Convex Approximation and Optimization with Applications in Magnitude Filter Design and Radiation Pattern Synthesis

Peter William Kassakian

Electrical Engineering and Computer Sciences
University of California at Berkeley

Technical Report No. UCB/EECS-2006-64
http://www.eecs.berkeley.edu/Pubs/TechRpts/2006/EECS-2006-64.html

May 18, 2006

Convex Approximation and Optimization with Applications in Magnitude Filter Design and Radiation Pattern Synthesis

by

Peter William Kassakian

B.A. (Massachusetts Institute of Technology) 1995
M.S. (Massachusetts Institute of Technology) 1999

A dissertation submitted in partial satisfaction
of the requirements for the degree of

Doctor of Philosophy

in

Electrical Engineering and Computer Science

in the

GRADUATE DIVISION

of the

UNIVERSITY OF CALIFORNIA, BERKELEY

Committee in charge:

Professor Laurent El Ghaoui, Chair
Professor David Wessel
Professor Michael Gastpar

Spring 2006

The dissertation of Peter William Kassakian is approved:

Chair

Date

Date

Date

University of California, Berkeley

Spring 2006

Convex Approximation and Optimization with Applications in Magnitude Filter Design and Radiation Pattern Synthesis

by

Peter William Kassakian

ABSTRACT

Convex Approximation and Optimization with Applications in
Magnitude Filter Design and Radiation Pattern Synthesis
by
Peter William Kassakian
Doctor of Philosophy in Electrical Engineering and Computer Science
University of California, Berkeley
Professor Laurent El Ghaoui, Chair

Using convex optimization to help solve nonconvex problems in engineering is an area of intense research activity. In this thesis we study a specific nonconvex optimization problem called magnitude least-squares that has applications primarily in magnitude filter design. Solving the problem is difficult because of the existence of many local minima. We study it in depth, deriving methods for its approximate solution, proving equivalences among differing formulations, relating it to other well-studied problems, and proving estimates of the quality of the solutions obtained using the methods. We discover structure in the problem that distinguishes it from some more general problems of the same algebraic form. The structure is related to the fact that the variables in the problem are complex-valued. We exploit this structure when proving bounds on the quality of solutions obtained using semidefinite relaxation.

In addition to a detailed and generally abstract study of this specific optimization problem, we solve several practical problems in signal processing. Some of the application examples serve to illustrate the applicability of the magnitude least-squares problem, and include multidimensional magnitude filter design, magnitude filter design for nonlinearly delayed tapped filters, and spatial filtering using arbitrarily positioned array elements. We also present several application examples that illustrate the modeling capabilities of convex optimization. We use least-squares techniques to reason about the capabilities of clustered

arrays of loudspeakers to accurately synthesize radiation patterns. We also provide an elegant convex optimization-based procedure for designing linear-phase audio equalizers.

DEDICATION

Dedicated to Mom, Dad, and Meg

ACKNOWLEDGMENTS

I would like to thank David Wessel and the Center for New Music and Audio Technologies for financially supporting me from my first term at U.C. Berkeley to my last. I would also like to express my gratitude to Meyer Sound Laboratories for consistently funding this research. Support for the project was also provided by a UC Discovery Grant in Digital Media from the University of California's Industry-University Cooperative Research Program (IUCRP).

I would like to thank my advisors Laurent, Michael, David, and Lieven, all of whom provided invaluable assistance both in terms of technical expertise and research strategy. I would also like to acknowledge the Center for New Music and Audio Technologies and Meyer Sound Laboratories for allowing me space and time to tackle some topics that did not immediately admit direct practical value. I truly appreciate having been given that opportunity. I would like to thank Richard Andrews, Edmund Campion, and the other researchers in the lab, Adrian Freed, Rimas Avizienis, Perrin Meyer, Brian Vogel, Ed Berdahl, Michael Zbyszynski, Matt Wright, Psyche Loui, Ali Momeni, Roberto Morales, and John MacCallum.

Finally I'd like to express my deep gratitude to Ruth Gjerde and Mary Byrnes.

Contents

Chapter 1

Introduction

1.1 Mathematical Modeling and Optimization

Consider how people use calculators. They typically have a collection of numbers that represent quantities of some important items. They have in mind relations between these quantities and a quantity they're seeking. Before opening the drawer for the calculator, and possibly without being conscious of it, a person will perform a miniature analysis of the numerical task to determine if the calculator is capable of completing it. For example, if the calculation requires raising a number to a fractional power, and the calculator does not possess that button, then maybe there is no point in trying to use the calculator. Perhaps the problem must first be algebraically rearranged, or maybe no amount of rearranging would allow for its calculation on that calculator.

Often this analysis takes place long before we're contemplating reaching for the calculator. We've been faced with a scenario in which we suspect that the conditions we observe have implications that we should be able to infer. It could be as simple as observing the items in a supermarket shopping cart and recognizing that it should be possible to estimate the total cost based on the costs of the individual items. People generally know which life scenarios are amenable to mathematical modeling suitable for calculation on a calculator. They lean away from attempting to cast a scenario in mathematical terms if they think they have a poor chance of calculating it. Every once in a while, a person will be motivated enough to approximate the problem to increase the odds that he or she can obtain a helpful

answer.

Everyone is unconsciously making mathematical modeling decisions based both on the importance of the problem, and on the likelihood they can solve it. The same is true in engineering, but the calculator is more elaborate. For an engineer, the motivation to cast a scenario in mathematical terms is very strong, and the algebraic skill very high, but nevertheless, he or she is involved in the same decision-making and modeling process as a calculator owner.

In this thesis we will solve practical problems using mathematical optimization. But we use the problems as means to learn about mathematical modeling decisions, and the ramifications these decisions have on our ability to solve the problems. To facilitate this goal, we will narrow the study to several interrelated problems, comparing different solution ideas, and paying close attention to how the modeling and formulation affects the solutions obtained.

In particular we wish to avoid modeling a problem in such a way that our calculator cannot arrive at a solution in a timely manner. If we're lucky, this won't be an issue. But most of the time, even armed with the ultimate calculator, we face many obstacles on the road from real-life scenario to numerical solution. The first, sometimes overlooked obstacle is identifying when and how the scenario is amenable to mathematical description. If we are able to find a precise way to describe the scenario in mathematical terms, then we have made a start. We have a description, but not necessarily a solution method.

Many methods exist for solving many classes of mathematical problem. Some work well on some classes of problem, but are not applicable for others. Some work very well on some problem instances, but are sensitive and may not work all the time. Some ideas do not work well on any problem. This thesis is focused on employing numerical solution methods that are very reliable and very straightforward. The methods mirror the reliability we would expect from a calculator.

Although there are plenty of very reliable methods for other classes of problem, by and large, such reliable methods apply to problems that are *convex*. Our aim will be to model

scenarios not just in mathematical terms, but in such a way that the resulting problem is convex, and hence can be solved using our calculator of choice: convex optimization.

To achieve this aim, our modeling strategy must negotiate the conflicting requirements that the model be both accurate, and convex. This is impossible in most cases. We must make a compromise somewhere. To keep the study of these compromises organized and well-defined, we will adopt the following philosophy. When faced with a real-life scenario, we will first seek to model the situation in mathematical terms as accurately as we can. Next, we will determine if the the resulting problem is convex. If so, we solve. If not, we will search for a new modeling scheme, or a set of algebraic manipulations that result in a convex problem. We will see a very interesting and nontrivial example of such a manipulation in Chapter 4, involving spectral factorization. If we cannot discover a convex formulation, we will choose an accurate mathematical description, and perform a very well-defined approximation that will result in a convex problem. The answer obtained from the approximate problem, of course will be approximate.

For comparison, let's consider an alternative philosophy. Suppose we seek an accurate model of the scenario, and attempt to solve numerically without worrying about convexity. There are many examples of successful employment of this philosophy; in fact it may be the dominant philosophy in engineering. But it has the primary drawback that we must understand a great deal about the inner workings of the solution method before we can be confident that the solution obtained is meaningful to us. Depending on the solution method, we may obtain different answers. Not worrying about convexity leads to worry about the numerical solution algorithm. In metaphorical terms, this philosophy leads to becoming involved in the design of the calculator.

The most useful, widespread, well-understood, and successful engineering solutions are the ones that possess convex formulations. These include linear regression, least-mean-square (LMS) adaptive filtering, minimum variance distortionless response (MVDR) beamforming, and one-dimensional filter design to name a few in signal processing. When faced with a new problem, it is of critical importance that, at the very least, we recognize whether

the problem is equivalent in mathematical form to these well-studied problems.

It is the aim of this thesis not just to find answers to problems arising in signal processing, but rather to frame the problems with respect to other well-studied problems. The well-studied problems we will pay attention to are the convex problems. In doing so, we will gain a better understanding of the problem – why it is easy or hard, ways to measure the quality of the solutions, and it's similarity to other problems we've seen. We also gain insight from a rich duality theory which plays a special role in the analysis of convex problems.

1.2 Preview of Thesis

The application area of the problems we will study in the coming chapters is signal processing, for use especially in audio, acoustics, and music. We will touch upon several different applications, treating some lightly, the purpose of inclusion being to provide an example or to aid in the explanation of a concept, and treating a few in great depth. As mentioned previously, we will frame all our problems in relation to convex optimization. Several of the problems we study are novel, the formulations and solutions of which form academic and practical contributions in the engineering community. Arguably more valuable is the analysis of the relations among the problems that we are able to form using the convex optimization philosophy.

In the next chapter, we will introduce convex optimization in the context of two practical problems that arose as a part of an ongoing project at the Center for New Music and Audio Technologies (CNMAT). The first problem concerns synthesizing radiation patterns using a clustered array of loudspeakers. We can determine which patterns the array is capable of synthesizing by casting the scenario abstractly and geometrically in infinite dimensions. We will use least-squares to reason about radiation pattern synthesis. Furthermore, we can use the formulation to answer questions about electronically controlling and rotating the patterns. The second problem we will look at is the design of a linear-phase equalizer/limiter for use with the clustered loudspeaker array. We will see that our aware-

ness of convex optimization allows us to quickly and easily design a relatively complicated system of filters. This chapter can be understood as a warm-up for the heart of the thesis which begins in Chapter 3.

Chapter 3 introduces a class of problem that we will focus on for the remainder of the thesis, investigating formulations, approximation methods, solution methods, and applications. The problem is called the magnitude least-squares (MLS) problem. In this chapter we will motivate its study by considering a diverse set of applications, all of which can be formulated as magnitude least-squares problems. We will also consider the relation of the MLS problem to several important recently developed approaches to one-dimensional filter design, and recognize that the MLS problem can be regarded as a more general problem in some aspects. We will review the methods for one-dimensional filter design, in Chapter 4, and recognize that the MLS problem can be viewed as a form of approximate polynomial factorization.

In Chapter 5 we will consider several potential ways to solve the magnitude least-squares problem, including local methods that do not rely on convexity. As alluded to earlier, these methods have the disadvantage that they do not provide us much insight into the structure of the problem, and may result in suboptimal local solutions. We will continue by considering ideas for convexly approximating the problem, using duality concepts.

Chapter 6 forms the core of the thesis. We will prove important facts pertaining to the accuracy of solutions obtained using the convex approximations. We will begin the discussion by reviewing a paper written by Yurii Nesterov [Nes97] in which he generalizes, in several ways, a scheme originally discovered by Goemans and Williamson in [GW95]. Not only does the scheme work very well in practice, but importantly, it carries with it a provably good estimate of the quality of the obtained solutions. We show how we can directly apply Nesterov's extension to the magnitude least-squares problem, and investigate other similar ideas, one of which improves the quality estimate, inspired by the work of Shuzhong Zhang and Yongwei Huang [ZH04]. The reason for the existence of the better quality estimate is the complex-valued structure of our problem. We will present novel insight into

this phenomenon by proving quality bounds for a related complex-valued problem in a very direct way. The bounds we derive here represent the most important and deepest contribution of the dissertation. It complements the work in [ZH04] where they derive a superior bound through careful evaluation of integrals.

In the final chapters we will solve problems using the techniques, and discuss the practical performance of the methods in relation to the proved quality bounds. We will also consider a set of problems that are closely related to the magnitude least-squares problem, along with the application that inspire their formulation, carefully identifying which of these can be solved with fixed relative accuracy in the same way as described in Chapter 6, and which cannot. Finally, in Chapter 9 we will summarize the contribution of this dissertation to the engineering community, and provide suggestions for further research.

Chapter 2

Convex Optimization Problems in Acoustics

2.1 Introduction

In this chapter we will introduce the philosophy of using modeling and optimization to solve real-world problems by examining some problems in acoustics, signal processing, and music. We will seek formulations that result in problems that are convex, and consequently are straightforward to solve numerically. The two problems discussed in this chapter arose as part of a "real-world" acoustics research project at the Center for New Music and Audio Technologies (CNMAT).

We will first discuss an interesting problem of characterizing the set of radiation patterns accurately synthesizable using a clustered array of loudspeakers. This analysis involves constructing a set of problems problem that can be formulated very accurately using a convex model. We will describe the real-world challenge in nonmathematical terms, provide the model, explain the solution, and present specific results. We will then consider another practical problem that again, we can solve exactly using convex optimization: the design of a linear-phase audio crossover/equalizer network for use with loudspeakers.

2.2 Characterization of Clustered Loudspeaker Arrays

Consider an array of closely spaced, independent loudspeakers. Such systems have been designed and constructed at the Center for New Music and Audio Technologies, see Fig-

ure 2.1. The primary aim in the development of these arrays was to create a sound synthesis system capable of reproducing complicated radiation patterns, reminiscent of those naturally generated by acoustic musical instruments. It has been speculated that the dynamic and nontrivial radiation patterns produced by acoustic instruments give the listener experiences that are fuller and more exciting than those generated by electronic instruments amplified through mono or stereo speakers [OW01].

Figure 2.1: **Dodecahedral Speaker Array.**

The dodecahedral system pictured in Figure 2.1 was designed so that each loudspeaker could be independently controlled, each with its own isolated enclosure of air within the

interior of the cabinet. The requirement that this chamber of air be one liter, as suggested by the loudspeaker specifications, was the chief factor determining the dimensions of the array. These dimensions, in turn, determine which radiation patterns we can accurately synthesize.

After experimenting with straightforward methods of generating nontrivial patterns with the array, such as running sets of speakers in opposite phase to create directional cancellation effects, it became clear that we needed to know more about the theoretical capabilities of the array. For example, to what extent could this array be used to synthesize the radiation pattern of a piano? How much better is the synthesis if we allow ourselves the use of high-order filters as opposed to sign (in phase/out of phase) filters? If we can synthesize a radiation pattern that's oriented in a particular direction, does that mean there's a way to electronically rotate the pattern smoothly?

We are able to answer some of these questions by formulating a series of least-squares fitting problems. The solutions to these convex problems, along with the fitting error, give us vital information about the sets of radiation patterns that are synthesizable and controllable using the array. The work is documented in detail in an Audio Engineering Society (AES) paper entitled "Characterization of Spherical Loudspeaker Arrays" [KW04].

2.2.1 Radiation Patterns

To begin our discussion of the dodecahedral loudspeaker array, it's useful to review some simple ideas about radiation patterns. A sound source produces sound by creating a physical disturbance in the air. Sound waves radiate outward from the source decreasing in intensity as they travel further from the source. Once far enough from the source, the amplitudes and phases of the waves relative to that of the sound source, as functions both of frequency and of three-dimensional spatial angle settle into a fixed pattern, known as the far-field radiation pattern. This is the acoustic entity we would like to synthesize and control using the array.

Because loudspeaker arrays consist of a finite number of drivers in a fixed geometry,

the far-field radiation patterns they can produce are limited to a subset of all the conceivable patterns. For example, a two speaker system could never reproduce the complicated radiation pattern generated by a violin. A single loudspeaker has a fixed radiation pattern at every frequency; we have no means to control it. An arbitrary vibrating body could conceivably possess an arbitrarily complicated radiation pattern at each frequency. We wish to find which types of patterns can be well-approximated using arrays of loudspeakers.

It is convenient to parameterize the set of radiation patterns in terms of spherical harmonics. The spherical harmonics are solutions to the Helmholtz equation, (see Chapter 10 of [Bla00] and also [Ber93]), and span the entire space of physically realizable far-field patterns. Said another way, any physically realistic spatial response can be represented by its spherical harmonic expansion as a function of frequency. The decomposition has intuitive appeal, indexing the spatial complexity or "spatial frequency" of radiation patterns. Our characterization of the sets of patterns that can or cannot be generated with the array will be with respect to these convenient functions. We will see that the spherical harmonics possess other useful and meaningful properties as well.

2.2.2 Physical System

The loudspeaker array consists of independent loudspeakers arranged in a fixed geometry. The important elements of the system are (i) the total pattern produced as a function of frequency, (ii) the individual loudspeaker patterns as functions of frequency, and (iii) the frequency responses of the linear filters applied to the loudspeaker signals. We will assign variables to these quantities.

Assume that the array consists of N loudspeakers. The physical system is linear and a frequency domain block diagram is shown in Figure 2.2. We can calculate the output of the system as a linear combination of N individual loudspeaker patterns. An individual loudspeaker pattern is the output function response of a single array element (evaluated at

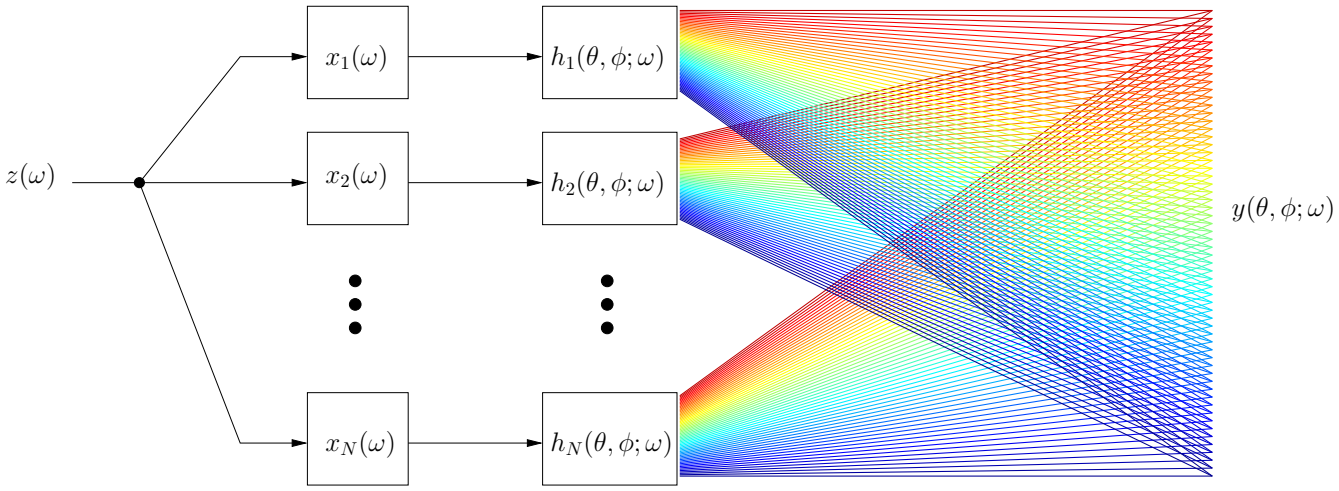


Figure 2.2: **Linear System Model.** The input is the Fourier transform of the input signal evaluated at ω (a complex scalar), and the output is a pattern (a complex-valued function of spherical angles (θ, ϕ)).

a frequency ω). We have a frequency domain linear system relation,

$$y(\theta, \phi; \omega) = z(\omega) \sum_{n=1}^{N} x_n(\omega) h_n(\theta, \phi; \omega),$$

where each $h_n(\theta, \phi; \omega)$ is an individual loudspeaker function, $y(\theta, \phi; \omega)$ is the resultant output function, $z(\omega)$ is the Fourier transform of the monophonic input signal evaluated at ω, and $x_n(\omega)$ is the frequency response of the nth filter evaluated at ω.

In abstract terms, the patterns are complex functions defined on S^2, the unit sphere. Suppose f and g are two patterns. It isn't difficult to show that the function

$$\langle f, \overline{g} \rangle = \int_{S^2} f(s) \overline{g(s)} \mathrm{d}s$$

defines an inner-product on the vector space of complex functions on S^2. This integral in spherical coordinates is

$$\langle f, \overline{g} \rangle = \int_0^{2\pi} \int_0^{\pi} f(\theta, \phi) \overline{g(\theta, \phi)} \sin(\theta) \mathrm{d}\theta \mathrm{d}\phi$$

The integral squared error between pattern f and g then corresponds to

$$\begin{aligned} \text{Error} &= \int_{S^2} |f(s) - g(s)|^2 \, \mathrm{d}s \\ &= \langle f - g, \overline{f - g} \rangle \\ &= \|f - g\|^2 \end{aligned} \tag{2.1}$$

Given an arbitrary pattern g (at a fixed frequency), the gains that minimize the error as defined in Equation (2.1) between a target pattern g and a pattern f realizable by the array, can be calculated as the solution to a linear least squares problem. The system is infinite in the sense that the output is a function, but the least-squares problem is finite and only depends on a Gram matrix involving the desired output, and the individual loudspeaker functions. The least-squares problem is one of the simplest and most widely applied convex optimization problems. For a good derivation and explanation of least-squares problems involving function spaces, see either [Che82] or [Lue69].

Evaluation of the entries of the Gram matrix requires integration however, and if the individual loudspeaker patterns are derived from measured data, the integrals must be approximated. We will approximate the integrals by sums of squared error evaluated at points uniformly randomly sampled on S^2.

By sampling the frequency axis we can compute optimal responses $\{x_1(\omega_i), ..., x_N(\omega_i)\}$ at each fixed frequency ω_i. This analysis seeks to determine the patterns that are approximatable under the best possible conditions, i.e., when we are able to specify a linear filter of arbitrarily large order for each loudspeaker signal, rendering the sets of responses at different frequencies independently adjustable. We are seeking to characterize the physical capabilities of the array with respect to its ability to accurately synthesize radiation patterns.

2.2.3 Continuous and Sampled Least-Squares

For notational simplicity, we will suppress dependence on ω, remembering that ω indexes a family of least-squares problems. Given a desired pattern d, we have that the optimal

approximation in the span of $\{h_1, ..., h_N\}$ can be written as

$$d^\star = \sum_{n=1}^{N} x_n^\star h_n,$$

where the following equations are satisfied:

$$\sum_{n=1}^{N} x_n^\star \langle h_i, \overline{h_n} \rangle = \langle h_i, \overline{d} \rangle, \qquad i = 1, ..., N. \tag{2.2}$$

Equations (2.2) are the normal equations. In matrix form, they are

$$Hx^\star = b \tag{2.3}$$

where $H \in \mathbb{C}^{N \times N}$ and $b \in \mathbb{C}^N$ are defined as

$$H_{ji} = \langle h_i, \overline{h_j} \rangle \qquad i, j = 1, ..., N$$
$$b_i = \langle h_i, \overline{d} \rangle \qquad i = 1, ..., N$$

A spatially sampled version of the above problem is derived by first uniformly randomly sampling M points from S^2 and creating the vectors

$$[h_1(\theta_1, \phi_1), ..., h_1(\theta_M, \phi_M)]^T$$
$$\vdots$$
$$[h_N(\theta_1, \phi_1), ..., h_N(\theta_M, \phi_M)]^T, \quad \text{and}$$
$$[d(\theta_1, \phi_1), ..., d(\theta_M, \phi_M)]^T.$$

As M increases, the sampled functions approximate more and more closely the continuous functions

$$h_1(\theta, \phi)$$
$$\vdots$$
$$h_N(\theta, \phi) \quad \text{and}$$
$$d(\theta, \phi).$$

The finite vector inner product $h_i^H h_j$ then becomes an approximation to $\frac{M}{4\pi}\langle h_i, \overline{h_j}\rangle$. Define the matrix $A \in \mathbb{C}^{M\times N}$, and $d_s \in \mathbb{C}^M$as follows

$$A = \begin{bmatrix} h_1(\theta_1,\phi_1) & \cdots & h_N(\theta_1,\phi_1) \\ \vdots & \vdots & \vdots \\ h_1(\theta_M,\phi_M) & \cdots & h_N(\theta_M,\phi_M) \end{bmatrix},$$

$$d_s = \begin{bmatrix} d(\theta_1,\phi_1) \\ \vdots \\ d(\theta_M,\phi_M) \end{bmatrix}$$

Then the sampled version of Equation (2.3) is seen to be

$$A^H A x^\star = A^H d_s. \tag{2.4}$$

Equation (2.4) can be constructed and solved easily, the dominating factor in the computation time being the lookup and/or calculation of the sampled functions A and d_s.

To conclude this section on least-squares problems, we want to emphasize that the solution to the normal equations is a linear function of the desired pattern d. This implies that once we know the optimal solution vectors $x_1^\star, ..., x_r^\star$ for several desired responses $d_1, ..., d_r$, we immediately know the solution for any linear combination $\sum \lambda_i d_i$. It is $\sum \lambda_i x_i^\star$.

This linearity is important because it allows us to draw conclusions about the entire uncountable set of radiation patterns by solving problems involving only the countable set of spherical harmonics. It also sometimes allows us to bound errors across sets of patterns defined by continuous control signals, like rotation, by solving finite sets of optimization problems. We will now look with more detail at the spherical harmonics.

2.2.4 Spherical Harmonics

The spherical harmonics provide a convenient and natural countable parameterization of the the set of acoustic radiation patterns. By computing the best array-achievable approximations to each spherical harmonic, we can immediately determine the best approximation to a linear combination of spherical harmonics via the same linear combination of solutions, as mentioned in the previous section.

The functions are parameterized by their *degree* l, and *order* m. Figure 2.3 shows the complex functions plotted from degree 0 to degree 2. The magnitude is represented as distance from the origin, and phase as a color gradient.

For example the "dipole" corresponds to $l = 1$, $m = 0$. Other rotated orientations of this dipole function can be calculated by forming a weighted combination of patterns drawn from the $l = 1$ grouping. In fact, synthesis of any function composed of $l = 1$ patterns, oriented in any direction, is possible using only the $l = 1$ patterns. This is because the linear subspaces spanned by harmonics with the same degree are invariant with respect to rigid rotation through spatial angles θ and ϕ. See [SC94], and refer to [KFR03] for an application of this principle in computer graphics. For example, given a pattern that is in the subspace generated by the five degree 2 basis functions, any rotation of that pattern also possesses a spherical harmonic expansion consisting only of degree 2 patterns. Since we wish to control patterns in arbitrary rotational orientations, this decomposition into rotationally invariant subspaces is especially useful.

Another important property of the spherical harmonics is their orthonormality. Specifically, with respect to the inner product defined above, if f is a spherical harmonic,

$$\langle f, \overline{f} \rangle = 1, \tag{2.5}$$

and if f and g are distinct spherical harmonics,

$$\langle f, \overline{g} \rangle = 0. \tag{2.6}$$

2.2.5 Uniform Error

Given a particular target pattern, the error as a function of frequency can be determined numerically by discretizing the frequency axis, and solving a set of problems of the form described previously. We would also like to determine how the error generally varies as a function of spherical harmonic degree. In this section we will discuss the computation of uniform upper and lower bounds on the normalized error, given a *subspace* of desired patterns.

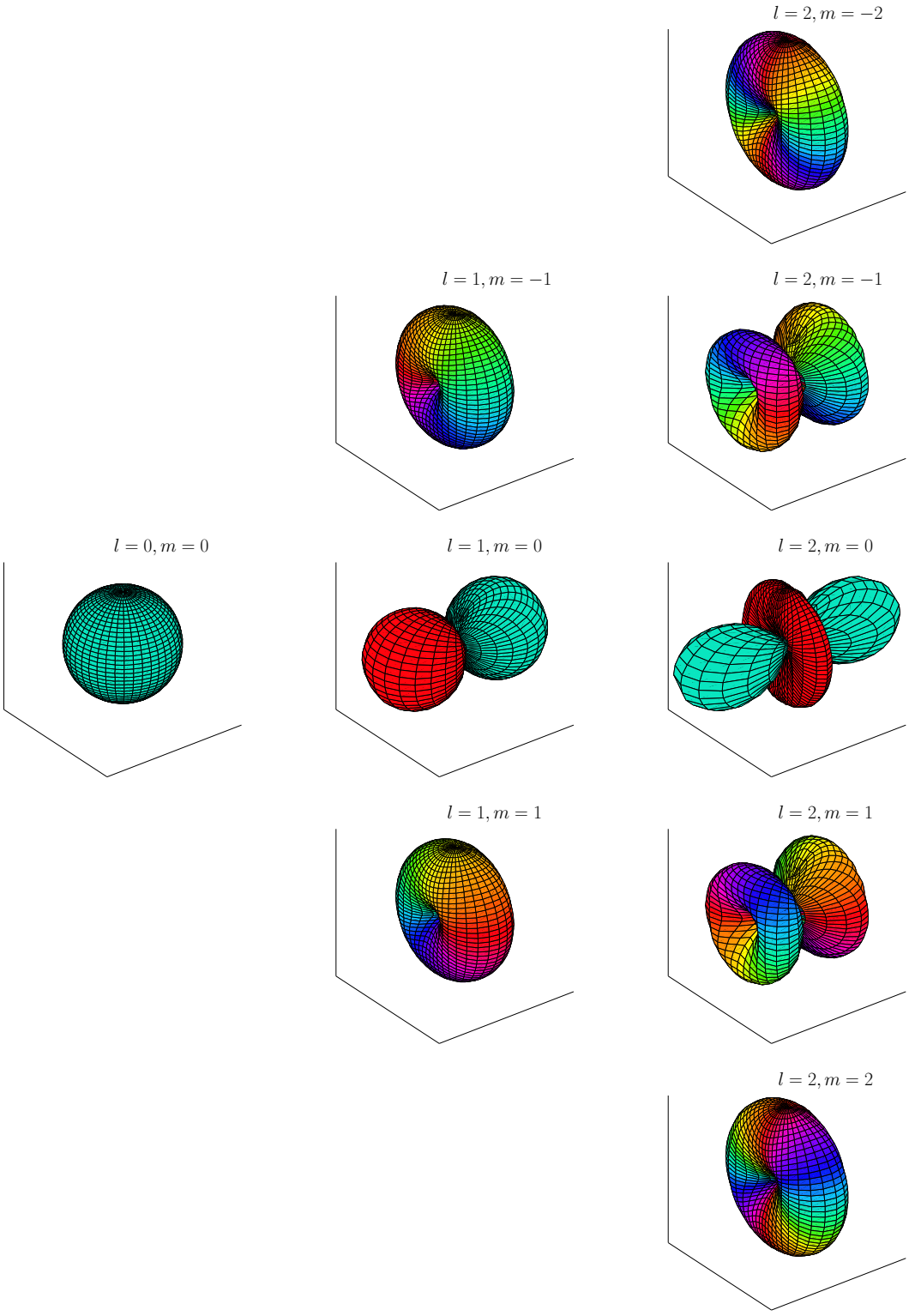


Figure 2.3: **Spherical harmonics to degree 2**

The issue we are addressing is that while it is straightforward to calculate the best-fitting error given a particular target pattern, we wish to say something more general about the error associated with the entire space of patterns. Because of the simplicity of the abstract space we have defined, we have enough information to generalize our understanding of the error. For each harmonic degree, we will compute normalized errors for each of the generating basis functions (i.e., for all the harmonic orders associated with the degree), and will also compute upper and lower bounds on the error given *any* unity norm target pattern in the generated subspace.

Let A be the data matrix, let B be an $M \times D$ matrix, the columns of which are responses corresponding to spherical harmonics of a given degree. Let $X^\star$ be an $N \times D$ matrix, the columns of which are the optimal loudspeaker signal gains associated with the D target signals. Let λ be a vector consisting of D complex control weights. Because of the linearity of the normal equations, we have that for any λ,

$$\min_x \|Ax - B\lambda\| = \|AX^\star\lambda - B\lambda\| \tag{2.7}$$

Since the columns of B are spherical harmonics, they are orthonormal, and the linear combination $B\lambda$ has norm

$$\|B\lambda\| = \|\lambda\|. \tag{2.8}$$

Given a subspace spanned by spherical harmonics, the largest approximation error associated with unity gain patterns in the subspace is given by

$$\begin{aligned} &\max_{\|\lambda\|=1} \|AX^\star\lambda - B\lambda\| \\ =&\max_{\|\lambda\|=1} \|(AX^\star - B)\lambda\| \\ =&\sigma_{\max}(AX^\star - B), \end{aligned} \tag{2.9}$$

where $\sigma_{\max}(AX^\star - B)$ is the maximum singular value of the matrix $AX^\star - B$. Similarly, the smallest approximation error associated with unity gain patterns in the subspace is given

by

$$\begin{aligned}
&\min_{\|\lambda\|=1} \|AX^\star\lambda - B\lambda\| \\
=&\min_{\|\lambda\|=1} \|(AX^\star - B)\lambda\| \\
=&\sigma_{\min}(AX^\star - B),
\end{aligned} \tag{2.10}$$

where $\sigma_{\min}(AX^\star - B)$ is the minimum singular value.

Geometrically, the set of individual loudspeaker functions generate a linear subspace of achievable responses. Given a particular desired pattern, the best approximation is its projection onto this subspace. Equations (2.9) and (2.10) are related to the *angles* between subspaces generated by the columns of A and B. Intuitively, as the angles between the subspaces increases, the normalized error increases accordingly. If one regarded the error in the least-squares optimization problem as a distance between a point and a subspace, the error we are considering is a generalization. It's the cosine of the principle angle between subspaces.

We will now show that the errors defined in (2.9) and (2.10) do correspond to a geometric angle. Let $Q_A = [Q_{A1}, Q_{A2}]$ and R_A be the QR decomposition of matrix A. Then Q is orthonormal with the columns of Q_{A1} spanning the range of A and the columns of Q_{A2} spanning the nullspace $A^\perp$ of A. See [GL96], page 228 for properties of the QR decomposition and also see [TI97]. As treated on pages 239–240 of [GL96], the error associated with the general full rank least-squares problem $\min \|Ax - b\|$ in terms of the QR decomposition of A is

$$\text{error} = \|Q_{A2}b\|. \tag{2.11}$$

Applying this to problems (2.9) and (2.10), yields

$$\begin{aligned}
&\max_{\|\lambda\|=1} \|AX^\star\lambda - B\lambda\| \\
=&\max_{\|\lambda\|=1} \|Q_{A2}B\lambda\| \\
=&\sigma_{\max}(Q_{A2}B) \\
\stackrel{\text{def}}{=}&\cos(\theta_{\max}),
\end{aligned} \tag{2.12}$$

where the last equality follows essentially by the definition of the principle angles between the subspaces spanned by Q_{A2} and (orthonormal) matrix B, i.e., the nullspace of A and the range of B. Similarly,

$$\begin{aligned} &\min_{\|\lambda\|=1} \|AX^{\star}\lambda - B\lambda\| \\ &= \cos(\theta_{\min}), \end{aligned} \tag{2.13}$$

Given a subspace, in particular, a constant degree subspace, upper and lower bounds on the associated least-squares error can be calculated by solving the singular value decomposition problems described above. This is useful because, often it is the case that control operations can be identified with particular subspaces of patterns. The most clear example of this is rigid rotation which, due to the properties of the spherical harmonics, is associated with fixed-degree subspaces. Other examples are control strategies involving electronic fading among different sets of patterns. For more applications involving angles between linear subspaces see pages 405-410 of [BV04].

