Math. Progr. Ser. A **93**(3), 331–359 (2002)
6. A.T. Chottera, G.A. Jullien, A linear programming approach to recursive digital filter design with linear phase. IEEE Trans. Circ. Syst. CAS **29**(3), 139–149 (1982)
7. W.-S. Lu, S.-C. Pei, C.-C. Tseng, A weighted least-squares method for the design of stable 1-D and 2-D IIR digital filters. IEEE Trans. Signal Process. **46**, 1–10 (1998)
8. M.C. Lang, Least-squares design of IIR filters with prescribed magnitude and phase response and a pole radius constraint. IEEE Trans. Signal Process. **48**(11), 3109–3121 (2000)
9. A. Tarczynski, G.D. Cain, E. Hermanowicz, M. Rojewski, A WISE method for designing IIR filters. IEEE Trans. Signal Process. **49**(7), 1421–1432 (2001)
10. W.-S. Lu, Design of stable minimax IIR digital filters using semidefinite programming, in *IEEE International Symposium Circuit System (ISCAS)*, vol. 1 (Geneva, Switzerland 2000), pp. 355–358
11. A. Jiang, H.K. Kwan, IIR digital filter design with new stability constraint based on argument principle. IEEE Trans. Circ. Syst. I **56**(3), 583–593 (2009)
12. K.E. Steiglitz, L.E. McBride, A technique for the identification of linear systems. IEEE Trans. Auto. Control AC **10**, 461–464 (1965)
13. X. Lai, Z. Lin, Minimax design of IIR digital filters using a sequential constrained least-squares method. IEEE Trans. Signal Proc. **58**(7), 3901–3906 (2010)
14. A. Jiang, H.K. Kwan, Minimax design of IIR digital filters using SDP relaxation technique. IEEE Trans. Circ. Syst. I **57**(2), 378–390 (2010)
15. R.C. Nongpiur, D.J. Shpak, A. Antoniou, Improved design method for nearly linear-phase IIR filters using constrained optimization. IEEE Trans. Signal Proc. **61**(4), 895–906 (2013)
16. R. Niemistö, B. Dumitrescu, Simplified procedures for Quasi-Equiripple IIR filter design. IEEE Signal Proc. Lett. **11**(3), 308–311 (2004)
17. A.T. Chottera, G.A. Jullien, Design of Two-dimensional recursive digital filters using linear programming. IEEE Trans. Circ. Syst. CAS **29**(12), 817–826 (1982)

18. J.-H. Lee, Y.-M. Chen, A new method for the design of two-dimensional recursive digital filters. IEEE Trans. Acoust. Speech Signal Proc. **36**(4), 589–598 (1988)
19. Z. Lin, L.T. Bruton, N.R. Bartley, Design of highly selective two-dimensional recursive fan filters by relaxing symmetry constraints. Electron. Lett. **24**(22), 1361–1362 (1988)
20. P.K. Rajan, M.N.S. Swamy, Quadrantal symmetry associated with two-dimensional digital transfer functions. IEEE Trans. Circ. Syst. CAS **25**(6), 341–344 (1978)
21. W.S. Lu, T. Hinamoto, Optimal design of IIR digital filters with robust stability using conic-quadratic-programming updates. IEEE Trans. Signal Proc. **51**(6), 1581–1592 (2003)

Chapter 9
Optimization with the Atomic Norm

Abstract Sparse representations have shown great potential in achieving parsimony, with applications in various signal processing topics. Typically, sparse representations are made with the help of an overcomplete basis, called also dictionary. The case of infinite dictionaries depending on a few parameters with continuous values has been tackled recently. The atomic norm is the mathematical notion that helps finding sparse representations in this case. Of particular interest are the dictionaries formed on the basis of trigonometric polynomials, because they are connected with elementary signal processing problems such as identifying the frequencies of a sum of sinusoids (called also line spectrum estimation) or finding the direction of arrival (DOA) of radio or sound sources, using linear arrays of sensors. It turns out that optimization problems with this instance of atomic norm are intimately connected with the Bounded Real Lemma for trigonometric polynomials and can be reduced through it to SDP problems. This chapter presents the basic atomic norm optimization problem and its solution via BRL, together with several extensions such as the matrix or 2D case, solved by the appropriate BRL forms. Besides line spectrum and DOA estimation, the important and fertile problem of super-resolution is discussed. The presentation skips the underlying theory that guarantees the success of the atomic norm approach, whose main hypothesis is that the frequencies of the sinusoids are sufficiently well separated, and insists on the optimization problems that effectively give the solution.

9.1 Sparse Representations

Traditionally, discrete-time signals are represented or analyzed with the help of orthogonal bases, like those given by the Discrete Fourier, Discrete Cosine, Walsh–Hadamard, or other transforms. Nonorthogonal bases, such as those given by some wavelets transforms, have also found various uses. Departing from this trend, the latest 20 years have seen an increased popularity of overcomplete bases or frames, used especially in the context of sparse representations. Let $\boldsymbol{A} \in \mathbb{R}^{m \times M}$ a matrix

B. Dumitrescu, *Positive Trigonometric Polynomials and Signal Processing Applications*, Signals and Communication Technology,
DOI 10.1007/978-3-319-53688-0_9

with more columns than rows, i.e., $M > m$, which can be seen as a redundant basis for $\mathbb{R}^m$. Such a matrix is often called a dictionary and its columns $\boldsymbol{a}_i$, $i = 1 : M$, are called atoms; typically, all atoms have norm equal to 1. A signal $\boldsymbol{x} \in \mathbb{R}^m$ has a sparse representation if it can be expressed as the linear combination

$$\boldsymbol{x} = \sum_{i \in \mathcal{I}} c_i \boldsymbol{a}_i \tag{9.1}$$

of a few atoms, where the index set $\mathcal{I}$ represents the support of the signal and $|\mathcal{I}| = s$, with $s \ll M$ and $s < m$. An overcomplete dictionary offers many more possibilities of sparse representations than a usual basis and, in particular, many more subspaces of dimension s where the representations can lie. Many applications in, for example, compression, denoising, and image processing have shown that such dictionaries allow parsimonious representations and clear benefits with respect to traditional transforms.

The basic problem in this context is to find the sparse representation (9.1) of a given signal $\boldsymbol{x}$ (assuming that such a representation indeed exists). Even if s is known, this is essentially an NP-hard problem, since all combinations of s atoms are candidates. There are many heuristics for solving the problem, some having guaranteed success depending on the properties of the dictionary. Obtaining the representation is more likely if the atoms are well spread on all directions (the scalar product of two atoms is small) and s is small.

A particularly successful idea (although one of the most computationally demanding) is to solve the optimization problem

$$\begin{aligned} \min_{\boldsymbol{c} \in \mathbb{R}^M} \quad & \|\boldsymbol{c}\|_1 \\ \text{s.t.} \quad & \boldsymbol{A}\boldsymbol{c} = \boldsymbol{x} \end{aligned} \tag{9.2}$$

Here, the ℓ_1 norm is defined by $\|\boldsymbol{c}\|_1 = \sum_i |c_i|$.

The problem (9.2) can be seen as a convex relaxation of the minimization of the ℓ_0 "norm" $\|\boldsymbol{c}\|_0$, which is the number of nonzero elements of the vector $\boldsymbol{c}$. (Note that this is not a norm since it is not homogeneous.) Under certain conditions, the solution of the convex (linear programming, in fact) problem (9.2) is indeed the sparsest representation of $\boldsymbol{x}$. Some variations of the problem, such as the well-known lasso, allow the recovery of noisy representations.

The topic of this chapter is the extension of sparse representations to dictionaries with an infinite number of atoms. In this case, instead of being the columns of a matrix $\boldsymbol{A}$, the atoms have a parametric form, depending on one or more real parameters. So, the vector $\boldsymbol{c}$ from (9.2) would have infinite size, case in which the ℓ_1 norm is not defined. A new tool is necessary, and this is the atomic norm, presented in the next section.

9.2 Atomic Norms

We replace the matrix dictionary $\boldsymbol{A}$ with a set of atoms $\mathcal{A}$, also called dictionary. This set may be finite, although the infinite case is the most interesting. We assume that $\mathcal{A}$ is centrally symmetric about the origin, i.e., $\boldsymbol{a} \in \mathcal{A}$ implies $-\boldsymbol{a} \in \mathcal{A}$, and that all atoms are extreme points of the convex hull of $\mathcal{A}$, i.e., no atom lies in the convex hull of the other atoms.

Definition 9.1 The *atomic norm* associated with $\mathcal{A}$ is

$$\|\boldsymbol{x}\|_{\mathcal{A}} = \inf\{t \geq 0 \mid \boldsymbol{x} \in t \cdot \mathrm{conv}(\mathcal{A})\}, \tag{9.3}$$

where $\mathrm{conv}(\mathcal{A})$ is the convex hull of $\mathcal{A}$. ■

It can be proved that

$$\|\boldsymbol{x}\|_{\mathcal{A}} = \inf\left\{\sum_{\boldsymbol{a}\in\mathcal{A}} c_{\boldsymbol{a}} \mid \boldsymbol{x} = \sum_{\boldsymbol{a}\in\mathcal{A}} c_{\boldsymbol{a}}\boldsymbol{a},\ \text{with } c_{\boldsymbol{a}} \geq 0, \forall \boldsymbol{a} \in \mathcal{A}\right\}. \tag{9.4}$$

Indeed, for a short justification, consider $z \in \mathrm{conv}(\mathcal{A})$, but $z \notin \mathcal{A}$, and z on the boundary of $\mathrm{conv}(\mathcal{A})$. By directly using the representation tz in (9.3), it results that $\|z\|_{\mathcal{A}} = t = 1$. However, by the definition of $\mathrm{conv}(\mathcal{A})$, the vector z can be expressed as the convex combination of some atoms, $z = \sum_i c_i \boldsymbol{a}_i$, $\sum_i c_i = 1$. Then, the triangle inequality gives $\|z\|_{\mathcal{A}} \leq \sum_i c_i \|\boldsymbol{a}_i\|_{\mathcal{A}} = 1$, since $\|\boldsymbol{a}_i\|_{\mathcal{A}} = 1$ by definition. So, in a linear combination of vectors from $\mathrm{conv}(\mathcal{A})$, one can always replace a vector that is not an atom with a linear combination of atoms without increasing the sum of the coefficients.