2.2.6 Results

In what follows we will compute approximation errors associated with spherical arrays. The characterization of the arrays in terms of these errors is dependent on the geometry and the individual loudspeaker patterns. The characterizations could be used to assist in the design of an array, or provide guidance when using an existing array. The analysis tells us which patterns can be synthesized accurately and which cannot.

The data in the A matrices is generated by simulating a realistic measurement scenario. Constructing synthetic A matrices allows for a controlled comparison among arrays having differing geometric layout.

2.2.7 Uniform Error Example

It's necessary to provide an example to illustrate the distinction between the uniform error as discussed previously, and error associated with specific target patterns. The example is

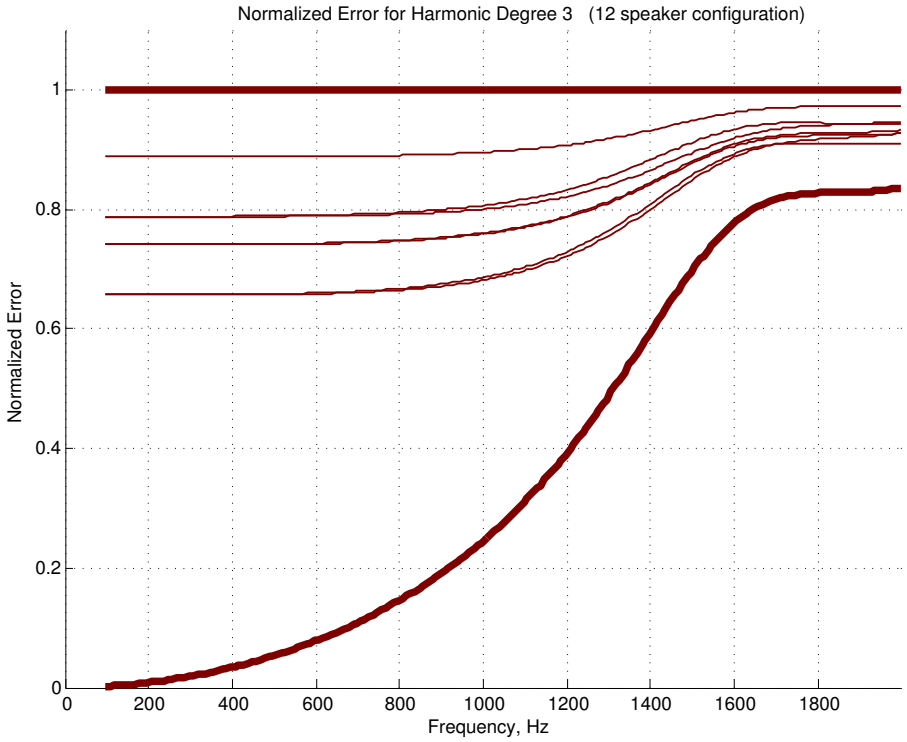


Figure 2.4: **Error associated with the degree 3 subspace.** Upper and lower bounds are plotted (bold) along with errors for the seven degree 3 spherical harmonics.

for a dodecahedral array of diameter approximately 37 cm. The least-squares problem is solved across frequency for each of the seven degree 3 spherical harmonic target patterns, and the error is plotted along with the upper and lower bounds, calculated as described. This example shows that the gap between the bounds can be large. By looking only at the error associated with the spherical harmonic targets, it's not obvious that the there is a wide error spread across the subspace. The upper bound is nearly 1, indicating that there exists a particular pattern, that is a unity-norm combination of the seven basis functions, that is extremely difficult for the array to synthesize. Similarly, there is a pattern that is well-synthesizable, at least at low frequencies. The error associated with spherical harmonic basis patterns lies in between. We can determine the extreme patterns by looking at the singular vectors in the singular value decomposition used to calculate the uniform errors. Intuitively, the existence of extreme patterns is caused by the fact that the array has a

particular fixed geometry and orientation. Certain pattern orientations are more or less "aligned" with the geometry of the array. Of course, the situation is subtle, and involves not only the geometry of the array, but the individual loudspeaker patterns as well. The physical system is not perfectly spherically symmetric; it is dodecahedrally symmetric.

2.2.8 Uniform Error Characterizations

To compare arrays of differing geometries, plots of the error bounds as functions of both frequency and spherical harmonic degree is natural. As the spherical harmonic degree increases, the complexity of the patterns increases. We will compare the uniform error bounds for four spherical arrays. The four arrays are i) a cubic array (6 elements), ii) a dodecahedral array (12 elements), iii) an icosahedral array (20 elements), and iv) a sixty element array (like the icosahedral array, but with 3 elements per triangular face). The comparisons are made holding the diameters of the arrays fixed. It should be noted that scaling the dimensions of an array has a very predictable effect on the array's frequency/harmonic degree characteristic chart. It simply rescales the frequency axis. The configurations for the four arrays are shown in Table 2.2.8.

Configuration	Number of Drivers	Driver Spacing	Diameter
Cubic	6	18 cm	26 cm
Dodecahedral	12	14 cm	26 cm
Icosahedral	20	9.3 cm	26 cm
Sixty-Sided	60	4.6 cm	26 cm

Table 2.1: Array configurations.

2.2.9 Discussion

Investigating the plots reveals that as the number of drivers in the array increases, more complex patterns can be reproduced. This makes intuitive sense because an array with a large number of drivers has more degrees of freedom for control. The charts allow us to quickly assess the radiation pattern synthesis capabilities of a given array or proposed array

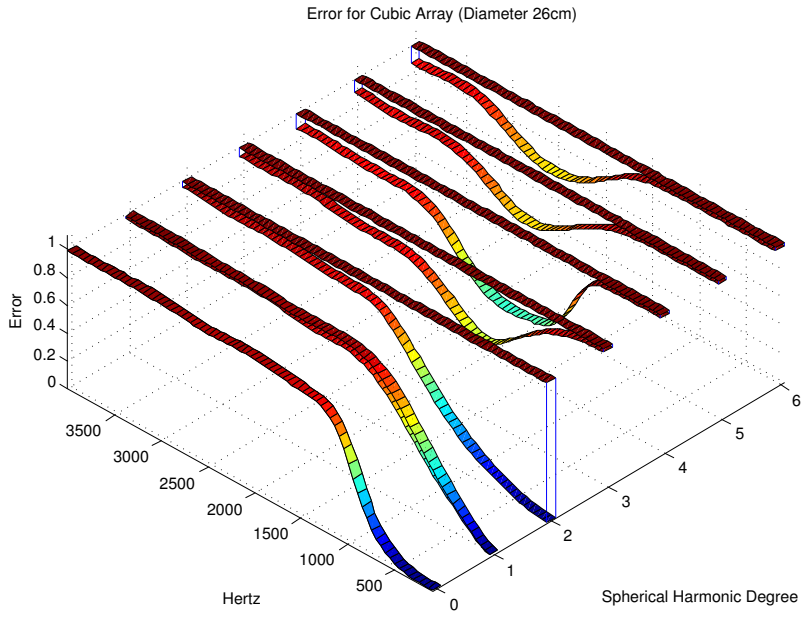


Figure 2.5: **Error bounds for cubic array.**

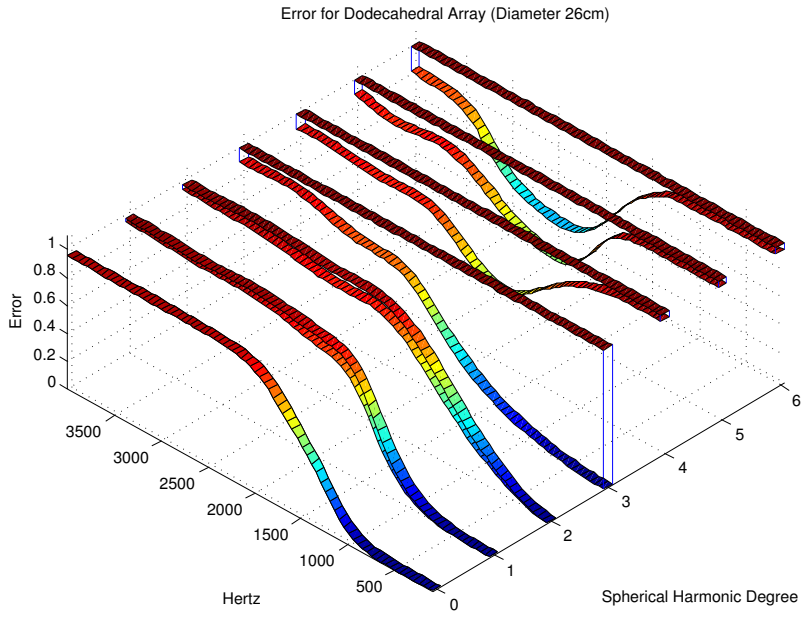


Figure 2.6: **Error bounds for dodecahedral array.**

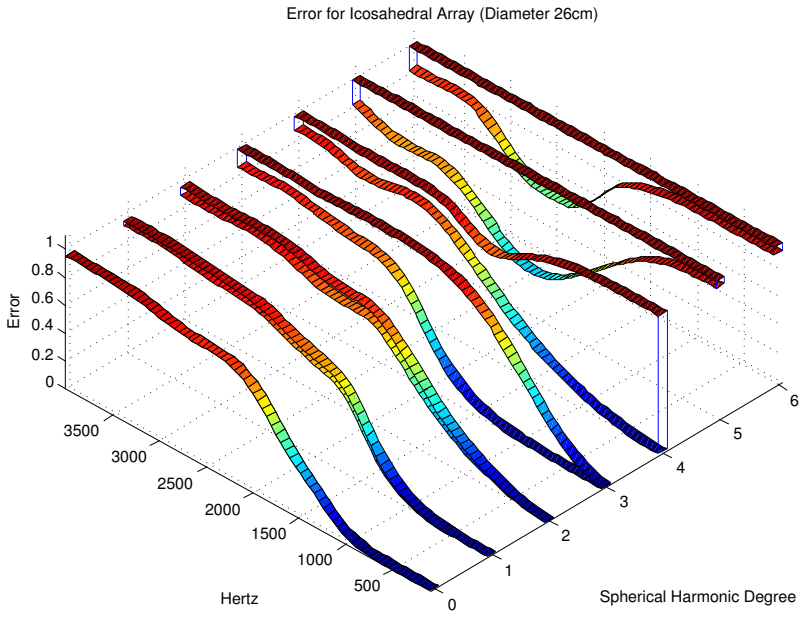


Figure 2.7: **Error bounds for icosahedral array.**

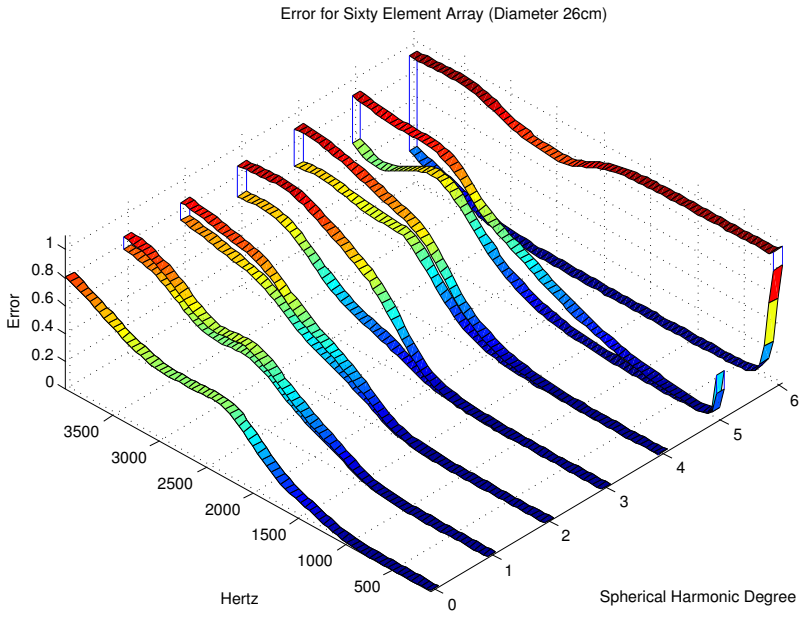


Figure 2.8: **Error bounds for sixty element array.**

design. They also guide us as to which patterns are electronically controllable, and which are not.

2.2.10 Summary

The optimization problem employed to help generate the characterization charts is convex; it's a least-squares problem in infinite dimensions. Viewing the radiation patterns as elements of an abstract vector space allowed us to draw specific conclusions very quickly and elegantly about the behavior of these arrays. Since the goal of this work was to provide practical answers as part of a larger research project, the least-squares optimization problem itself plays an understated role as a simple mathematical tool.

2.3 Design of Linear-Phase Equalizer

In the previous section, we used least-squares, arguably the simplest of all nontrivial optimization problems to answer questions about the radiation patterns produced by loudspeaker arrays. In this section we will discuss an application of convex optimization where the optimization problem itself takes a more central role. We will use it to design a set of linear-phase filters that will be used to split a signal into minimally overlapping bands suitable for signal equalization or frequency-domain limiting.

Limiting the power of a signal that's driving a loudspeaker is critical to protect the speaker from damage. The simplest type of limiting is performed by clipping the input waveform to a maximum amplitude. When engaged, this form of limiting produces audible and generally unpleasant sounding artifacts. A method that results in less noticeable artifacts is that of frequency domain limiting. In this case, we split the signal into multiple bands, limiting the signal in each band independently. It works particularly well if we expect the signals to have peaks in the frequency domain, which is often the case with sound signals. See Figure 2.9 for a diagram of this type of system.

We would like to design the filters so that with no limiting, i.e., when the signal is small enough, the resulting system passes the signal with very little alteration. Additionally, we

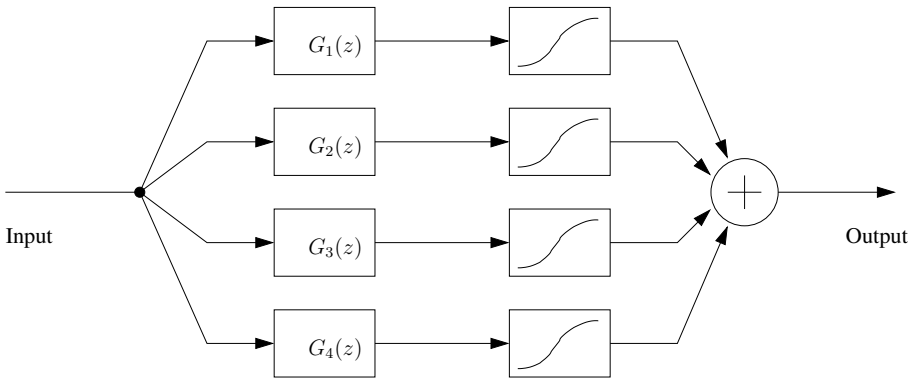


Figure 2.9: **Limiter**

desire the filters to be linear-phase and such that they split the input signal into minimally overlapping bands. A final aim is to make the design method flexible. We would like the procedure to be general enough to allow for arbitrary specification of the number of bands, the frequency cutoffs, and the orders of the filters. We will show that we are able to form a convex model to achieve these goals.

2.3.1 Model

We want the filters to be such that the system is always strictly linear-phase, even if the limiters are engaged. The most straightforward way of ensuring this is to require that the filters, $G_1(z), G_2(z), ..., G_n(z)$ all be symmetric. The frequency response of such a filter g centered at the origin is

$$G(j\omega) = \begin{cases} g_0 + 2\sum_{i=1}^{(n-1)/2} g_i \cos(i\omega) & \text{if } n \text{ is odd} \\ 2\sum_{i=1}^{n/2} g_i \cos\left((i-\frac{1}{2})\omega\right) & \text{if } n \text{ is even} \end{cases}$$

Note that the frequency responses are real since they are symmetric (around 0). Ultimately the filters we will implement will necessarily be causal. Delaying the designed symmetric filters to be causal has a benign effect on their frequency responses; the delay generates a

linear phase factor. Define the following parameters for the design:

m	: the number of bands, i.e., the number of filters
$\omega_0^c, \omega_1^c, ..., \omega_m^c$	: the cutoff frequencies
$n_1, n_2, ..., n_m$	: the filter orders
l	: the number of frequency response discretization points
$\omega_1, \omega_2, ..., \omega_l$	: samples in frequency
$\lambda \in [0, 1]$	: a constant controlling trade-off between frequency selectivity of filters, and reconstruction accuracy in the absence of limiting.

The way we will enforce a nearly distortionless response in the absence of limiting is by discretizing the frequency axis into l samples, and forming a convex constraint on the deviation between the total system response and the ideal distortionless response (unity):

$$\delta \geq \sum_{i=1}^{l} \left(\sum_{k=1}^{m} G_k(j\omega_i) - 1 \right)^2$$

We will encourage frequency selectivity of the filters forming the following constraints,

$$\begin{aligned} G_1\left(j\frac{\omega_1^c + \omega_0^c}{2}\right) &= 1 \\ G_2\left(j\frac{\omega_2^c + \omega_1^c}{2}\right) &= 1 \\ &\vdots \\ G_m\left(j\frac{\omega_m^c + \omega_{m-1}^c}{2}\right) &= 1, \end{aligned}$$

and by minimizing the out-of-band energy for each filter:

$$e_k \geq \sum_{\omega_i \notin [\omega_{k-1}^c, \omega_k^c]} (G_k(j\omega_i))^2 \qquad \forall k = 1, 2, ..., m.$$

The final convex optimization problem is formed as follows:

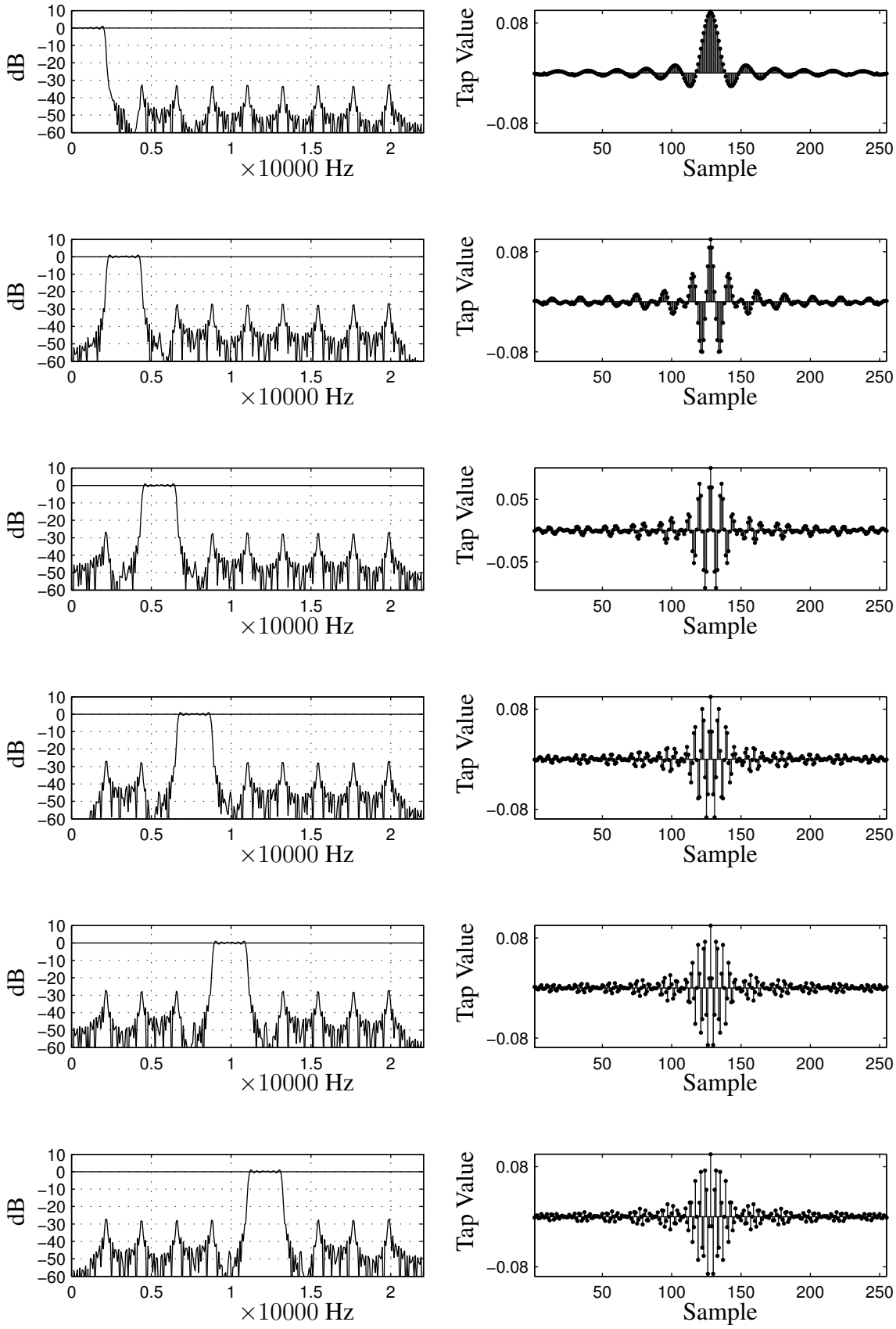
$$\begin{aligned} \min \quad & \lambda\delta + (1-\lambda)\sum_{k=1}^{m} e_k \\ \text{s.t.} \quad & G_k\left(j\frac{\omega_k^c + \omega_{k-1}^c}{2}\right) = 1 \qquad \forall \quad k = 1, 2, ..., m \\ & e_k \geq \sum_{\omega_i \notin [\omega_{k-1}^c, \omega_k^c]} |G_k(j\omega_i)|^2 \qquad \forall \quad k = 1, 2, ..., m \\ & \delta \geq \sum_{i=1}^{l} \left(\sum_{k=1}^{m} G_k(j\omega_i) - 1\right)^2 . \end{aligned} \tag{2.14}$$

The optimization problem above can be solved easily using off-the-shelf convex optimization routines. We will now present some designs using our approach.

2.3.2 Ten-Band Equalizer

The first example is a linearly-spaced, ten-band filterbank that could be used for graphic equalization, limiting, or signal monitoring. We split the frequency spectrum into 10 linearly spaced bands, and set the number of taps for each filter to be 256. We will adjust λ so that the total frequency response deviates from unity by no more than 1dB. Thus we will solve the problem multiple times, converging on an appropriate λ efficiently by bisection. The plots show significant frequency selectivity of the filters, and a total response that is nearly flat (and linear-phase). In our setup, we are minimizing the out-of-band energy of the filters, but we could just as well minimize the out-of-band peak power; it too would also result in a convex problem. In this application however, minimizing the energy may be more appropriate.

We see that allocating 256 taps per filter results in a system that achieves good frequency selectivity, and a total response that deviates from unity by no more than 1dB. In the next section, we will design a filter network for use with a particular loudspeaker system. In that system, the frequency cutoffs are determined by the characteristics of the loudspeakers.


dB
10
0
−10
−20
−30
−40
−50
−60
0
0.5
1
1.5
2
×10000 Hz
Tap Value
0.08
−0.08
0.05
−0.05
50
100
150
200
250
Sample

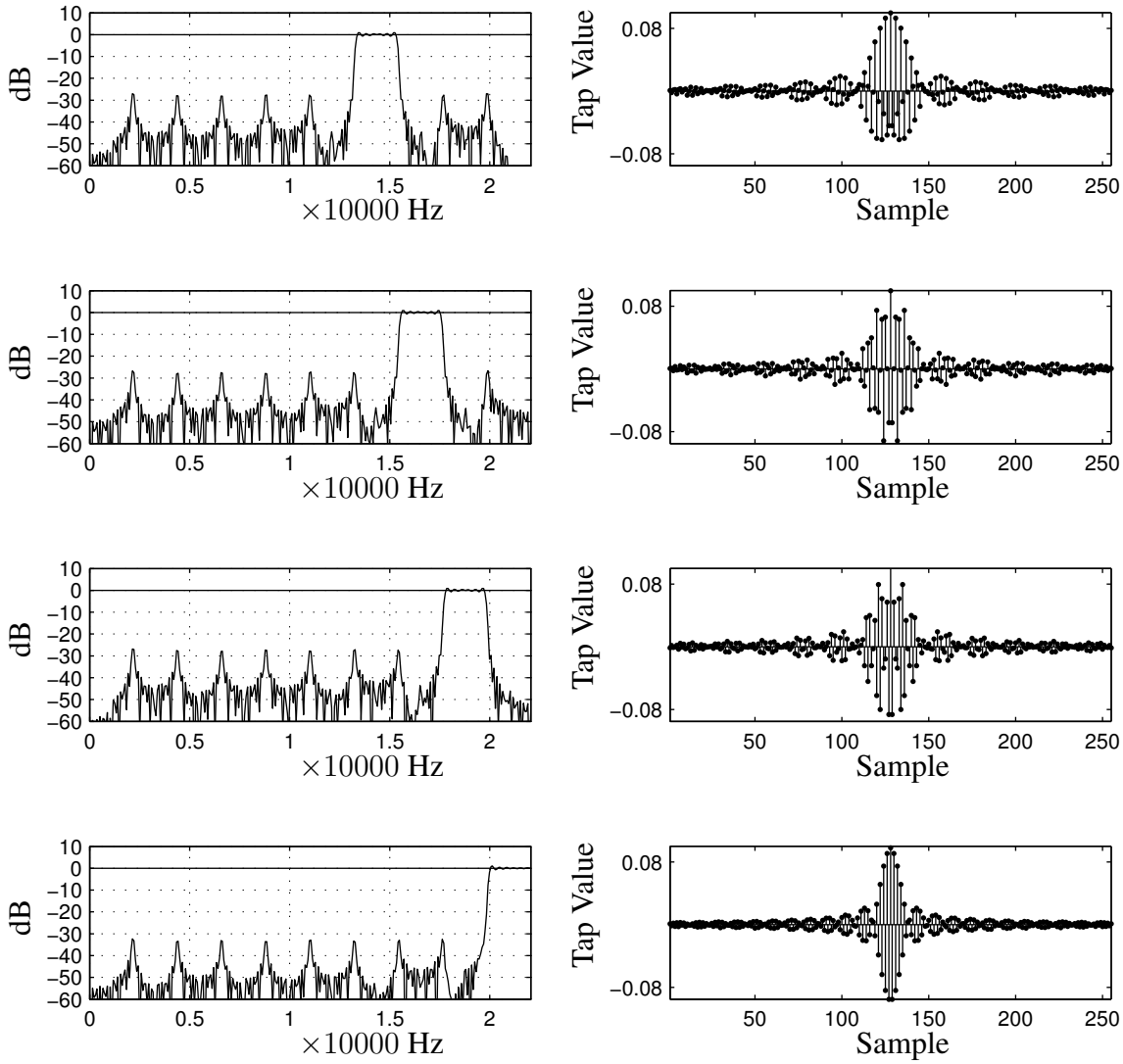


Figure 2.10: **Frequency responses and filters for the ten-band equalizer.**

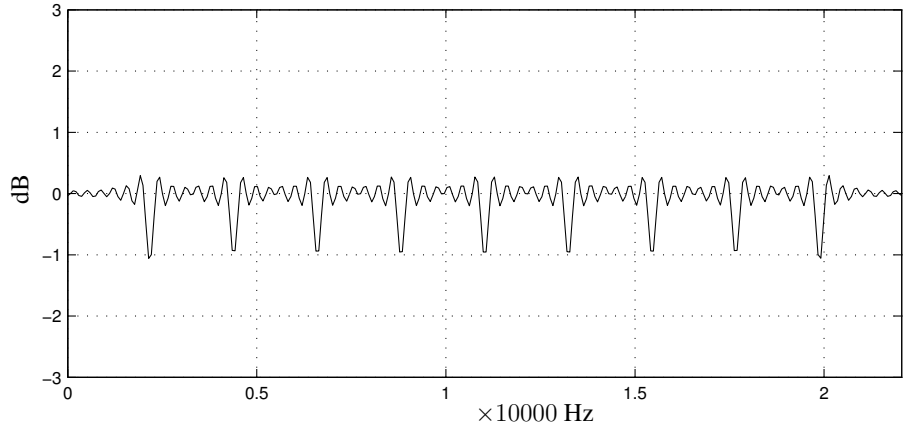


Figure 2.11: **Total response of the ten-band equalizer.** The response is nearly flat.

2.3.3 Three-Band Equalizer

The previous example served to illustrate the type of designs we're seeking. In this section, we solve an important practical problem related to loudspeaker protection. The loudspeakers we wish to protect are tweeters that are part of a large spherical array. Being tweeters, there is a frequency ω_{high} above which the speakers are very efficient and are able to handle a lot of power. If the tweeters were part of a conventional three-way loudspeaker system, we would want to design a cross-over network with cross-over at ω_{high} with the intention of directing lower frequencies to another loudspeaker. Since our speakers are part of a large array, we have a different intention. We want to extend the frequency range of the all-tweeter system by allowing the individual speakers to work together (in phase) at lower, midrange frequencies.

We wish to design a set of filters that split a loudspeaker audio signal into three bands. The lowest frequency channel will be diverted away from the loudspeaker and sent to a subwoofer. The remaining two channels will feed the tweeters, but will be subjected to different limiting thresholds. The tweeters can handle large signals in the highest band, but we must carefully limit the midrange signal. In this section we are not concerned with the design of the limiting function or threshold, but are focusing on the three-band filter network.

The cutoff frequencies are derived from inspecting plots of measured frequency responses of the tweeters. They are

$$\omega_0^c = 0 \quad \text{Hz}$$

$$\omega_1^c = 400 \quad \text{Hz}$$

$$\omega_2^c = 2000 \quad \text{Hz}$$

$$\omega_3^c = 22050 \quad \text{Hz}$$

We specify the filter lengths to be 300 taps. The figures show the frequency responses and filters for the three-band design. We see again that we can achieve good frequency separation, and a nearly flat response. The fact that we have a flexible way to quickly

design filters allows us to audition several different designs for differing parameters. If, for example, we desire even more frequency selectivity, or a flatter response, we can increase the order of the filters accordingly. Conversely, we can see what the optimal performance is for fixed filter orders. We have not just designed a set of filters; we have designed a design procedure.

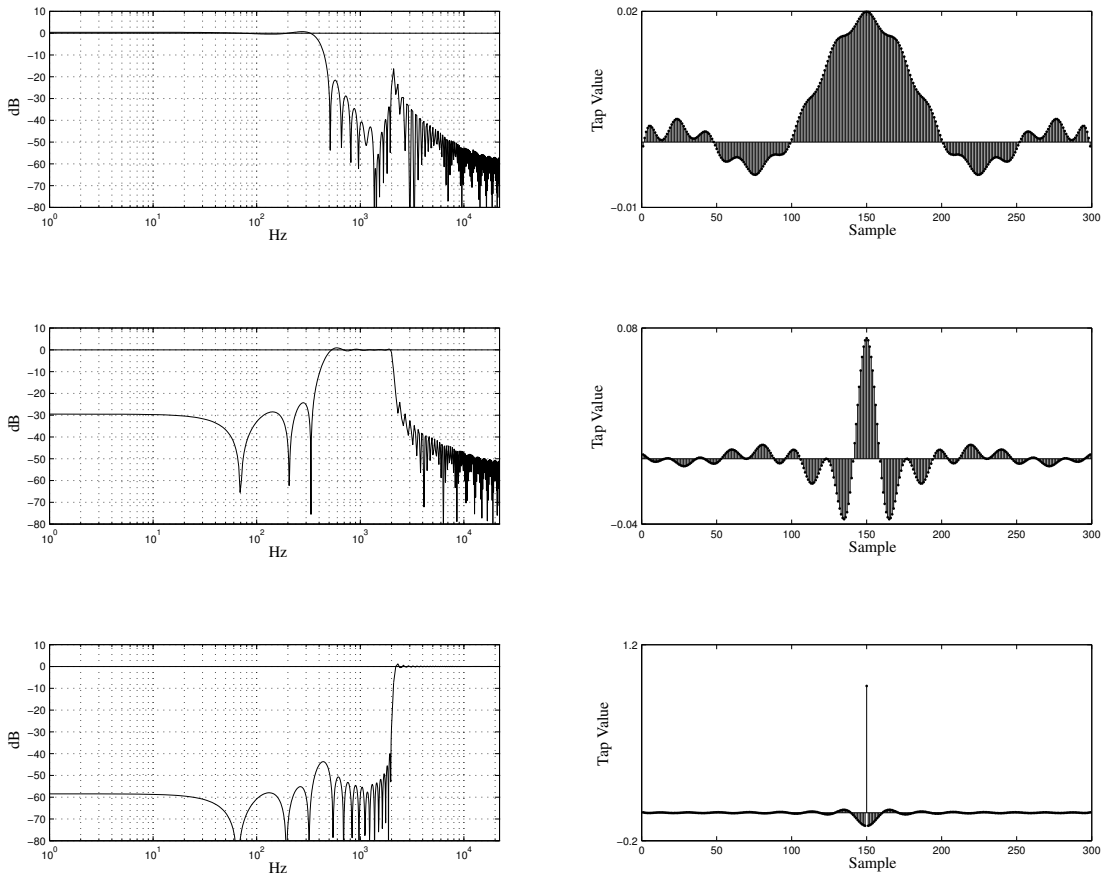


Figure 2.12: **Frequency responses and filters for the three-band equalizer for use with an array of tweeters.**

2.4 Conclusion

In this chapter, we solved two practical problems using convex optimization as a primary tool. The chapter serves to show that very practical problems can admit convex formulations, which allows us to solve them with ease. In the next chapter, we will introduce a nonconvex problem that we will focus on for the remainder of the dissertation.

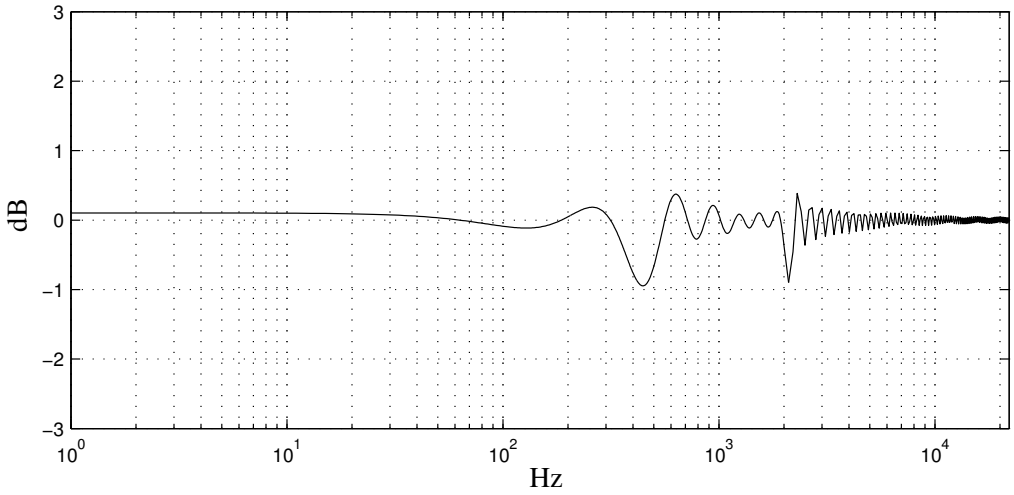


Figure 2.13: **Total Response of the three-band equalizer.** The response is nearly flat.

2.5 Acknowledgment

The authors wish to thank Meyer Sound Laboratories for use of their anechoic chamber, their collaboration with respect to finding solutions to the technical problems described in this chapter, and their continued financial support of this work. Support for this project was also provided by a UC Discovery Grant in Digital Media from the University of California's Industry-University Cooperative Research Program (IUCRP). We also want to specifically thank John Meyer, Perrin Meyer, Laurent El-Ghaoui, Adrian Freed, Rimas Avizienis, Ed Berdahl, David Wessel, and Lieven Vandenberghe.