The dual atomic norm is

$$\|\boldsymbol{h}\|_{\mathcal{A}}^{*} \overset{\Delta}{=} \sup_{\|\boldsymbol{x}\|_{\mathcal{A}}\leq 1} \mathrm{Re}[\boldsymbol{x}^H \boldsymbol{h}]. \tag{9.5}$$

We note that the unit ball $\|\boldsymbol{x}\|_{\mathcal{A}} \leq 1$ in the atomic norm is $\mathrm{conv}(\mathcal{A})$. Since the atoms are extremal points of this unit ball, it follows that

$$\|\boldsymbol{h}\|_{\mathcal{A}}^{*} = \sup_{\boldsymbol{a}\in\mathcal{A}} \mathrm{Re}[\boldsymbol{a}^H \boldsymbol{h}]. \tag{9.6}$$

There are several sets of atoms that can be useful in signal processing applications and not only. A simple and not really relevant example is the ℓ_1 norm, which is the atomic norm if the set $\mathcal{A}$ is made of the unit vectors (this is in fact the case of the finite dictionary $\boldsymbol{A} = \boldsymbol{I}$); note that this interpretation is not helpful in the context of problem (9.2), since it would mean that $|\boldsymbol{x}| = \boldsymbol{c}$; a signal is usually sparse in some transform domain, not in the standard basis. Another example is that of the nuclear norm in the context of low-rank matrices; the atoms are rank-1 matrices with unit Frobenius norm. In the context of this book, the most interesting example, which will be discussed in detail in the remainder of this section, is that of atoms that are extracted from the canonical basis for trigonometric polynomials, on the unit circle.

9.2.1 Canonical Polynomial Basis Atoms

Trigonometric polynomials are naturally connected with the set

$$\mathcal{A} = \left\{ \boldsymbol{a}(\omega, \varphi) = \mathrm{e}^{j\varphi} \boldsymbol{\psi}(\omega) \mid \omega, \varphi \in [-\pi, \pi] \right\}, \tag{9.7}$$

where $\boldsymbol{\psi}(\omega) = [1 \ \mathrm{e}^{j\omega} \ \ldots \ \mathrm{e}^{jn\omega}]$, like in (2.1). This is clearly a parametric dictionary since the atoms depend on two parameters that take continuous values. A sparse linear combination of vectors from $\mathcal{A}$ is in fact a sum of sinusoids. Indeed, if

$$\boldsymbol{x} = \sum_i c_i \boldsymbol{a}(\omega_i, \varphi_i), \quad c_i \geq 0, \tag{9.8}$$

is such a linear combination, then an element of the vector $\boldsymbol{x}$ has the expression

$$x_\ell = \sum_i c_i \mathrm{e}^{j\varphi_i} \mathrm{e}^{j\ell\omega_i} \tag{9.9}$$

and hence the signal is a sum of discrete-time sinusoids of frequencies ω_i with complex coefficients $c_i \mathrm{e}^{j\varphi_i}$. Estimating the frequencies of the sinusoids, given a number of noisy measurements, is a common signal processing problem, often named line spectrum estimation. Before discussing specific problems, let us study the atomic norm in this case.

The dual norm is

$$\|\boldsymbol{h}\|_{\mathcal{A}}^* = \sup_{\omega,\varphi} \mathrm{Re}[\boldsymbol{h}^H \boldsymbol{a}(\omega, \varphi)] = \sup_{\omega,\varphi} \left| \mathrm{e}^{-j\varphi} \sum_{k=0}^{n} h_k \mathrm{e}^{-jk\omega} \right| = \max_{|z|=1} \left| \sum_{k=0}^{n} h_k z^{-k} \right|. \tag{9.10}$$

The first equality is (9.6), the second is obtained from the definition (9.7), and the third is a simple conclusion. So, the dual norm is the maximum of a causal trigonometric polynomial on the unit circle, i.e., the H_∞ norm of the FIR system whose coefficients are in the vector $\boldsymbol{h}$, see (4.41). By dualizing the dual atomic norm and using the BRL from Corollary 4.27, which holds with nonstrict inequality in the unidimensional case, we can express the atomic norm as

$$\begin{aligned} \|\boldsymbol{x}\|_{\mathcal{A}} = \sup_{\|\boldsymbol{h}\|_{\mathcal{A}}^* \leq 1} \mathrm{Re}[\boldsymbol{x}^H \boldsymbol{h}] = \ & \max_{\boldsymbol{h}, \boldsymbol{Q}} \ \mathrm{Re}[\boldsymbol{x}^H \boldsymbol{h}] \\ & \text{s.t.} \ \ \delta_k = \mathrm{tr}[\boldsymbol{\Theta}_k \boldsymbol{Q}], \ k = 0 : n, \\ & \qquad \begin{bmatrix} \boldsymbol{Q} & \boldsymbol{h} \\ \boldsymbol{h}^H & 1 \end{bmatrix} \succeq 0 \end{aligned} \tag{9.11}$$

We have thus obtained an SDP formulation for computing the atomic norm. The exact decomposition (9.8) can be then computed as shown by the following theorem.

Theorem 9.2 *Let (9.8) be the linear combination of atoms for which the minimum is attained in (9.4). The frequencies ω_i from (9.8) are those for which $|H(\omega_i)| = 1$, where $H(z)$ is the causal trigonometric polynomial whose coefficients are given by the vector $\boldsymbol{h}$ which is the solution of (9.11).*

Proof Let $\mathcal{O}$ be the set of frequencies $\tilde{\omega}_i$ for which $|H(\tilde{\omega}_i)| = |\boldsymbol{a}(\tilde{\omega}_i, \varphi)^H \boldsymbol{h}| = 1$ (the value of φ is arbitrary). If $\omega \notin \mathcal{O}$, then $|H(\omega)| < 1$. If $\boldsymbol{h}$ is the solution of (9.11), then $\|\boldsymbol{x}\|_{\mathcal{A}} = \text{Re}[\boldsymbol{x}^H \boldsymbol{h}]$. Noting that (9.8) means also $\|\boldsymbol{x}\|_{\mathcal{A}} = \sum_i c_i$, it results that

$$\|\boldsymbol{x}\|_{\mathcal{A}} = \text{Re}[\boldsymbol{x}^H \boldsymbol{h}] = \sum_i c_i |\boldsymbol{a}(\omega_i, \varphi_i)^H \boldsymbol{h}| = \sum_i c_i |H(\omega_i)| \le \sum_i c_i.$$

The equality is attained only if $\omega_i \in \mathcal{O}$. ■

So, the frequencies defining the optimal atoms from (9.8) are the extrema of $|H(\omega)|$ and can be computed, for example, as the roots on the unit circle of the symmetric polynomial $R(z) = 1 - H^*(z^{-1})H(z)$; note that these roots are double, since they are also extreme points. With the frequencies ω_i available, computing the coefficients from (9.8) can be trivially done by solving a least squares linear system (having in principle an exact solution). Denoting s the number of frequencies, the system is

$$\left[\boldsymbol{a}(\omega_1, 0) \ \ldots \ \boldsymbol{a}(\omega_s, 0) \right] \hat{\boldsymbol{c}} = \boldsymbol{x}. \tag{9.12}$$

Expressing the coefficients of the solution as $\hat{c}_i = c_i e^{j\varphi_i}$, with $c_i > 0$, we completely retrieve the information from (9.8).

We will see later when the representation (9.8) is necessary. For the moment, let us give a simple example of computation.

Example 9.3 Let $n = 31$ and $\boldsymbol{x} = 1.5\boldsymbol{a}(0.2\pi) - 2\boldsymbol{a}(0.3\pi) + 2.5\boldsymbol{a}(0.6\pi)$. The phase φ is ignored here, but it is clear that the coefficients become positive if $\varphi_2 = \pi$ and $\varphi_1 = \varphi_3 = 0$. By solving (9.11), we obtain the expected result $\|\boldsymbol{x}\|_{\mathcal{A}} = 6$, i.e., the sum of the (positive) coefficients. The graph of $|H(\omega)|$ is shown in Fig. 9.1. It is visible that the maximum value is 1 and it is attained in exactly three points, namely the frequencies defining the atoms with which the signal has been created. ■

This example might suggest that the above method for computing the atomic norm can recover any sparse linear combination of atoms. This is true only if the minimum distance between two distinct frequencies is larger than a bound, as will be later explained in Sect. 9.3.3. If the frequencies are not well separated, the program (9.11) still gives the atomic norm, but the condition $|H(\omega)| = 1$ will possibly give a different set of frequencies, that corresponding to the *optimal atoms*. For instance, if in Example 9.3 we change the last frequency from 0.6π into 0.32π, then the atomic norm is 5.1718 (not 6 as we might hastily assume) and the optimal representation

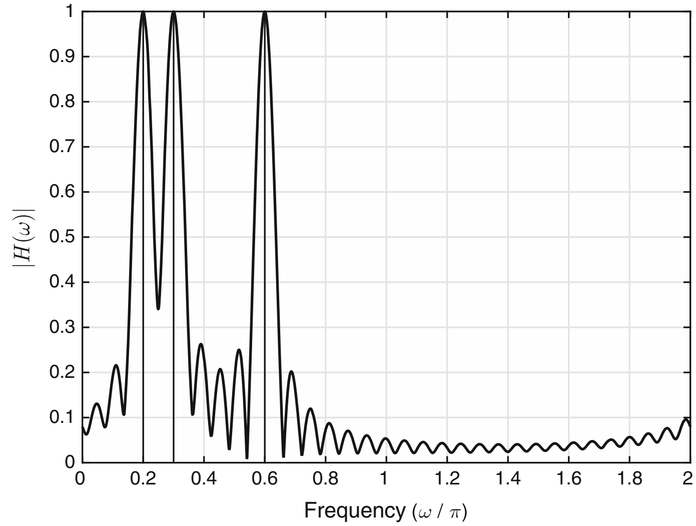


Fig. 9.1 Graph of $|H(\omega)|$ from Example 9.3. The vertical lines mark the true frequencies

(9.8) has 31 atoms. Since the degree of $H(z)$ is $n = 31$, this is actually the maximum number of atoms that can appear. So, in this case, the computation of the atomic norm does not lead to the sparsest representation with the considered dictionary.

This result is natural: One cannot always recover a sparse representation by simply finding (or, in a more general context, minimizing) the atomic norm. Finding the sparsest representation is an NP-hard problem, and the minimization of the atomic norm is only a well-founded heuristic, not an unconditionally guaranteed tool.

Before presenting applicative uses of the atomic norm, let us note that, since (9.11) is an SDP problem, its dual has the same value; hence, the atomic norm can be computed also with

$$\begin{aligned} \|\boldsymbol{x}\|_{\mathcal{A}} = \min_{\zeta,\boldsymbol{\lambda}} \ & \tfrac{1}{2}(\lambda_0 + \zeta) \\ \text{s.t.} \ & \begin{bmatrix} \mathrm{Toep}(\lambda_0, \ldots, \lambda_n) & \boldsymbol{x} \\ \boldsymbol{x}^H & \zeta \end{bmatrix} \succeq 0 \end{aligned} \tag{9.13}$$

The proof is presented in Sect. 9.5.2. In view of Theorem 2.14, the presence of a Toeplitz matrix in the dual is no surprise. Relation (9.13) can be proved directly, without the help of the dual, see [1].

9.2.2 Matrix Atoms

A generalization of the atomic set (9.7) to matrices is interesting when modeling several sums of sinusoids with the same frequencies but with different phases. Let us consider a set $\mathcal{A}$ with $(n + 1) \times \ell$ matrix atoms

$$\boldsymbol{A}(\omega, \boldsymbol{b}) = \boldsymbol{\psi}(\omega) \cdot \boldsymbol{b}^H, \quad \|\boldsymbol{b}\| = 1. \tag{9.14}$$

We can easily check that all atoms have the same 2-norm and no atom lies in the convex hull of the others. So, the atomic norm can be defined as in (9.3). Similarly to (9.10), the dual norm is

$$\begin{aligned} \|\boldsymbol{H}\|_{\mathcal{A}}^* &= \sup_{\omega, \|\boldsymbol{b}\|=1} \operatorname{Re}\left\{\operatorname{tr}[\boldsymbol{H}^H \boldsymbol{A}(\omega, \boldsymbol{b})]\right\} = \sup_{\omega, \|\boldsymbol{b}\|=1} \left|\boldsymbol{b}^H \boldsymbol{H}^H \boldsymbol{\psi}(\omega)\right| \\ &= \max_{|z|=1} \|\boldsymbol{H}(z)\| = \max_{\omega \in [-\pi, \pi]} \sigma_{\max}[\boldsymbol{H}(\omega)] \end{aligned} \tag{9.15}$$

where (with an abuse of notation) $\boldsymbol{H}(z) = \boldsymbol{H}^H \boldsymbol{\psi}(z)$ is a causal polynomial with $\ell \times 1$ coefficients. So, again the BRL for trigonometric polynomials applies—this time in the form of Theorem 4.32. Since the coefficients of the polynomial concatenated like in (4.47) form a long vector, in view of Remark 4.33, it is better to work with the transposed polynomial; in this case, the block column of coefficients (4.47) is exactly the matrix $\boldsymbol{H}$ whose dual norm is sought. So, similar to (9.11), we can express the atomic norm as

$$\begin{aligned} \|\boldsymbol{X}\|_{\mathcal{A}} = \sup_{\|\boldsymbol{H}\|_{\mathcal{A}}^* \le 1} \operatorname{Re}\left\{\operatorname{tr}[\boldsymbol{X}^H \boldsymbol{H}]\right\} = \; & \max_{\boldsymbol{H}, \boldsymbol{Q}} \operatorname{Re}\left\{\operatorname{tr}[\boldsymbol{X}^H \boldsymbol{H}]\right\} \\ & \text{s.t. } \; \delta_k = \operatorname{tr}[\boldsymbol{\Theta}_k \boldsymbol{Q}], \; k = 0:n, \\ & \qquad \begin{bmatrix} \boldsymbol{Q} & \boldsymbol{H} \\ \boldsymbol{H}^H & \boldsymbol{I}_\ell \end{bmatrix} \succeq 0 \end{aligned} \tag{9.16}$$

We proceed similarly to the scalar case. Let

$$\boldsymbol{X} = \sum_i c_i \boldsymbol{\psi}(\omega_i) \cdot \boldsymbol{b}_i^H, \quad c_i \ge 0, \; \|\boldsymbol{b}_i\| = 1, \tag{9.17}$$

be the linear combination of atoms for which the minimum is attained in (9.4).

Theorem 9.4 *The frequencies ω_i from (9.17) are those for which $\|\boldsymbol{H}(\omega_i)\| = 1$, where $\boldsymbol{H}(z)$ is the causal trigonometric polynomial whose coefficients are given by the rows of the matrix $\boldsymbol{H}$ which is the solution of (9.16). Moreover, it results that $\boldsymbol{b}_i = \boldsymbol{H}(\omega_i)$.*

Proof Let $\mathcal{O}$ be the set of frequencies $\tilde{\omega}_i$ for which $\|\boldsymbol{H}(\tilde{\omega}_i)\|^2 = 1$. If $\omega \notin \mathcal{O}$, then $\|\boldsymbol{H}(\omega)\| < 1$. Reminding that (9.8) means also $\|\boldsymbol{X}\|_{\mathcal{A}} = \sum_i c_i$, it results for the optimal $\boldsymbol{H}$ that

$$\|\boldsymbol{X}\|_{\mathcal{A}} = \operatorname{Re}\left\{\operatorname{tr}[\boldsymbol{X}^H \boldsymbol{H}]\right\} = \sum_i c_i \operatorname{Re}|\boldsymbol{b}_i^H \boldsymbol{H}(\omega_i)| \le \sum_i c_i.$$

The equality is attained only if $\omega_i \in \mathcal{O}$ and $\boldsymbol{b}_i = \boldsymbol{H}(\omega_i)$. ■

Writing explicitly the scalar polynomials, we denote $\boldsymbol{H}(z) = [H_1(z) \ldots H_\ell(z)]^T$. Computing the frequencies from $\mathcal{O}$ amounts to finding the zeros of the polynomial

$$R(z) = 1 - \boldsymbol{H}^H(z^{-1})\boldsymbol{H}(z) = 1 - \sum_{k=1}^{\ell} H_k^*(z^{-1})H_k(z).$$

The degree of the polynomial is n. Once the frequencies are available, finding the coefficients c_i and the vectors $\boldsymbol{b}_i$ from (9.17) is a least squares problem proposed to the reader in **P** 9.2.

Example 9.5 We take now $n = 15$ and consider two signals that are sums of sinusoids with the same frequencies, but with different coefficients. The first is $\boldsymbol{x}_1 = 1.5\boldsymbol{a}(0.2\pi) - 2\boldsymbol{a}(0.3\pi) + 2.5\boldsymbol{a}(0.6\pi)$, like in Example 9.3. The second is $\boldsymbol{x}_2 = 2\boldsymbol{a}(0.2\pi, 0.5\pi) + \boldsymbol{a}(0.3\pi, 0.25\pi) - 1.5\boldsymbol{a}(0.6\pi, 1.8\pi)$. We denote $\boldsymbol{X} = [\boldsymbol{x}_1\, \boldsymbol{x}_2]$. By solving (9.16), we obtain $\|\boldsymbol{X}\|_{\mathcal{A}} = 7.65$, a value that can be explained if we express $\boldsymbol{X}$ using atoms (9.14):

$$\boldsymbol{X} = c_1\boldsymbol{A}(0.2\pi, \boldsymbol{b}_1) + c_2\boldsymbol{A}(0.3\pi, \boldsymbol{b}_2) + c_3\boldsymbol{A}(0.6\pi, \boldsymbol{b}_3), \tag{9.18}$$

with $c_i > 0$. By identifying the coefficients of the first atom with those of $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$ for the first frequency

$$c_1\boldsymbol{b}_1^H = [1.5\ 2e^{j0.5\pi}]$$

and taking into account that $\|\boldsymbol{b}_1\| = 1$, we get $c_1 = \|[1.5\ 2e^{j0.5\pi}]\|$, etc. The sum $c_1 + c_2 + c_3$ is then the atomic norm of $\boldsymbol{X}$. (Again, we must be aware that this happens only if the frequencies defining $\boldsymbol{X}$ are well separated.)

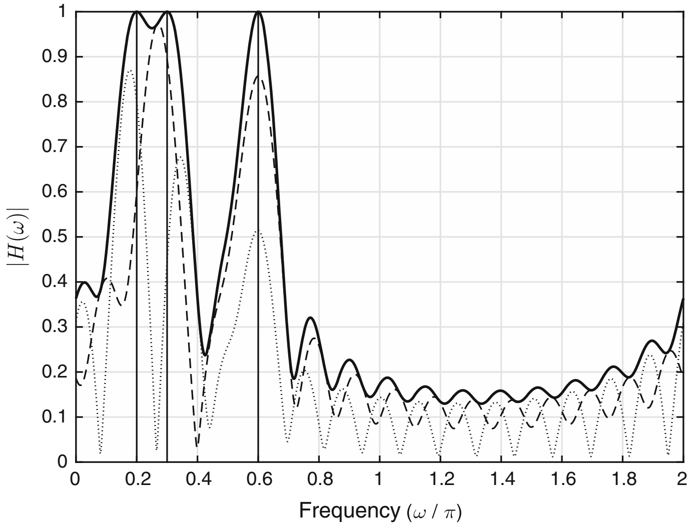


Fig. 9.2 Graph of $\|\boldsymbol{H}(\omega)\|$ (solid line), $|H_1(\omega)|$ (dashed line), and $|H_2(\omega)|$ (dotted line), for Example 9.5. The vertical lines mark the true frequencies

The graph of $\|H(\omega)\|$ is shown in Fig. 9.2, together with those of $|H_1(\omega)|$ and $|H_2(\omega)|$, where $\boldsymbol{H}(\omega) = [H_1(\omega)\ H_2(\omega)]^T$. The maximum is attained in the three frequencies defining our signals. ■

9.3 The Atomic Norm at Work

The facts and examples from the previous section suggest that the atomic norm is a good ingredient when sparse linear combinations of atoms are desired. This section presents several applications featuring the atom set (9.7) or its matrix form with atoms (9.14).

9.3.1 Line Spectral Estimation

The problem here is to recover a sum of s sinusoids from noisy measurements. The model is thus

$$\hat{\boldsymbol{x}} = \sum_{i=1}^{s} c_i \boldsymbol{a}(\omega_i, \varphi_i) + \boldsymbol{v}, \tag{9.19}$$

where $\boldsymbol{v}$ is a vector whose elements are Gaussian noise of zero mean and unknown variance. The number s of sinusoids is not known but is assumed to be small. Having only the measurements vector $\hat{\boldsymbol{x}}$, we want to find the frequencies ω_i. The name of line spectrum is intuitive, since the spectrum of the clean signal is concentrated in exactly s frequencies. (An image of the spectrum, without figuring the amplitudes, which should be equal to c_i, is given by the vertical lines from Fig. 9.1).

Since the atomic norm induces sparsity, we can try to find the line spectrum by solving the optimization problem

$$\min_{\boldsymbol{x}} \tfrac{1}{2}\|\boldsymbol{x} - \hat{\boldsymbol{x}}\|_2^2 + \gamma \|\boldsymbol{x}\|_{\mathcal{A}} \tag{9.20}$$

That is, we aim to find a signal that is near from the measured one and is a sparse sum of sinusoids. Here, $\gamma > 0$ is a trade-off parameter that roughly quantifies the relative importance of sparsity and distance from the measured signal.