Chapter 3

Introduction to Magnitude Least-Squares

3.1 Introduction

This chapter begins our discussion of the magnitude least-squares (MLS) problem. The problem is motivated by problems in magnitude filter design, but possesses application in other areas as well. It is a nonconvex problem that we will convexly approximate using semidefinite duality. We will begin this study by stating the abstract problem, and then describing the application areas for such a problem. The purpose of the chapter is to motivate our interest in the problem in terms of its applicability in engineering problems. It should be noted, however, that the problem is interesting in the abstract. The solution method we will employ will carry with it a provably good solution accuracy estimate, improving our perception of the method as being better than a simple heuristic.

3.2 Problem Statement

The magnitude least-squares problem is formulated as follows. Given a nonnegative real vector b, and a complex-valued matrix $A \in \mathbb{C}^{m\times n}$, we wish to find a vector $x \in \mathbb{C}^n$ that solves

$$\min_x \quad \| \, |Ax| - b \, \|_2^2 \equiv \sum_{i=1}^{m} (|A_i x| - b_i)^2, \tag{MLS}$$

where each A_i is a row of the matrix A. The problem can be understood to be a nonconvex, nonlinear, nondifferentiable variation of the standard linear least-squares problem,

$$\min_x \quad \| Ax - b \|_2^2 \equiv \sum_{i=1}^{m} (A_i x - b_i)^2. \tag{LS}$$

We interpret the goal to be one of fitting. In the applications we consider, the optimal fit is used to meet design ideals, to estimate parameters from noisy measurements, or more abstractly, to approximate functions. In all cases, it is the magnitude of the image of the complex matrix A that is the important quantity. In the next sections, we will identify the applications that benefit from a solution to this problem.

3.3 Applications for MLS

In what follows, we will not discuss the problem modeling and formulation steps required to cast the application problems in terms of the magnitude least-squares problem. Rather, we wish to use this chapter as a motivation and a preview for a more in-depth study of the problem. The following problems can be solved using our method for solving the MLS problem.

3.3.1 One-Dimensional Magnitude Filter Design

Though the design of linear filters for use in signal processing often involves fitting both magnitude and phase responses to design ideals, there are many cases where it is the magnitude response of the filter that is much more important. This is the case, for example, when we wish to attenuate the magnitude spectra of a signal over a frequency range as sharply as possible. We would prefer to trade phase distortion for better magnitude characteristics.

The design of a one-dimensional filter with desirable magnitude is a very well-studied problem in signal processing. Elegant methods exist for solving many variants of the problem. The existence of so many methods is due to the great applicability of the problem, motivating many people to study it. But the success of the methods is due in part to the existence of a mathematical theorem that is specific to one-dimensional polynomials: the

spectral factorization theorem. Though not explicitly employed in every method, its existence makes the problem precisely expressible as a convex optimization problem, and thus the one-dimensional magnitude fitting problem is solved with ease.

Even though the problem is easy, and can be solved without the approximation technique we will soon discuss, it represents the simplest in the class of problems our method can handle. It is the core problem. Our goal is to study a method that can solve a wider class of problem. One simple one-dimensional variant that does not benefit from the spectral factorization theorem is the design of an FIR filter with nonuniformly spaced filter taps. To solve this problem, we will cast it as a magnitude least-squares problem, and solve using convex approximation.

3.3.2 Multidimensional Magnitude Filter Design

Multidimensional magnitude filter design also cannot be expressed perfectly in convex terms. Like one-dimensional filter design, this problem has been studied at length over the last four decades. But many of the methods that work for one-dimensional filter design, particularly the ones that directly rely on spectral factorization, break down when applied to multidimensional filter design. Our magnitude least-squares formulation is applicable to these problems, in any finite number of dimensions. Additionally, it is applicable in the design of multidimensional nonuniformly spaced tap filters.

3.3.3 Static Magnitude Beamforming with Arbitrary Element Layout

Related to multidimensional magnitude filter design is magnitude beamforming. In beamforming, we have an device consisting of an array of elements. The elements could be sensors, for example, an antenna or microphone array, or they could be sources, like a clustered array of loudspeakers. Arrays have response characteristics that, in their most general form, are functions of frequency and three other dimensions. The system can be interpreted as a four-dimensional (spatial) filter. We can apply our method in a straightforward way to these problems, and the formulation can handle the even more important issue of arbitrary

element layout. The calculation of optimal static spatial filters can aid in the design of the layout of the array elements. An array can be used to enhance reception of signals coming from known directions and in known frequency bands. If it is an array of sources, then MLS can be used to optimally synthesize magnitude radiation patterns.

3.3.4 Approximate Factorization of Polynomials

An indirect consequence of the success of this method is that we can view it as a form of approximate spectral factorization. This is because the method surmounts the factorization obstacle present in higher dimensions. Ironically, of all the one-dimensional filter design methods, our method is most closely related to the ones that *do* rely on spectral factorization. The reason is clear if one considers that these are the methods that cast the one-dimensional filter design problem explicitly in convex terms.

We can use approximate spectral factorization, or more generally, approximate polynomial factorization as a direct substitute for the spectral factorization theorem in situations where a one-dimensional method breaks down in many dimensions. It results only in an approximate answer, of course, but often the solution obtained is of high quality. We can use this technique to design multidimensional filters in the minimax sense, and compare the solutions to others, obtained using more conventional means.

3.4 Discussion

The problem we will study for the remainder of the thesis is of interest to us for two general reasons. The first is that the problem is not convex but can be approximated and solved using convex optimization. The fact that we can prove a lower bound on the solution accuracy obtained using the scheme distinguishes the method from most other heuristic methods for solving nonconvex problems. The second reason we choose to study the problem is that it possesses several important practical applications, some of which we've mentioned above. Our hope is that understanding as much as possible about this one problem with respect to its applicability, solution accuracy, and relation to other problems will provide information

about convex approximation that will be meaningful in other areas not directly related to the magnitude least-squares problem.

Chapter 4

One Dimensional Magnitude Filter Design

4.1 Introduction

If we are interested in designing a one-dimensional FIR filter with prescribed magnitude response, we can do it using one of many different classical design techniques. They all work quite well. Such methods include window design, Parks-McClellan (see [OSB98] chapter 7), and methods based on spectral factorization (see [WBV97] [DLS02] [Alk03] [AV02]). We will focus on the design methods that are based on spectral factorization.

These methods are of interest to us for two reasons, the first being that they elegantly solve a special case of the magnitude least-squares (MLS) problem. The second is that the methods based on spectral factorization serve as fascinating examples of nonconvex problems that are reformulated in a very nontrivial way so as to be solved as convex optimization problems. They dramatically emphasize the point that the existence of a nonconvex formulation of a problem does not imply that the problem cannot be solved exactly using convex optimization. Hence it is often useful to rethink one's formulations and spend effort investigating whether a seemingly nonconvex problem could be reformulated in a convex way.

4.1.1 Frequency Response of an FIR Filter

The frequency response of an FIR filter h can be calculated as

$$H(e^{j\omega}) = \sum_{i=0}^{m-1} h_i e^{-j\omega i}. \tag{4.1}$$

It is almost always the case that the understood purpose of the filter is to change the frequency spectrum of the input signal, and design criteria are usually specified in terms of the frequency response of the filter. Sometimes the phase response is equally as important as the magnitude response. For example, a good audio equalizer might be designed such that the filters have linear phase so that time-domain peaks and other perceptible features of the time-domain signal are more likely to be preserved. In many cases however, the phase response of the filter is less important than the magnitude response. Dropping conditions on phase will always yield a more desirable magnitude response.

We see from (4.1) that the (complex) frequency response of the filter is a linear function of the coefficients h_i, $i = 1, 2, ..., m-1$. Because of this, we can easily solve a variety of design problems where the complex frequency response $H(e^{j\omega})$ enters into the problem convexly (it is either convexly constrained, or a convex function of it is a term in the objective function). The simplest of these problems would be where we specify a complex target response b, and try to find the filter h that minimizes the integral squared error between the filter's frequency response and the target response. This is formulated as follows.

Sample the frequency space at $n >> 0$ points. Call those frequencies ω_i, $i = 1, 2, ..., n$. Form a matrix A with each entry A_{ik} being the frequency response due to filter tap k evaluated at frequency ω_i, so that $A_{ik} = e^{-j\omega_i k}$. Assuming we desire a filter with real-valued taps, we can solve the problem as a linearly constrained complex least-squares problem:

$$\begin{aligned} \min_{h} \quad & \|Ah - b\|_2 \\ \text{s.t.} \quad & \mathrm{Im}\{h_i\} = 0, \qquad i = 1, 2, ..., m. \end{aligned} \tag{4.2}$$

In this formulation, we have sampled frequency space, effectively approximating the integral by a sum. In practice, this is not a serious matter because problem (4.2) can be solved

very efficiently for extremely large values of n (it even has a closed-form, algebraic solution). In fact, if integrals between the desired frequency response and the columns of A are easily calculated, then the continuous problem can be solved exactly, also as a finite linearly constrained least-squares problem.

Problem (4.2) is not as useful in practice as it may seem at first glance. The real and imaginary parts of the frequency response are far less meaningful than magnitude and phase, but the magnitude and phase are not linear functions of the decision variables. The operators $\mathrm{Re}\{\cdot\}$ and $\mathrm{Im}\{\cdot\}$ are linear whereas magnitude and phase are not.

Let's consider an example where we design a filter using (4.2). We will attempt to design a 10-tap FIR filter that best matches in the squared error sense, a high-pass, linear phase target response. Though it is usual to inspect the frequency response of a filter by creating magnitude and phase plots, we will instead create a three-dimensional plot to emphasize the fact that the least-squares fit occurs in the complex domain. The axes of the plot are normalized frequency (radians/π), real part of the response, and imaginary part of the response. On such a plot, an ideal high-pass filter with linear phase generates a line with zero magnitude at low frequencies, and a helix of magnitude 1 at high frequencies. A plot such as this allows us to visualize geometrically the way the designed filter matches the ideal response. The least-squares fit in the complex domain can be understood to be the sum of the squared pointwise distances from the ideal helical-like curve to the achieved 10-tap filter response. Figure 4.1.1 shows these two curves with (blue) lines representing error.

Hence problem (4.2) finds a solution that weights (essentially) equally the filter's deviation from the ideal response both in magnitude and phase. In the next example we will see how we can find a filter that is the solution to an optimization problem that disregards error in phase, in favor of a better fit in magnitude.

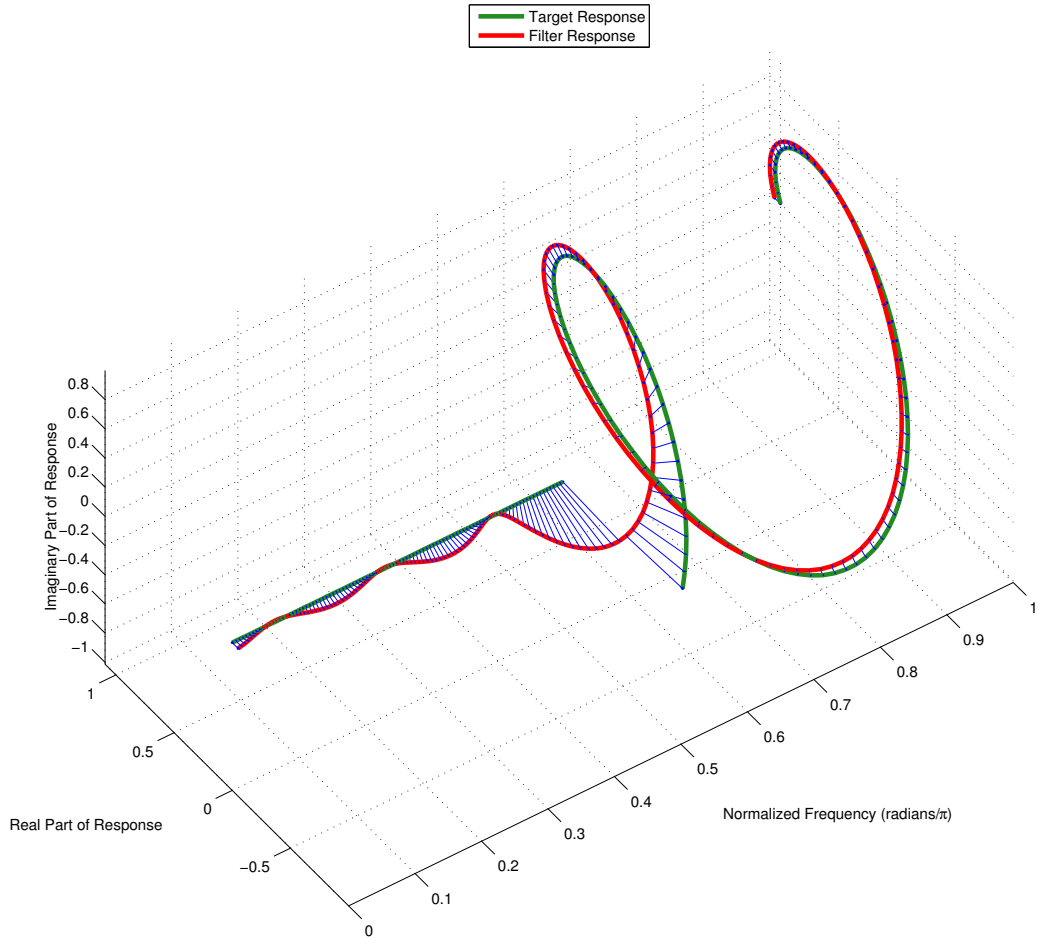


Figure 4.1: **Least-squares fit in complex domain.**

4.1.2 Magnitude Fitting via Spectral Factorization

The problem we're interested in this section is more natural and arguably more useful than problem (4.2). We will first consider the basic problem, and later consider several useful extensions where we drop the least-squares fit in favor of piecewise spectral mask constraints as discussed in [WBV97]. The crucial, and fascinating, ingredient that allows us to solve these problems is the use of a one-dimensional spectral factorization theorem.

Problem (4.2) is convex, and therefore straightforward to solve, yielding solutions that are guaranteed to be globally optimal. But consider the one-dimensional magnitude (squared) least-squares filter design problem:

$$\min_{h\in\mathbb{R}^m} \sum_{i=1}^{n} \left(\left| H(e^{j\omega_i}) \right|^2 - b_i^2 \right)^2$$

$$\equiv \min_{h\in\mathbb{R}^m} \sum_{i=1}^{n} \left(\left| \sum_{k=0}^{m-1} h_k e^{-j\omega_i k} \right|^2 - b_i^2 \right)^2 \tag{4.3}$$

where b is assumed to consist of a vector of nonnegative real numbers. This error function represents deviation between the squared magnitude of an FIR filter and that of a nonnegative real target. The problem is not convex as stated in (4.3), but we will see that using spectral factorization, we can cast the problem as a convex optimization problem.

The reason that the above problem is not convex revolves around the fact that the decision variables h enter into the objective function as squared variables. We can see their relation to the filter's autocorrelation sequence by expanding (4.3):

$$\begin{aligned}
&\left| \sum_{k=0}^{m-1} h_k e^{-j\omega_i k} \right|^2 \\
&= \sum_{k=0}^{m-1} h_k e^{-j\omega_i k} \sum_{l=0}^{m-1} h_l e^{j\omega_i l} \\
&= \sum_{k=0}^{m-1} \sum_{l=0}^{m-1} h_k e^{-j\omega_i k} h_l e^{j\omega_i l} \\
&= \sum_{k=0}^{m-1} \sum_{l=0}^{m-1} h_k h_l e^{-j\omega_i (k-l)} \\
&= \sum_{t=-m+1}^{m-1} \sum_{l=0}^{m-1} h_l h_{l+t} e^{-j\omega_i t}, \qquad (4.4) \\
&\quad \text{where we define } \; h_{l+t} = 0 \; \text{ if } \; l+t \notin \{1, 2, ..., m-1\}
\end{aligned}$$

The sums

$$r_t \equiv \sum_{l=0}^{m-1} h_l h_{l+t} \tag{4.5}$$
$$h_{l+t} = 0 \quad \text{if} \quad l+t \notin \{1, 2, ..., m-1\}$$

are the autocorrelation coefficients of the filter h. We can express the objective function

(4.3) in terms of the autocorrelation coefficients and form a new optimization problem:

$$\min_{r} \sum_{i=1}^{n} \left(\sum_{t=-m+1}^{m-1} r_t e^{-j\omega_i t} - b_i^2 \right)^2 \tag{4.6}$$
$$\text{s.t.} \quad r \text{ is the autocorrelation sequence of } h.$$

The objective function is now a convex function of variables r, but what about the constraint that r be the autocorrelation sequence of a length m FIR filter? Initially this looks as though it would not be a convex constraint because of the quadratic relation between r and h, but it turns out that, because of the spectral factorization theorem, it is convex.

The spectral factorization theorem (see [WBV97]), states that there exists an $r \in R^{2m-1}$ such that it is the autocorrelation sequence of $h \in R^m$ if and only if the Fourier transform $R(e^{j\omega})$ of r is nonnegative for all $\omega \in [0, \pi]$ (note that $R(e^{j\omega})$ is always real because r is always symmetric). And if this is the case, there exist several efficient methods for calculating a filter h that does indeed possess the autocorrelation sequence r.

The constraint that $R(e^{j\omega}) \geq 0$ is readily seen to be convex by observing that $R(e^{j\omega})$ is a linear function of r. It consists of an infinite number of pointwise constraints, however. From a practical point of view, we could finely sample $[0, \pi]$ and reduce the semi-infinite constraint to a large finite set of linear nonnegativity constraints. It turns out, however, that the semi-infinite constraint can be represented *exactly* using a finite set of linear matrix inequalities. The reason for this originates in the theory of positive polynomials studied by Leopold Fejér, David Hilbert and many others. We will employ a theorem of this sort taken from [HN02] (their Theorem 2.1).

The theorem implies that a trigonometric polynomial r with real coefficients is nonnegative on $[0, \pi]$ if and only if there exists a positive semidefinite Hermitian matrix Y such that

$$r_k = \sum_{i-j=k} Y_{ij}, \quad k = 0, 1, ..., n \tag{4.7}$$
$$\text{where } Y_{ij} = 0 \text{ for } i, j \text{ outside their definition range.}$$

These are linear constraints on the diagonal sums of a positive semidefinite matrix Y. Our

final, exact convex formulation of (4.3) is the following:

$$
\begin{aligned}
\min_{r} \quad & \sum_{i=1}^{n} \left(\sum_{t=-m+1}^{m-1} r_t e^{-j\omega_i t} - b_i^2 \right)^2 \\
\text{s.t.} \quad & r_i = r_{-i}, \qquad i = 1, 2, ..., m-1 \\
& \begin{bmatrix} Y_{11} & Y_{12} & \cdots & Y_{1m} \\ Y_{21} & Y_{22} & \cdots & Y_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ Y_{m1} & Y_{m2} & \cdots & Y_{mm} \end{bmatrix} \succeq 0 \\
& r_0 = Y_{11} + Y_{22} + \ldots + Y_{mm} \\
& r_1 = Y_{12} + Y_{23} + \ldots + Y_{(m-1)m} \\
& \vdots \\
& r_{m-1} = Y_{1m}
\end{aligned}
\tag{4.8}
$$

We recover the filter coefficients h using spectral factorization.

4.1.3 High-Pass Filter Example

We can find a high-pass filter using (4.8) that outperforms (in terms of magnitude fit) the filter found by solving (4.2). Inspecting the situation in three dimensions like Figure 4.1.1, we see that the target function for the magnitude least-squares problem (4.8) is not a one-dimensional path like it is for the complex least-squares problem, but rather a cylinder representing points of equal magnitude (a target magnitude of 1 for high frequencies). The magnitude least-squares fit is shown in Figure 4.1.3.

Figure 4.1.3 shows the magnitude responses of filters found by complex least-squares and by magnitude least-squares. Of course we see that the response of the filter found by magnitude least-squares has a more desirable magnitude fit. The response of the filter found by complex least-squares has phase response that is closer to its linear phase target.

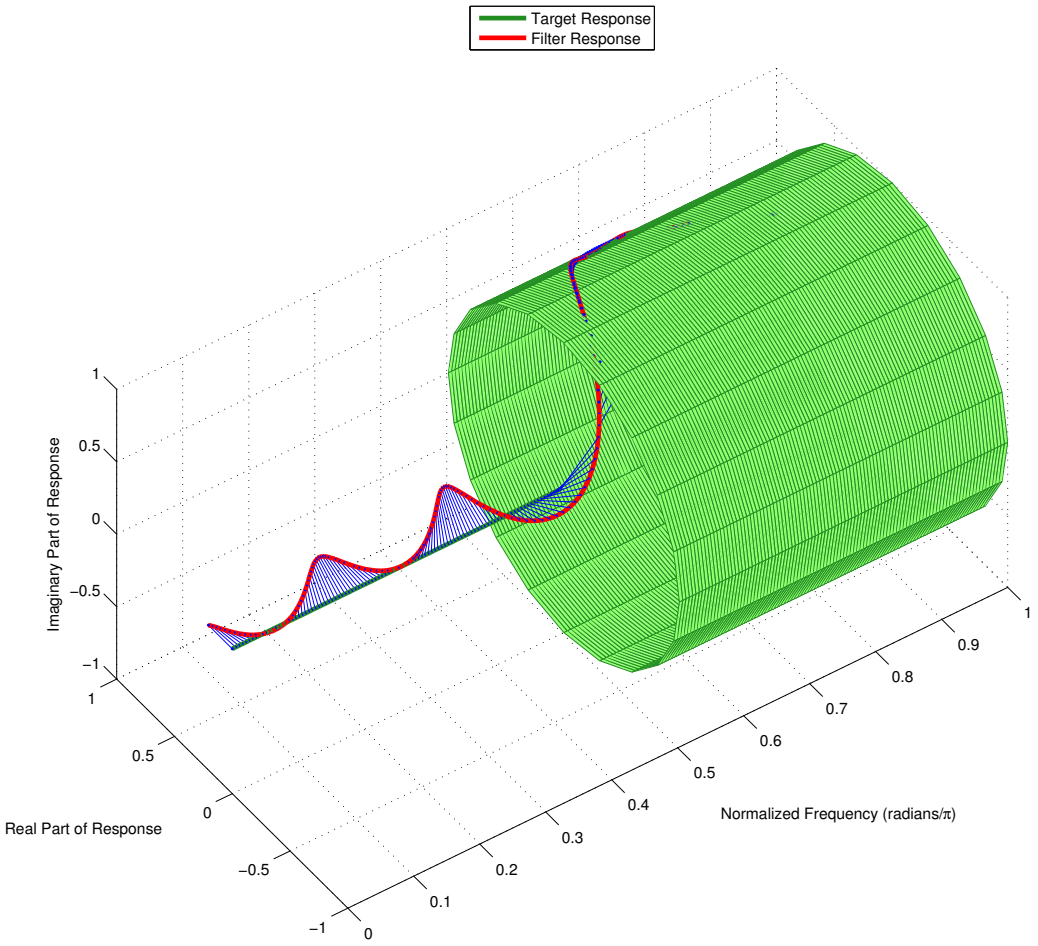


Figure 4.2: **Magnitude least-squares fit in complex domain.**

4.2 Discussion

The fact that we can reformulate convexly and solve problem (4.3) using semidefinite programming and the spectral factorization theorem may not seem too impressive considering the many other methods for one-dimensional filter design. But it is very interesting from a theoretical point of view. Since we are able to obtain an exact solution by solving a convex optimization problem, it means that we have exploited structure in the geometry of the original nonconvex problem.

The way we arrived at the convex formulation was to cast the objective function in terms of squared variables (the autocorrelation coefficients) instead of the filter coefficients. This technique can be generalized, and along with corresponding positivity results like the one

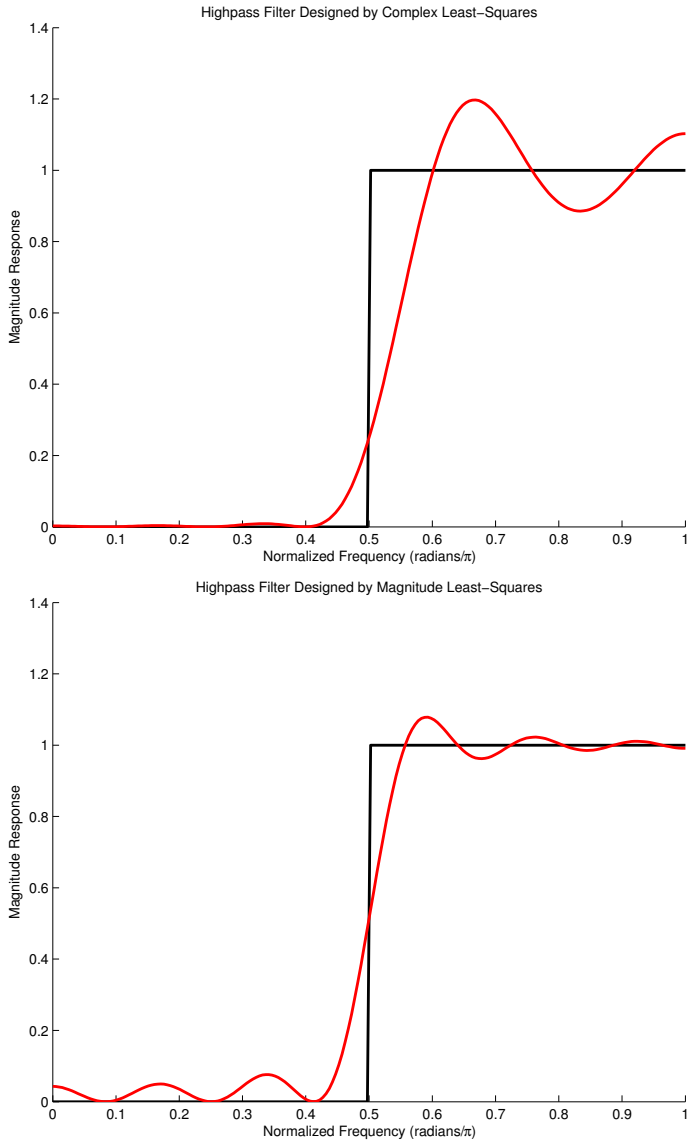


Figure 4.3: **Magnitude least-squares fit versus complex least-squares fit**

encountered in our spectral factorization theorem, is a topic of intense theoretical research. See for example [Par01] [PPP02] [Las01] [GW95]. The solution of nonconvex and/or difficult problems through clever convexification is a very exciting area in both mathematics and engineering. In the chapters that follow, we will look at the solution of the magnitude least-squares problem in many dimensions, as an extension of the technique described in this chapter.

Chapter 5

MLS Problem and Proposed Solution Methods

5.1 Introduction

In this chapter, we will look at the magnitude least-squares problem in the abstract, and consider several different ways of solving it. The exploration of the different methods for solving a problem is one of the most interesting and exciting parts of engineering and mathematics. If the problem were convex, then our success in solving it exactly and efficiently would be nearly guaranteed. We would, in that case, compare methods based on convenience either in formulation, ease of implementation, or computational resource requirements. Because the problem is not convex, we will discuss methods that are not guaranteed to find globally optimal solutions. We will argue that the methods we propose are very good ones, however. The solutions we obtain achieve objectives that are provably close to the globally minimal values. These methods are based on semidefinite relaxations.

We will discuss two variations on the magnitude least-squares problem, and derive methods for solving them. For each variation, we discuss a local method, and a semidefinite relaxation method. This gives us four distinct approaches to solving the general problem (actually two methods for each of two separate, but very similar problems). The best, most accurate method of solving the problem is to solve using the semidefinite relaxation with the intention of using the obtained point as a starting point for the local method. This approach guarantees us a locally optimal point that is provably close to the globally optimal

point in objective value. We will introduce and formulate the four methods, leaving the quality estimate proofs to Chapter 6.

This chapter provides a lot of information about the MLS problem, viewing it from several different angles. For example, when deriving a gradient based method, we are forced to inspect the differentiability of the objective function. Another method we will discuss will find local minima using a variable exchange method. This again gives us insight into the underlying geometry of the problem. The semidefinite relaxation methods are based on duality and provide yet another (more abstract) geometric viewpoint.

5.2 A Reformulation

Here we will review the magnitude least-squares problem and reformulate it in a way that will be more convenient when discussing methods to find solutions. Recall that the MLS problem is

$$\min_{x\in\mathbb{C}^n} \quad \| \, |Ax| - b \, \|_2^2 \equiv \sum_{i=1}^{m} (|A_i x| - b_i)^2, \tag{MLS}$$

where A is a complex matrix in general, and $b \in \mathbb{R}^m$ is a positive real vector. It's useful to prove a fact that will allow us to reformulate this problem in a way that doesn't involve the absolute value function. We will use the formulation to derive the variable exchange method, and use it again later in forming a convex dual relaxation.

Theorem T1: The magnitude least-squares problem (MLS) is equivalent to the following:

$$\begin{aligned} \min_{x,z} \quad & \sum_{i=1}^{m} |A_i x - b_i z_i|^2 \\ \text{s.t.} \quad & |z_i| = 1. \end{aligned} \tag{Theorem T1}$$

The theorem makes intuitive sense, for the introduced unity-modulus variables z_i are related to the fact that in the magnitude least-squares problem, the phases in the complex vector Ax are irrelevant. The variables z_i could be thought to represent those phases. To prove Theorem T1, we will first prove two lemmas.

Lemma L1: Given $y \in \mathbb{C}$, we can express its modulus as:

$$|y| = \max_{|z|=1} \mathrm{Re}\{y\bar{z}\} \qquad \text{(Lemma L1)}$$

Proof of Lemma L1: If $|y| = 0$, then any $|z| = 1$ maximizes (Lemma L1), and the statement is true. On the other hand, by the Cauchy-Schwartz inequality, we have that for any $y, z \in \mathbb{C}$,

$$|y\bar{z}| \leq |y||z|,$$

so that

$$\max_{|z|=1} |y\bar{z}| \leq |y||z^\star| = |y|, \qquad (5.1)$$

since $z^\star$, the maximizer of (5.1) is such that $|z^\star| = 1$. Since $|y| \neq 0$ we see that $z^\star = y/|y|$ maximizes (5.1), achieving the upper bound of $|y|$. Because $\mathrm{Re}\{y\bar{z}\} \leq \sqrt{\mathrm{Re}\{y\bar{z}\}^2 + \mathrm{Im}\{y\bar{z}\}^2} = |y\bar{z}|$, we have that

$$\max_{|z|=1} \mathrm{Re}\{y\bar{z}\} \leq \max_{|z|=1} |y\bar{z}|.$$

But $\mathrm{Re}\{y\bar{z}^\star\} = \mathrm{Re}\{|y|\} = |y| \geq \max_{|z|=1} |y\bar{z}|$ by (5.1), so

$$\max_{|z|=1} \mathrm{Re}\{y\bar{z}\} \geq \max_{|z|=1} |y\bar{z}|.$$

Hence,

$$\max_{|z|=1} \mathrm{Re}\{y\bar{z}\} = \max_{|z|=1} |y\bar{z}| = |y|.$$

This proves Lemma L1. ■

Lemma L2: Given $y \in \mathbb{C}$ and $a \in \mathbb{R}$, $a \geq 0$, the following holds:

$$(|y| - a)^2 = \min_{|z|=1} |y - az|^2 \qquad \text{(Lemma L2)}$$

Proof of Lemma L2: Let y and $a \geq 0$ be fixed variables in $\mathbb{C}$ and $\mathbb{R}$ respectively. We

have

$$\begin{aligned}
(|y| - a)^2 &= y^2 - 2a|y| + a^2 \\
&= y^2 - 2a\left(\max_{|z|=1} \mathrm{Re}\{y\bar{z}\}\right) + a^2 && \text{(by Lemma L1)} \\
&= y^2 - (\max_{|z|=1} 2a\mathrm{Re}\{y\bar{z}\}) + a^2 && \text{(since } a \geq 0\text{)} \\
&= y^2 + \min_{|z|=1}(-2a\mathrm{Re}\{y\bar{z}\}) + a^2 \\
&= \min_{|z|=1}\left(y^2 - 2a\mathrm{Re}\{y\bar{z}\} + a^2\right) \\
&= \min_{|z|=1} |y - az|^2,
\end{aligned}$$

proving Lemma L2. ■

Proof of Theorem T1: Using Lemma L2, we can rewrite each term of the sum in (MLS) as

$$(|A_i x| - b_i)^2 = \min_{|z_i|=1} |A_i x - b_i z_i|^2.$$

Substituting, we have that (MLS) can be written as

$$\begin{aligned}
\min_{x} \quad & \sum_{i=1}^{m} \min_{z_i} |A_i x - b_i z_i|^2 \\
\text{s.t.} \quad & |z_i| = 1.
\end{aligned} \tag{5.2}$$

Since each z_i enters into the sum in one and only one term, we can defining a vector $z \in \mathbb{C}^m$ with components z_i, and write (5.2) as

$$\begin{aligned}
\min_{x,z} \quad & \sum_{i=1}^{m} |A_i x - b_i z_i|^2 \\
\text{s.t.} \quad & |z_i| = 1,
\end{aligned} \tag{MLS.Z}$$

which proves Theorem T1. ■

5.3 Local Methods

5.3.1 Variable Exchange Method

The formulation (MLS.Z) now can form the basis for an iterative variable exchange method for finding a local minimizer. For fixed z, the minimizing x can be found easily as the

solution to a linear least-squares problem. In turn, for fixed x, the optimal z is obvious: z_i equals a complex number with modulus 1, with phase equal to that of $A_i x$. If we start with a feasible x and z, then holding one set of variables fixed and solving for the other can only improve the objective function. We know the objective function can never be negative, so by alternating the procedure, we will eventually find at least a local minimizer for the magnitude least-squares problem. The procedure is as follows:

Step 1: Choose solution tolerance $\epsilon > 0$.

Step 2: Choose initial x either randomly, or by any other procedure.

Step 3: For all i, set z_i as the unity modulus complex number with phase equal to that of $A_i x$.

Step 4: Hold z fixed, and find a new x as the solution to the unconstrained linear least-squares problem.

Step 5: Repeat steps 3 and 4 until the decrease in objective function has diminished to within ϵ.

In the next section we discuss a gradient-based local method for solving a very slightly modified version of the magnitude least-squares problem. The modification makes the objective function everywhere differentiable. Like the variable exchange method, the gradient-based method is only capable of finding local solutions.

5.3.2 Gauss-Newton Method

Here we introduce a method that is particularly well-suited to our nonlinear least-squares problem, called the Gauss-Newton method (see [BV04] p. 520). The Gauss-Newton method uses the gradient of the objective function to progressively take downhill steps, eventually discovering a locally minimum point. What distinguishes the Gauss-Newton method from other local methods like the Newton method or the steepest descent method is the exact choice of downhill direction (called the search direction), and the manner in which this is calculated.

In our problem, the search direction can be calculated very efficiently as the solution to a linear least-squares problem. This is the motivation behind using Gauss-Newton (as opposed to the Newton or steepest descent methods).