Using the expression (9.11) of the atomic norm, the problem (9.20) is equivalent to

$$\begin{array}{ll} \min\limits_{\boldsymbol{x},\boldsymbol{h}} & \tfrac{1}{2}\|\boldsymbol{x} - \hat{\boldsymbol{x}}\|_2^2 - \mathrm{Re}[\boldsymbol{x}^H \boldsymbol{h}] \\ \text{s.t.} & \|\boldsymbol{h}\|_{\mathcal{A}}^* \le \gamma \end{array} = \begin{array}{ll} \min\limits_{\boldsymbol{x},\boldsymbol{h},\boldsymbol{Q}} & \tfrac{1}{2}\|\boldsymbol{x} - \hat{\boldsymbol{x}}\|_2^2 - \mathrm{Re}[\boldsymbol{x}^H \boldsymbol{h}] \\ \text{s.t.} & \gamma^2 \delta_k = \mathrm{tr}[\boldsymbol{\Theta}_k \boldsymbol{Q}],\ k = 0 : n, \\ & \begin{bmatrix} \boldsymbol{Q} & \boldsymbol{h} \\ \boldsymbol{h}^H & 1 \end{bmatrix} \succeq 0 \end{array} \tag{9.21}$$

The first form is obtained by incorporating the multiplication with γ in the polynomial defined by $\boldsymbol{h}$; hence, the norm bound is changed from 1 into γ. The second form is obtained as usual with the BRL for trigonometric polynomials. At optimality, the value of $\text{Re}[\boldsymbol{x}^H\boldsymbol{h}]$ is maximum for the optimal $\boldsymbol{x}$, under the constraint $\|\boldsymbol{h}\|_{\mathcal{A}}^* \le \gamma$. So, like in the atomic norm computation (9.11), the frequencies where $|H(\omega)|$ attains its maximum value, i.e., now γ, are the optimal ones. We can thus retrieve (approximations of) the s frequencies from (9.19). The (approximated) coefficients c_i can be found by solving (9.12), using the optimal $\boldsymbol{x}$.

Example 9.6 We take again $n = 31$ and $\boldsymbol{x} = 1.5\boldsymbol{a}(0.2\pi) - 2\boldsymbol{a}(0.3\pi) + 2.5\boldsymbol{a}(0.6\pi)$. The noise $\boldsymbol{v}$ in (9.19) has variance $\sigma^2 = 0.01$. With $\gamma = 4$, the graph of $|H(\omega)|$ obtained by solving (9.21) is shown in Fig. 9.3. The frequencies where $|H(\omega)| = \gamma$ are very near from the true frequencies. Of course, typically it is necessary to solve the problem for several values of γ to have the confirmation of a plausible solution. If γ is too small the solution is usually not sparse. If γ is too large, it is possible that the computed frequencies are farther away from the true ones or even that the number of frequencies is smaller than the true s. ■

The solution presented above is satisfactory for n relatively small, at most a few hundreds. Otherwise, solving the SDP problem (9.21) takes too much time (if it can be solved); also, finding the maxima of a trigonometric polynomial of such order may become ill-conditioned and time-consuming. A lower complexity alternative using the matrix atomic norm is as follows.

Instead of using a single long signal (9.19), we split it into ℓ segments (that may be overlapped or not)

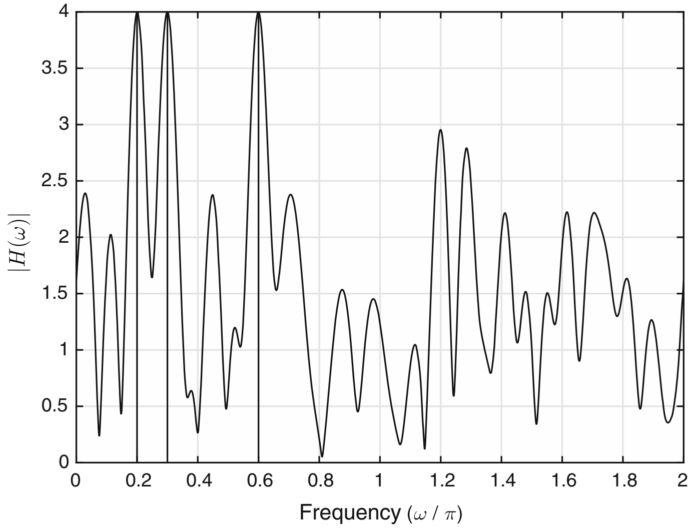


Fig. 9.3 Graph of $|H(\omega)|$ for line spectrum estimation, Example 9.6

$$\hat{\boldsymbol{x}}_k = \sum_{i=1}^{s} c_i \boldsymbol{a}(\omega_i, \varphi_{ik}) + \boldsymbol{v}_k, \quad k = 1 : \ell. \tag{9.22}$$

The coefficients and the frequencies are the same, but the phases are different for for each segment. We then join the segments in the matrix

$$\hat{\boldsymbol{X}} = [\hat{\boldsymbol{x}}_1 \; \hat{\boldsymbol{x}}_2 \; \ldots \; \hat{\boldsymbol{x}}_\ell]. \tag{9.23}$$

To keep the same notation as in the previous section, the size of this matrix is $(n+1) \times \ell$.

Using the matrix atomic norm defined by the atoms (9.14), we attempt finding the line spectrum by solving

$$\min_{\boldsymbol{X}} \; \tfrac{1}{2}\|\boldsymbol{X} - \hat{\boldsymbol{X}}\|_F^2 + \gamma \|\boldsymbol{X}\|_{\mathcal{A}} \tag{9.24}$$

Using the expression (9.16) of the atomic norm and reasoning like for (9.21), the problem (9.24) is equivalent to

$$\begin{array}{lll} \min\limits_{\boldsymbol{X},\boldsymbol{H}} \; \tfrac{1}{2}\|\boldsymbol{X} - \hat{\boldsymbol{X}}\|_F^2 - \mathrm{Re}\left\{\mathrm{tr}[\boldsymbol{X}^H \boldsymbol{H}]\right\} & = & \min\limits_{\boldsymbol{X},\boldsymbol{H},\boldsymbol{Q}} \; \tfrac{1}{2}\|\boldsymbol{X} - \hat{\boldsymbol{X}}\|_F^2 - \mathrm{Re}\left\{\mathrm{tr}[\boldsymbol{X}^H \boldsymbol{H}]\right\} \\ \text{s.t.} \;\; \|\boldsymbol{H}\|_{\mathcal{A}}^* \le \gamma & & \text{s.t.} \;\; \gamma^2 \delta_k = \mathrm{tr}[\boldsymbol{\Theta}_k \boldsymbol{Q}], \; k = 0 : n, \\ & & \qquad \begin{bmatrix} \boldsymbol{Q} & \boldsymbol{H} \\ \boldsymbol{H}^H & 1 \end{bmatrix} \succeq 0 \end{array} \tag{9.25}$$

Note that the size of the matrix $\boldsymbol{Q}$ is that of a segment; hence, for signals of the same length, it is much cheaper to solve the matrix version (9.25) than the scalar version (9.21). For a proper choice of γ, the graph of $\|\boldsymbol{H}(\omega)\|$ is similar to that from Fig. 9.3.

9.3.2 Direction of Arrival Estimation

A basic problem in radar is to estimate the direction of a source using an uniform linear array (ULA), as shown in Fig. 9.4. The array has $n+1$ identical sensors; the distance between two successive sensors is d. The source emits a sinusoidal signal of known frequency f and is far from the ULA. Hence, the lines of propagation between the source and the sensors can be considered parallel and they make an angle θ with the direction of the array. The dashed line marks a wave front: all its points are at equal distance from the source. The velocity of the waves is v and is known. The inter-sensor distance d is smaller than half the wavelength. Numbering the sensors from one end to the other and taking sensor 0 (the leftmost in the figure) as reference, the delay with which a wave hits sensor k is

$$\tau_k = \frac{kd\cos\theta}{v}. \tag{9.26}$$

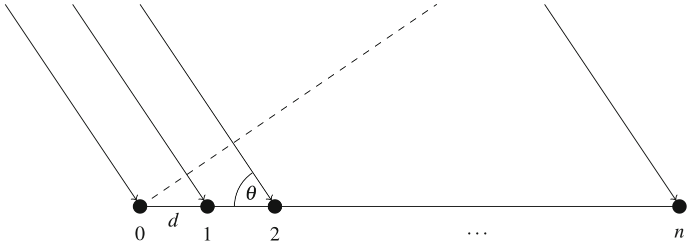


Fig. 9.4 Uniform linear array and impinging waves (from a single source)

A snapshot $\boldsymbol{x}$ is a vector of measurements of the signals received by the sensors at the same time t. At sensor k, the signal is

$$x_k = c\mathrm{e}^{jf(t-\tau_k)} = c\mathrm{e}^{jft}\mathrm{e}^{-jk\omega}, \quad \text{with } \omega = \frac{f\,d\cos\theta}{v}, \tag{9.27}$$

+ where c is a coefficient that depends on the power of the source. Assuming now that there are s sources that emit simultaneously from different angles θ_i, $i = 1 : s$, the measurement model is exactly (9.19), that is a sum of sinusoids with frequencies

$$\omega_i = \frac{f\,d\cos\theta_i}{v}. \tag{9.28}$$

The direction of arrival (DOA) problem consists of estimating the angles θ_i given the noisy snapshot $\hat{\boldsymbol{x}}$. Obviously, the solution from the previous section applies ad litteram. After solving (9.21) and estimating the frequencies ω_i, the DOAs result as

$$\cos\theta_i = \frac{v\omega_i}{fd}. \tag{9.29}$$

Of course, a more reliable estimation is used if several snapshots are used. In this case, the model is (9.22), and the SDP problem (9.25) is solved and again the DOAs result from (9.29).

Several variations in the DOA theme can be made, among which is the problem of DOA estimation with nonuniform arrays, in particular with ULA with missing sensors, see **P** 9.5.

9.3.3 Super-Resolution

Viewed as a general concept, super-resolution is the recovery of high frequency details of a signal from low-frequency information and comes in different forms in medical imaging, spectroscopy, microscopy, and astronomy, among others. We present here an abstract formulation that is basically equivalent in outcome with the atomic norm problem treated previously.

The primary signal is a sum of spikes

$$\xi(t) = \sum_i c_i \delta_{t_i}(t), \tag{9.30}$$

where δ_τ is a Dirac impulse located at time τ and the spikes temporal positions t_i are in the interval [0,1]. The coefficients c_i may be complex. The length of the interval is taken equal to 1 only for convenience; similar results hold for any finite interval. The Fourier coefficients of the signal are

$$x_k = \int_0^1 \mathrm{e}^{-j2\pi kt}\xi(t)dt = \sum_i c_i \mathrm{e}^{-j2\pi kt_i}, \quad k \in \mathbb{Z}. \tag{9.31}$$

One can immediately see that the Fourier coefficients are obtained by a linear combination of atoms from (9.7) similar with (9.8), with $\omega_i = 2\pi t_i$. The coefficients c_i can be easily made positive by the introduction of a phase factor, but the focus here is different. Assume that we have the Fourier coefficients corresponding to low frequencies, i.e., $|k| \le f_c$, where f_c, the cutting frequency, is an integer. Hence, the signal $\boldsymbol{x}$ from (9.31) has length $n + 1 = 2f_c + 1$.

The super-resolution problem consists of retrieving ξ, which has a significant high-frequency content, from only the above Fourier coefficients corresponding to low frequencies. The solution is related to the total variation norm. Considering partitions $\bigcup_l \mathcal{I}_i$ of [0,1] in a countable number of disjoint measurable subsets $\mathcal{I}_l$ (e.g., intervals, in our case), the total variation of a complex measure ν on [0,1] is

$$\|\nu\|_{\mathrm{TV}} = \sup \sum_{l=1}^{\infty} |\nu(\mathcal{I}_l)|, \tag{9.32}$$

where the supremum is taken over all possible partitions. For the function ξ defined in (9.30), the total variation norm is in fact $\|\xi\|_{\mathrm{TV}} = \sum_i |c_i|$. This norm is clearly a generalization of the ℓ_1-norm from the discrete case to the real line.

So, by analogy with (9.2), the optimization problem whose solution is, under certain conditions detailed below, the sum of spikes, has the simple form

$$\begin{aligned} &\min_{\tilde{\xi}} \ \|\tilde{\xi}\|_{\mathrm{TV}} \\ &\text{s.t. } \ \mathcal{F}\tilde{\xi} = \boldsymbol{x} \end{aligned} \tag{9.33}$$

where the map $\mathcal{F}$ collects the $2f_c + 1$ low-frequency Fourier coefficients of the signal $\tilde{\xi}(t)$. One can see that problem (9.33) is equivalent to minimizing the atomic norm of $\boldsymbol{x}$ using the atoms (9.7). Hence, by solving (9.11) and then finding the roots of the trigonometric polynomial associated with the solution, we can retrieve the frequencies ω_i and thus the temporal positions t_i of the spikes from (9.30).

Beyond the similarity with the atomic norm approach, the line of thought based on total variation has lead to an important result regarding the recovery of the sum of spikes.

Theorem 9.7 ([2]) *If the distance $|t_i - t_j|$ between any two temporal positions $t_i \neq t_j$ is larger than $2/f_c$ and $f_c \geq 128$, then the unique solution of (9.33) is (9.30).*

Besides the technical condition $f_c \geq 128$, this theorem says that if the temporal positions t_i, or the frequencies ω_i, are not too close, then they can be exactly recovered by solving an SDP problem and finding the roots of a polynomial. This is a remarkable result, different from all previous methods. Note, however, that it stands for noiseless signals that have the exact form (9.30).

9.4 Generalizations to 2D

The extension of the results from Sect. 9.2.1 to the multivariate case is relatively straightforward. For the ease of exposition, we confine the discussion to 2D, which is representative for the differences with respect to the univariate case.

The 2D generalization of the set (9.7) is made of atoms

$$\boldsymbol{a}(\boldsymbol{\omega}, \varphi) = \mathrm{e}^{j\varphi} \boldsymbol{\psi}(\omega_2) \otimes \boldsymbol{\psi}(\omega_1) = \mathrm{e}^{j\varphi} \boldsymbol{\psi}(\boldsymbol{\omega}), \tag{9.34}$$

where $\boldsymbol{\omega} = (\omega_1, \omega_2) \in [-\pi, \pi]^2$ is the 2D frequency, φ is the (unique) phase variable and $\boldsymbol{\psi}(\boldsymbol{\omega})$ is defined in (3.24).

Similarly to (9.10), the dual norm is

$$\|\boldsymbol{h}\|_{\mathcal{A}}^{*} = \sup_{\boldsymbol{\omega}, \varphi} \operatorname{Re}[\boldsymbol{h}^H \boldsymbol{a}(\boldsymbol{\omega}, \varphi)] = \max_{\boldsymbol{\omega}, \varphi} \left| \boldsymbol{h}^H \boldsymbol{\psi}(\boldsymbol{\omega}) \right| = \max_{|z|=1} |H(z)|, \tag{9.35}$$

where $H(z)$ is the causal polynomial (3.27). So, again, we are led to the BRL from Corollary 4.27. This time, unfortunately, there is no equivalence between the SDP formulation and the polynomial boundedness. However, for practical purposes, we can approximate the atomic norm with

$$\begin{aligned} \|\boldsymbol{x}\|_{\mathcal{A}} = \sup_{\|\boldsymbol{h}\|_{\mathcal{A}}^{*} \leq 1} \operatorname{Re}[\boldsymbol{x}^H \boldsymbol{h}] \gtrsim \max_{\boldsymbol{h}, \boldsymbol{Q}} \ & \operatorname{Re}[\boldsymbol{x}^H \boldsymbol{h}] \\ \text{s.t.} \ & \delta_{\boldsymbol{k}} = \operatorname{tr}[\boldsymbol{\Theta}_{\boldsymbol{k}} \boldsymbol{Q}], \ \boldsymbol{k} \in \mathcal{H}, \\ & \begin{bmatrix} \boldsymbol{Q} & \boldsymbol{h} \\ \boldsymbol{h}^H & 1 \end{bmatrix} \succeq 0 \end{aligned} \tag{9.36}$$

where $\mathcal{H}$ is a halfspace. Of course, the approximation can be better (from below) for a higher relaxation degree, which means a larger matrix $\boldsymbol{Q}$ and hence higher complexity.

All the other developments from Sect. 9.2.1 apply as follows: After solving (9.36), the frequencies $\boldsymbol{\omega}_i$ that define the signal

$$\boldsymbol{x} = \sum_i c_i \boldsymbol{a}(\boldsymbol{\omega}_i, \varphi_i), \quad c_i \geq 0, \tag{9.37}$$

are found by solving the polynomial equation $|H(\boldsymbol{\omega})| = 1$. This is somewhat more difficult than in the univariate case, but still feasible. The coefficients c_i are again computed via a least squares problem.

The approximation induced by the relaxation can be checked following the general rules from Sect. 3.5.3, which take here a much simpler form. The atomic norm of the signal, as computed above, is $\sum_i c_i$. If this sum is equal to the value of the SDP problem (9.36), then we have obtained an optimality certificate. (Note however that another condition must be met: the least squares should have an exact solution, which means that the computed signal (9.37) is identical with the original signal. Again, all these procedures work under the condition that the frequencies $\boldsymbol{\omega}_i$ are sufficiently far one from another.)

Example 9.8 Let $n_1 = n_2 = 7$, so the signal has 64 samples. We take the signal

$$\boldsymbol{x} = 1.5\boldsymbol{a}(0.2\pi, 0.7\pi) - 2\boldsymbol{a}(0.3\pi, -0.5\pi) + 2.5\boldsymbol{a}(0.6\pi, 0.1\pi),$$

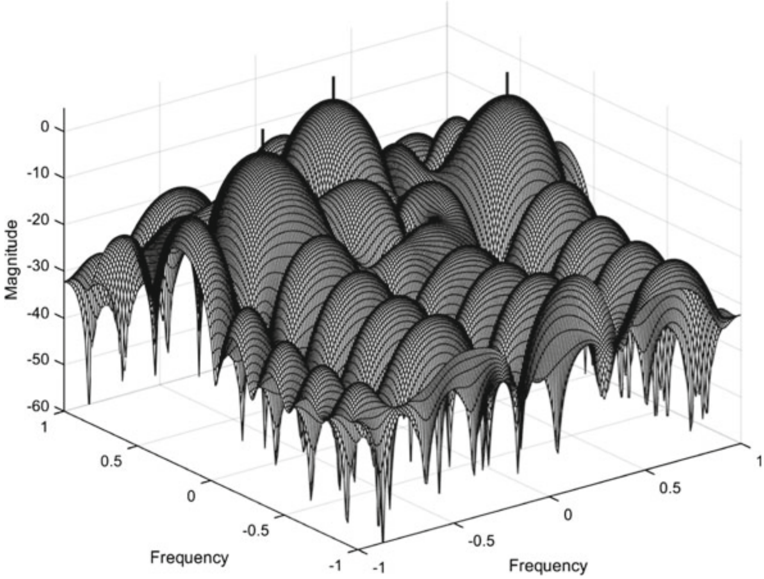


Fig. 9.5 Graph of $|H(\boldsymbol{\omega})|$ from Example 9.8. The small vertical lines mark the true frequencies

where only the two frequencies are used to define an atom (phase is not specified). By solving (9.36), we obtain the expected result $\|\boldsymbol{x}\|_{\mathcal{A}} = 6$. The graph of $|H(\boldsymbol{\omega})|$ is shown in Fig. 9.5, where the value 1 is attained for the correct frequencies. ■

9.5 Details and Other Facts

9.5.1 *Sums of Real Exponentials*

It is tempting to extend the results presented in this chapter to real or hybrid polynomials. Let us consider the former case. The atoms have the form $\boldsymbol{\psi}(t) = [1\ t\ t^2\ \dots\ t^n]^T$. A condition like $t \in [T_1, T_2]$ is imposed for having only atoms with bounded norm; however, the atoms have not equal norm. Consider the problem of recovering a signal

$$\boldsymbol{x} = \sum_{i=1}^{s} c_i \boldsymbol{\psi}(t_i), \tag{9.38}$$

where $t_i > 0$ are given values. For example, such a signal is a sum of real fading exponentials sampled on a uniform grid of time instants starting from zero; if τ is the grid step and $\lambda_i > 0$ is the exponent of the i–the exponential, then $t_i = e^{-\lambda_i \tau}$, $T_1 = 0$, $T_2 = 1$.

Using the set $\mathcal{A}$ of atoms $\boldsymbol{\psi}(t)$, with $t \in [T_1, T_2]$, and proceeding like in (9.10), the dual norm is

$$\|\boldsymbol{p}\|_{\mathcal{A}}^* = \sup_{t \in [T_1, T_2]} \boldsymbol{p}^T \boldsymbol{\psi}(t) = \max_{t \in [T_1, T_2]} \left| \sum_{k=0}^{n} p_k t^k \right|. \tag{9.39}$$

So, the dual norm is the maximum of a polynomial on an interval. We denote $P(t)$ the polynomial with coefficients p_k, $k = 0 : n$. Going back to the atomic norm by dualization of the dual norm, we obtain, similarly to (9.11)

$$\|\boldsymbol{x}\|_{\mathcal{A}} = \max_{\|\boldsymbol{p}\|_{\mathcal{A}}^* \le 1} \boldsymbol{x}^T \boldsymbol{p} = \max_{|P(t)| \le 1,\ t \in [T_1, T_2]} \boldsymbol{x}^T \boldsymbol{p} \tag{9.40}$$

The constraint of the rightmost optimization problem can be expressed as a BRL, whose explicit form is left to the reader. For inspiration, see problem **P** 4.8; the BRL for hybrid polynomials can be immediately written for a real polynomial that is positive on an interval. The solution of the resulting SDP problem allows forming the polynomial $P(t)$ and the condition $|P(t)| = 1$ gives the values t_i from (9.38).