The objective function in the magnitude least-squares problem (MLS) is not everywhere differentiable, which has the potential to cause problems in our gradient method. For this reason, we will attempt to solve what is actually a different (but very closely related) problem. Instead of solving (MLS), we will find a local minimizer of

$$\min_{x\in\mathbb{C}^n} \quad \sum_{i=1}^{m}(|A_i x|^2 - b_i^2)^2 \equiv \sum_{i=1}^{m}(|a_i^H x|^2 - b_i^2)^2, \tag{MSLS}$$

where for notational clarity, we've introduced the variables $a_i \equiv A_i^H$, which are the Hermitian transposes of the rows of the matrix A. We can call this variant of the magnitude least-squares problem the *magnitude-squared least-squares problem*.

The objective function is real, but it is a function of complex variables. To avoid any confusion arising as a result of conflicting definitions of what's meant by "complex gradient", we will rewrite (MSLS) in terms of the real and imaginary parts of the complex variables. From there we can find the necessary gradients and derive the Gauss-Newton method.

Let the real and imaginary parts of variables a_i and x be notated by a_{iR}, a_{iI}, x_R, and x_I. Then it isn't hard to show that

$$|a_i^H x|^2 = \begin{bmatrix} x_R \\ x_I \end{bmatrix}^T \begin{bmatrix} a_{iR} & -a_{iI} \\ a_{iI} & a_{iR} \end{bmatrix} \begin{bmatrix} a_{iR}^T & a_{iI}^T \\ -a_{iI}^T & a_{iR}^T \end{bmatrix} \begin{bmatrix} x_R \\ x_I \end{bmatrix},$$

so that (MSLS) is equivalent to the following which only involves real variables:

$$\min_{x_R, x_I \in \mathbb{R}^n} \sum_{i=1}^{m} \left(\begin{bmatrix} x_R \\ x_I \end{bmatrix}^T \begin{bmatrix} a_{iR} & -a_{iI} \\ a_{iI} & a_{iR} \end{bmatrix} \begin{bmatrix} a_{iR}^T & a_{iI}^T \\ -a_{iI}^T & a_{iR}^T \end{bmatrix} \begin{bmatrix} x_R \\ x_I \end{bmatrix} - b_i^2 \right)^2.$$

Again, for notational convenience and clarity, let

$$\mathcal{A}_i \equiv \begin{bmatrix} a_{iR} & -a_{iI} \\ a_{iI} & a_{iR} \end{bmatrix}, \text{ and let}$$
$$u \equiv \begin{bmatrix} x_R \\ x_I \end{bmatrix}.$$

The problem is then

$$\min_{u\in\mathbb{R}^{2n}} \sum_{i=1}^{m} (u^T \mathcal{A}_i \mathcal{A}_i^T u - b_i^2)^2. \tag{5.3}$$

Let $f(u) = \sum f_i(u)^2$ denote the objective function in (5.3). The gradient of $f(u)$ is

$$\begin{aligned}\nabla f(u) &= 2\sum_{i=1}^{m} f_i(u)\nabla f_i(u) \\ &= 4\sum_{i=1}^{m} (u^T \mathcal{A}_i \mathcal{A}_i^T u - b_i^2)\mathcal{A}_i \mathcal{A}_i^T u.\end{aligned}$$

The Hessian of $f(u)$ is

$$\begin{aligned}\nabla^2 f(u) &= 2\sum_{i=1}^{m} \nabla f_i(u)\nabla f_i(u)^T + f_i(u)\nabla^2 f_i(u) \\ &= 4\sum_{i=1}^{m} 2\mathcal{A}_i \mathcal{A}_i^T u u^T \mathcal{A}_i \mathcal{A}_i^T + (u^T \mathcal{A}_i \mathcal{A}_i^T u - b_i^2)\mathcal{A}_i \mathcal{A}_i^T. \end{aligned} \tag{5.4}$$

The Newton search direction is the direction along which lies the minimizer of the second-order approximation of the objective function (see [BV04]). This direction is given by

$$\Delta u_{nt} = -\nabla^2 f(u)^{-1} \nabla f(u)$$

The Gauss-Newton method keeps only the first term of (5.4) an approximation to the Hessian in the calculation of a downward search direction. The Gauss-Newton search direction then is

$$\begin{aligned}\Delta u_{gn} &= \left(8\sum_{i=1}^{m} \mathcal{A}_i \mathcal{A}_i^T u u^T \mathcal{A}_i \mathcal{A}_i^T\right)^{-1} \left(4\sum_{i=1}^{m} (u^T \mathcal{A}_i \mathcal{A}_i^T u - b_i^2)\mathcal{A}_i \mathcal{A}_i^T u\right) \\ &= \frac{1}{2}\left(\sum_{i=1}^{m} \mathcal{A}_i \mathcal{A}_i^T u u^T \mathcal{A}_i \mathcal{A}_i^T\right)^{-1} \left(\sum_{i=1}^{m} (u^T \mathcal{A}_i \mathcal{A}_i^T u - b_i^2)\mathcal{A}_i \mathcal{A}_i^T u\right)\end{aligned}$$

Upon further inspection, we can see that Δu_{gn} is actually the solution to the linear least-squares problem

$$\Delta u_{gn} = \underset{y}{\operatorname{argmin}} \frac{1}{2}\sum_{i=1}^{m} \left(u^T \mathcal{A}_i \mathcal{A}_i y_i - (u^T \mathcal{A}_i \mathcal{A}_i^T u - b_i^2)\right)^2.$$

The Gauss-Newton direction can therefore be calculated much more efficiently than the Newton direction for this problem. We can solve the problem by taking steps in the direction Δu_{gn}, of a size determined by using a backtracking line search. The backtracking line search is a simple, efficient procedure that ensures that we always will progress downhill, avoiding steps that are too large. Eventually the progress downhill slows to within a prescribed tolerance and we have obtained a locally minimum solution.

5.4 Semidefinite Relaxations

In the preceding section, we derived local methods for both the magnitude least-squares problem (MLS), and the very similar magnitude-squared least-squares problem (MSLS). Now we will derive (convex) semidefinite programs that are relaxations of these two problems. The solutions obtained by these methods can be used as starting points for the local methods. In Chapter 7 we will look closely at how much improvement we can gain using these combinations of methods.

5.4.1 Relaxation for MLS

We will start with the magnitude least-squares problem (MLS) and derive a semidefinite relaxation using duality. After calculating the dual, we will interpret it in terms of the original nonconvex primal problem. This will allow us to see in a more direct way that indeed it is a relaxation, the solutions of which, therefore, must underestimate the global minimum of (MLS).

As proved above, (MLS) is equivalent to (MLS.Z). Defining a diagonal matrix B with diagonal entries b_i, (MLS.Z) can be written in matrix notation as

$$\begin{aligned}
&\min_x \min_z \quad \|Ax - Bz\|^2 \quad \text{s.t.} \quad |z_i| = 1, \quad i = 1, 2, ..., m \\
&= \min_z \left(\min_x \|Ax - Bz\|^2 \right) \quad \text{s.t.} \quad |z_i| = 1, \quad i = 1, 2, ..., m \\
&= \min_z \|A(A^\dagger Bz) - Bz\|^2 \quad \text{s.t.} \quad |z_i| = 1, \quad i = 1, 2, ..., m,
\end{aligned}$$

where $A^\dagger$ is the pseudo-inverse of A, which equals $(A^H A)^{-1} A^H$ if A has full rank. Define

$W \in \mathbb{C}^{m \times m}$ as

$$W \equiv (AA^{\dagger}B - B)^H (AA^{\dagger}B - B),$$

we see that $W \succeq 0$ and that (MLS) is equivalent to the problem

$$\begin{aligned} \min_{z} \quad & z^H W z \\ \text{s.t.} \quad & |z_i| = 1, \quad i = 1, 2, ..., m \end{aligned} \tag{MLS.A}$$

Problem (MLS.A) can be expressed as a quadratic objective function with quadratic equality constraints:

$$\begin{aligned} \min_{z} \quad & z^H W z \\ \text{s.t.} \quad & |z_i|^2 = 1, \quad i = 1, 2, ..., m, \end{aligned}$$

Introducing a vector of real Lagrange multipliers, we can derive the dual. Define the notation $D(x)$ to be the diagonal matrix with entries x_i, the elements of x.

$$\begin{aligned} L(z, \xi) &\equiv z^H W z + \sum_{i=1}^{m} \xi_i (1 - |z_i|^2) \\ &= z^H W z - z^H D(\xi) z + 1_m^T \xi \\ &= z^H (W - D(\xi)) z + 1_m^T \xi. \end{aligned}$$

Now, minimizing over z we can calculate the dual function:

$$\begin{aligned} g(\xi) &\equiv \min_{z} L(z, \xi) \\ &= \begin{cases} -\infty & \text{if } W - D(\xi) \not\succeq 0 \\ 1_m^T \xi & \text{if } W - D(\xi) \succeq 0. \end{cases} \end{aligned}$$

Maximizing the dual function yields the Lagrangian dual of (MLS.A):

$$\begin{aligned} \max_{\xi} \quad & 1_m^T \xi \\ \text{s.t.} \quad & W - D(\xi) \succeq 0 \end{aligned} \tag{MLS.D}$$

We can now find a dual problem associated with this convex optimization problem, to arrive at a formulation that can easily be seen to be a relaxation of (MLS.A). The most

efficient way to compute this is to appeal to the concepts of conic duality, by introducing a positive semidefinite matrix of multipliers $Z \succeq 0$. Define the notation $d(X)$ to be the vector consisting of the diagonal elements of X.

$$\begin{aligned} L(\xi, Z) &\equiv 1_m^T \xi + \operatorname{Tr}\left(Z(W - D(\xi))\right) \\ &= (1_m - d(Z))^T \xi + \operatorname{Tr}\left(ZW\right). \end{aligned}$$

Maximizing over ξ yields the dual function,

$$g(Z) = \begin{cases} \infty & \text{if } 1_m - d(Z) \neq \vec{0} \\ \operatorname{Tr}\left(ZW\right) & \text{if } 1_m - d(Z) = \vec{0}. \end{cases}$$

So the dual of (MLS.D) is

$$\begin{aligned} \min_{Z} \quad & \operatorname{Tr}\left(ZW\right) \\ \text{s.t.} \quad & Z_{ii} = 1, \quad i = 1, 2, ..., m, \\ & Z \succeq 0. \end{aligned} \tag{MLS.P}$$

By expressing (MLS.A) in a similar form, we can see that (MLS.P) can be derived directly from (MLS.A) by relaxing a rank constraint. Equation (MLS.A) can be expressed as

$$\begin{aligned} \min_{z} \quad & \operatorname{Tr}\left(zz^H W\right) \\ \text{s.t.} \quad & |z_i| = 1, \quad i = 1, 2, ..., m. \end{aligned}$$

By using the fact that a positive semidefinite matrix can be factored as $Z = zz^H$ if and only if it is rank one, we have

$$\begin{aligned} \min_{Z} \quad & \operatorname{Tr}\left(ZW\right) \\ \text{s.t.} \quad & Z_{ii} = 1, \quad i = 1, 2, ..., m, \\ & Z \succeq 0 \\ & \operatorname{rank}(Z) = 1. \end{aligned} \tag{5.5}$$

Thus, we see that (MLS.P) can be derived directly by relaxing the rank constraint.

To summarize, we have derived a primal-dual pair, (MLS.P) and (MLS.D), associated with our main problem (MLS.A). Since (MLS.P) is strictly feasible, (take $Z = I$, the identity), the Slater condition is satisfied for this pair, and strong duality holds, i.e., the solutions to these dual optimization problems are equal. Furthermore, we know that this solution underestimates the solution to (MLS.A) since (MLS.P) is a relaxation.

5.4.2 Solving MLS via the Relaxation

The relaxation (MLS.P) can be efficiently solved using semidefinite programming solvers such as SeDuMi [Stu99]. The variables in (MLS.P) do not obviously correspond to the variables in the original nonconvex problem (MLS.A) (equivalent to (MLS)), however. How does solving this relaxation help us to solve (MLS.A) then?

First, the optimal objective of (MLS.P) must always be less than or equal to the objective of (MLS.A). If we had obtained a feasible solution to (MLS.A) (found for example using the a local method like the variable exchange method described above), then we could compare it's obtained objective to that of the lower bound. If the differential were small, we would know that even if we hadn't discovered the exact global minimizer, we were close. We can, however, use the solution of (MLS.P) to help us find such a solution.

The optimal variables in (MLS.P) form a semidefinite matrix $Z^\star$. If $Z^\star$ has rank one, then it is feasible (and optimal) also for the rank-constrained problem (5.5), and we need only factor it to find the optimal for (MLS.A). That is usually too much to hope for. More often, the answer $Z^\star$ will not be rank one. In this case, as we will discuss at length in the next chapter, we will be able to find good feasible solutions to (MLS.A) by considering $Z^\star$ to be a covariance matrix of a normal probability distribution. We can randomly sample from the distribution, obtaining vectors which we can scale in a straightforward way to satisfy the equality constraints in (MLS.A). We will be able to prove that solutions obtained in this way are guaranteed to possess objectives that are close in value to the global optimal.

5.4.3 Relaxation for MSLS

Just as the semidefinite relaxation for (MLS) can be interpreted as the problem that results from discarding the rank constraint in the problem (5.5), we can form a rank-constrained semidefinite programming problem equivalent to (MSLS). We can rewrite (MSLS) as

$$\begin{aligned} \min_{X \succeq 0} \quad & \sum_{i=1}^{m} \left(\mathrm{Tr}\left(a_i a_i^H X \right) - b_i^2 \right)^2 . \\ \text{s.t.} \quad & \mathrm{Rank}\,(X) = 1 \end{aligned} \tag{5.6}$$

A semidefinite relaxation of (5.6) then can easily be seen to be

$$\min_{X \succeq 0} \quad \sum_{i=1}^{m} \left(\mathrm{Tr}\left(a_i a_i^H X \right) - b_i^2 \right)^2 . \tag{MSLS.P}$$

Like we did with (MLS) we can derive this result using duality. To do so, it's convenient to define a operator that converts a matrix of size $n \times n$ to a column vector of size $n^2 \times 1$. Call this operator $\overrightarrow{[\cdot]}$. We can express (5.6) as

$$\begin{aligned} \min_{X} \quad & \sum_{i=1}^{m} \left(\mathrm{Tr}\left(a_i a_i^H X \right) - b_i^2 \right)^2 . \\ \text{s.t.} \quad & X = x x^H \end{aligned} \tag{MSLS.A}$$

We can derive the dual by introducing the complex matrix Y of Lagrange multipliers and forming the Lagrangian,

$$L(x, X, Y) \equiv \quad \sum_{i=1}^{m} \left(\mathrm{Tr}\left(a_i a_i^H X \right) - b_i^2 \right)^2 + \mathrm{Tr}\left(Y x x^H \right) - \mathrm{Tr}\left(Y X \right) . \tag{5.7}$$

Define $\mathcal{A}$ and c as follows:

$$\mathcal{A} \equiv \begin{bmatrix} \overrightarrow{\left[a_1 a_1^H\right]}^H \\ \overrightarrow{\left[a_2 a_2^H\right]}^H \\ \vdots \\ \overrightarrow{\left[a_m a_m^H\right]}^H \end{bmatrix} , \qquad c \equiv \overrightarrow{\left[\sum_{i=1}^{m} 2 b_i^2 a_i a_i^H \right]}$$

Then (5.7), the Lagrangian can be expressed in matrix/vector notation as

$$L(x, X, Y) = \overrightarrow{[X]}^H \mathcal{A}^H \mathcal{A} \overrightarrow{[X]} - (c + \overrightarrow{[Y]})^H \overrightarrow{[X]} + x^H Y x + \sum_{i=1}^{m} b_i^4$$

Minimizing over x and X yields the dual function,

$$\begin{aligned} g(Y) &\equiv \min_{x,X} L(x, X, Y) \\ &= \begin{cases} -\infty & \text{if} \quad Y \not\succeq 0 \\ -\overrightarrow{[Y]}^H (4\mathcal{A}^H\mathcal{A})^{-1}\overrightarrow{[Y]} + c^H(2\mathcal{A}^H\mathcal{A})^{-1}\overrightarrow{[Y]} & \\ \qquad -c^H(4\mathcal{A}^H\mathcal{A})^{-1}c + \sum_{i=1}^m b_i^4 & \text{if} \quad Y \succeq 0 \end{cases} \end{aligned}$$

The dual problem is therefore

$$\begin{aligned} \max_{Y} \quad & -\overrightarrow{[Y]}^H (4\mathcal{A}^H\mathcal{A})^{-1}\overrightarrow{[Y]} + c^H(2\mathcal{A}^H\mathcal{A})^{-1}\overrightarrow{[Y]} - c^H(4\mathcal{A}^H\mathcal{A})^{-1}c + \sum_{i=1}^m b_i^4 \quad \text{(MSLS.D)} \\ \text{s.t.} \quad & Y \succeq 0 \end{aligned}$$

The dual of the dual problem is derived by introducing a positive semidefinite matrix Z corresponding to the constraint in (MSLS.D):

$$L(Y, Z) = -\overrightarrow{[Y]}^H (4\mathcal{A}^H\mathcal{A})^{-1}\overrightarrow{[Y]} + c^H(2\mathcal{A}^H\mathcal{A})^{-1}\overrightarrow{[Y]} + \overrightarrow{[Z]}^H\overrightarrow{[Y]} - c^H(4\mathcal{A}^H\mathcal{A})^{-1}c + \sum_{i=1}^m b_i^4$$

Maximizing this quadratic function with respect to Y will reveal the dual function of the dual problem:

$$g(Z) = \begin{cases} \infty & \text{if} Z \not\succeq 0 \\ \frac{1}{4}\left(c^H(2\mathcal{A}^H\mathcal{A})^{-1} + \overrightarrow{[Z]}^H\right)(4\mathcal{A}^H\mathcal{A})\left((2\mathcal{A}^{\mathcal{A}})^{-1}c + \overrightarrow{[Z]}\right) & \\ \qquad -c^H(4\mathcal{A}^H\mathcal{A})^{-1}c + \sum_{i=1}^m b_i^4 & \text{if} Z \succeq 0 \end{cases}$$

Finally we can reduce and obtain the dual of the dual problem:

$$\begin{aligned} \min_{Z} \quad & \overrightarrow{[Z]}^H\mathcal{A}^H\mathcal{A}\overrightarrow{[Z]} + \overrightarrow{[Z]}^H c + \sum_{i=1}^m b_i^4 \\ \text{s.t.} \quad & Z \succeq 0 \\ \equiv \min_{Z} \quad & \sum_{i=1}^m \left(\text{Tr}\left(a_i a_i^H Z\right)\right)^2 + \sum_{i=1}^m \text{Tr}\left(2b_i^2 a_i a_i^H Z\right) + \sum_{i=1}^m b_i^4 \\ \text{s.t.} \quad & Z \succeq 0 \\ = \min_{Z} \quad & \sum_{i=1}^m \left(\text{Tr}\left(a_i a_i^H Z\right) - b_i^2\right)^2 \quad \text{(MSLS.P)} \\ \text{s.t.} \quad & Z \succeq 0 \end{aligned}$$

This is exactly the relaxation we discovered by ignoring the rank constraint in (5.6).

5.4.4 Solving MSLS via the Relaxation

In exactly the same way as we can use the relaxation (MLS.P) to help us find solutions for (MLS), we can use (MSLS.P) to help us find solutions for (MSLS). Again, since we know that (MSLS.P) is a relaxation, a solution to it provides us with a lower bound on the globally optimal objective for (MSLS). This has the potential to let us know whether a solution obtained with a local method, such as Gauss-Newton, has objective approaching that of the global minimum. We can also use an optimal solution $Z^\star$ to (MSLS.P) to help us discover a good solution to (MSLS) or a good starting point for a local method.

This would entail considering the positive semidefinite matrix $Z^\star$ to be a covariance matrix for a normally distributed random vector. Randomly sampling the distribution provides us with candidate solution vectors for (MSLS). In contrast to (MLS.P), the relaxation (MSLS.P) does not possess equality constraints, so any vector is feasible for (MSLS), simplifying the process of obtaining a primal feasible point. We will prove that sampling from this distribution is markedly better than sampling from some other distribution.

5.5 Summary

In this chapter we have defined four methods for solving two similar problems related to magnitude fitting. The two problems are the magnitude least-squares problem (MLS) and the magnitude-squared least-squares problem (MSLS). For each problem we have derived both a local method and a semidefinite relaxation method. In the next chapter we will analyze the semidefinite relaxation methods in more detail, explaining how to use the relaxations to obtain good solutions to the original nonconvex problems (MLS) and (MSLS). We will justify the random sampling approaches touched on in this chapter by proving guarantees that the solutions obtained in this way will be of high quality. In combination with the local methods, we are armed with a methodology that performs well in practice and also possesses theoretical structure and meaning.

Chapter 6

Obtaining Primal Solutions from Relaxations and Quality Estimates

6.1 Introduction

In the last chapter, we derived methods for solving the magnitude least-squares (MLS) and magnitude-squared least-squares problems (MSLS). In this chapter we will investigate more completely how to use the semidefinite relaxation solutions to help solve these non-convex problems. The technique will involve considering the positive semidefinite solution matrices to be covariance matrices of random variables, and sampling from those distributions to find primal feasible points. We will justify the idea by proving facts about the relative quality of the solutions obtained.

The important and now famous paper by Goemans and Williamson [GW95] detailed this approach and used it to approximately solve an NP-complete problem MAX CUT with fixed relative accuracy. The MAX CUT problem is a nonconvex maximization problem that has many applications in operations research and engineering. It can be comprehended abstractly as a graph partitioning problem. They prove that by using the method we will discuss, which relies on semidefinite relaxation, one can be guaranteed of obtaining a solution with objective

$$\nu_{\text{gw}} \geq .87856\mu^*_{\text{gw}} \geq 0.$$

where ν_{gw} is the solution obtained using this method, and ν^* is the global maximum, which

for that problem, is always nonnegative. This guarantee has theoretical appeal. The method is even more exciting because of observations that the solutions obtained in this way are often far closer to the true optimal than what is guaranteed.

Subsequent to Goemans and Williamson's work, researchers discovered many extensions to the idea, and applied the principles to other similar problems. Indeed, that is what we will do in this chapter for the magnitude least-squares and magnitude-squared least-squares problems. An important extension to the work was derived by Yurii Nesterov in [Nes98], where he applied the same idea in an elegant way to general quadratic problems with constraints on squared variables. We will start with a review of his proof and will be able to apply the result directly to the magnitude least-squares problem. We will later show that an even better bound exists, due in part to the structure of the complex variables in our problem. This will rely on recent work published by Shuzhong Zhang and Yongwei Huang in [ZH04]. Finally we will derive a novel bound for the magnitude-squared least-squares problem that, similar to Zhang and Huang's work, uses the complex structure of the problem to improve the bound calculated for the real version of the problem.

6.2 Nesterov's Proof

The clarity and elegance of the proof published in [Nes98] and [Nes97] is worth repeating here. It will serve to convey the key ideas of these kinds of proofs, and help to clarify and condense the other results in this the chapter.

The problems Nesterov is concerned with are the following:

$$f^*(A) \equiv \max_x \quad x^T A x \qquad \text{(6.0A)}$$
$$\text{s.t.} \quad x_i = \pm 1, \quad i = 1, 2, ..., n,$$

and

$$f_*(A) \equiv \min_x \quad x^T A x \qquad \text{(6.1A)}$$
$$\text{s.t.} \quad x_i = \pm 1, \quad i = 1, 2, ..., n,$$

where A is an arbitrary symmetric $(n \times n)$-matrix.

Note the similarity of (6.1A) to our magnitude least-squares equivalent problem (MLS.A). The differences between the two problems are first that the variables in (MLS.A) are complex, whereas in Nesterov's problem they are real, and second that the matrix W in (MLS.A) is positive semidefinite, whereas A above is arbitrary (but symmetric). We will show that these differences allow us to prove a slightly stronger result for (MLS.A).

Just as we derived the semidefinite relaxation for (MLS.A) in the last chapter, we can see that semidefinite relaxations for (6.0A) is

$$\begin{aligned} s^*(A) \equiv \max_x \quad & \mathrm{Tr}\,(AX) && \text{(6.2P)} \\ \text{s.t.} \quad & X \succeq 0 \\ & d(X) = \mathbf{1} \end{aligned}$$

$$\begin{aligned} = \min_\xi \quad & \mathbf{1}^T \xi && \text{(6.2D)} \\ \text{s.t.} \quad & D(\xi) \succeq A \end{aligned}$$

and the relaxation for (6.1A) is

$$\begin{aligned} s_*(A) \equiv \min_x \quad & \mathrm{Tr}\,(AX) && \text{(6.3P)} \\ \text{s.t.} \quad & X \succeq 0 \\ & d(X) = \mathbf{1} \end{aligned}$$

$$\begin{aligned} = \max_\xi \quad & \mathbf{1}^T \xi && \text{(6.3D)} \\ \text{s.t.} \quad & D(\xi) \preceq A \end{aligned}$$

where the function $D(\cdot)$ maps a vector to a diagonal matrix with diagonal entries equal to the entries in the vector, and the boldfaced notation $\mathbf{1}$ represents a vector of all ones of appropriate dimension (it's in $\mathbb{R}^n$). Now we will recount Nesterov's proof as laid out in [Nes97].

We first note that because (6.2P) and (6.3P) are relaxations of (6.0A) and (6.1A), we have the following relationship among the optimal values:

$$s_*(A) \leq f_*(A) \leq f^*(A) \leq s^*(A)$$

As explained by Nesterov, the goal is to make that relation more precise. He starts by imposing the assumption that A is positive semidefinite. Assume that x^* is a solution to (6.0A). Denote $I^* = \{i : x_I^* = 1\}$. Let $V = \begin{bmatrix} v_1 & v_2 & ... & v_n \end{bmatrix}$ be an $(n \times n)$-matrix. For $a \in \mathbb{R}^n$, denote $\sigma(a) \in \mathbb{R}^n$ as the vector with components $\mathrm{sgn}(a_i)$, and define $\mathrm{sgn}(0) = 0$.

Lemma L3: If $A \succeq 0$,

$$\begin{aligned} f^*(A) = \max_{u,v} \quad & (\sigma(V^T u))^T A \sigma(V^T u) \\ \text{s.t.} \quad & \|v_i\| = 1 \qquad i = 1, 2, ..., n \\ & \|u\| = 1. \end{aligned} \tag{Lemma L3}$$

Proof: Denote the right hand side of the above equation by f. Taking $x = \sigma(V^T u)$ we have $x_i = \pm 1$, so that $f \leq f^*(A)$. On the other hand, fix an arbitrary $u \in \mathbb{R}^n$. For $i \in I^*$, choose $v_i = u$ and for $i \notin I^*$, choose $v_i = -u$. Then $\sigma(V^T u) = x^*$ so that $f \geq f^*(A)$, proving the lemma. ■

Let $E_u(f(u))$ denote the average value of $f(u)$ on the n-dimensional unit sphere.

Lemma L4:

$$\begin{aligned} f^*(A) = \max_{V} \quad & E_u((\sigma(V^T u))^T A \sigma(V^T u)) \\ \text{s.t.} \quad & \|v_i\| = 1 \qquad i = 1, 2, ..., n \end{aligned} \tag{Lemma L4}$$

Proof: The average value on the sphere cannot exceed the maximum value, which is given by the right hand side of (Lemma L3). So we have

$$\begin{aligned} f^*(A) \geq \max_{V} \quad & E_u((\sigma(V^T u))^T A \sigma(V^T u)) \ . \\ \text{s.t.} \quad & \|v_i\| = 1 \qquad i = 1, 2, ..., n \end{aligned}$$

On the other hand, expanding the right hand side of (Lemma L4), we have

$$E_u\left((\sigma(V^T u))^T A \sigma(V^T u)\right) \qquad = \qquad \sum_{i=1}^{n} \sum_{i=1}^{n} a_{ij} E_u\left(\mathrm{sgn}(v_i^T u)\mathrm{sgn}(v_j^T u)\right)$$

Let $y \in \mathbb{R}^n, \|y\| = 1$. If we choose $v_i = y$ for $i \in I^*$, and $v_i = -y$ for $i \notin I^*$, then

$$\begin{aligned} E_u\left(\mathrm{sgn}(v_i^T u)\mathrm{sgn}(v_j^T u)\right) &= \begin{cases} 1, & (i, j \in I^*) \quad \text{or} \quad (i, j \notin I^*) \\ -1, & \text{otherwise} \end{cases} \\ &\equiv x_i^* \cdot x_j^*. \end{aligned}$$

Thus,

$$f^*(A) \leq \max_V \quad E_u\left((\sigma(V^T u))^T A \sigma(V^T u)\right).$$
$$\text{s.t.} \quad \|v_i\| = 1 \qquad i = 1, 2, ..., n$$

proving the lemma ■

Note here that the truth (and the proof) of Lemma L4 does not depend on the expectation $E_u(\cdot)$ being defined as a the average value over a sphere. The probability density function of u can be arbitrary. We will see that a spherically symmetric density function is what is required for the next result to be meaningful. Density functions that are spherically symmetric include the uniform density function defined on a sphere (the one suggested by Nesterov), but also include a normal distribution with identity covariance matrix, defined over all of $\mathbb{R}^n$.

Nesterov uses the notation $[\cdot]$ to apply functions of single variables to the entries of matrices. For example, $\arcsin[X]$ is the matrix whose entries are $\arcsin(x_{ij})$, and $[X]^k$ is the matrix whose entries are x_{ij}^k.

Theorem T2:

$$f^*(A) = \max_X \quad \frac{2}{\pi}\text{Tr}\left(A \cdot \arcsin[X]\right) \qquad \text{(Theorem T2)}$$
$$\text{s.t.} \quad X \succeq 0$$
$$d(X) = \mathbf{1}$$

Proof: First, we will establish that the feasible sets of (Theorem T2) and (Lemma L4) are in one-to-one correspondence. If X is feasible for (Theorem T2), then we have that $X \succeq 0$, $d(X) = \mathbf{1}$. For any such X, we can choose $V = X^{1/2}$ and it will be feasible for (Lemma L4), i.e., each column v_i will be such that $\|v_i\| = 1$. Likewise, given a V feasible for (Lemma L4), the matrix $X = V^T V$ is positive semidefinite, with diagonal entries equal to 1. Thus, Theorem T2 is true if the objective functions are equal, that is, if

$$E_u\left((\sigma(V^T u))^T A \sigma(V^T u)\right) = \frac{2}{\pi}\text{Tr}\left(A \cdot \arcsin[X]\right) \qquad (6.4)$$

Thinking geometrically, we can see that

$$E_u\left(\text{sgn}(v_i^T u)\cdot\text{sgn}(v_j^T u)\right) = 1 - 2\mathbf{Pr}\left(\text{sgn}(v_i^T u)\neq\text{sgn}(v_j^T u)\right)$$

The probability that these dot products have opposite sign is easily computed because the probability density function is spherically symmetric. Nesterov refers to Lemma 1.2 in [GW95]:

$$E_u\left(\text{sgn}(v_i^T u)\cdot\text{sgn}(v_j^T u)\right) = 1 - \frac{2}{\pi}\arccos(v_i^T v_j) = \frac{2}{\pi}\arcsin(v_i^T v_j)$$

The result follows because the expectation operator is linear. ■

We are now in a position to prove the quality bound:

Theorem T3:

$$f^*(A) \geq \frac{2}{\pi}\max_X \quad \text{Tr}\,(A\cdot X) \quad \equiv \quad \frac{2}{\pi}s^*(A) \qquad \text{(Theorem T3)}$$
$$\text{s.t.} \quad X \succeq 0$$
$$d(X) = \mathbf{1}$$

Proof: It isn't hard to show that $\arcsin[X] \geq X$, for any X with entries x_{ij} such that $|x_{ij}| \leq 1$, (as is the case for a feasible X in (Theorem T3)). Since $A \succeq 0$, the result follows immediately from Theorem T2. ■

Thus, given the nonconvex quadratic problem (6.0A), we have a way of solving it that guarantees us a solution that is greater than a fraction $(2/\pi)$ of the global maximum. The way we generate this solution is suggested by the proof: we solve the semidefinite relaxation obtaining a positive semidefinite X^*, randomly sample u multiple times from a spherically symmetric probability distribution, calculate $x = \sigma(X^{1/2}u)$, and choose the candidate x that achieves the largest objective. We have proved that the mean value of the objective will be greater than $2/\pi$ times that of the global maximum, so the approach of choosing the winner is also guaranteed to be bounded below (and will most likely be even closer to the true maximum value). Note that the procedure for getting the random candidate x is equivalent to sampling from a normal distribution with covariance matrix X, and then rounding (applying the function $\sigma(\cdot)$).

We will continue and complete the review of Nesterov's ideas by attending to the definiteness of the A matrix. This is particularly important for us because the magnitude least-squares problem is a minimization problem with a positive semidefinite A matrix. The proof above applies to the maximization case.

Theorem T4: For an indefinite A, the following relations hold:

$$s_*(A) \le f_*(A) \le s_{1-\alpha}(A) \le s_\alpha(A) \le f^*(A) \le s^*(A) \quad \text{(Theorem T4)}$$

where $\alpha = \dfrac{2}{\pi}$ and $s_\beta \equiv \beta s^*(A) + (1-\beta)s_*(A)$ for $\beta \in [0,1]$

Proof: Because the feasible vectors in (6.0A) and (6.1A) have entries with absolute value 1, the following holds for any $\xi \in \mathbb{R}^n$:

$$f^*(A + D(\xi)) = f^*(A) + \mathbf{1}^T\xi$$
$$f_*(A + D(\xi)) = f_*(A) + \mathbf{1}^T\xi$$
$$s^*(A + D(\xi)) = s^*(A) + \mathbf{1}^T\xi$$
$$s_*(A + D(\xi)) = s_*(A) + \mathbf{1}^T\xi$$

Denote ξ^* to be the solution to relaxation (6.2D) and ξ_* the solution to (6.3D). Because ξ^* and ξ_* are feasible, we have

$$D(\xi^*) - A \succeq 0$$
$$\mathbf{1}^T\xi^* = s^*(A)$$
$$A - D(\xi_*) \succeq 0$$
$$\mathbf{1}^T\xi_* = s_*(A).$$

Using these facts and (Theorem T3), we have

$$\begin{aligned} s^*(A) - f_*(A) &= \mathbf{1}^T\xi^* - f_*(A) \\ &= \mathbf{1}^T\xi^* + f_*(-A) \\ &= f^*(D(\xi^*) - A) \\ &\geq \frac{2}{\pi}s^*(D(\xi^*) - A) \\ &= \frac{2}{\pi}(\mathbf{1}^T\xi^* - s_*(A)) \\ &= \frac{2}{\pi}(s^*(A) - s_*(A)). \end{aligned}$$

Similarly, we have that

$$f^*(A) - s_*(A) \geq \frac{2}{\pi}(s^*(A) - s_*(A)).$$

Combining the two results proves the theorem. ■

This completes our review of the Nesterov's paper [Nes97]. We now will state one last result from [Nes98] that will be important when we consider applying the result above to the complex magnitude least-squares problem. Using very straightforward logic, we can extend the ideas presented above to derive quality bounds for the following more general problems:

$$\begin{aligned} m^*(A) \equiv \max_x \quad & x^T A x \\ \text{s.t.} \quad & B[x]^2 \leq c \end{aligned} \tag{6.5}$$

$$\begin{aligned} m_*(A) \equiv \min_x \quad & x^T A x \quad , \\ \text{s.t.} \quad & B[x]^2 \leq c \end{aligned} \tag{6.6}$$

where B is an arbitrary matrix and c is a vector, defining linear constraints on the squared variables $[x]^2$. Note that we recover problems (6.0A) and (6.1A) when $B = I$, the identity, and $c = \mathbf{1}$, a vector of all ones. The two problems (6.5) and (6.6) admit the semidefinite

relaxation primal and dual pairs,

$$\begin{array}{rlcll} s^*(A) \equiv & \max\limits_X & \mathrm{Tr}\,(AX) & = & \min\limits_\xi \quad c^T\xi \\ & \text{s.t.} & X \succeq 0 & & \text{s.t.} \quad d(B^T\xi) - A \succeq 0 \\ & & Bd(X) \leq c & & \xi \geq 0 \end{array}$$

$$\begin{array}{rlcll} s_*(A) \equiv & \min\limits_X & \mathrm{Tr}\,(AX) & = & \max\limits_\xi \quad c^T\xi \\ & \text{s.t.} & X \succeq 0 & & \text{s.t.} \quad d(B^T\xi) + A \succeq 0 \\ & & Bd(X) \leq c & & \xi \geq 0 \end{array}$$

We have the same quality guarantee for these more general problems as we did for (6.0A) and (6.1A). We refer the reader to [Nes98] for proofs of these relations (they follow easily using the same logic as above). The result is precisely that stated in (Theorem T4), namely that

$$s_*(A) \leq m_*(A) \leq s_{1-\alpha}(A) \leq s_\alpha(A) \leq m^*(A) \leq s^*(A),$$

$$\text{where} \quad \alpha = \frac{2}{\pi} \quad \text{and} \quad s_\beta \equiv \beta s^*(A) + (1-\beta)s_*(A) \quad \text{for} \quad \beta \in [0,1].$$

We also have a straightforward way of obtaining solutions that satisfy this bound.