Although apparently straightforward, the extension to real polynomials has a major drawback: it is very ill-conditioned, as real polynomials tend to be. Some numerical experiments have shown that one can recover the original atoms, i.e., the values t_i, only for signals whose length is at most 20–30 in the noiseless case. In the noisy case, the computed t_i may be quite far from the true values. Hence, the practical appeal of this approach seems very limited.

9.5.2 Proof of (9.13)

The Lagrangian function associated with (9.11) is the real part of (the arguments are obvious, hence ignored)

$$\begin{aligned} L &= -\boldsymbol{x}^H \boldsymbol{h} - \sum_{k=0}^{n} \lambda_k \left(\delta_k - \operatorname{tr}[\boldsymbol{\Theta}_k \boldsymbol{Q}]\right) - \operatorname{tr}\begin{bmatrix} \boldsymbol{Z} & \boldsymbol{u} \\ \boldsymbol{u}^H & 1 \end{bmatrix}\begin{bmatrix} \boldsymbol{Q} & \boldsymbol{h} \\ \boldsymbol{h}^H & 1 \end{bmatrix} \\ &= -(\boldsymbol{x} + 2\boldsymbol{u})^H \boldsymbol{h} - \lambda_0 + \operatorname{tr}\left[\left(\sum_{k=0}^{n} \lambda_k \boldsymbol{\Theta}_k - \boldsymbol{Z}\right)\boldsymbol{Q}\right] - \zeta \end{aligned}$$

To obtain the Lagrange dual function, we minimize the above with respect to the primal variables. The minimum is finite only if $2\boldsymbol{u} = -\boldsymbol{x}$ and

$$\boldsymbol{Z} = \sum_{k=0}^{n} \lambda_k \boldsymbol{\Theta}_k.$$

After imposing $\begin{bmatrix} \boldsymbol{Z} & \boldsymbol{u} \\ \boldsymbol{u}^H & 1 \end{bmatrix} \succeq 0$ and scaling with 2, the maximization of the dual function is in fact (9.13).

9.6 Bibliographical and Historical Notes

There is a large body of literature on sparse representations. One can start reading with [3].

The atomic norm was introduced in [4] as a convex optimization framework for a series of problems involving sparse representations, low-rank matrices and tensors, orthogonal matrices, and permutation matrices. It was employed in [1] for compressed sensing off the grid, using an infinite dictionary with atoms built from the canonical trigonometric polynomial basis; the same paper contains a probabilistic estimation of the number of measurements that are necessary to recover a sum of sines, with a frequency separation condition similar to that from Theorem 9.7. These ideas were applied for line spectral estimation with atomic norm denoising [5]. The extension to the matrix case, more precisely to spectral estimation with multiple measurement vectors, was given in [6].

The ideas regarding super-resolution come from [2], where the problem of finding the full spectrum from low-frequency information was solved for a sum of spikes. This line of research was continued for noisy signals in [7].

The application of the atomic norm approach for estimating the direction of arrival was proposed in [8]. Sparse representations techniques for DOA were initiated in [9] long before, but using a finite dictionary obtained by uniformly sampling the infinite one.

Extensions to the multidimensional case were given in [10, 11], with the problem of estimating a two-dimensional spectrum as immediate target. In [12], similar tools were used for an extension on the sphere.

Several other developments have been lately proposed. Among them is super-resolution with prior knowledge [13], for example, using weighted atomic norm with different weights on different frequency intervals. See problem **P** 9.3. The extension of line spectrum estimation with arbitrary (not equidistant) time samples is discussed in [14]. The dual problem is expressed with prolate spheroidal wave functions instead of polynomials. In [15], algorithms for estimating a sum of complex exponentials convolved with unknown waveforms are given.

There is also intensive research on algorithms not based on SDP (and hence not using the BRL for trigonometric polynomials), in order to overcome the high complexity for large problems. Among the main results are a greedy algorithm for atomic norm regularization [16] or the appeal to nonconvex measures instead of the atomic norm [17, 18], leading to optimization problems solved through iterative methods, for example, based on reweighted atomic norm minimization.

The developments for real polynomials from Sect. 9.5.1 are not present in the literature, most likely due to their poor numerical behavior.

Problems

P 9.1 Let $\boldsymbol{x} \in \mathbb{C}^n$ be a sequence and $\boldsymbol{f} \in \mathbb{C}^n$ the coefficients of its discrete Fourier transform. Show that $\|\boldsymbol{x}\|_{\mathcal{A}} \le \sum_i |f_i|$, where $\mathcal{A}$ is the set (9.7).

P 9.2 Assume that we know the matrix $\boldsymbol{X}$ and the frequencies ω_i from the optimal atomic decomposition (9.17). Propose a method to compute the coefficients $c_i > 0$ and the vectors $\boldsymbol{b}_i$, $\|\boldsymbol{b}_i\| = 1$. Hint: Denote $\tilde{\boldsymbol{b}}_i = c_i \boldsymbol{b}_i$ and form a linear least squares problem with $\tilde{\boldsymbol{b}}_i$ as unknowns.

P 9.3 (Weighted atomic norm.) We modify the atoms from (9.7) by introducing weights. So, an atom is now $\boldsymbol{a}(\omega, \varphi) = \mu(\omega) e^{j\varphi} \boldsymbol{\psi}(\omega)$, where $\mu(\omega) > 0$ is a piecewise constant function. One can interpret weights as prior information that the frequencies appearing in the decomposition of a signal are more likely in an interval (where the weights hence have a low value) than in another (where the weights are higher). Having the frequency split $[-\pi, \pi] = \bigcup_{i=1}^{m} \mathcal{J}_i$, where $\mathcal{J}_i$ are intervals, and the weighting function $\mu(\omega) = d_i$, for $\omega \in \mathcal{J}_i$, show that the atomic norm can be expressed as

$$\begin{aligned}\|\boldsymbol{x}\|_{\mathcal{A}} = \max_{\boldsymbol{h}} \ & \mathrm{Re}[\boldsymbol{x}^H \boldsymbol{h}] \\ \text{s.t.} \ & |H(\omega)| \le d_i, \ \forall \omega \in \mathcal{J}_i, \ i = 1 : m\end{aligned}$$

where $H(z)$ is the causal trigonometric polynomial whose coefficients vector is $\boldsymbol{h}$. Use the BRL on intervals to express this relation as an SDP problem. Show how to compute the frequencies that characterize the atoms appearing in the atomic decomposition of $\boldsymbol{x}$.

P 9.4 (Explicit 2D atoms.) The elements of a matrix $\boldsymbol{X}$ are a superposition of s sinusoids, i.e., are given by

$$x_{k_1,k_2} = \sum_{i=1}^{s} c_i e^{j(k_1\omega_{1i} + k_2\omega_{2i})}.$$

Show that

$$\boldsymbol{X} = \sum_{i=1}^{s} c_i \boldsymbol{\psi}(\omega_{1i})\boldsymbol{\psi}(\omega_{2i})^T = [\boldsymbol{\psi}(\omega_{11}) \ \ldots \ \boldsymbol{\psi}(\omega_{1s})] \cdot \mathrm{diag}(c_1, \ldots, c_s) \cdot [\boldsymbol{\psi}(\omega_{21}) \ldots \ \boldsymbol{\psi}(\omega_{2s})]^T.$$

The first equality shows the explicit construction of a 2D signal as a linear combination of a few matrix atoms. Denoting $\boldsymbol{x} = \mathrm{vec}(\boldsymbol{X})$, show that the above relations are equivalent with (9.37).

P 9.5 (ULA with missing sensors.) A particular array for DOA estimation is that obtained from a long ULA by removing some of the sensors; with proper choice of the removed sensors positions, the resulting array is only slightly inferior to the full ULA, but obviously cheaper. Translating the DOA problem with such an array to the sum of sinusoids (9.19), the interpretation is that some samples of the signal $\hat{\boldsymbol{x}}$ are missing. Show how to modify the SDP problem (9.21) such that the DOAs (or the frequencies) can be estimated.

P 9.6 Generalize the 2D line spectrum estimation problem to the matrix atomic norm case and find the SDP problem that is similar to (9.36).

References

1. G.G. Tang, B.N. Bhaskar, P. Shah, B. Recht, Compressed sensing off the grid. IEEE Trans. Inf. Theory **59**(11), 7465–7490 (2013)
2. E.J. Candes, C. Fernandez-Granda, Towards a mathematical theory of super-resolution. Commun. Pure Appl. Math. **67**(6), 906–956 (2014)
3. A.M. Bruckstein, D.L. Donoho, M. Elad, From sparse solutions of systems of equations to sparse modeling of signals and images. SIAM Rev. **51**(1), 34–81 (2009)
4. V. Chandrasekaran, B. Recht, P.A. Parrilo, A.S. Willsky, The convex geometry of linear inverse problems. Found. Comput. Math. **12**(6), 805–849 (2012)

5. B.N. Bhaskar, G. Tang, B. Recht, Atomic norm denoising with applications to line spectral estimation. IEEE Trans. Signal Proc. **61**(23), 5987–5999 (2013)
6. Y. Li, Y. Chi, Off-the-grid line spectrum denoising and estimation with multiple measurement vectors. IEEE Trans. Signal Proc. **64**(5), 1257–1269 (2016)
7. E.J. Candes, C. Fernandez-Granda, Super-resolution from Noisy data. J. Fourier Anal. Appl. **19**(6), 1229–1254 (2013)
8. A. Xenaki, P. Gerstoft, Grid-free compressive beamforming. J. Acoust. Soc. Am. **137**(4), 1923–1935 (2015)
9. D. Malioutov, M. Cetin, A.S. Willsky, A sparse signal reconstruction perspective for source localization with sensor arrays. IEEE Trans. Signal Proc. **53**(8), 3010–3022 (2005)
10. W. Xu, J.F. Cai, K.V. Mishra, M. Cho, A. Kruger, Precise semidefinite programming formulation of atomic norm minimization for recovering D-dimensional ($D \geq 2$) off-the-grid frequencies, in *Information Theory and Applications Workshop* (2014)
11. Y. Chi, Y. Chen, Compressive two-dimensional harmonic retrieval via atomic norm minimization. IEEE Trans. Signal Proc. **63**(4), 1030–1042 (2015)
12. T. Bendory, S. Dekel, A. Feuer, Super-resolution on the sphere using convex optimization. IEEE Trans. Signal Proc. **63**(9), 2253–2262 (2015)
13. K.V. Mishra, M. Cho, A. Kruger, W.Y. Xu, Spectral super-resolution with prior knowledge. IEEE Trans. Signal Proc. **63**(20), 5342–5357 (2015)
14. K. Mahata, M.M. Hyder, Frequency estimation from arbitrary time samples. IEEE Trans. Signal Proc. **64**(21), 5634–5643 (2016)
15. D. Yang, G. Tang, M.B. Wakin, Super-resolution of complex exponentials from modulations with unknown waveforms. IEEE Trans. Inf. Theory **62**(10), 5809–5830 (2016)
16. N. Rao, P. Shah, S. Wright, Forward-backward greedy algorithms for atomic norm regularization. IEEE Trans. Signal Proc. **63**(21), 5798–5811 (2015)
17. J. Fang, F. Wang, Y. Shen, H. Li, R.S. Blum, Super-resolution compressed sensing for line spectral estimation: an iterative reweighted approach. IEEE Trans. Signal Proc. **64**(18), 4649–4662 (2016)
18. Z. Yang, L. Xie, Enhancing sparsity and resolution via reweighted atomic norm minimization. IEEE Trans. Signal Proc. **64**(4), 995–1006 (2016)

Appendix A: Semidefinite Programming

We present here some basic facts about semidefinite programming (SDP) and the more general semidefinite-quadratic-linear programming (SQLP). For more information, we recommend [1–4].

An SDP problem can be expressed in two standard forms. The inequality form is

$$\begin{aligned} \max_{\boldsymbol{y}\in\mathbb{R}^m} \ & \boldsymbol{b}^T\boldsymbol{y} \\ \text{s.t.} \ & \boldsymbol{Z} = \boldsymbol{A}_0 - \textstyle\sum_{i=1}^m y_i\boldsymbol{A}_i \succeq 0 \end{aligned} \tag{A.1}$$

where the matrices $\boldsymbol{A}_i \in \mathbb{R}^{n\times n}$, $i = 0 : m$, are symmetric. The objective is linear and the constraint is a linear matrix inequality (LMI), so the optimization problem (A.1) is convex. The equality form is

$$\begin{aligned} \min_{\boldsymbol{X}\in\mathbb{R}^{n\times n}} \ & \mathrm{tr}[\boldsymbol{A}_0\boldsymbol{X}] \\ \text{s.t.} \ & \mathrm{tr}[\boldsymbol{A}_i\boldsymbol{X}] = b_i, \quad i = 1 : m \\ & \boldsymbol{X} \succeq 0 \end{aligned} \tag{A.2}$$

This is the problem dual to (A.1). If the Slater condition holds (there exists $\boldsymbol{y} \in \mathbb{R}^m$ such that $\boldsymbol{Z} \succ 0$, i.e., the set described by the LMI constraint of (A.1) has an interior point), then the solutions of (A.1) and (A.2) exist and satisfy the equalities $\boldsymbol{b}^T\boldsymbol{y} = \mathrm{tr}[\boldsymbol{A}_0\boldsymbol{X}]$ (i.e., the optimal values of the two problems are equal) and $\boldsymbol{Z}\boldsymbol{X} = 0$ (called complementarity condition).

The problem (A.1) may have several LMI constraints $\boldsymbol{Z}_k \succeq 0$, case in which the dual problem (A.2) has the same number of matrix variables $\boldsymbol{X}_k$. (This situation corresponds to a single block diagonal LMI and matrix variable $\boldsymbol{X}$.) SDP problems in equality form are often formulated with nonsymmetric constant matrices $\boldsymbol{A}_i$, since $\mathrm{tr}[\boldsymbol{A}_i\boldsymbol{X}] = \mathrm{tr}[(\boldsymbol{A}_i + \boldsymbol{A}_i^T)/2 \cdot \boldsymbol{X}]$, i.e., their symmetrization is trivial. The variable vector $\boldsymbol{y}$ may be complex, case in which the objective of (A.1) is $\mathrm{Re}[\boldsymbol{b}^T\boldsymbol{y}]$ and the positive semidefinite matrix variable $\boldsymbol{X}$ of (A.2) is also complex (and Hermitian).

Interior point algorithms solve typically both problems (A.1) and (A.2) (and are called primal-dual algorithms). The algorithms are iterative, and the complexity of an

B. Dumitrescu, *Positive Trigonometric Polynomials and Signal Processing Applications*, Signals and Communication Technology,
DOI 10.1007/978-3-319-53688-0

iteration is $O(n^2m^2)$; the number of iterations can be regarded as a constant (depends lightly on m and n). The algorithms offer a certificate of optimality—the value of the gap between the computed values of (A.1) and (A.2).

Besides standard LMIs, SQLP problems contain other types of convex constraints that are particular cases of LMIs but can be treated more efficiently. An SQLP problem in inequality form is

$$\begin{aligned} \max_{\boldsymbol{y}\in\mathbb{R}^m} \ & \boldsymbol{b}^T\boldsymbol{y} \\ \text{s.t.} \ & \boldsymbol{A}_0 - \textstyle\sum_{i=1}^m y_i\boldsymbol{A}_i \succeq 0 \\ & \|\boldsymbol{d} - \boldsymbol{D}^T\boldsymbol{y}\|_2 \le \gamma - \boldsymbol{e}^T\boldsymbol{y} \\ & \boldsymbol{f} - \boldsymbol{F}^T\boldsymbol{y} \ge 0 \end{aligned} \tag{A.3}$$

The first constraint is a standard LMI. The second constraint has a second-order cone (SOC) form, with $\boldsymbol{D}\in\mathbb{R}^{m\times p}$, $\boldsymbol{d}\in\mathbb{R}^p$, $\boldsymbol{e}\in\mathbb{R}^m$, $\gamma\in\mathbb{R}$. The third constraint is linear, with $\boldsymbol{f}\in\mathbb{R}^q$, $\boldsymbol{F}\in\mathbb{R}^{m\times q}$; the vector inequality is understood elementwise. The dual of (A.3) has the equality form

$$\begin{aligned} \min_{\boldsymbol{X},\boldsymbol{x},\xi,\tilde{\boldsymbol{x}}} \ & \mathrm{tr}[\boldsymbol{A}_0\boldsymbol{X}] + \boldsymbol{d}^T\boldsymbol{x} + \gamma\xi + \boldsymbol{f}^T\tilde{\boldsymbol{x}} \\ \text{s.t.} \ & \begin{bmatrix} \mathrm{tr}[\boldsymbol{A}_1\boldsymbol{X}] \\ \vdots \\ \mathrm{tr}[\boldsymbol{A}_m\boldsymbol{X}] \end{bmatrix} + \boldsymbol{D}\boldsymbol{x} + \boldsymbol{e}\xi + \boldsymbol{F}\tilde{\boldsymbol{x}} = \boldsymbol{b} \\ & \boldsymbol{X}\succeq 0, \ \|\boldsymbol{x}\|_2 \le \xi, \ \tilde{\boldsymbol{x}} \ge 0 \end{aligned} \tag{A.4}$$

The variables are $\boldsymbol{X}\in\mathbb{R}^{n\times n}$, $\boldsymbol{x}\in\mathbb{R}^p$, $\xi\in\mathbb{R}$, $\tilde{\boldsymbol{x}}\in\mathbb{R}^q$. The SOC dual variables are related by $\|\boldsymbol{x}\|_2\le\xi$. This is equivalent with the LMI

$$\begin{bmatrix} \xi\boldsymbol{I} & \boldsymbol{x} \\ \boldsymbol{x}^T & \xi \end{bmatrix} \succeq 0,$$

but it is much more efficient to use algorithms specific to SOC than to treat the SOC constraint as an LMI. The linear elementwise nonnegative variable is $\tilde{\boldsymbol{x}}$.

SQLP has properties similar to those of SDP (equal values at optimality if Slater's condition holds, complementarity relation, etc.). The complexity of solving an SQLP problem is mainly dictated by the SDP part.

There are a number of free libraries for solving SQLP (or only SDP) problems, among which we can cite SeDuMi [5] (used exclusively for the examples from this book), SDPT3 [6] and SDPA [7]. A friendlier and more general environment for convex optimization is provided by CVX [8].

References

1. L. Vandenberghe, S. Boyd, Semidefinite programming. SIAM Rev. **38**(1), 49–95 (1996)
2. J.P. Haeberly, M.V. Nayakkankuppam, M.L. Overton, Mixed semidefinite-quadratic-linear programs, in *Recent Advances in LMI Methods for Control*, eds. by L. El Ghaoui, S.I. Niculescu (SIAM, 2000), pp. 41–54
3. M.S. Lobo, L. Vandenberghe, S. Boyd, H. Lebret, Application of second-order cone programming. Linear Alg. Appl. **284**, 193–228 (1998)
4. S. Boyd, L. Vandenberghe, *Convex Optimization* (Cambridge University Press, Cambridge, 2004)
5. J.F. Sturm. Using SeDuMi 1.02, a Matlab toolbox for optimization over symmetric cones. Optim. Methods Softw. **11**, 625–653 (1999). http://sedumi.ie.lehigh.edu
6. K.C. Toh, M.J. Todd, R.H. Tütüncü, SDPT3 – a Matlab software package for semidefinite

   programming. Optim. Methods Softw. **11**, 545–581 (1999). http://www.math.nus.edu.sg/~mattohkc/sdpt3.html
7. K. Fujisawa, M. Fukuda, M. Kojima, K. Nakata, Numerical evaluation of the SDPA (SemiDefinite programming algorithm), in *High Performance Optimization*, eds. by H. Frenk, K. Roos, T. Terlaky, S. Zhang (Kluwer Academic Press, 2000), pp. 267–301. http://grid.r.dendai.ac.jp/sdpa
8. M. Grant, S. Boyd, *CVX: Matlab Software for Disciplined Convex Programming, version 2.1* (2014). http://cvxr.com/cvx

Appendix B: Spectral Factorization

Given a nonnegative trigonometric polynomial $R(z)$ defined as in (1.1), a spectral factorization algorithm computes the causal trigonometric polynomial $H(z)$ such that (1.11) holds. The existence of a solution is ensured by Theorem 1.1. We present here several algorithms and comment on their properties. We consider implicitly polynomials with real coefficients; for many algorithms, the extension to complex coefficients is immediate. More reading on this topic can be found, e.g., in [1–3] and the articles cited as references for each method given below.

B.1 Root Finding

As discussed in Remark 1.2, a spectral factor can be found by computing the roots of $R(z)$, or rather the $2n$ roots of $\tilde{R}(z) = z^n R(z)$. Several root finding algorithms are used and compared in [4, 5]. While the convergence speed and the complexity of root finding methods may be different, there are more important issues occurring in spectral factorization.