6.3 Application of Nesterov's Bound to MLS

The magnitude least-squares problem is equivalent to a complex-valued version of (6.1A), as we showed in the last chapter:

$$\begin{array}{ll} \min\limits_z & z^H W z \\ \text{s.t.} & |z_i| = 1, \quad i = 1, 2, ..., m \end{array} \qquad \text{(MLS.A)}$$

In order to apply Nesterov's bound to this problem, we must express it as a function of real variables. See Appendix A where we show that there is such a formulation. We also prove in Appendix A that the semidefinite relaxations associated with each are also equivalent to each other (a nontrivial fact).

The equivalent real-valued formulation of (MLS.A) is

$$\begin{aligned} \min_{z_R,\, z_I} \quad & \begin{bmatrix} z_R \\ z_I \end{bmatrix}^T \begin{bmatrix} W_R & -W_I \\ W_I & W_R \end{bmatrix} \begin{bmatrix} z_R \\ z_I \end{bmatrix} \\ \text{s.t.} \quad & (z_R)_i^2 + (z_I)_i^2 = 1, \quad i = 1, 2, ..., m \end{aligned} \tag{MLS.AReal}$$

where z_R and z_I are real-valued decision variables that correspond to the real and imaginary parts of z and likewise W_R and W_I are the real and imaginary parts of W in (MLS.A). We see that this problem can immediately be expressed in the same form as (6.6):

$$\begin{aligned} \min_{z_R,\, z_I} \quad & \begin{bmatrix} z_R \\ z_I \end{bmatrix}^T \begin{bmatrix} W_R & -W_I \\ W_I & W_R \end{bmatrix} \begin{bmatrix} z_R \\ z_I \end{bmatrix} \\ \text{s.t.} \quad & \begin{bmatrix} I & I \\ -I & -I \end{bmatrix} \begin{bmatrix} [z_R]^2 \\ [z_I]^2 \end{bmatrix} \leq \begin{bmatrix} \mathbf{1} \\ -\mathbf{1} \end{bmatrix} \end{aligned} \tag{6.7}$$

Thus the bound applies. Nesterov's scheme for solving the problem involves the following steps. First, we solve the semidefinite relaxation of (6.7):

$$\begin{aligned} \min_{Z_1,\, Z_2,\, Z_3} \quad & \operatorname{Tr}\left(\begin{bmatrix} Z_1 & Z_2 \\ Z_2^T & Z_3 \end{bmatrix} \begin{bmatrix} W_R & -W_I \\ W_I & W_R \end{bmatrix} \right) \\ \text{s.t.} \quad & (Z_1)_{ii} + (Z_3)_{ii} = 1, \quad i = 1, 2, ..., m \\ & \begin{bmatrix} Z_1 & Z_2 \\ Z_2^T & Z_3 \end{bmatrix} \succeq 0 \end{aligned} \quad . \tag{MLS.PReal}$$

After obtaining a solution

$$Y^* \equiv \begin{bmatrix} Z_1^* & Z_2^* \\ Z_2^{T*} & Z_3^* \end{bmatrix} \succeq 0,$$

we factor it uniquely as follows:

$$Y^* = \begin{bmatrix} Z_1^* & Z_2^* \\ Z_2^{T*} & Z_3^* \end{bmatrix} = \begin{bmatrix} \sqrt{[D(Z_1^*)]} & 0 \\ 0 & \sqrt{[D(Z_3^*)]} \end{bmatrix} \begin{bmatrix} X_1 & X_2 \\ X_2^T & X_3 \end{bmatrix} \begin{bmatrix} \sqrt{[D(Z_1^*)]} & 0 \\ 0 & \sqrt{[D(Z_3^*)]} \end{bmatrix},$$

where the operator $D(\cdot)$ applied to a matrix is defined to be the diagonal matrix with entries taken from the diagonal of the argument, i.e., if M is a matrix with diagonal entries m_{ii}, then

$$D(M) \equiv \begin{bmatrix} m_{11} & 0 & \cdots & 0 \\ 0 & m_{22} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & m_{nn} \end{bmatrix}.$$

Note that this factorization implies that $d(X_1) = d(X_3) = \mathbf{1}$. The next step is to sample K times from a normal probability distribution

$$u_k \sim \mathcal{N}\left(0, \begin{bmatrix} X_1 & X_2 \\ X_2^T & X_3 \end{bmatrix}\right), \quad k = 1, 2, ..., K.$$

With each sample, we form a candidate solution vector using the sign-rounding idea present in the proofs above, and the reapplication of the squareroot factor:

$$\begin{bmatrix} (z_R)_k \\ (z_I)_k \end{bmatrix} = \begin{bmatrix} \sqrt{[D(Z_1^*)]} & 0 \\ 0 & \sqrt{[D(Z_3^*)]} \end{bmatrix} \operatorname{sgn}([u_k]), \quad k = 1, 2, ..., K$$

The candidate solution vectors $(z_R)_k$ and $(z_I)_k$, if chosen in this way, are feasible for (MLS.AReal). We can evaluate the objective function (MLS.AReal) for each sample. Assuming we have chosen K large enough we are guaranteed to discover a feasible solution with objective satisfying the bounds. The final solution vector can then be easily converted back into complex terms.

We can interpret the sign-rounding step as choosing one of the four quadrants of the complex plane to locate each complex decision variable. The diagonal entries of Z_1 and Z_3 determine phase-offset angles within the quadrant for the variables. This procedure works well in practice, and possesses the quality bound proved above. In the next section we will introduce a slightly improved (but very similar) method of rounding that carries with it a corresponding improvement in quality estimate.

6.4 Zhang/Huang Rouding

We were able to apply Nesterov's result to the magnitude least-squares problem, after reformulating it as a real quadratic minimization problem with constraints on squared variables. In the paper [ZH04], Shuzhong Zhang and Yongwei Huang prove a result that improves Nesterov's bound for complex problems like our formulation (MLS.A). In Nesterov's proof above, it is easy to calculate the expectation (6.4). In the case where the variables in the problem are complex, calculating the corresponding expectation is much more difficult. Nevertheless, the authors in [ZH04] are able to do just that through very clever and careful

evaluation of complicated integrals. They are concerned with the problem

$$g^*(A) \equiv \max_z \quad z^H A z \qquad \text{(Z)}$$
$$\text{s.t.} \quad |z_i| = 1, \quad i = 1, 2, ..., m$$

where A is positive semidefinite. Our magnitude least-squares equivalent problem (MLS.A) is a minimization problem. Zhang and Huang are able to prove that using a different rounding scheme the following bound holds:

$$g^*(A) \geq \frac{\pi}{4} \max_X \quad \text{Tr}\,(A \cdot X) \quad \equiv \quad \frac{\pi}{4} r^*(A),$$
$$\text{s.t.} \quad X \succeq 0$$
$$d(X) = \mathbf{1}$$

where all the variables in the problem are complex-valued. In Appendix A we prove that this relaxation involving complex variables, derived from the complex-valued formulation (Z) is actually equivalent to the real-valued relaxation derived from the real-valued equivalent problem. The ingredient that makes Zhang and Huang's bound better than Nesterov's bound is their superior method of rounding the random samples.

The maximization and minimization versions of (Z), though not touched on in [ZH04], are related in the same way as for Nesterov's problems. We have

Theorem T5: For an indefinite A, the following relations hold:

$$r_*(A) \leq g_*(A) \leq r_{1-\alpha}(A) \leq r_\alpha(A) \leq g^*(A) \leq r^*(A) \qquad \text{(Theorem T5)}$$
$$\text{where} \quad \alpha = \frac{\pi}{4} \quad \text{and} \quad r_\beta \equiv \beta r^*(A) + (1-\beta) r_*(A) \quad \text{for} \quad \beta \in [0, 1]$$

Proof: The result follows by exactly the same logic as Theorem T3. ■

The rounding scheme that makes this possible is simple. Instead of converting the problem to one with real-valued variables, with the intention of rounding the randomly generated samples to ± 1 as is generally applicable for the wider class of problems discussed by Nesterov, we keep the variables complex and project them to the nearest point on the complex unit circle. In Nesterov's proof, the operator $\sigma(x)$ was synonymous with $\text{sgn}([x])$. The Zhang/Huang rounding could be considered an extended definition, so that

$\sigma(z) \equiv z/|z|$. It should be noted that both methods of rounding perform well in practice. The two rounding schemes do differ and we are able to prove the slightly stronger result, which is of theoretical interest.

6.5 A Quality Proof for MSLS

In this section we present quality bounds for the magnitude-squared least-squares problem (MSLS). This problem is a variant of the magnitude least-squares problem and was discussed in the last chapter. Recall that it is

$$\mu = \min_{x \in \mathbb{C}^n} \quad \sum_{i=1}^{m} (|a_i^H x|^2 - b_i^2)^2 \tag{MSLS}$$

and admits the semidefinite relaxation,

$$\begin{aligned} \nu_* = \min_{X} \quad & \sum_{i=1}^{m} \left(\mathrm{Tr}\left(a_i a_i^H X\right) - b_i^2 \right)^2 \\ \text{s.t.} \quad & X \succeq 0 \end{aligned} \tag{MSLS.P}$$

In analyzing the magnitude least-squares equivalent problem (MLS.A) in the last section, we discovered that the bound for the complex-valued version of Nesterov's problem admitted a better bound, as proved by Zhang and Huang. The reason this is true is that the complex-valued problem (MLS.AReal) possesses more structure. Please refer to Appendix A for a discussion of the structure associated with complex variables and how that structure carries through in formulating relaxations. In the present section we will present novel results related to (MSLS), and show that in this problem too, the complex structure allows us to prove a tighter bound. We will use many of the same concepts used by Nesterov in his proof. The analysis by Zhang and Huang of the complex version of (6.0A) required careful evaluation of a complicated complex-valued integral. For the magnitude-squared least-squares problem discussed here, we will be able to reason about the structure and associated bounds directly by expanding the problem into real and imaginary parts. More elegantly, we can arrive at the same result by keeping the variables complex.

The derivation discussed in this section, along with the proofs discussed in Appendix A

provide us insight into why the complex versions of these problems have provably tighter semidefinite relaxations.

How does the optimal objective ν_* of (MSLS.P) compare to the true optimal μ that solves (MSLS)? We know that the $\nu_* \leq \mu$ because (MSLS.P) is a relaxation of (MSLS) which is exactly equivalent to the rank-constrained problem

$$\begin{aligned} \mu = \min_{X \succeq 0} \quad & \sum_{i=1}^{m} ((\mathrm{Tr}\left(a_i a_i^H X\right))^2 - b_i^2)^2. \\ \text{s.t.} \quad & \mathrm{Rank}\,(X) = 1. \end{aligned}$$

Now we will derive another semidefinite programming problem that will provide for us an upper bound on μ. The solution of (MSLS.P) provides us with an optimal complex semidefinite matrix X_*. We could obtain a primal feasible solution to (MSLS) by sampling from a complex normal distribution with zero mean and covariance matrix X_*. Given an arbitrary positive semidefinite matrix X, let's look at the expected value

$$\begin{aligned} \gamma(X) &= E \sum_{i=1}^{m} (u^H a_i a_i^H u - b_i^2)^2. \\ u &\sim \mathcal{N}(0, X) \end{aligned} \tag{6.8}$$

where u is a complex random variable sampled from a normal distribution with zero mean and covariance matrix X. Expand (6.8) as

$$\begin{aligned} \gamma(X) &= \sum_{i=1}^{m} E(u^H a_i a_i^H u - b_i^2)^2. \\ &= \sum_{i=1}^{m} E((u^H a_i a_i^H u)^2 - 2b_i^2 u^H a_i a_i^H u + b_i^4) \end{aligned} \tag{6.9}$$

Let $y_i = a_i^H u$. Since u consists of jointly Gaussian random variables, y_i is also Gaussian (and scalar). We know that $Ey_i = 0$ and that

$$\begin{aligned} E|y_i|^2 &= E u^H a_i a_i^H u \\ &= E\mathrm{Tr}\left(a_i a_i^H u u^H\right) \\ &= \mathrm{Tr}\left(a_i a_i^H E u u^H\right) \\ &= \mathrm{Tr}\left(a_i a_i^H X\right) \end{aligned}$$

We can express (6.9) in terms of y_i:

$$\begin{aligned}\gamma(X) &= \sum_{i=1}^{m} E(|y_i|^4 - 2b_i^2|y_i|^2 + b_i^4) \\ &= \sum_{i=1}^{m} E|y_i|^4 - 2b_i^2 E|y_i|^2 + b_i^4 \\ &= \sum_{i=1}^{m} E|y_i|^4 - 2b_i^2 E|y_i|^2 + b_i^4\end{aligned}$$

The expectation $E|y_i|^2$ is just the variance $\sigma_y^2 = \mathrm{Tr}\left(a_i a_i^H X\right)$ of y_i. The expectation $E|y_i|^4$ is the fourth moment of the complex-valued, normally distributed random variable y_i and is only a function of the mean and variance. It is $E|y_i|^4 = 2\sigma_y^4 = 2(\mathrm{Tr}\left(a_i a_i^H X\right))^2$ (see [Ree62]). Note that for a real-valued normally distributed random variable r, the fourth central moment is, in fact, different. It is $Er^4 = 3\sigma_r^4$. We will see that this is the key fact that makes the complex valued problem possess a superior quality bound. After we have completed the derivation, we will also derive the result without invoking the theorem in [Ree62], but derive it by expanding into real and imaginary parts and investigating the structures of the matrices. We have that

$$\gamma(X) = \sum_{i=1}^{m} 2(\mathrm{Tr}\left(a_i a_i^H X\right))^2 - 2b_i^2 \mathrm{Tr}\left(a_i a_i^H X\right) + b_i^4$$

The optimal value ν^* of the convex semidefinite programming problem

$$\nu^* = \min_{X \succeq 0} \quad \sum_{i=1}^{m} 2(\mathrm{Tr}\left(a_i a_i^H X\right))^2 - 2b_i^2 \mathrm{Tr}\left(a_i a_i^H X\right) + b_i^4, \qquad \text{(UPPER)}$$

must be an upper bound for μ since sampling u from the optimal X^* of (UPPER), substituting and evaluating the primal objective function must at some point minorize $\gamma(X^*)$ which equals ν^*.

From this we have two semidefinite programming problems one representing a lower bound on the objective of the magnitude-squared least-squares problem, and one representing an upper bound. By looking at the duals of these two problems, we can see that their optimal objectives are functions of one another. In fact, it holds that

$$\nu^* - \sum_{i=1}^{m} b_i^4 = \frac{1}{2}\left(\nu_* - \sum_{i=1}^{m} b_i^4\right). \qquad (6.10)$$

The above is a simple consequence of the differing quadratic coefficients in the very similar semidefinite programs. This motivates us to disregard the inconsequential constant term $\sum b_i^4$ in the original primal problem (MSLS). All of the arguments still hold. Define the following:

$$\tilde{\mu} = \min_{x \in \mathbb{C}^n} \sum_{i=1}^{m} (|a_i^H x|^2 - b_i^2)^2 - b_i^4, \tag{6.11}$$

$$\tilde{\nu}_* = \min_{X \succeq 0} \quad \sum_{i=1}^{m} (\mathrm{Tr}\left(a_i a_i^H X\right))^2 - 2b_i^2 \mathrm{Tr}\left(a_i a_i^H X\right) \tag{6.12}$$

$$\tilde{\nu}^* = \min_{X \succeq 0} \quad \sum_{i=1}^{m} 2(\mathrm{Tr}\left(a_i a_i^H X\right))^2 - 2b_i^2 \mathrm{Tr}\left(a_i a_i^H X\right) \tag{6.13}$$

The values $\tilde{\nu}^*$, and $\tilde{\nu}_*$ must both be nonpositive because $X = 0$ is feasible for (6.12) and (6.13). By the same expectation based arguments presented above, we can establish the following (all equivalent) bounds on these nonpositive values:

$$\tilde{\nu}_* \leq \quad \tilde{\mu} \quad \leq \frac{1}{2}\tilde{\nu}_* \tag{BOUND1}$$

$$2\tilde{\nu}^* \leq \quad \tilde{\mu} \quad \leq \tilde{\nu}^*. \tag{BOUND2}$$

Reformulating the bounds in the same style as Goemans and Williamson, we have proved that we have a method of obtaining a solution with objective equal to $\nu_{\mathrm{msls}} \equiv (1/2)\tilde{\nu}_*$ such that

$$\nu_{\mathrm{msls}} \leq \quad \alpha\bar{\mu} \quad \leq 0 \tag{BOUND3}$$

$$\text{where} \quad \alpha = \frac{1}{2}.$$

We see that the bound ($\alpha = 0.5$) is not as tight as it is for Goemans and Williamson's MAX CUT problem ($\alpha = 0.878$), nor is it as good as the bounds obtained for Nesterov's problem ($\alpha = 0.637$) or for the complex-valued problem studied by Zhang and Huang ($\alpha = 0.785$), which is inherited by the magnitude least-squares equivalent problem (MLS.A). Nevertheless it is an achievement to be able to prove the bound above, as it places this

method in a small class of methods that possess such guarantees, improving our perception of it as being something different than a simple heuristic.

We also see from the above analysis that the complex problem (MSLS) possesses a bound that is better than its real counterpart. This is a consequence of the differing fourth moments associated with complex and real normal distributions. The result parallels the improvement seen in the complex counterpart of Nesterov's problem as studied by Zhang and Huang. We will investigate the details of this next.

6.6 MSLS Quality Bound Derived by Complex to Real Conversion

We can derive the result in the previous section by expanding the problem in real and imaginary parts. This allows us to see just how the structure associated with the complex variables is responsible for the improved bound. We will begin by expanding (MSLS) in real and imaginary parts.

$$\mu = \min_{x\in\mathbb{C}^n} \quad \sum_{i=1}^{m}(|a_i x|^2 - b_i^2)^2 \tag{MSLS}$$

$$= \min_{x_R,x_I\in\mathbb{R}^n} \quad \sum_{i=1}^{m}\left(\begin{bmatrix}x_R\\ x_I\end{bmatrix}\begin{bmatrix}(a_i)_R & -(a_i)_I\\ (a_i)_I & (a_i)_R\end{bmatrix}\begin{bmatrix}(a_i)_R^T & (a_i)_I^T\\ -(a_i)_I^T & (a_i)_R^T\end{bmatrix}\begin{bmatrix}x_R\\ x_I\end{bmatrix} - b_i^2\right)^2 \tag{MSLS.Real}$$

A semidefinite relaxation for (MSLS.Real) is the following:

$$\begin{aligned}\nu_{r*} = \min_{X_1,X_2,X_3} \quad & \sum_{i=1}^{m}\left(\mathrm{Tr}^2\left(\begin{bmatrix}(a_i)_R & -(a_i)_I\\ (a_i)_I & (a_i)_R\end{bmatrix}\begin{bmatrix}(a_i)_R^T & (a_i)_I^T\\ -(a_i)_I^T & (a_i)_R^T\end{bmatrix}\begin{bmatrix}X_1 & X_2\\ X_2^T & X_3\end{bmatrix}\right) - b_i^2\right)^2.\\ \text{s.t.} \quad & \begin{bmatrix}X_1 & X_2\\ X_2^T & X_3\end{bmatrix} \succeq 0\end{aligned} \tag{MSLS.PReal}$$

Because of the block structure of the coefficient matrices

$$\begin{bmatrix}(a_i)_R & -(a_i)_I\\ (a_i)_I & (a_i)_R\end{bmatrix}\begin{bmatrix}(a_i)_R^T & (a_i)_I^T\\ -(a_i)_I^T & (a_i)_R^T\end{bmatrix},$$

we can, using the same ideas proved in Appendix A, reformulate the relaxation so that the decision variables possess the same block structure. Said a different way, without loss of

generality, we can assume that

$$\begin{bmatrix} X_1 & X_2 \\ X_2^T & X_3 \end{bmatrix} \quad \text{is such that}$$
$$2X_R \equiv X_1 = X_3 = X_1^T = X_3^T \quad \text{and}$$
$$2X_I \equiv X_2 = -X_2^T$$

Thus, (MSLS.PReal) can be rewritten as

$$\begin{aligned} \nu_{r*} = \min_{X_R, X_I} \quad & \sum_{i=1}^{m} \left(\mathrm{Tr}^2 \left(\begin{bmatrix} (a_i)_R & -(a_i)_I \\ (a_i)_I & (a_i)_R \end{bmatrix} \begin{bmatrix} (a_i)_R^T & (a_i)_I^T \\ -(a_i)_I^T & (a_i)_R^T \end{bmatrix} \begin{bmatrix} \frac{1}{2}X_R & \frac{1}{2}X_I \\ -\frac{1}{2}X_I & \frac{1}{2}X_R \end{bmatrix} \right) - b_i^2 \right)^2. \\ \text{s.t.} \quad & \begin{bmatrix} X_R & X_I \\ -X_I & X_R \end{bmatrix} \succeq 0 \end{aligned}$$

In line with the logic in the previous section we will sample a random vector $\begin{bmatrix} u_R^T & u_I^T \end{bmatrix}^T$ from real-valued normal distribution, and calculate the expectation of the objective function evaluated at $\begin{bmatrix} u_R^T & u_I^T \end{bmatrix}^T$. We will calculate

$$\gamma(X_R, X_I) = E \sum_{i=1}^{m} \left(\begin{bmatrix} u_R \\ u_I \end{bmatrix} \begin{bmatrix} (a_i)_R & -(a_i)_I \\ (a_i)_I & (a_i)_R \end{bmatrix} \begin{bmatrix} (a_i)_R^T & (a_i)_I^T \\ -(a_i)_I^T & (a_i)_R^T \end{bmatrix} \begin{bmatrix} u_R \\ u_I \end{bmatrix} - b_i^2 \right)^2 \tag{6.14}$$
$$\begin{bmatrix} u_R \\ u_I \end{bmatrix} \sim \mathcal{N} \left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \frac{1}{2} \begin{bmatrix} X_R & X_I \\ -X_I & X_R \end{bmatrix} \right)$$

Define the (2×1)-vectors y_i as

$$y_i \equiv \begin{bmatrix} (y_i)_R \\ (y_i)_I \end{bmatrix} \equiv \begin{bmatrix} (a_i)_R^T & (a_i)_I^T \\ -(a_i)_I^T & (a_i)_R^T \end{bmatrix} \begin{bmatrix} u_R \\ u_I \end{bmatrix}$$

The random vectors y_i are distributed normally with zero mean and covariance matrix

$$\begin{aligned} E \begin{bmatrix} (y_i)_R^2 & (y_i)_R(y_i)_I \\ (y_i)_R(y_i)_I & (y_i)_I^2 \end{bmatrix} &= E \begin{bmatrix} (a_i)_R^T & (a_i)_I^T \\ -(a_i)_I^T & (a_i)_R^T \end{bmatrix} \begin{bmatrix} u_R \\ u_I \end{bmatrix} \begin{bmatrix} u_R \\ u_I \end{bmatrix}^T \begin{bmatrix} (a_i)_R & -(a_i)_I \\ (a_i)_I & (a_i)_R \end{bmatrix} \\ &= \begin{bmatrix} (a_i)_R^T & (a_i)_I^T \\ -(a_i)_I^T & (a_i)_R^T \end{bmatrix} E \left(\begin{bmatrix} u_R \\ u_I \end{bmatrix} \begin{bmatrix} u_R \\ u_I \end{bmatrix}^T \right) \begin{bmatrix} (a_i)_R & -(a_i)_I \\ (a_i)_I & (a_i)_R \end{bmatrix} \\ &\equiv \frac{1}{2} \begin{bmatrix} (a_i)_R^T & (a_i)_I^T \\ -(a_i)_I^T & (a_i)_R^T \end{bmatrix} \begin{bmatrix} X_R & X_I \\ -X_I & X_R \end{bmatrix} \begin{bmatrix} (a_i)_R & -(a_i)_I \\ (a_i)_I & (a_i)_R \end{bmatrix} \end{aligned} \tag{6.15}$$

Because of the symmetries of the matrices, we can reduce (6.15) to

$$\begin{aligned} Ey_iy_i^T &= \begin{bmatrix} E((y_i)_R^2) & E((y_i)_R(y_i)_I) \\ E((y_i)_R(y_i)_I) & E((y_i)_I^2) \end{bmatrix} \\ &= \frac{1}{2} \begin{bmatrix} \begin{bmatrix} (a_i)_R \\ (a_i)_I \end{bmatrix}^T \begin{bmatrix} X_R & X_I \\ -X_I & X_R \end{bmatrix} \begin{bmatrix} (a_i)_R \\ (a_i)_I \end{bmatrix} & 0 \\ 0 & \begin{bmatrix} (a_i)_R \\ (a_i)_I \end{bmatrix}^T \begin{bmatrix} X_R & X_I \\ -X_I & X_R \end{bmatrix} \begin{bmatrix} (a_i)_R \\ (a_i)_I \end{bmatrix} \end{bmatrix}. \end{aligned}$$

From this, we can compute several important expectations that enter into the evaluation of (6.14). The first is that

$$E\left((y_i)_R^2 + (y_i)_I^2\right) = \begin{bmatrix}(a_i)_R \\ (a_i)_I\end{bmatrix}^T \begin{bmatrix}X_R & X_I \\ -X_I & X_R\end{bmatrix} \begin{bmatrix}(a_i)_R \\ (a_i)_I\end{bmatrix}.$$

The evaluation of the second expectation will require an identity proved in [Pap91], namely that for a bivariate normally distributed random vector $\begin{bmatrix}z_1 & z_2\end{bmatrix}^T$ with covariance matrix $\begin{bmatrix}E(z_1^2) & E(z_1 z_2) \\ E(z_1 z_2) & E(z_2^2)\end{bmatrix}$, we can express the fourth central moments as

$$\begin{aligned} E(z_1^4) &= 3E^2(z_1^2) \\ E(z_2^4) &= 3E^2(z_2^2) \\ E(z_1^2 z_2^2) &= 2E(z_1^2)E(z_2^2) + 4E^2(z_1 z_2) \end{aligned} \tag{6.16}$$

Applying this to y_i, we have that

$$E\left(\left((y_i)_R^2 + (y_i)_I^2\right)^2\right) = 2\left(\begin{bmatrix}(a_i)_R \\ (a_i)_I\end{bmatrix}^T \begin{bmatrix}X_R & X_I \\ -X_I & X_R\end{bmatrix} \begin{bmatrix}(a_i)_R \\ (a_i)_I\end{bmatrix}\right)^2.$$

We see here that if the components of y_i were the real and imaginary parts of a complex number, we have essentially derived Reed's lemma about the fourth moment of a complex normal distribution. We have shown that for our special case,

$$E|y_i|^4 = 2E^2(|y_i|^2),$$

which differs from the real identity (6.16) proved in Papoulis [Pap91]. Finally we have the framework to evaluate the expected objective function (6.14).

$$\begin{aligned} \gamma(X_R, X_I) = \sum_{i=1}^{m} \; & 2\mathrm{Tr}^2\left(\begin{bmatrix}(a_i)_R & -(a_i)_I \\ (a_i)_I & (a_i)_R\end{bmatrix} \begin{bmatrix}(a_i)_R^T & (a_i)_I^T \\ -(a_i)_I^T & (a_i)_R^T\end{bmatrix} \begin{bmatrix}X_R & X_I \\ -X_I & X_R\end{bmatrix}\right) \\ & - 2b_i^2 \mathrm{Tr}\left(\begin{bmatrix}(a_i)_R & -(a_i)_I \\ (a_i)_I & (a_i)_R\end{bmatrix} \begin{bmatrix}(a_i)_R^T & (a_i)_I^T \\ -(a_i)_I^T & (a_i)_R^T\end{bmatrix} \begin{bmatrix}X_R & X_I \\ -X_I & X_R\end{bmatrix}\right) + b_i^4 \end{aligned}$$

As in our (more elegant) complex derivation of the quality estimate, the solution to the problem

$$\begin{aligned}\nu_r^* \equiv \min_{X_R, X_I} \quad & \sum_{i=1}^{m} \; 2\mathrm{Tr}^2\left(\begin{bmatrix}(a_i)_R & -(a_i)_I \\ (a_i)_I & (a_i)_R\end{bmatrix}\begin{bmatrix}(a_i)_R^T & (a_i)_I^T \\ -(a_i)_I^T & (a_i)_R^T\end{bmatrix}\begin{bmatrix}X_R & X_I \\ -X_I & X_R\end{bmatrix}\right) \\ & \quad - 2b_i^2\mathrm{Tr}\left(\begin{bmatrix}(a_i)_R & -(a_i)_I \\ (a_i)_I & (a_i)_R\end{bmatrix}\begin{bmatrix}(a_i)_R^T & (a_i)_I^T \\ -(a_i)_I^T & (a_i)_R^T\end{bmatrix}\begin{bmatrix}X_R & X_I \\ -X_I & X_R\end{bmatrix}\right) + b_i^4 \\ \text{s.t.} \quad & \begin{bmatrix}X_R & X_I \\ -X_I & X_R\end{bmatrix} \succeq 0\end{aligned}$$

represents an upper bound for (MSLS.Real), and is related to μ and ν_{r*} in exactly the same way as in (6.10), namely that

$$\nu_r^* - \sum_{i=1}^{m} b_i^4 = \frac{1}{2}\left(\nu_{r*} - \sum_{i=1}^{m} b_i^4\right). \tag{6.17}$$

This yields the same quality estimate as derived previously.

6.7 Discussion

We have shown that in the magnitude-squared least-squares problem, we can prove a quality bound in the same spirit as Nesterov's proof for quadratic problems with constraints on squared variables. The complex structure of our problem allows us to prove a tighter bound than what is provable for a real-valued problem of the same form. This phenomenon was also seen in Zhang and Huang's analysis of the complex version of Nesterov's problem. Zhang and Huang's proof required the evaluation of complicated integrals, obscuring intuition about the reasons for the superior bound. Our analysis of the more simple (because the problem doesn't have constraints) magnitude-squared least-squares problem reveals that it is the differing fourth central moments between real-valued and complex-valued normal distributions that could be interpreted to be the reason.

The bounds proved in this chapter not only are of theoretical interest, but their existence suggests that perhaps the methods will work well in practice. In the final chapters of this dissertation, we will compare solution methods, and look at practical applications of the magnitude least-squares and magnitude-squared least-squares problems.

Chapter 7

Comparisons of Methods for MLS and MSLS

7.1 Introduction

In this short chapter, we will solve randomly generated instances of both the magnitude least-squares (MLS) problem and the magnitude-squared least-squares (MSLS) problem. In the previous chapters we have examined the problems in great detail, proving much about the expected performance of, in particular, the semidefinite programming methods.

The information presented in this chapter is condensed and important. It shows that our proved quality estimates appear to be valid, and it gives us an intuitive feel for how the methods perform in practice. We will see specifically that the best method for the solution of these problems is to use the semidefinite relaxation to find a starting point for a local method. The charts presented also show that a locally found solution with a random starting point is usually suboptimal, as is a semidefinite relaxation solution without additional local minimization.

7.2 Presentation and Analysis of the Data

The charts shown in Figures 7.1 and 7.2 show the results of solving 100 randomly generated problem instances using the methods under discussion. For each problem class (either MLS or MSLS), we calculate five values: the upper and lower bounds present in the quality proofs presented in Chapter 6, the objective achieved by a solution found with the local

method using a random starting point, the objective achieved by a solution found using the semidefinite relaxation and rounding scheme derived from the quality proofs, and finally the objective achieved by a solution found using both methods in conjunction, i.e., using the semidefinite programming solution as a starting point for the local method.

For visual clarity, we have sorted the random trials in decreasing order of the objective attained by the semidefinite relaxation method. This allows us to see clearly the relationships among the values.

We see that the upper and lower bounds do indeed hold. This is reassuring considering that we rigorously proved this fact. We see though that the upper bound in both cases (MLS and MSLS) is not particularly tight. The solutions found by all three methods (local, SDP, and both in conjunction) are much closer to the lower bound than to the upper. This doesn't necessarily suggest that the bound isn't tight in general. There could exist certain outlying problem instances whose solutions attain the upper bound. Discovering such an example would be interesting and worth exploration. If such an example were found, it would necessarily prove that the bounds we've calculated are actually the best possible. On the other hand, it is not hard to construct a problem instance that achieves the lower bound.

We have inspected many such charts for different problem dimensions. The charts presented are representative of the typical behaviors. As the number of variables approaches the dimension of b in the problems, the objectives become very small representing an error-free magnitude least-square fit. When this happens, the underlying geometry becomes simpler resulting in fewer local minima. All the methods (but not the upper bound necessarily) tend to find the correct solution. A similar phenomenon occurs if the number of variables is very very small compared to the dimension of b.

For the majority of cases, there exist local minima as is evident by inspecting the figures. The solution obtained by using both methods in conjunction is always guaranteed to be superior to the SDP solution, but we see that almost without fail, it finds a better solution that the one obtained using a randomly generated starting point. In the MLS case, more often than not the local solution is significantly worse than the solution obtained using the

semidefinite relaxation alone. The opposite is true in the MSLS case. But in both cases, the best approach is clearly to use both the SDP and the local method in conjunction.

7.3 Summary

We have inspected the behavior of the methods discussed with respect to randomly generated problem instances. This approach gives us an intuitive feel for the performance of the methods. It is clearly the case that (a) the bounds we proved in Chapter 6 are valid, and (b) the best optimization strategy with respect to finding the best solution for these problems is to use the semidefinite relaxation as a means to find a starting point for a local method.

The randomly generated problem instances may or may not be representative of problem instances that we encounter in practical engineering problems. Problems derived from applications are notorious for possessing structure that's absent from random problem instances. Nevertheless, the data collected from such experiments is of value and does give us a richer understanding of the problem.

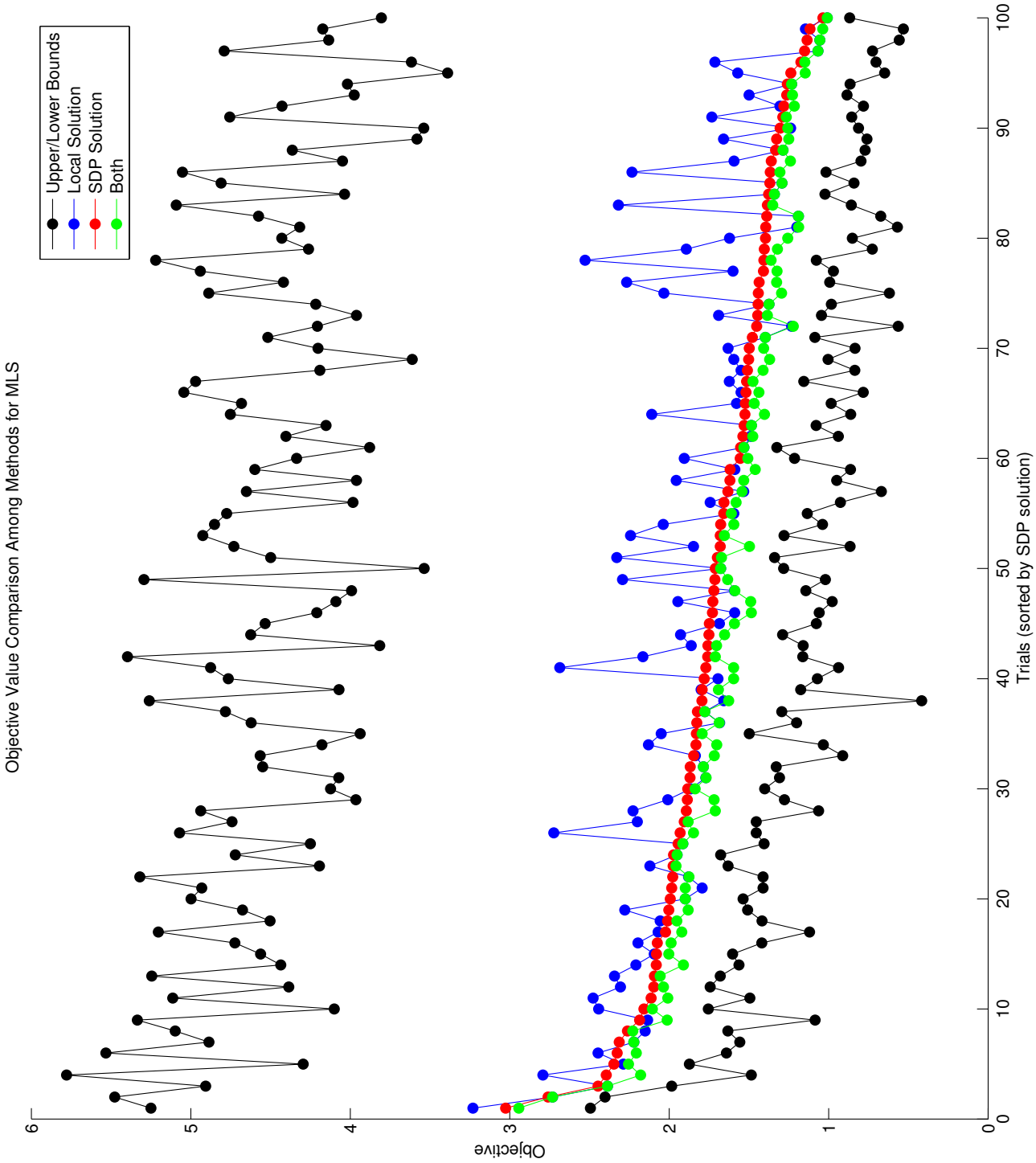


Figure 7.1: Achieved objectives using the methods for randomly generated MLS problem instances ($A \in \mathbb{C}^{50\times 10}$)

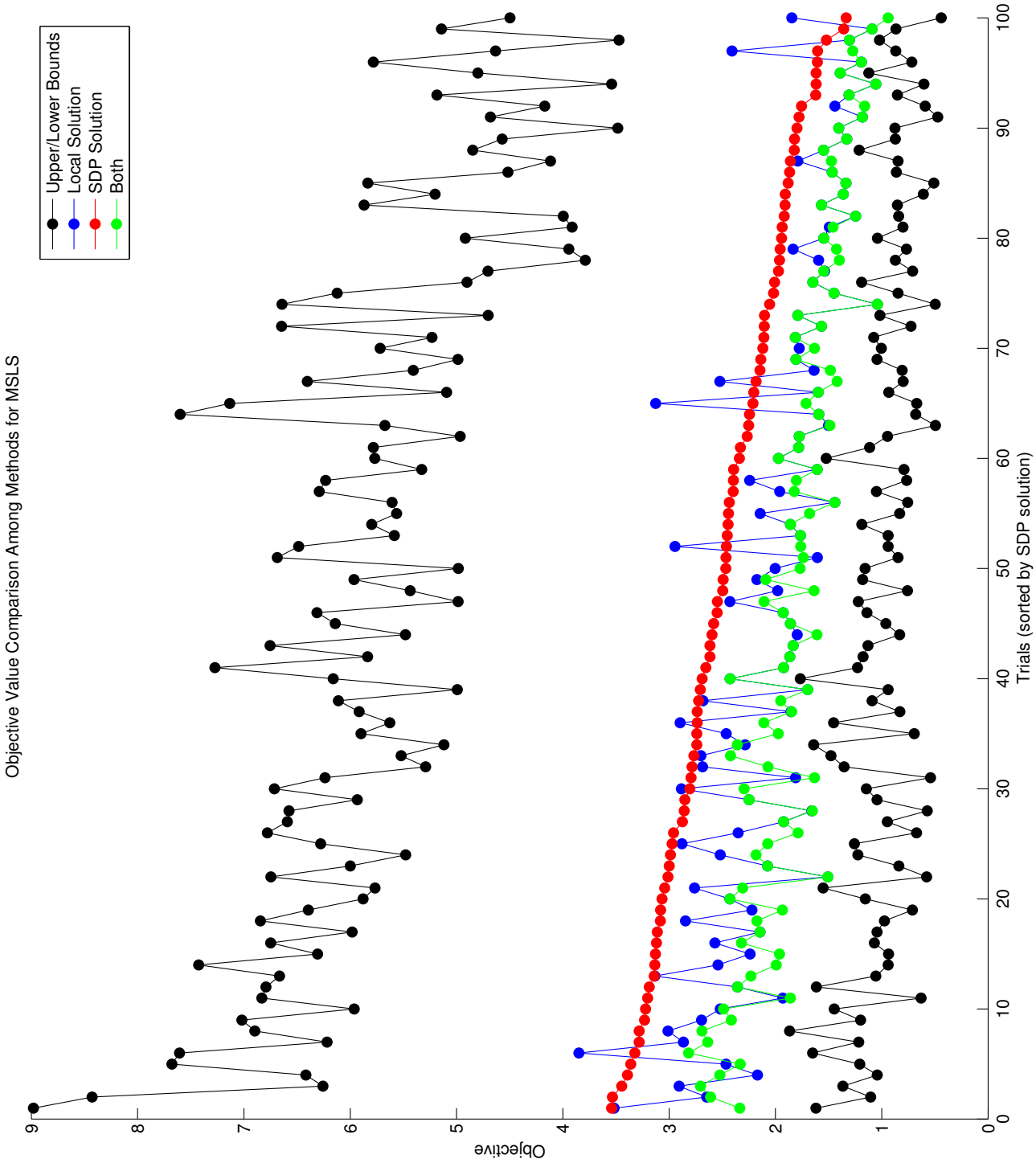


Figure 7.2: Achieved objectives using the methods for randomly generated M2LS problem instances ($A \in \mathbb{C}^{50\times 10}$)

Chapter 8

Application Examples and Problem Variations

8.1 Introduction

In this final technical chapter, we will solve several engineering problems using the the methods described. In the course of doing this we will also discuss slight variations on the magnitude least-squares problem that we can solve using the same or similar methods. These variations include solving a magnitude least-squares problem subject to linear constraints on the variables (including a very useful constraint that the variables be real-valued), and solving the problem with a constraint on total or average power.

8.2 Multidimensional Filter Design

The first example represents the initial and primary motivation for studying this problem: the design of multidimensional filters. As mentioned previously, the design of multidimensional filters is significantly more difficult than that of one-dimensional filters.

Figure 8.1 shows the magnitude of a complex two-dimensional filter of size 15×15, designed to have a magnitude response that resembles a pyramid. The target response was sampled on 31×31 equally spaced grid, making the dimensions of the A matrix 961×225. We can see that the achieved pattern matches the target closely. The design procedure for this example was as simple as could be, armed with the magnitude least-squares machinery, and would have been difficult using classical methods.

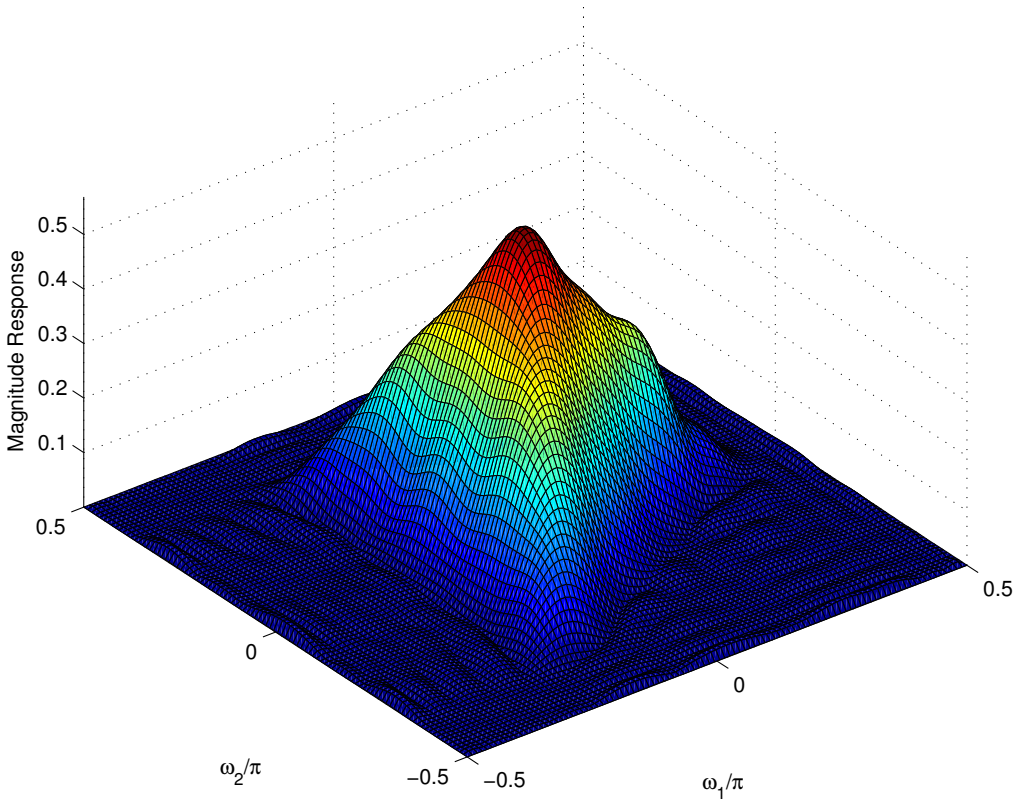


Figure 8.1: Magnitude frequency response of 15×15 two–dimensional filter.

8.3 Real-Valued Multidimensional Filter Design

The filter designed in Section 8.2 had complex-valued taps. Sometimes complex variables are natural, as is the case with the design of spatial filters, but more often we desire the filter taps to have real coefficients. We can solve this problem as well the more general version: magnitude least-squares subject to linear constraints on the variables:

$$\begin{aligned} \min_{x \in \mathbb{C}^n} \quad & \sum_{i=1}^{m} \left(|A_i x| - b_i \right)^2 \\ \text{s.t.} \quad & Dx = f \end{aligned} \tag{8.1}$$

The critical element that allows us to solve this variation is the fact that the solution to a linear least-squares problem subject to linear constraints is linear in b. The problem

$$\begin{aligned} p^\star = \min_{x} \quad & \|Ax - b\| \\ \text{s.t.} \quad & Dx = f \end{aligned} \tag{8.2}$$

has solution $x^\star$ satisfying

$$\begin{bmatrix} A^H A & D^H \\ D & 0 \end{bmatrix} \begin{bmatrix} x^\star \\ \nu^\star \end{bmatrix} = \begin{bmatrix} 2A^H b \\ f \end{bmatrix}.$$

This result can be derived easily using optimality conditions. See [BV04]. We can use the same algebraic tricks we employed when reformulating the standard magnitude least-squares problem, that allowed us to cast the problem as a quadratic subject to constraints on squared variables. We first introduce a vector c of decision variables each constrained to have modulus 1:

$$\begin{aligned} p^\star = \min_x \quad & \|Ax - Bc\| \\ \text{s.t.} \quad & Dx = f \\ & |c_i|^2 = 1 \quad \forall i \end{aligned} \tag{8.3}$$

Because of the above result, we have an analytic solution for $x^\star$ that is linear in c. The variable $x^\star$ satisfies

$$x^\star = \begin{bmatrix} I & 0 \end{bmatrix} \begin{bmatrix} 2A^H A & D^H \\ D & 0 \end{bmatrix}^{-1} \begin{bmatrix} 2A^H B & 0 \\ 0 & f \end{bmatrix} \begin{bmatrix} c \\ 1 \end{bmatrix}, \tag{8.4}$$

$$\text{assuming} \quad \begin{bmatrix} 2A^H A & D^H \\ D & 0 \end{bmatrix} \quad \text{has full rank.} \tag{8.5}$$

Substituting into the objective function, we have the formulation

$$\begin{aligned} p^\star = \min_c \quad & \begin{bmatrix} c \\ 1 \end{bmatrix}^H W \begin{bmatrix} c \\ 1 \end{bmatrix}, \\ \text{s.t.} \quad & |c_i|^2 = 1 \quad \forall i \end{aligned} \tag{8.6}$$

$$\begin{aligned} \text{where } W \equiv & \begin{bmatrix} 2B^T A & 0 \\ 0 & f^H \end{bmatrix} \begin{bmatrix} 2A^H A & D^H \\ D & 0 \end{bmatrix}^{-1} \begin{bmatrix} A^H A & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} 2A^H A & D^H \\ D & 0 \end{bmatrix}^{-1} \begin{bmatrix} 2A^H B & 0 \\ 0 & f \end{bmatrix} \\ & - \begin{bmatrix} 2B^T A & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} 2A^H A & D^H \\ D & 0 \end{bmatrix}^{-1} \begin{bmatrix} 2A^H B & 0 \\ 0 & f \end{bmatrix} + \begin{bmatrix} B^T B & 0 \\ 0 & 0 \end{bmatrix} \end{aligned}$$

The most important problems of this type are where we constrain the variable x to be real-valued. The formulation of a problem of this type simply requires us to first cast (8.2) in terms of the real and imaginary parts of x. We have shown in Appendix A how to do this in general. We will now design a nonuniformly-spaced-tap FIR filter with real coefficients.

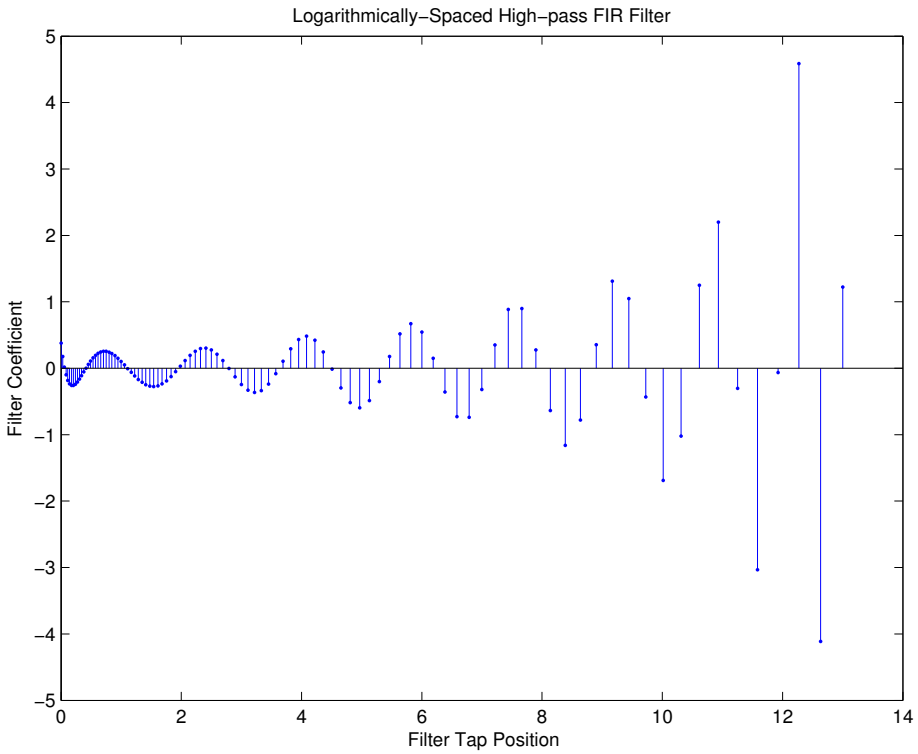


Figure 8.2: A logarithmically-spaced FIR filter

The methods for designing one-dimensional filters with magnitude specifications that rely on spectral factorization cannot be applied to the case of nonuniformly-spaced filters. This is what motivates us to formulate the problems as magnitude least-squares problems.

Figure 8.2 and Figure 8.4 show logarithmically spaced FIR filters that has been designed using our technique to have a high-pass characteristic and a low-pass characteristic. Figure 8.3 and Figure 8.5 show the resultant frequency responses. It is interesting to inspect the solution – as the spacing among the taps increases, so does the magnitude of the taps. It seems that the summed energy per unit length of the filter remains nearly constant.

8.4 Approximate Polynomial Factorization

The magnitude least-squares formulation can be used to solve a variety of polynomial factorization problems. One very common use of this is spectral factorization in signal processing. In that case, the spectral factorization theorem, discussed in Chapter 4, helps us

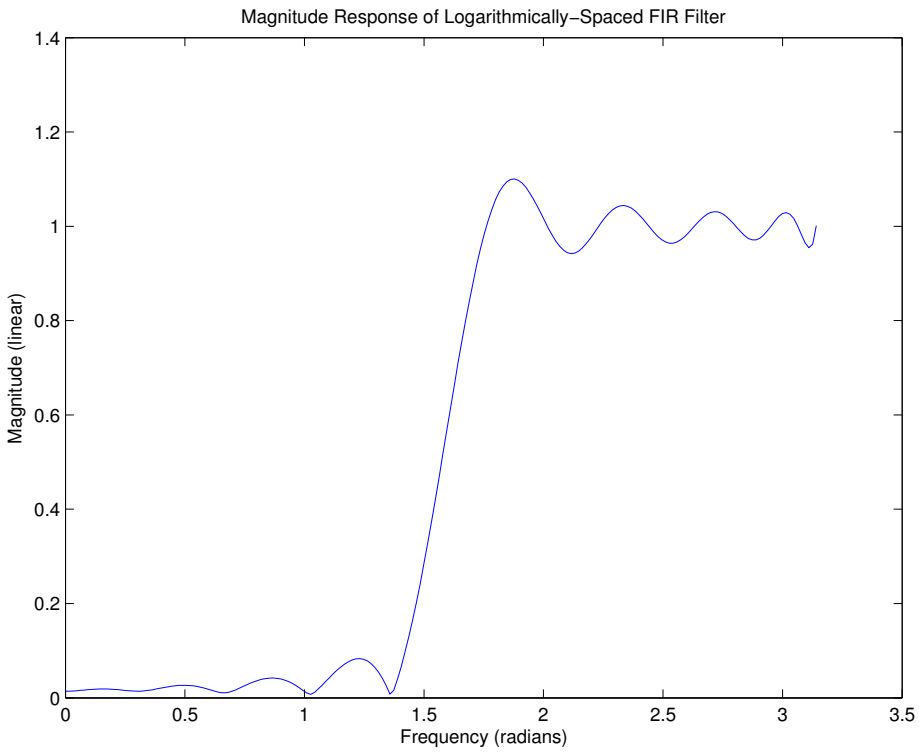


Figure 8.3: The high-pass response of a logarithmically-spaced FIR filter

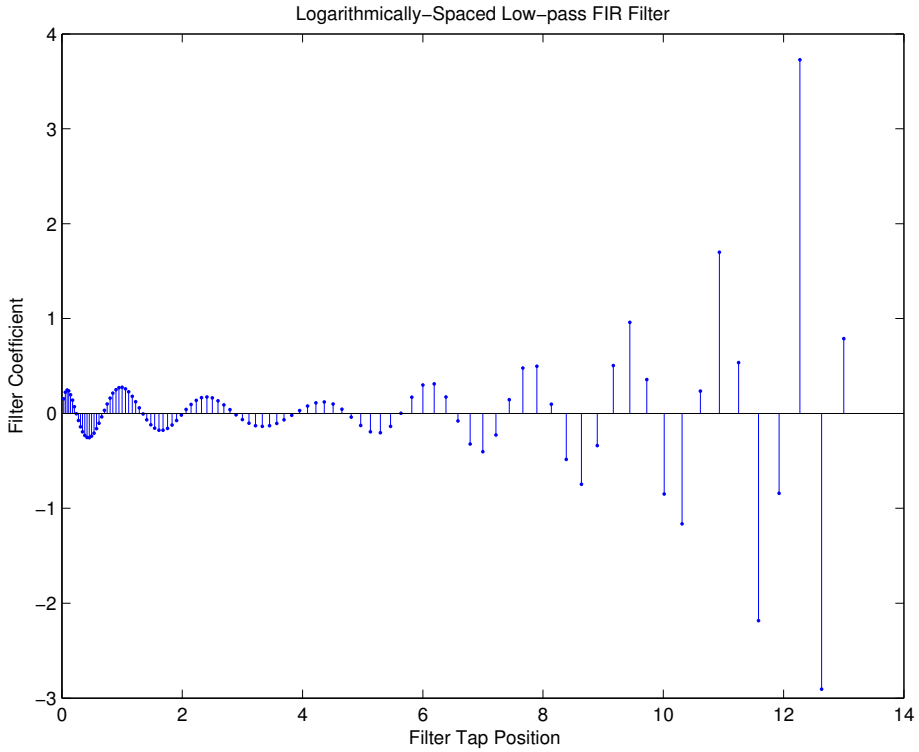


Figure 8.4: A logarithmically-spaced FIR filter

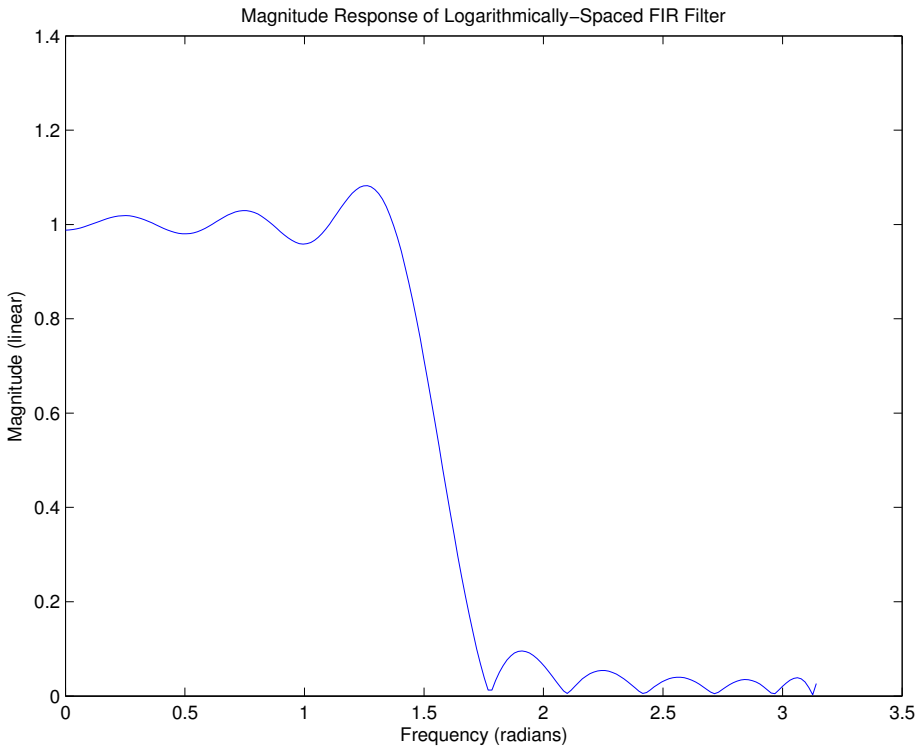


Figure 8.5: The low-pass response of a logarithmically-spaced FIR filter

determine when a factorization exists and helps us calculate the spectral factor. For polynomials of dimension greater than one, however, there is no such theorem. We can use the magnitude least-squares formulation to find a polynomial whose square closely matches (in the least-squares sense) that of a higher degree polynomial. This can be interpreted as a form of approximate polynomial factorization.

This topic is related to a field of study called *sums of squares programming*. The field has attracted the attention of many theoreticians and engineers because of the deep mathematical ideas it involves, and the numerous practical applications associated with it [Par01] [PPP02] [Las01]. The main idea is that the coefficients of a general polynomial, be it multidimensional, multivariate, or both, can be constrained to make the polynomial everywhere nonnegative by using semidefinite programming. A nonnegative polynomial, in turn, can be factored into a sum of (smaller) squared polynomials. Many engineering problems, particularly in the area of control systems and signal processing, are naturally formulated in these terms. Filter design is one such problem.

The factorization problem we are solving, in contrast to the sum of squares factorization, is more difficult. Given a higher-order polynomial p, we seek to find a single squared polynomial term q^2 that approximates $|p|$ in the least-squares sense. The sum of squares problem can be understood to be the discovery of a *sum* of squared terms $\sum q_i^2$ that approximate $|p|$.

8.4.1 Perfect Factorization

If the polynomial p does possess a perfect, single term factorization, then we can use MLS to find the factors. Take for example the trigonometric polynomial

$$\begin{aligned} p(e^{j\omega_1}, e^{j\omega_2}) = \; & 2.87 + 0.84\cos(\omega_1) + 1.06\cos(\omega_2) + 2.6\cos(\omega_1 - \omega_2) \\ & - 0.26\cos(2\omega_1 - \omega_2) + 0.2\cos(2\omega_2 - \omega_1) - 0.02\cos(2(\omega_1 - \omega_2)). \end{aligned} \tag{8.7}$$

It isn't obvious whether if this polynomial is expressible as the squared magnitude of another polynomial. We can find out by hypothesizing that there exists a polynomial $q(e^{j\omega_1}, e^{j\omega_2})$ such that $|q(e^{j\omega_1}, e^{j\omega_2})|^2 = p(e^{j\omega_1}, e^{j\omega_2})$ for all ω_1 and ω_2. Clearly this cannot ever be the case if the polynomial p is anywhere negative or anywhere nonreal. These are not sufficient conditions however. A graph of the polynomial p is shown in Figure 8.4.1. We hypothesize that there is a polynomial q of the form

$$q(e^{j\omega_1}, e^{j\omega_2}) = x_1 + x_2 e^{j\omega_1} + x_3 e^{j\omega_2} + x_4 e^{2j\omega_1} + x_5 e^{j(\omega_1+\omega_2)} + x_6 e^{2j\omega_2} \tag{8.8}$$

that is a factor of the polynomial p. We can set up a magnitude least-squares problem to help us answer the question. We form a matrix A where each column corresponds to a coefficient of q evaluated on a densely sampled grid of points in the domain of the polynomial ($(\omega_1, \omega_2) \in [-\pi, \pi] \times [-\pi, \pi]$). The target vector b for the problem consists of (the squareroots of) samples of the polynomial p.

If we solve this problem for the above polynomial, we indeed discover that p is factorable as

$$p(e^{j\omega_1}, e^{j\omega_2}) = |0 + 1.0e^{j\omega_1} + 1.3e^{j\omega_2} - 0.1e^{2j\omega_1} + 0.4e^{j(\omega_1+\omega_2)} + 0.1e^{2j\omega_2}|^2$$

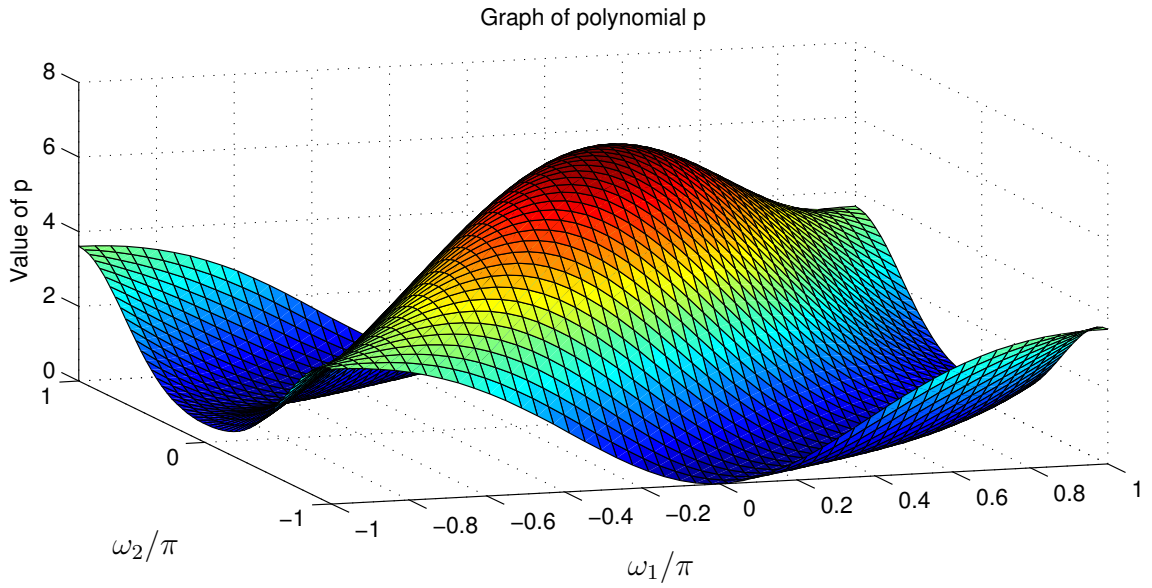


Figure 8.6: The polynomial p of Equation (8.7)

The answer is not unique. For example, if q is a solution, then so is $-q$. Most of the time the magnitude least-squares procedure will find the correct answer, but as is the case in most MLS problems, there can be several locally optimal points. Sometimes, if we are unlucky, we will discover a suboptimal solution.

8.4.2 Approximate Factorization

Even more useful and interesting than the problem of finding exact polynomial factorizations when they exist is the idea of finding approximate factorizations when they don't exist. The magnitude least-squares algorithm does just that. If the target polynomial is not factorable exactly, the magnitude least-squares algorithm will return a polynomial of specified dimension and degree that approximates the target. This result fits nicely with the work [PPP02] on sums of squares programming because that framework provides for us a way to constrain polynomials to be positive and real, which are necessary conditions for a target vector in a magnitude least-squares problem.

Another similar situation where the magnitude least-squares problem is useful is when we have inexact measurements of a polynomial, or repeated noisy measurements. The magnitude least-squares problem formulation lets us simultaneously estimate and factor

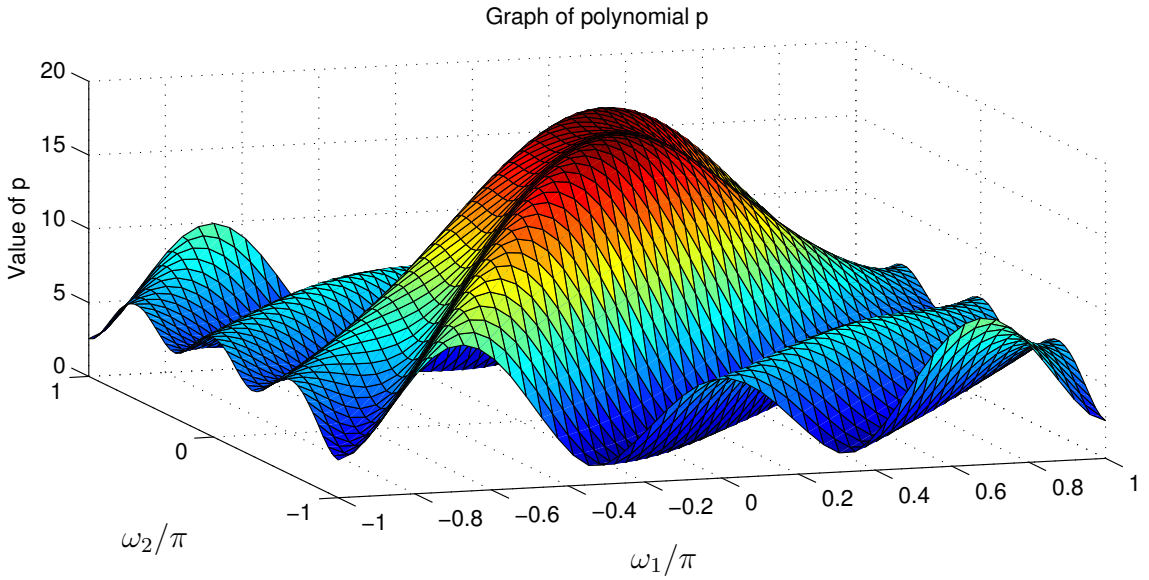


Figure 8.7: Polynomial p to factor

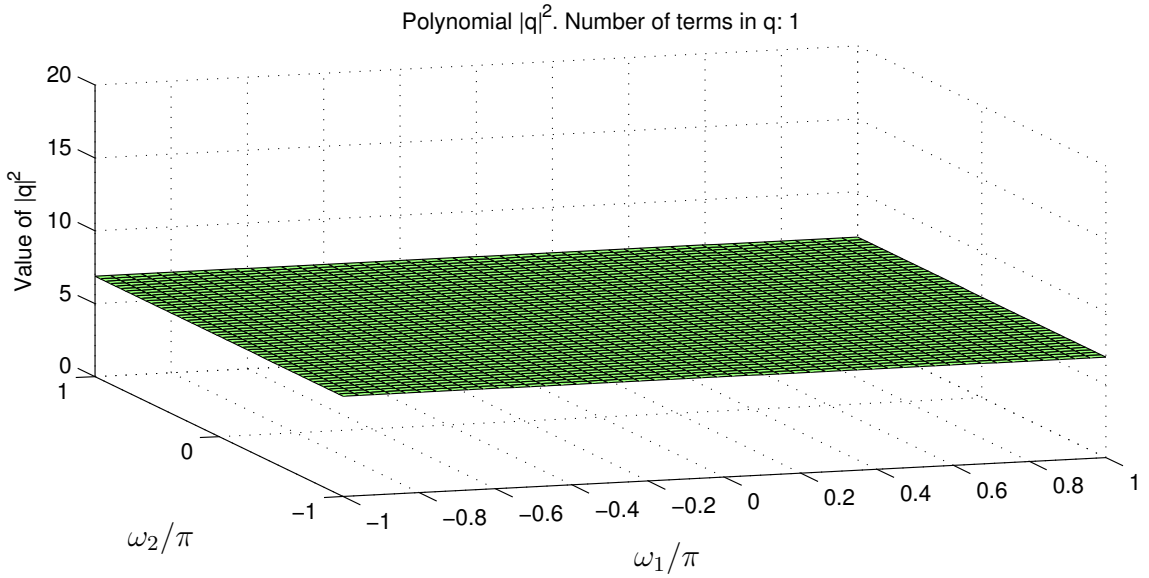


Figure 8.8: Single term factorization (constant)

the polynomial.