It is often assumed that the zeros on the unit circle have exactly double multiplicity (which is a fair assumption in practice). These are single roots of $\tilde{R}'(z) = d\tilde{H}(z)/dz$ and thus can be computed more accurately. The identification of zeros on the unit circle is made by associating two zeros of $\tilde{R}(z)$ with a zero of $\tilde{R}'(z)$. This procedure works well if the double zeros are well separated.

Even more delicate is the formation of $H(z)$ from the zeros inside the unit circle. An heuristic called Leja ordering specifies the order in which the elementary factors $z - z_k$ have to be multiplied, depending on the values of the roots z_k. Even so, the process of computing the coefficients of a polynomial from its zeros is prone to instability.

We conclude by not recommending this algorithm, unless the polynomial is short, and thus, the results are safe. However, this algorithm has the advantage of choosing easily the roots for the spectral factors and so it becomes interesting, e.g., when

B. Dumitrescu, *Positive Trigonometric Polynomials and Signal Processing Applications*, Signals and Communication Technology,
DOI 10.1007/978-3-319-53688-0

approximately linear-phase spectral factors are sought; these factors have roots both inside and outside the unit circle.

B.2 Newton–Raphson Algorithm

The relation $R(z) = H(z)H(z^{-1})$ can be seen as the system of quadratic equations

$$r_k = \boldsymbol{h}^T \boldsymbol{\Theta}_k \boldsymbol{h}, \; k = 0 : n. \tag{B.1}$$

(See relations (1.17), (2.27).)

The Newton–Raphson method for solving a nonlinear system $\boldsymbol{f}(\boldsymbol{x}) = 0$, with $\boldsymbol{x}, \boldsymbol{f}(\boldsymbol{x}) \in \mathbb{R}^m$ is based on the iteration

$$\boldsymbol{x}_{i+1} = \boldsymbol{x}_i - \boldsymbol{f}'(\boldsymbol{x})^{-1} \boldsymbol{f}(\boldsymbol{x}), \tag{B.2}$$

where i is the iteration number and $\boldsymbol{f}'(\boldsymbol{x})$ is the Jacobian matrix of the function $\boldsymbol{f}$, evaluated in $\boldsymbol{x}$.

In the case of the system (B.1), the nonlinear function is

$$\boldsymbol{f}(\boldsymbol{h}) = \begin{bmatrix} \vdots \\ \boldsymbol{h}^T \boldsymbol{\Theta}_k \boldsymbol{h} - r_k \\ \vdots \end{bmatrix}$$

and its Jacobian is

$$\boldsymbol{f}'(\boldsymbol{h}) = \begin{bmatrix} h_0 & h_1 & \dots & h_n \\ h_1 & h_2 & \iddots & 0 \\ \vdots & \iddots & \iddots & 0 \\ h_n & 0 & \dots & 0 \end{bmatrix} + \begin{bmatrix} h_0 & h_1 & \dots & h_n \\ 0 & h_0 & \ddots & h_{n-1} \\ \vdots & \ddots & \ddots & \vdots \\ 0 & 0 & \dots & h_0 \end{bmatrix}. \tag{B.3}$$

It results that the main task in an iteration (B.2) is to solve a linear system of equations. Moreover, due to the Toeplitz-plus-Hankel form of the matrix (B.3), fast algorithms are possible. Finally, an appealing trait of the method is that initialization with $\boldsymbol{h}_0 = [1 \; 0 \; \dots \; 0]^T$, i.e., with a trivial polynomial of degree 1, makes the Newton–Raphson iteration converge to the minimum-phase spectral factor. (Actually, this always happens when the initial polynomial is minimum-phase itself.) More details on this method can be found in [5–7] and the references therein.

The method is relatively fast, but the convergence is slower when there are zeros on the unit circle. The matrix (B.3) tends to be ill-conditioned as the iterative process advances.

B.3 Factorization of Banded Toeplitz Matrices

In Sect. 1.3, we have argued that the banded Toeplitz matrices $\boldsymbol{R}_m$ defined in (1.28) are positive semidefinite. If they are positive definite, which means that $R(z)$ has no zeros on the unit circle, then there exist a Cholesky factorization $\boldsymbol{R}_m = \boldsymbol{L}_m \boldsymbol{L}_m^H$, where $\boldsymbol{L}_m$ is lower triangular. Moreover, the matrix $\boldsymbol{L}_m$ is also banded. It can be proved that, as $m \to \infty$, the nonzero elements on the last row of $\boldsymbol{L}_m$ tend to be equal to $[h_n \ \ldots \ h_0]$, i.e., to the coefficients of the spectral factor $H(z)$. This is the base of the Bauer spectral factorization method [8]. Faster algorithms, taking advantage of the Toeplitz structure, are given, e.g., in [9].

B.4 Hilbert Transform Method

For a minimum-phase system, in our case the spectral factor $H(z)$, the magnitude of the frequency response determines completely the response. The phase is given by

$$\arg H(\omega) = -\mathcal{H}\{\log |H(\omega)|\}, \tag{B.4}$$

where $\mathcal{H}$ is the Hilbert transform, i.e., the linear system whose frequency response is

$$\mathcal{H}(\theta) = \begin{cases} -j, & \text{for } \theta \in (0, \pi), \\ 0, & \text{for } \theta = 0, \pi, \\ j, & \text{for } \theta = (-\pi, 0). \end{cases} \tag{B.5}$$

The relation (B.4) can be implemented using several FFT transforms, along the following steps.

1. Compute the log magnitude

$$x_\ell = \log |H(\omega_\ell)| = \tfrac{1}{2} \log R(\omega_\ell), \tag{B.6}$$

on a grid of $N \gg n$ points $\omega_\ell = 2\pi\ell/N$, $\ell = 0 : N-1$. The FFT can be used for an efficient calculation of $R(\omega_\ell)$. The number of points for the FFT can be taken as a power of two, the smallest larger than, e.g., $20n$.

2. Compute the Hilbert transform (B.4), using FFT twice. Let

$$X_i = FFT(x_\ell), \ \ i = 0 : N-1,$$

be the discrete Fourier transform of the sequence x_ℓ. In the transform domain, the application of the Hilbert transform (B.5) is equivalent to the operations

$$X_i = \begin{cases} -jX_i, & \text{for } i = 1 : N/2 - 1, \\ 0, & \text{for } i = 0, N/2, \\ jX_i, & \text{for } i = N/2 + 1 : N - 1. \end{cases}$$

Going back to the original domain by $y_\ell = IFFT(X_i)$, we obtain the phase response on the frequency grid.

3. The frequency response of the spectral factor is $H(\omega_\ell) = \mathrm{e}^{x_\ell - jy_\ell}$.

4. The coefficients of $H(z)$ result by applying IFFT (and retaining only the first $n+1$ coefficients; the others should be approximately zero). If N is a multiple of $n+1$, then an FFT of order $n+1$ can be used.

The idea of this method is attributed to Kolmogorov.

B.5 Polynomials with Matrix Coefficients

Many spectral factorization methods generalize to polynomials with matrix coefficients by simply replacing scalars with matrices. For the methods from this appendix, this is true for the Cholesky factorization and the Newton–Raphson methods (although for the latter we are not aware of any confirming experiments). Also, the two algorithms from Sect. 2.6 generalize directly (see [10] and [11]). For example, when the coefficients are $\kappa \times \kappa$ matrices, the problem (2.45) must be replaced with

$$\begin{aligned} \max_{\boldsymbol{Q}} \ & \operatorname{tr} \boldsymbol{Q}_{00} \\ \text{s.t.} \ & \mathrm{TR}[\boldsymbol{\Theta}_{\kappa k} \boldsymbol{Q}] = \boldsymbol{R}_k, \quad k = 0 : n \\ & \boldsymbol{Q} \succeq 0 \end{aligned} \tag{B.7}$$

where $\boldsymbol{Q}_{00}$ is the $\kappa \times \kappa$ upper left block of the matrix variable $\boldsymbol{Q}$ and $\boldsymbol{R}_k$ are the coefficients of the polynomial to be factorized. The solution $\boldsymbol{Q}^\star$ of (B.7) has rank κ and thus can be written as (through, e.g., eigenvalue decomposition or Cholesky factorization with pivoting)

$$\boldsymbol{Q}^\star = \begin{bmatrix} \boldsymbol{H}_0^H \\ \vdots \\ \boldsymbol{H}_n^H \end{bmatrix} \begin{bmatrix} \boldsymbol{H}_0 \ldots \boldsymbol{H}_n \end{bmatrix},$$

where $\boldsymbol{H}_k$ are the coefficients of the desired spectral factor.

References

1. S.P. Wu, S. Boyd, L. Vandenberghe, FIR filter design via spectral factorization and convex optimization, in *Applied and Computational Control, Signals and Circuits*, ed. by B. Datta (Birkhauser, 1997), pp. 51–81

2. A.H. Sayed, T. Kailath, A survey of spectral factorization methods. Numer. Lin. Alg. Appl. **8**, 467–496 (2001)
3. P.P. Vaidyanathan, *Multirate Systems and Filter Banks* (Prentice Hall, New Jersey, 1993)
4. M. Lang, B.C. Frenzel, Polynomial root finding. IEEE Signal Process. Lett. **1**, 141–143 (1994)
5. H.J. Orchard, A.N. Willson Jr., On the computation of a minimum-phase spectral factor. IEEE Trans. Circ. Syst. I **50**(3), 365–375 (2003)
6. G.T. Wilson, Factorization of the covariance generating function of a pure moving average process. SIAM J. Numer. Anal. **6**(1), 1–7 (1969)
7. C.J. Demeure, T.M. Mullis, A Newton-Raphson method for moving-average spectral factorization using the euclid algorithm. IEEE Trans. Acoust. Speech Sign. Proc. **38**(10), 1697–1709 (1990)
8. F.L. Bauer, Ein Direktes Iterationsverfahren zur Hurwitz-Zerlegung eines Polynom. Arch. Elektr. Übertragung **9**(6), 285–290 (1955)
9. J. Rissanen, Algorithms for triangular decomposition of block Hankel and Toeplitz matrices with applications to factoring positive matrix polynomials. Math. Comput. **27**(121), 147–154 (1973)
10. J.W. McLean, H.J. Woerdeman, Spectral factorizations and sums of squares representations via semidefinite programming. SIAM J. Matrix Anal. Appl. **23**(3), 646–655 (2002)
11. B.D.O. Anderson, K.L. Hitz, N.D. Diem, Recursive algorithm for spectral factorization. IEEE Trans. Circ. Syst. **21**(6), 742–750 (1974)

Index

B. Dumitrescu, *Positive Trigonometric Polynomials and Signal Processing Applications*, Signals and Communication Technology,
DOI 10.1007/978-3-319-53688-0