Figure 8.7 shows the magnitude of a complex polynomial p that does not possess a single polynomial factorization. We can use our procedure to find polynomials q whose squares approximate p for increasing degrees. Figures 8.8–8.12 show the increasingly good approximations as we increase the dimension and orders of the polynomial q.

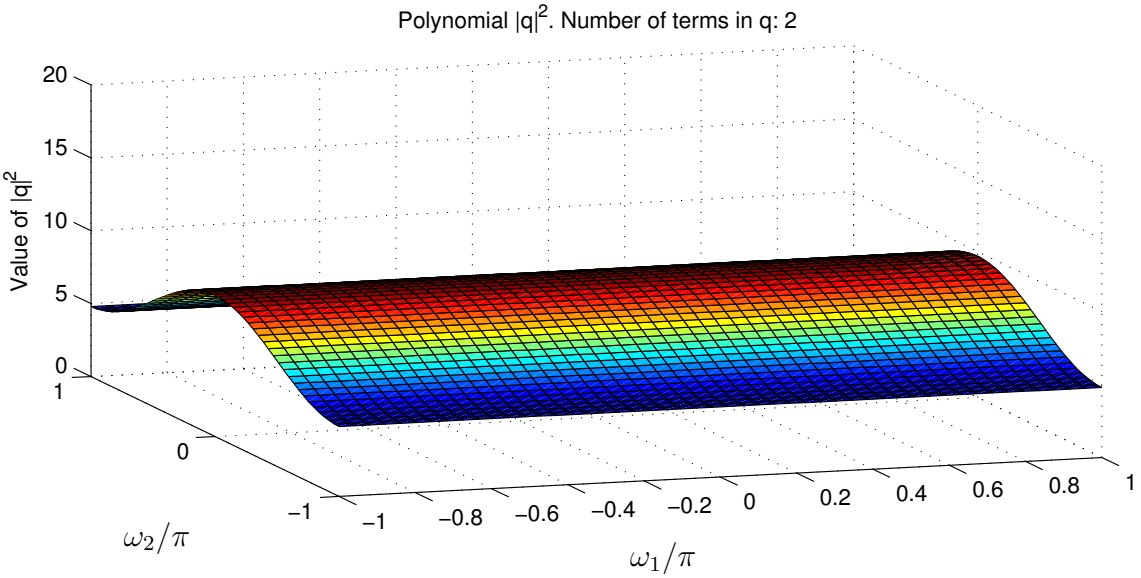


Figure 8.9: Magnitude of two term factor

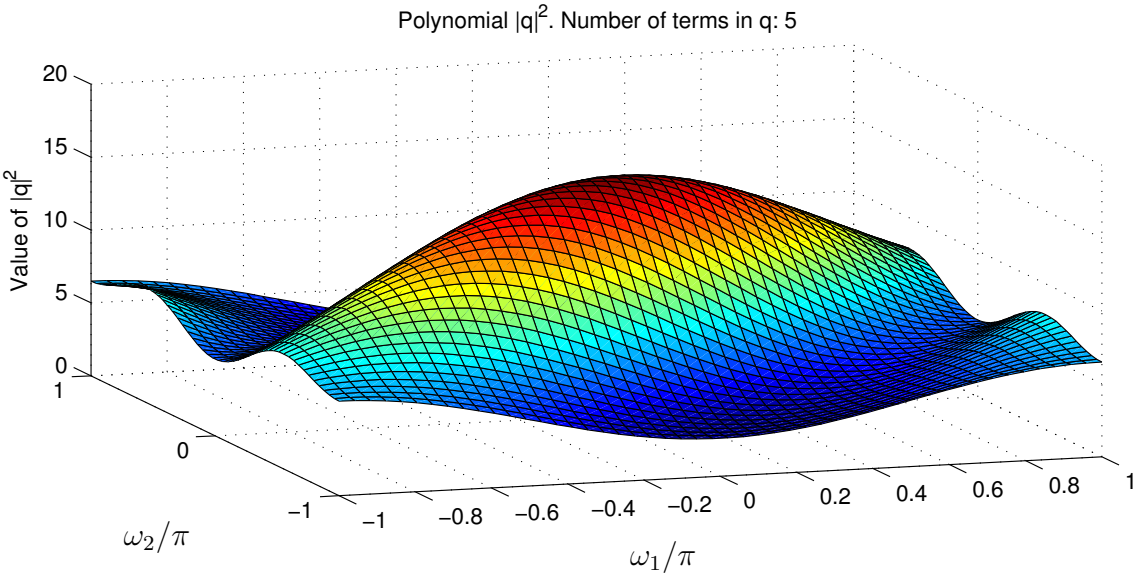


Figure 8.10: Magnitude of five term factor

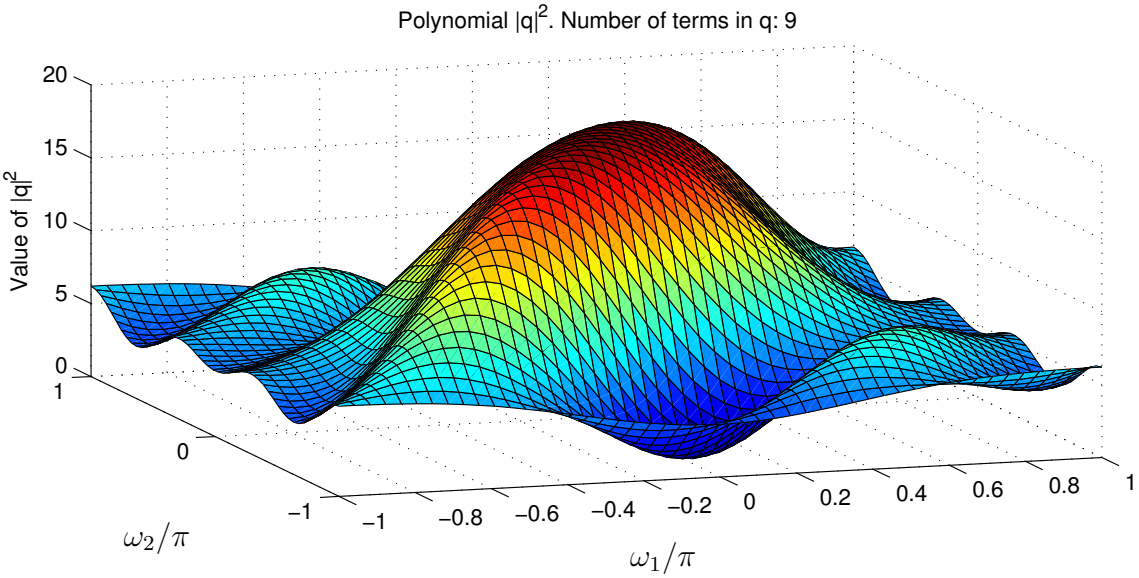


Figure 8.11: Magnitude of nine term factor

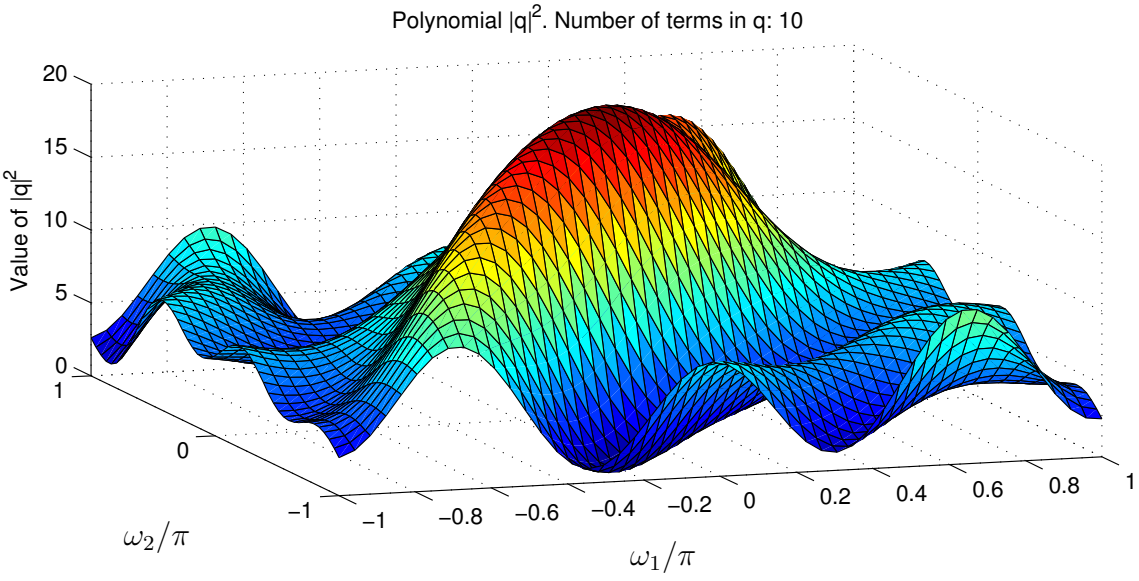


Figure 8.12: Magnitude of ten term factor

8.5 Radiation Pattern Synthesis

This example will demonstrate another variation of the magnitude least-squares problem that was motivated by the loudspeaker work at the Center for New Music and Audio Technologies (CNMAT). The loudspeaker system depicted in Figure 2.2 in Chapter 2 is used primarily for amplifying electronic instruments. Because the system possesses many loudspeakers (twelve), we can control the radiation patterns that the array produces.

A particularly pleasing (in the author's opinion) effect that can be created with the system is that of acoustic nulling. If we design the radiation patterns across frequency such that the listener is situated in an acoustic null, then he or she will predominantly hear the sound reflecting off the walls and other structures in the room. This can have an enveloping quality, and the effect could conceivably be used compositionally by the performer.

To achieve this effect, it is not enough to generate a null in exactly one direction. Ideally, we would like there to be a solid angle over which the radiated magnitude is zero or very close to zero. This would allow the listener to move around, and has the potential to include a larger audience. For this example, we will show one way of attacking the problem using an average power constrained version of MLS.

We are given a layout of array elements with ideal uniform individual radiation patterns, and wish to find complex coefficients that when applied to the signals at the elements, yield a desirable radiation pattern at a given frequency. Our criteria for a desirable pattern are that the total power of the pattern be equal to a constant, that the response be close to zero over a specified solid angle, and that at other angles, the magnitude response be constant. This is a different problem than the classical beamforming problem of nulling in different directions subject to unity responses in other particular directions. Our problem simultaneously holds the total power of the response to be a constant, while balancing the desire for a zero response in the solid angle with the desire for a uniform response in other directions. In the musical application above, this would be to ensure that the overall loudness of the array remains constant, and is spread evenly in directions other than the nulling angles.

In abstract terms, the problem can be formulated as

$$\begin{aligned} \min_{x} \quad & \sum_{i=1}^{m} (|A_i x| - b_i)^2 \\ \text{s.t.} \quad & x^H D^H D x = p. \end{aligned} \tag{8.9}$$

where A and D have rows corresponding to responses in different directions and b is the desired response. For this application, A and D are equal, and the desired response b is zero for directions in the nulling angle, and equal to a constant for other directions. The constant is chosen to make the overall desired pattern have total power equal to unity. Note that the problem differs from the regular MLS problem because we will strictly enforce that the resultant response have unity power $p = 1$. We will forumulate this problem so that we can solve it as a semidefinite relaxation of similar form as the magnitude least-squares problem. The solution we find will possess the same quality estimates as Nesterov's problem, i.e., the reformulation will take the form of a quadratic objective and constraints on squared variables.

The trick is to transform coordinate systems. Let $U\Lambda U^H = D^H D$ be the eigenvalue decomposition of $D^H D$. Then we can define a new set of variables in a rotated coordinate system, $\tilde{x} = U^H x$. This means $x = U\tilde{x}$. We can now introduce the unity-modulus variables c (as discussed in Chapter 5), and express the problem in terms of $\tilde{x}$:

$$\begin{aligned} \min_{\tilde{x},c} \quad & \|AU\tilde{x} - Bc\| \\ \text{s.t.} \quad & \sum_{i=1}^{n} \lambda_i \tilde{x}_i^H \tilde{x}_i = 1 \\ & |c_i|^2 = 1 \quad \forall i. \end{aligned} \tag{8.10}$$

This finally can be written in the desired form, from which we can easily solve using the semidefinite relaxation and rounding method. We recover the variable x as $x = U\tilde{x}$.

$$\begin{aligned} \min_{\tilde{x},c} \quad & \begin{bmatrix} \tilde{x} \\ c \end{bmatrix}^H \begin{bmatrix} U^H A^H A U & -U^H A^H B \\ -B^T A U & B^T B \end{bmatrix} \begin{bmatrix} \tilde{x} \\ c \end{bmatrix} \\ \text{s.t.} \quad & \sum_{i=1}^{n} \lambda_i |\tilde{x}_i|^2 = 1 \\ & |c_i|^2 = 1 \quad \forall i. \end{aligned} \tag{8.11}$$

Figures 8.6–8.18 show the results of applying the technique. We performed the optimization for multiple solid angle widths, and at two different frequencies having wavelengths equal to the diameter of the array, and half that length (the radius). The array is two-dimensional and consists of six elements arranged hexagonally.

8.6 Discussion

In this chapter, we presented several application examples for the magnitude least-squares problem. The primary application for the problem is filter design, but we see that there are purely mathematical reasons to study the problem. The approximate factorization of polynomials is one such mathematical application. The ideas can be used in conjunction with sums-of-squares programming, which is a research topic that is currently being studied intensely by many researchers.

The generality of the method needs to be emphasized. The filter design problems to which the magnitude least-squares formulation applies can take many different forms. The filter responses need only be linear in the design coefficients. There are, however, many very useful filter design problems that cannot be solved using MLS. One very common, and very useful problem is that of designing a filter subject to upper and lower bounds on magnitude. This problem cannot be cast as an MLS problem. To solve those problems, we can sometimes consider the generalized autocorrelation coefficients to be variables, and use approximate spectral factorization to find the filters. This has the potential to help a great deal in finding a suitable starting point for a local method. We do not possess any theoretical bounds on the quality of solutions obtained using this procedure, which distinguishes from the examples presented in this chapter.

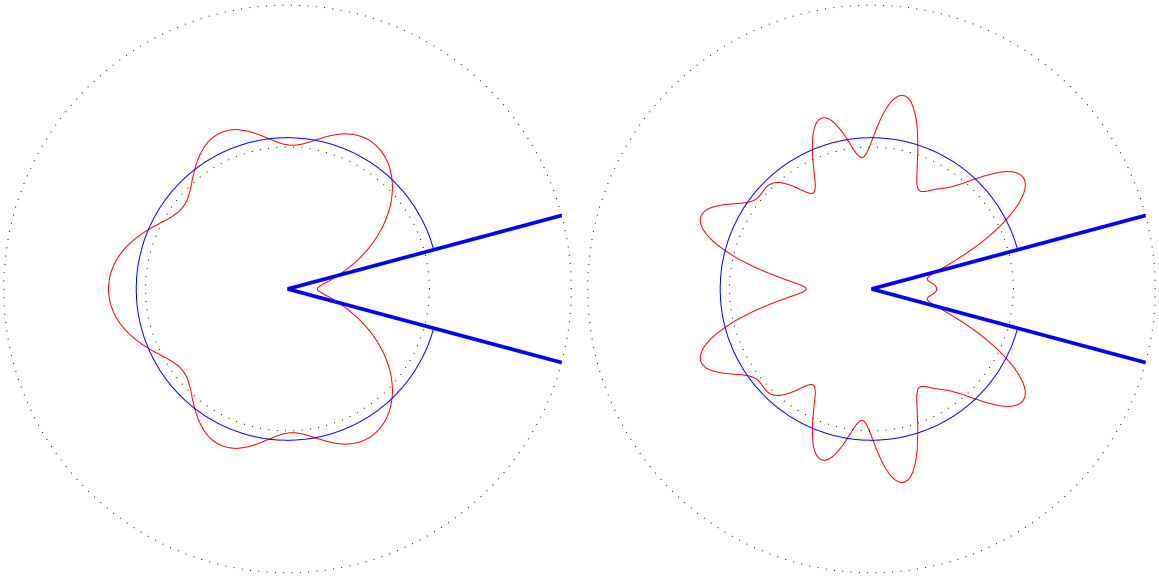


Figure 8.13: Polar responses with solid angle of 30° and wavelength equal to the diameter of the array (left) and radius of the array (right)

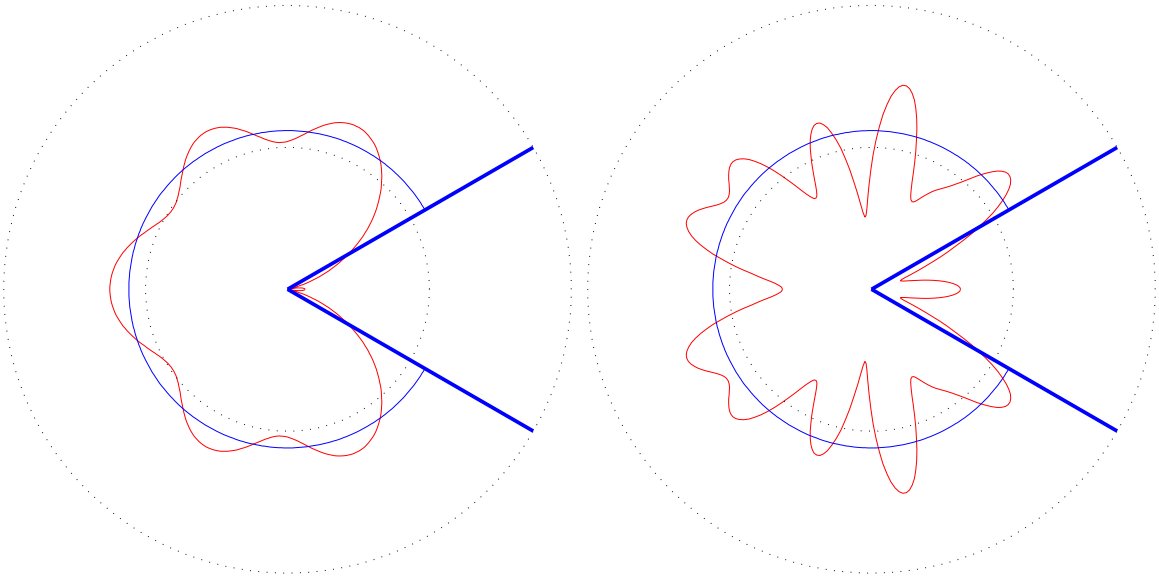


Figure 8.14: Polar responses with solid angle of 60° and wavelength equal to the diameter of the array (left) and radius of the array (right)

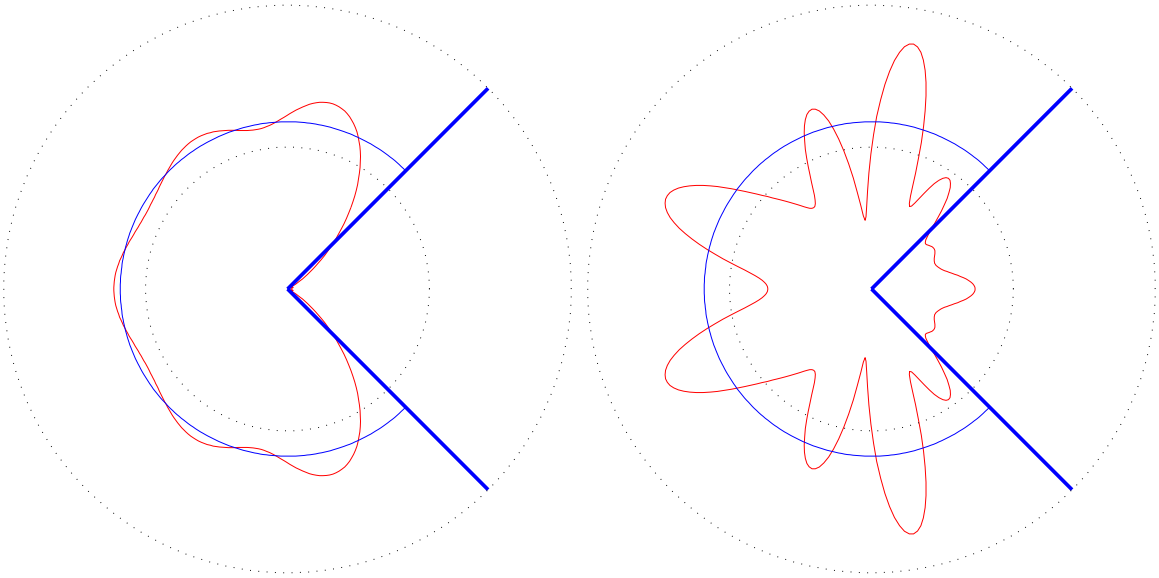


Figure 8.15: Polar responses with solid angle of $90°$ and wavelength equal to the diameter of the array (left) and radius of the array (right)

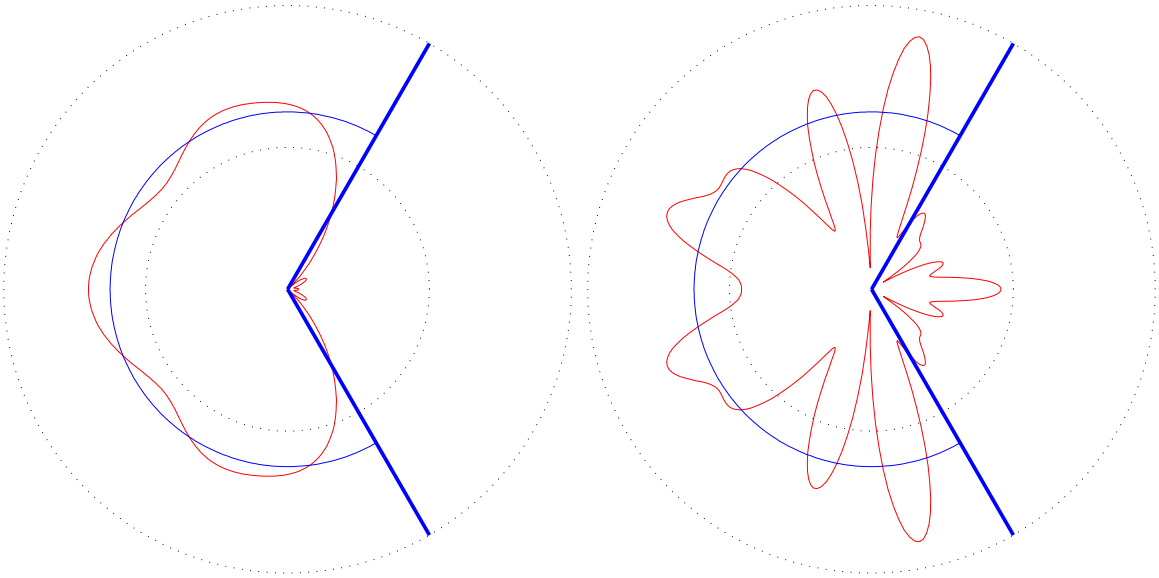


Figure 8.16: Polar responses with solid angle of $120°$ and wavelength equal to the diameter of the array (left) and radius of the array (right)

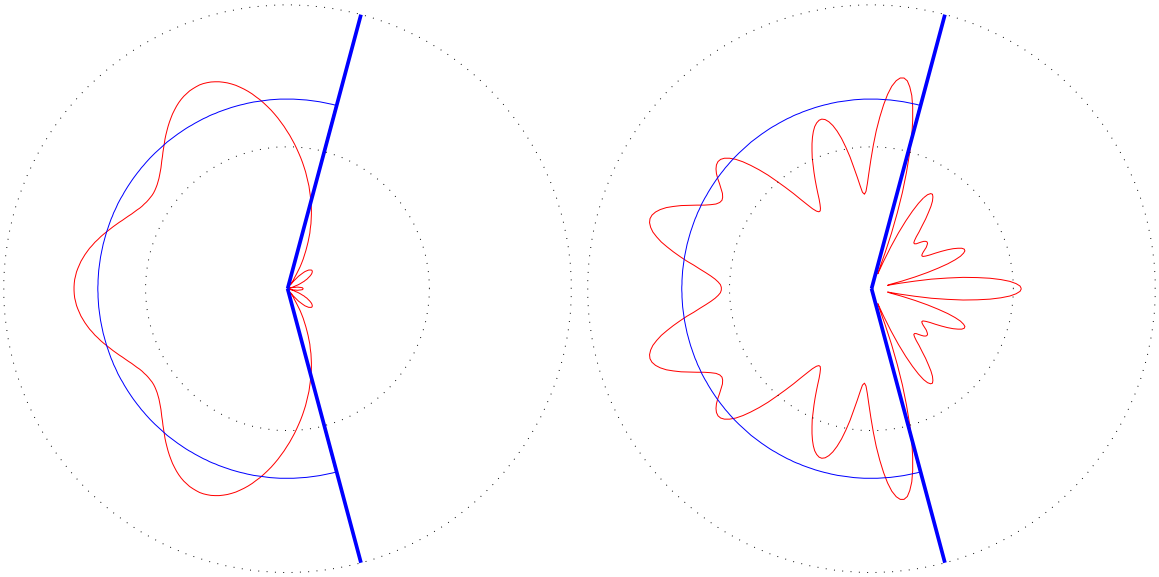


Figure 8.17: Polar responses with solid angle of 150° and wavelength equal to the diameter of the array (left) and radius of the array (right)

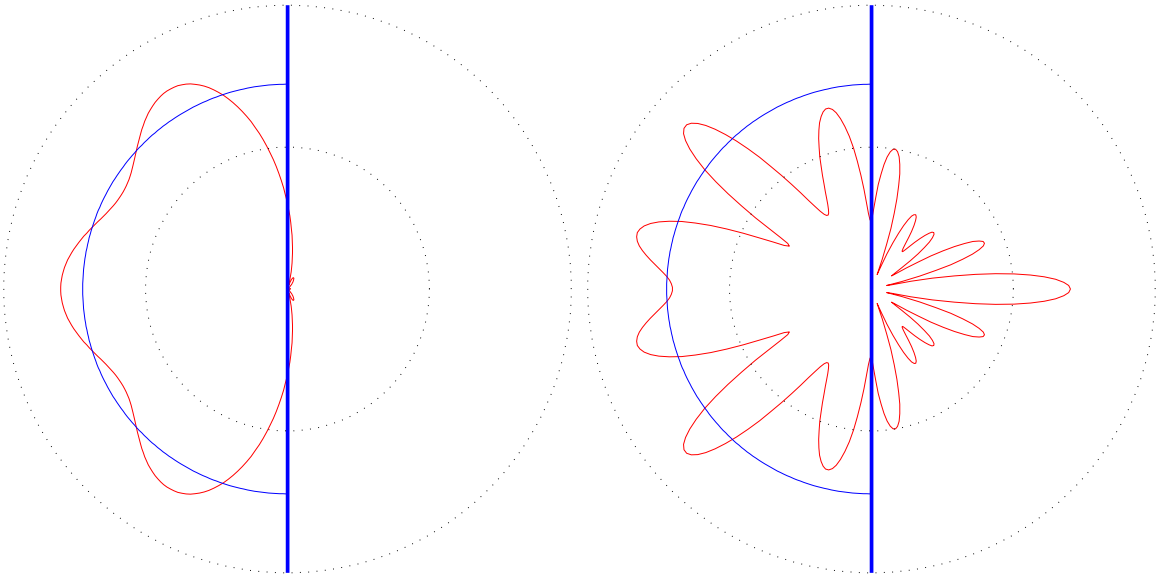


Figure 8.18: Polar responses with solid angle of 180° and wavelength equal to the diameter of the array (left) and radius of the array (right)

Chapter 9

Conclusion

This thesis has contributed to the engineering community in three ways. First, we have provided methods for solving several concrete, practical problems arising primarily in signal processing. Second, we have developed and intensely studied a specific optimization tool – the magnitude least-squares relaxation. The development of the tool has focused not only on its applicability to specific problems in engineering, but its performance as a general method for solving an abstract class of optimization problem. Third, we have discovered insight into complex-valued semidefinite relaxations in general. For the magnitude-squared least-squares problem, we have exploited structure associated with complex numbers to derive improved solution quality bounds. In this final chapter, we will review these three contributions and provide suggestions for further research.

9.1 Contributions

9.1.1 Solution of Practical Problems

In Chapter 2 we used convex optimization to help us elegantly answer questions about the performance capabilities of a clustered loudspeaker array. We also developed a novel linear-phase audio equalizer design method that is flexible and reliable. Both of these applications are important to an ongoing loudspeaker array project at the Center for New Music and Audio Technologies, sponsored by Meyer Sound Laboratories. In this thesis, these designs represent examples of the straightforward and powerful technique of convex

optimization.

In later chapters, we solved practical filter design problems using semidefinite relaxation and the magnitude least-squares (MLS) formulations. The problems we solved as examples in the thesis include multidimensional magnitude filter design, FIR filter design for nonlinearly delayed tapped filters, and spatial filtering using arbitrarily positioned array elements. All of the problems could be solved using essentially the same computational tool, developed in the heart of the thesis.

9.1.2 Development of the Optimization Tool

In Chapters 3–8 we developed and studied a method to solve a class of optimization problem called magnitude least-squares. The study of this problem originally was motivated by a desire to generalize methods of one-dimensional magnitude filter design that are based on convex optimization such as those in [WBV97], [Alk03], and [DLS02]. The magnitude least-squares problem finds application in the strightforward design examples discussed in Chapter 8, but can be applied in many other situations as well, like the design of banks of rational filters.

We have studied the problem in the abstract, paying close attention to the expected quality of solutions obtained using semidefinite relaxation. The depth and completeness of study adds to the literature on semidefinite relaxations and optimization in general. We proved that the magnitude least-squares problem falls within a broader class of problems treated by Nesterov [Nes97] and others [GW95][ZH04]. By carefully proving relations among the differing formulations and problem classes, we were able to discover structure in our problem that could be exploited in proving quality estimates. We also derived local methods (a Gauss-Newton method and a variable exchange method) for the problems which, when used in conjunction with semidefinite relaxation, yield very high-quality solutions to the problem.

9.1.3 Contribution to the Understanding of Complex Structure in SDP Relaxations

Perhaps the most exciting contribution of the thesis is our contribution to understanding complex structure in semidefinite relaxations. Inspired by the work of Zhang and Huang in [ZH04], where they prove that a complex-valued version of Nesterov's problem [Nes97] possesses superior quality bounds, we sought to show something similar for a simpler problem. We were successful in showing that the complex structure of the magnitude-squared least-squares problem allows us to improve on quality bound for the real-valued magnitude-squared least-squares problem. The key fact that caused the improvement is the differing fourth moments between real-valued and complex-valued normal distributions. We proved this both in the complex domain, and by converting to real formulations and investigating the block structure of the matrices. It seems clear that even more understanding of these problems could come by studying other types of structure, and how that structure manifests itself in the quality estimate.

9.2 Further Research

There are several directions for extended research on the topics presented in this dissertation. On practical side, the careful and novel application of optimization to practical problems in engineering always constitutes useful research. Particularly important and relevant to this thesis are discoveries of new formulations that cleverly convert nonconvex problems into ones that are convex. On the theoretical side, we believe there is a great deal to be learned from the study of structure in optimization problems, particularly in its relation to approximate solution methods and the quality of the approximations. For example, we speculate that there are other types of matrix block structures, similar to that associated with complex structure, that should result in improved semidefinite relaxation quality bounds.

In addition to these ideas, we feel that much benefit could come from the study of other specific difficult problems. The most value type of study would include solving the problem

with a comprehensive set of methods, including even more methods than we studied for the magnitude least-squares problem. Other methods could include neural networks, simulated annealing, genetic programming methods, and other novel ideas. There is a mathematical richness and elegance to convex optimization that is extremely seductive. There are many other optimization methods and the successful practicioner will attempt to understand as many of them as possible.

Appendix A

Real and Complex Relaxations

A.1 Introduction

In what follows, we will show that the semidefinite relaxation obtained for the complex magnitude least-squares problem (MLS) is equivalent to the relaxation obtained if we were to formulate the problem in real terms (using real and imaginary parts of the variables). This result is plausible, but not obvious (to the author). The fact that the scalars in the problem are complex imposes a certain structure on its corresponding real formulation. Bounds on the quality of the complex relaxation have been established by Shuzhong Zhang and Yongwei Huang of the Chinese University of Hong Kong [ZH04], and interestingly, are better than similar bounds associated with the same problem restricted to real-valued scalars [Nes97]. Therefore it is the complex structure of the problem that results in tighter semidefinite approximations.

A.2 Complex to Real Conversions and Semidefinite Relaxations

A.2.1 Primal Problem and its Complex Relaxation

The problem we are interested in is the nonconvex complex-valued optimization problem

$$\begin{aligned} \min_{z} \quad & z^H W z \\ \text{s.t.} \quad & |z_i| = 1, \quad i = 1, 2, ..., m, \end{aligned} \tag{CP}$$

where W is hermitian symmetric, ensuring that the objective function is real. As discussed previously, the magnitude least-squares problem (MLS) can be expressed as (CP), (though problem (CP) is more general). We can construct an equivalent optimization problem that is a complex semidefinite program with a rank constraint:

$$\begin{aligned} \min_{Z} \quad & \mathrm{Tr}\,(ZW) \\ \text{s.t.} \quad & Z_{ii} = 1, \quad i = 1, 2, ..., m \\ & Z \succeq 0 \\ & \text{Rank}\,(Z) = 1. \end{aligned} \tag{A.1}$$

Neglecting the rank constraint yields the "complex semidefinite relaxation" for this problem. It is

$$\begin{aligned} \min_{Z} \quad & \mathrm{Tr}\,(ZW) \\ \text{s.t.} \quad & Z_{ii} = 1, \quad i = 1, 2, ..., m \\ & Z \succeq 0 \end{aligned} \tag{CSDP}$$

We will consider two different approaches to generating real-valued relaxations for (CP), attempting to understand the algebraic and geometric relations between them. The first formulation is derived by converting (CP) to a real-valued nonconvex problem, and finding a real-valued semidefinite relaxation. The second is derived by converting the complex relaxation (CSDP) to an equivalent real-valued problem. In each, we will witness structure that is consequence of the problem originating in the complex domain. We will find that the two relaxations are indeed equivalent in terms of solutions obtained, but their feasible sets differ geometrically.

A.2.2 Complex to Real Conversion of the Primal Problem

Suppose we first convert (CP) to an equivalent optimization problem over the reals. Expanding in real and imaginary parts, we have

$$\begin{aligned} \min_{z_R,\, z_I} \quad & (z_R - jz_I)^T (W_R + jW_I)(z_R + jz_I) \\ \text{s.t.} \quad & (z_i)_R^2 + (z_i)_I^2 = 1, \quad i = 1, 2, ..., m. \end{aligned}$$

Since W is conjugate symmetric, we know that W_R is symmetric, W_I is skew symmetric, and that the objective function is real for any z. Collecting nonzero terms,

$$\begin{aligned} \min_{z_R,\, z_I} \quad & z_R^T W_R z_R - z_R^T W_I z_I + z_I^T W_R z_I + z_I^T W_I z_R \\ \text{s.t.} \quad & (z_i)_R^2 + (z_i)_I^2 = 1, \quad i = 1, 2, ..., m. \end{aligned} \tag{A.2}$$

We can now express the objective function in matrix notation as

$$\begin{aligned} \min_{z_R,\, z_I} \quad & \begin{bmatrix} z_R \\ z_I \end{bmatrix}^T \begin{bmatrix} W_R & -W_I \\ W_I & W_R \end{bmatrix} \begin{bmatrix} z_R \\ z_I \end{bmatrix} \\ \text{s.t.} \quad & (z_R)_i^2 + (z_I)_i^2 = 1, \quad i = 1, 2, ..., m. \end{aligned} \tag{RP1}$$

We can see that this real formulation parallels the complex (CP), but possesses block structure in the objective matrix and pairwise quadratic equality constraints. We now have an equivalent problem in $\mathbb{R}^{2m}$, in place of a complex problem in $\mathbb{C}^m$.

A.2.3 Relaxation Derived from Real Formulation

We can derive a dual problem for (RP1) by introducing Lagrange multipliers associated with each of the m equality constraints, and minimizing the associated Lagragian over z_R and z_I. From there, we can compute the associated semidefinite relaxation for (RP1) by again taking the dual (i.e., it is the "dual of the dual"). Alternatively, resulting in the same relaxation, we can express (RP1) as a semidefinite program with a rank constraint,

paralleling (A.1):

$$\begin{aligned} \min_{Z_1, Z_2, Z_3} \quad & \text{Tr}\left(\begin{bmatrix} Z_1 & Z_2 \\ Z_2^T & Z_3 \end{bmatrix}\begin{bmatrix} W_R & -W_I \\ W_I & W_R \end{bmatrix}\right) \\ \text{s.t.} \quad & (Z_1)_{ii} + (Z_3)_{ii} = 1, \quad i = 1, 2, ..., m \\ & \begin{bmatrix} Z_1 & Z_2 \\ Z_2^T & Z_3 \end{bmatrix} \succeq 0 \\ & \text{Rank}\left(\begin{bmatrix} Z_1 & Z_2 \\ Z_2^T & Z_3 \end{bmatrix}\right) = 1. \end{aligned}$$

Dropping the rank constraint yields the real-valued semidefinite program relaxation:

$$\begin{aligned} \min_{Z_1, Z_2, Z_3} \quad & \text{Tr}\left(\begin{bmatrix} Z_1 & Z_2 \\ Z_2^T & Z_3 \end{bmatrix}\begin{bmatrix} W_R & -W_I \\ W_I & W_R \end{bmatrix}\right) \\ \text{s.t.} \quad & (Z_1)_{ii} + (Z_3)_{ii} = 1, \quad i = 1, 2, ..., m \\ & \begin{bmatrix} Z_1 & Z_2 \\ Z_2^T & Z_3 \end{bmatrix} \succeq 0 \end{aligned} \qquad \text{(RSDP1)}$$

A.2.4 Complex to Real Conversion of the Complex Relaxation

How does the real relaxation (RSDP1) relate to the complex relaxation (CSDP)? To express (CSDP) in terms of real and imaginary parts, we need to use a lemma, proved as an exercise in [BV04].

Lemma L5: For $Z^H = Z \in \mathbb{C}^{n \times n}$,

$$Z \succeq 0 \quad \text{if and only if} \quad \begin{bmatrix} Z_R & -Z_I \\ Z_I & Z_R \end{bmatrix} \succeq 0, \qquad \text{(Lemma L5)}$$

where Z_R and Z_I are the real and imaginary parts of Z respectively.

We can then expand (CSDP) in real terms:

$$\begin{aligned} \min_{Z_R, Z_I} \quad & \text{Tr}\left((Z_R + jZ_I)(W_R + jW_I)\right) \\ \text{s.t.} \quad & (Z_R)_{ii} = 1, \quad i = 1, 2, ..., m \\ & \begin{bmatrix} Z_R & -Z_I \\ Z_I & Z_R \end{bmatrix} \succeq 0. \end{aligned}$$

Collecting terms and writing in matrix form produces a real-valued semidefinite relaxation,

different from (RSDP1):

$$\begin{aligned} \min_{Z_R, Z_I} \quad & \frac{1}{2}\mathrm{Tr}\left(\begin{bmatrix} Z_R & -Z_I \\ Z_I & Z_R \end{bmatrix}\begin{bmatrix} W_R & -W_I \\ W_I & W_R \end{bmatrix}\right) \\ \text{s.t.} \quad & (Z_R)_{ii} = 1, \quad i = 1, 2, ..., m \\ & \begin{bmatrix} Z_R & -Z_I \\ Z_I & Z_R \end{bmatrix} \succeq 0. \end{aligned} \tag{RSDP2}$$

Problem (RSDP2) differs from (RSDP1) in that the matrix constrained to be positive semidefinite in (RSDP2) has a block structure, whereas the positive semidefinite matrix in (RSDP1) is structurally unconstrained. The feasible sets are different because the number of decision variables is different. But are these optimization problems equivalent in some sense? Also, (RSDP2) was obtained by converting to real variables (CSDP). Could it have been obtained as a relaxation of a nonconvex real-valued problem, and if so, how does this problem relate to (RP1)? These are the questions we will address in the next section.

A.3 Relations Among Problem Formulations

For the purposes of this report we will consider two optimization problems to be equivalent if there is a well-defined and simple transformation that converts the optimal solution from one problem to the optimal solution for the other, and vice-versa. Intuitively this means that solving one of the problems immediately solves the other.

In what follows we will prove that (RSDP1) is equivalent to (RSDP2), and that they are related through a linear transformation. It is helpful to refer to Figure A.1 which is a diagram consisting of different problems, the relationships among which we will establish in this chapter.

A.3.1 Equivalence of the Relaxations

To prove that (RSDP1) is equivalent to (RSDP2), we first will need to prove a lemma.

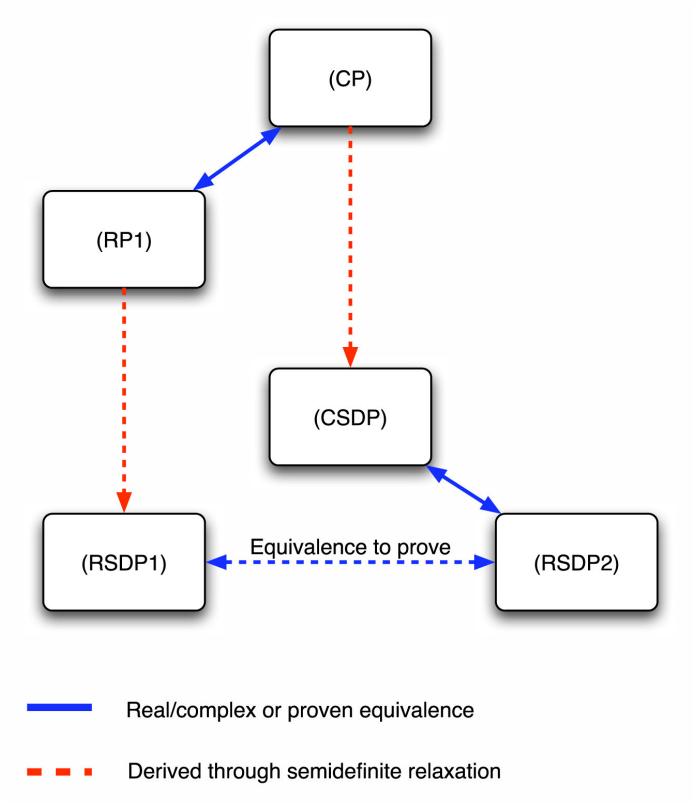


Figure A.1: Relationships among the problems.

Lemma L6:

$$\text{For } \begin{bmatrix} X_1 & X_2 \\ X_2^T & X_3 \end{bmatrix} \in \mathbb{R}^{2n \times 2n}, \quad \text{if } \begin{bmatrix} X_1 & X_2 \\ X_2^T & X_3 \end{bmatrix} \succeq 0, \quad \text{then } \begin{bmatrix} X_3 & -X_2 \\ -X_2^T & X_1 \end{bmatrix} \succeq 0.$$

(Lemma L6)

Proof: Let u and v be arbitrary vectors in $\mathbb{R}^n$. We have

$$\begin{bmatrix} u \\ v \end{bmatrix}^T \begin{bmatrix} X_3 & -X_2 \\ -X_2^T & X_1 \end{bmatrix} \begin{bmatrix} u \\ v \end{bmatrix} = v^T X_1 v + u^T X_3 u - 2u^T X_2^T v = \begin{bmatrix} -v \\ u \end{bmatrix}^T \begin{bmatrix} X_1 & X_2 \\ X_2^T & X_3 \end{bmatrix} \begin{bmatrix} -v \\ u \end{bmatrix} \geq 0.$$

This proves the lemma. ■

Theorem T6: Problem (RSDP1) is equivalent to Problem (RSDP2)

(Theorem T6)

.

Proof: Suppose we have calculated an optimal solution $\{Z_1^*, Z_2^*, Z_3^*\}$, for (RSDP1) with

objective ν^*. Assign

$$Z_R = Z_1^* + Z_3^* \tag{A.3}$$
$$Z_I = Z_2^{*T} - Z_2^*.$$

Since $\{Z_1^*, Z_2^*, Z_3^*\}$ is feasible for (RSDP1), we have that

$$\begin{bmatrix} Z_1^* & Z_2^{*T} \\ Z_2^* & Z_3^* \end{bmatrix} \succeq 0, \quad \text{and}$$
$$(Z_1^*)_{ii} + (Z_3^*)_{ii} = 1, \quad i = 1, 2, ..., m.$$

This means, by using (Lemma L6) and the convexity of the semidefinite cone, that

$$\begin{aligned} \begin{bmatrix} Z_R & -Z_I \\ Z_I & Z_R \end{bmatrix} &= \begin{bmatrix} Z_1^* + Z_3^* & Z_2^* - Z_2^{*T} \\ Z_2^{*T} - Z_2^* & Z_1^* + Z_3^* \end{bmatrix} \\ &= \begin{bmatrix} Z_1^* & Z_2^* \\ Z_2^{*T} & Z_3^* \end{bmatrix} + \begin{bmatrix} Z_3^* & -Z_2^{*T} \\ -Z_2^* & Z_1^* \end{bmatrix} \succeq 0. \end{aligned}$$

We also can see that

$$(Z_R)_{ii} = (Z_1^* + Z_3^*)_{ii} = (Z_1^*)_{ii} + (Z_3^*)_{ii} = 1, \quad i = 1, 2, ..., m.$$

Thus $\{Z_R, Z_I\}$ is feasible for (RSDP2).

The objective obtained by $\{Z_R, Z_I\}$ is

$$\begin{aligned} \mu &= \frac{1}{2}\mathrm{Tr}\left(\begin{bmatrix} Z_R & -Z_I \\ Z_I & Z_R \end{bmatrix}\begin{bmatrix} W_R & -W_I \\ W_I & W_R \end{bmatrix}\right) \\ &= \frac{1}{2}\mathrm{Tr}\left(\begin{bmatrix} Z_1^* + Z_3^* & Z_2^* - Z_2^{*T} \\ Z_2^{*T} - Z_2^* & Z_1^* + Z_3^* \end{bmatrix}\begin{bmatrix} W_R & -W_I \\ W_I & W_R \end{bmatrix}\right) = \nu^* \end{aligned}$$

Suppose there existed an optimal $\{Z_R^*, Z_I^*\}$ obtaining a strictly smaller objective $\mu^* < \mu = \nu^*$. Then

$$\mu^* = \frac{1}{2}\mathrm{Tr}\left(\begin{bmatrix} Z_R^* & -Z_I^* \\ Z_I^* & Z_R^* \end{bmatrix}\begin{bmatrix} W_R & -W_I \\ W_I & W_R \end{bmatrix}\right) < \nu^*$$

Assign

$$Z_1 = Z_3 = \frac{1}{2}Z_R^* \tag{A.4}$$
$$Z_2 = -\frac{1}{2}Z_I^*.$$

We have that $\{Z_1, Z_2, Z_3\}$ is feasible for (RSDP1):

$$\begin{bmatrix} Z_1 & Z_2 \\ Z_2^T & Z_3 \end{bmatrix} = \frac{1}{2}\begin{bmatrix} Z_R^* & -Z_I^* \\ Z_I^* & Z_R^* \end{bmatrix} \succeq 0,$$
$$(Z_1)_{ii} + (Z_3)_{ii} = \frac{1}{2}(2(Z_R^*)_{ii}) = 1, \quad i = 1, 2, ..., m.$$

The objective ν obtained by $\{Z_1, Z_2, Z_3\}$ is

$$\begin{aligned} \nu &= \mathrm{Tr}\left(\begin{bmatrix} Z_1 & Z_2 \\ Z_2^T & Z_3 \end{bmatrix}\begin{bmatrix} W_R & -W_I \\ W_I & W_R \end{bmatrix}\right) \\ &= \frac{1}{2}\mathrm{Tr}\left(\begin{bmatrix} Z_R^* & -Z_I^* \\ Z_I^* & Z_R^* \end{bmatrix}\begin{bmatrix} W_R & -W_I \\ W_I & W_R \end{bmatrix}\right) = \mu < \nu^*, \end{aligned}$$

which contradicts the fact that ν^* was the optimal objective for (RSDP1). Thus, $\{Z_R, Z_I\}$ is an optimal solution for (RSDP2), and we have defined a simple linear transformation (A.3) to obtain an optimal for solution for (RSDP2) given an optimal solution to (RSDP1).

Conversely, if we are given an optimal solution $\{Z_R^*, Z_I^*\}$ for (RSDP2), we can use the transformation defined in (A.4) to obtain a feasible point for (RSDP2) with the same objective. Using similar arguments as above, it is easy to verify that the feasible point obtained in this way must be an optimal solution for (RSDP1). Thus, given an optimal solution for one of either (RSDP1) or (RSDP2), we can easily obtain a solution for the other via the transformations (A.3), and (A.4). This establishes that these problems are equivalent. ■

We have discussed thus far five distinct optimization problems, related to one another through real/complex transformations, semidefinite relaxation, or proven equivalence, shown in Figure A.1. We can conclude as a result of (Theorem T6) and our definition of problem equivalence that (CSDP) is equivalent to (RSDP1) via (RSDP2). To display this fact, we can connect these two problems in the diagram, Figure A.2.

A.3.2 Equivalence of the Primal Problems

The five problems encountered so far are the complex-valued primal problem (CP), its complex-valued semidefinite relaxation (CSDP), a real-valued nonconvex problem (RP1) derived by expanding (CP) into real and imaginary parts, the semidefinite relaxation (RSDP1)

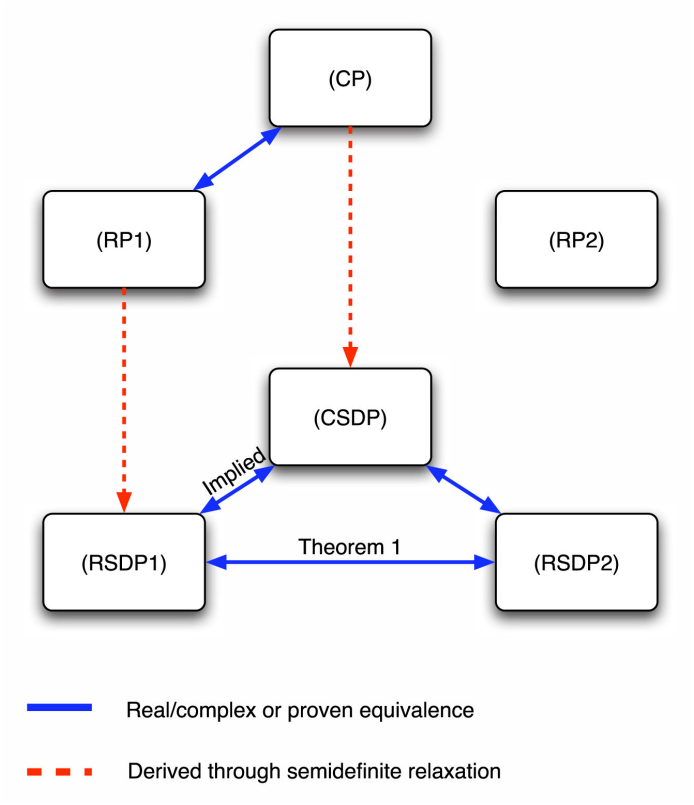


Figure A.2: Relationships among the problems with implied connection between (CSDP) and (RSDP1), and placeholder for (RP2).

of (RP1), and the relaxation (RSDP2) derived by expanding the complex variables in (CSDP) into real and imagainary parts. In this section we will introduce a sixth problem, (RP2). In Figure A.2 problem (RP2) is shown as being unrelated to the others, but we will define it so that it is a real-valued nonconvex problem from which (RSDP2) may be derived through relaxation. We then will prove that (RP2) is equivalent to (RP1) (and therefore equivalent to (CP)).

Problem (RP2) we will define as follows:

$$\begin{aligned} \min_{Z_R,\, Z_I} \quad & \frac{1}{2}\mathrm{Tr}\left(\begin{bmatrix} Z_R & -Z_I \\ Z_I & Z_R \end{bmatrix}\begin{bmatrix} W_R & -W_I \\ W_I & W_R \end{bmatrix}\right) \\ \text{s.t.} \quad & (Z_R)_{ii} = 1, \quad i = 1, 2, ..., m \\ & \begin{bmatrix} Z_R & -Z_I \\ Z_I & Z_R \end{bmatrix} \succeq 0. \\ & \mathrm{Rank}\left(\begin{bmatrix} Z_R & -Z_I \\ Z_I & Z_R \end{bmatrix}\right) = 2 \end{aligned} \tag{RP2}$$

Problem (RSDP2) is clearly a relaxation of of (RP2); it is obtained by ignoring the rank constraint. After proving some additional facts, we will be able to verify that (RSDP2) may be derived from (RP2) through Lagrangian duality.

With the goal of proving that (RP2) is equivalent to (RP1), we first establish a lemma and a theorem.

Lemma L7: Suppose $A, B \in \mathbb{R}^{n\times m}$ are such that $AA^T = BB^T$. Then there exists an orthonormal matrix $\Psi \in \mathbb{R}^{m\times m}$ such that $A = B\Psi$.

(Lemma L7)

Proof: First assume A and B each have full rank m. Since $AA^T = BB^T \equiv C$, the columns of C are in the range of both A and B. Since A and B are each full rank, there exist m nonzero eigenvalues of C, since the eigenvalues of C are the squares of the singular values of A and/or B. This means the columns of C span an m-dimensional subspace in $\mathbb{R}^n$. Since the column space of C is contained in both m-dimensional ranges of A and B, the ranges of A and B must be the same. Hence, the columns of A are in the range of B, meaning there exists $Q \in \mathbb{R}^{m\times m}$ such that $A = BQ$. But $AA^T = BQQ^TB^T = BB^T$, which implies $QQ^T = I$, since no column of Q is in the nullspace of B (otherwise a column of A would be zero, contradicting the fact that A has full rank.) Hence $\Psi \equiv Q$ is orthonormal.

Now, consider the case that A and B have rank $r < m$. They have the same rank because they share the same singular values. Let $A = U_A\Sigma V_A^T$ and $B = U_B\Sigma V_B^T$ be the singular value decompositions of A and B. Without loss of generality, assume it is the last $m - r$ columns of Σ that are zero. Then

$$\begin{bmatrix} A' & \mathbf{0} \end{bmatrix} = AV_A = U_A\Sigma \quad \text{and}$$
$$\begin{bmatrix} B' & \mathbf{0} \end{bmatrix} = BV_B = U_B\Sigma,$$

defining A' and B' to be the first r (nonzero) columns of AV_A and BV_B respectively. Ob-

serve that

$$C = AA^T = AV_AV_A^TA^T = \begin{bmatrix} A' & \mathbf{0} \end{bmatrix} \begin{bmatrix} A'^T \\ \mathbf{0}^T \end{bmatrix} = A'A'^T$$
$$= BB^T = BV_BV_B^TB^T = \begin{bmatrix} B' & \mathbf{0} \end{bmatrix} \begin{bmatrix} B'^T \\ \mathbf{0}^T \end{bmatrix} = B'B'^T.$$

The $n \times r$ matrices A' and B' are each full rank, so using the first part of the proof, we know there exists orthonormal $\Psi' \in \mathbb{R}^{r \times r}$ such that $A' = B'\Psi'$. Extend Ψ' to form an $m \times m$ orthonormal matrix

$$\Psi'' \equiv \begin{bmatrix} \Psi' & \mathbf{0} \\ \mathbf{0} & I \end{bmatrix}.$$

Now we can verify that

$$\begin{bmatrix} A' & \mathbf{0} \end{bmatrix} = \begin{bmatrix} B' & \mathbf{0} \end{bmatrix} \begin{bmatrix} \Psi' & \mathbf{0} \\ \mathbf{0} & I \end{bmatrix}$$
$$AV_A = BV_B \begin{bmatrix} \Psi' & \mathbf{0} \\ \mathbf{0} & I \end{bmatrix}$$
$$\Longrightarrow \quad A = BV_B \begin{bmatrix} \Psi' & \mathbf{0} \\ \mathbf{0} & I \end{bmatrix} V_A^T.$$

The matrix

$$\Psi \equiv V_B \begin{bmatrix} \Psi' & \mathbf{0} \\ \mathbf{0} & I \end{bmatrix} V_A^T$$

is orthonormal, being the product of three orthonormnal matrices, proving $A = B\Psi$. ■

Theorem T7: Let $X \in \mathbb{C}^{n \times n}$, and denote the real and imaginary parts of X as X_R, and X_I respectively.

$$X \succeq 0, \quad \text{Rank}\,(X) = 1 \quad \text{if and only if}$$
$$\begin{bmatrix} X_R & -X_I \\ X_I & X_R \end{bmatrix} \succeq 0, \quad \text{Rank}\left(\begin{bmatrix} X_R & -X_I \\ X_I & X_R \end{bmatrix} \right) = 2 \qquad \text{(Theorem T7)}$$

Proof: We know from (Lemma L5) that

$$X \succeq 0 \quad \text{if and only if} \quad \begin{bmatrix} X_R & -X_I \\ X_I & X_R \end{bmatrix} \succeq 0.$$

Now, suppose additionally that Rank $(X) = 1$. Then, since $X \succeq 0$, it must factor as

$$\begin{aligned} X = yy^H &= (y_R + jy_I)(y_R^T - jy_I^T) \\ &= (y_R y_R^T + y_I y_I^T) + j(y_I y_R^T - y_R y_I^T) \\ &\equiv X_R + jX_I \end{aligned}$$

This means

$$\begin{aligned} \begin{bmatrix} X_R & -X_I \\ X_I & X_R \end{bmatrix} &= \begin{bmatrix} y_R y_R^T + y_I y_I^T & -y_I y_R^T + y_R y_I^T \\ y_I y_R^T - y_R y_I^T & y_R y_R^T + y_I y_I^T \end{bmatrix} \\ &= \begin{bmatrix} y_R \\ y_I \end{bmatrix} \begin{bmatrix} y_R^T & y_I^T \end{bmatrix} + \begin{bmatrix} y_I \\ -y_R \end{bmatrix} \begin{bmatrix} y_I^T & -y_R^T \end{bmatrix}, \end{aligned}$$

implying

$$\text{Rank} \left(\begin{bmatrix} X_R & -X_I \\ X_I & X_R \end{bmatrix} \right) \leq 2.$$

Suppose

$$\text{Rank} \left(\begin{bmatrix} X_R & -X_I \\ X_I & X_R \end{bmatrix} \right) < 2.$$

The rank being strictly less than 2 means the factors above must be linearly related. In other words, there exists a real scalar t such that

$$\begin{bmatrix} y_R \\ y_I \end{bmatrix} = t \begin{bmatrix} y_I \\ -y_R \end{bmatrix}.$$

This, in turn, is only possible if $y_R \equiv y_I \equiv 0$. This means $X \equiv 0$, contradicting the assumption that Rank $(X) = 1$. Therefore

$$\text{Rank} \left(\begin{bmatrix} X_R & -X_I \\ X_I & X_R \end{bmatrix} \right) = 2.$$

Conversely, suppose Rank $\left(\begin{bmatrix} X_R & -X_I \\ X_I & X_R \end{bmatrix} \right) = 2$. Then there exists a factorization

$$\begin{bmatrix} X_R & -X_I \\ X_I & X_R \end{bmatrix} = \begin{bmatrix} a_1 \\ a_2 \end{bmatrix} \begin{bmatrix} a_1^T & a_2^T \end{bmatrix} + \begin{bmatrix} b_1 \\ b_2 \end{bmatrix} \begin{bmatrix} b_1^T & b_2^T \end{bmatrix} = \begin{bmatrix} a_1 a_1^T + b_1 b_1^T & a_1 a_2^T + b_1 b_2^T \\ a_2 a_1^T + b_2 b_1^T & a_2 a_2^T + b_2 b_2^T \end{bmatrix}. \quad \text{(A.5)}$$

This implies

$$X_R = a_1 a_1^T + b_1 b_1^T = a_2 a_2^T + b_2 b_2^T$$
$$= \begin{bmatrix} a_1 & b_1 \end{bmatrix} \begin{bmatrix} a_1^T \\ b_1^T \end{bmatrix} = \begin{bmatrix} a_2 & b_2 \end{bmatrix} \begin{bmatrix} a_2^T \\ b_2^T \end{bmatrix}.$$

Now, using (Lemma L7), we know there exists an orthonormal 2×2 matrix Ψ such that

$$\begin{bmatrix} a_1 & b_1 \end{bmatrix} = \begin{bmatrix} a_2 & b_2 \end{bmatrix} \Psi. \tag{A.6}$$

From (A.5), we also have

$$X_I = - \begin{bmatrix} a_1 & b_1 \end{bmatrix} \begin{bmatrix} a_2^T \\ b_2^T \end{bmatrix} = \begin{bmatrix} a_2 & b_2 \end{bmatrix} \begin{bmatrix} a_1^T \\ b_1^T \end{bmatrix}.$$

Substituting, we have

$$- \begin{bmatrix} a_2 & b_2 \end{bmatrix} \Psi \begin{bmatrix} a_2^T \\ b_2^T \end{bmatrix} = \begin{bmatrix} a_2 & b_2 \end{bmatrix} \Psi^T \begin{bmatrix} a_2^T \\ b_2^T \end{bmatrix} \tag{A.7}$$

For Ψ to be a 2×2 orthonormal matrix, it must possess the following form:

$$\Psi = \begin{bmatrix} \sin\theta & -\cos\theta \\ \cos\theta & \sin\theta \end{bmatrix}.$$

Combining with (A.7), we have

$$a_2 a_2^T \sin\theta - a_2 b_2^T \cos\theta + b_2 a_2^T \cos\theta + b_2 b_2^T \sin\theta$$
$$= -a_2 a_2^T \sin\theta + a_2 b_2^T \cos\theta - b_2 a_2^T \cos\theta - b_2 b_2^T \sin\theta. \tag{A.8}$$

We know that a_1, a_2, b_1, b_2 cannot all be identically zero because of our rank assumption (A.5). We can assume without loss of generality that either a_2 or b_2 possesses a nonzero element (if not, we can rewrite (A.7) in terms of a_1 and b_1). Looking at a diagonal element corresponding to this nonzero element in (A.8), we see that θ must be either 0 or π, implying

$$\Psi = \begin{bmatrix} 0 & -1 \\ 1 & 0 \end{bmatrix} \quad \text{or} \quad \Psi = \begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix}.$$

Using (A.6), we can write a_1 and b_1 in terms of a_2 and b_2:

$$(a_1 = -b_2 \quad \text{and} \quad b_1 = a_2) \quad \text{or}$$

$$(a_1 = b_2 \quad \text{and} \quad b_1 = -a_2)$$

This allows us to see that X_R and X_I are indeed the real and imaginary parts of a complex rank 1 matrix. In the first case, the complex factorization giving rise to X_R and X_I is

$$X_R = a_2 a_2^T + b_2 b_2^T = \mathrm{Re}\{(a_2 + jb_2)(a_2^T - jb_2^T)\}$$
$$X_I = -a_2 b_2^T + b_2 a_2^T = \mathrm{Im}\{(a_2 + jb_2)(a_2^T - jb_2^T)\}.$$

And in the second case,

$$X_R = a_2 a_2^T + b_2 b_2^T = \mathrm{Re}\{(a_2 + jb_2)(a_2^T - jb_2^T)\}$$
$$X_I = a_2 b_2^T - b_2 a_2^T = \mathrm{Im}\{(a_2 - jb_2)(a_2^T + jb_2^T)\}.$$

Either way, we have proved the matrix

$$X = X_R + jX_I \tag{A.9}$$

has rank 1, completing the proof of the theorem. ■

Corollary C1: Problem (RP2) is equivalent to Problem (CP), and therefore the three problems, (CP), (RP1), and (RP2) are all equivalent.

(Corollary C1)

Proof: Using (Theorem T7), we can see that

$$Z \text{ is feasible for (A.1) (and therefore (CP)) if and only if}$$
$$\begin{bmatrix} Z_R & -Z_I \\ Z_I & Z_R \end{bmatrix} \quad \text{is feasible for (RP2).}$$

It is easy to check that the objective functions match at every feasible point, establishing the equivalence. Since (CP) is equivalent to (RP1), we can also conclude that (RP1) is equivalent to (RP2). ■

This allows us to complete the diagram. Figure A.3 displays the relationships among the different formulations. We see that for this problem it does not matter in which order we derive the relaxation, the resulting problem will be the same.

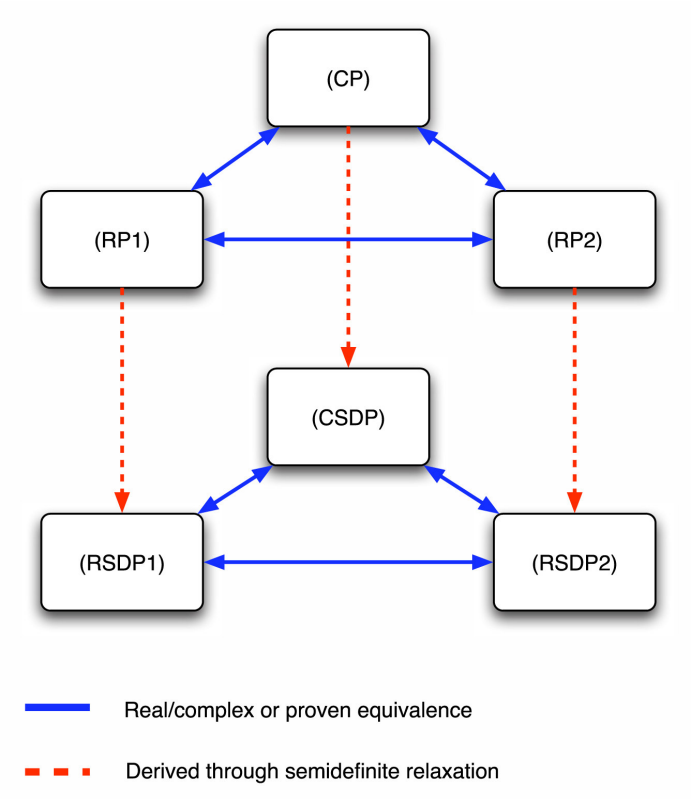


Figure A.3: Fully connected relationships among the problems.

Bibliography

[Alk03] Brien Forrest Alkire. *Convex optimization problems involving autocorrelation sequences*. PhD thesis, University of California Los Angeles, 2003.

[AV02] Brien Alkire and Lieven Vandenberghe. Convex optimization problems involving finite autocorrelation sequences. *Mathematical Programming, Series A*, 93(3):331–359, 2002.

[Ber93] Leo L. Beranek. *Acoustics*. Acoustical Society of America, 1993.

[Bla00] David T. Blackstock. *Fundamentals of Physical Acoustics*. John Wiley & Sons, Inc., 2000.

[BV04] Stephen Boyd and Lieven Vandenberghe. *Convex Optimization*. Cambridge University Press, 2004.

[Che82] E. W. Cheney. *Approximation Theory*. Chelsea, 2nd edition, 1982.

[DLS02] T. Davidson, Z. Luo, and J. Sturm. Linear matrix inequality formulation of spectral mask constraints. *IEEE Transactions on Signal Processing*, 50:2702–2715, November 2002.

[GL96] Gene H. Golub and Charles F. Van Loan. *Matrix Computations*. Johns Hopkins University Press, 3rd edition, 1996.

[GW95] M. Goemans and D. Williamson. Improved approximation algorithms for maximum cut and satisfiability problems using semidefinite programming. *Journal of Association for Computing Machinery*, 42(6):1115–1145, June 1995.

[HN02] Yvan Hachez and Yurii Nesterov. Optimization problems over nonnegative trigonometric polynomials with interpolation constraints. Technical report, CESAME – Université Catholique de Louvain, Av. Georges Lemaître 4-6, B-1348 Louvain-le-lNeuve, Belgium, February 2002.

[KFR03] Michael Kazhdan, Thomas Funkhouser, and Szymon Rusinkiewicz. Rotation invariant spherical harmonic representation of 3d shape descriptors. In *Proceedings of the Eurographics/ACM SIGGRAPH symposium on Geometry processing*, pages 156–164. Eurographics Association, 2003.

[KW04] Peter Kassakian and David Wessel. Characterization of loudspeaker arrays. In *Proceedings of the 117th Audio Engineering Society Convention*, San Francisco, CA, October 2004.

[Las01] J. Lasserre. Global optimization with polynomials and the problem of moments, 2001.

[Lue69] David G. Luenberger. *Optimization by Vector Space Methods*. John Wiley and Sons, Inc., 1969.

[Nes97] Yurii Nesterov. Quality of semidefinite relaxation for nonconvex quadratic optimization. Discussion Paper 9719, Center for Operations Research & Econometrics (CORE), Université Catholique de Louvain, 1997.

[Nes98] Yurii Nesterov. Semidefinite relaxation and nonconvex quadratic optimization. *Optimization Methods and Software*, 9:141–160, 1998.

[OSB98] Alan V. Oppenheim, Ronald W. Schafer, and John R. Buck. *Discrete-Time Signal Processing*. Prentice-Hall, Inc., 1998.

[OW01] Nicolas Misdariis Olivier Warusfel. Directivity synthesis with a 3D array of loudspeakers application for stage performance. In *Proceedings of the COST G-6 Conference on Digital Audio Effects*, 2001.

[Pap91] Athanasios Papoulis. *Probability, Random Variables, and Stochastic Processes*. WCB McGraw-Hill, 1991.

[Par01] P. Parrilo. Semidefinite programming relaxations for semialgebraic problems, 2001.

[PPP02] S. Prajna, A. Papachristodoulou, and P. A. Parrilo. *SOSTOOLS: Sum of squares optimization toolbox for MATLAB*, 2002.

[Ree62] I. S. Reed. On a moment theorem for complex gaussian processes. *IEEE Transactions on Information Theory*, 8(3):194–195, April 1962.

[SC94] Zhengwei Su and Philip Coppens. Rotation of real spherical harmonics. *Acta Crystallographica Section A*, A50:636–643, 1994.

[Stu99] J. F. Sturm. Using SeDuMi 1.02, a MATLAB toolbox for optimization over symmetric cones. *Optimization Methods and Software*, 11:625–653, 1999.

[TI97] Lloyd N. Trefethen and David Bau III. *Numerical Linear Algebra*. Society for Industrial and Applied Mathematics (SIAM), 1997.

[WBV97] S. Wu, S. Boyd, and L. Vandenberghe. FIR filter design via spectral factorization and convex optimization. *Applied Computational Control, Signal and Communications*, pages 1–33, 1997.

[ZH04] Shuzhong Zhang and Yongwei Huang. Complex quadratic optimization and semidefnite programming. Technical report, The Chinese University of Hong Kong, August 2004.