KTH Mathematics

A Convex Optimization Approach to Complexity Constrained Analytic Interpolation with Applications to ARMA Estimation and Robust Control

ANDERS BLOMQVIST

Doctoral Thesis
Stockholm, Sweden 2005

TRITA-MAT 05/OS/01
ISSN 1401-2294
ISRN KTH/OPT SYST/DA-05/01-SE
ISBN 91-7283-950-3

Department of Mathematics
KTH
SE-100 44 Stockholm
SWEDEN

Akademisk avhandling som med tillstånd av Kungl Tekniska högskolan framlägges till offentlig granskning för avläggande av teknologie doktorsexamen måndagen den 7 februari 2005 klockan 10.00 i Kollegiesalen, Administrationsbyggnaden, Kungl Tekniska Högskolan, Valhallavägen 79, Stockholm.

Tryck: Universitetsservice US AB

Till Elsa och Jan

Abstract

Analytical interpolation theory has several applications in systems and control. In particular, solutions of low degree, or more generally of low complexity, are of special interest since they allow for synthesis of simpler systems. The study of degree constrained analytic interpolation was initialized in the early 80's and during the past decade it has had significant progress.

This thesis contributes in three different aspects to complexity constrained analytic interpolation: theory, numerical algorithms, and design paradigms. The contributions are closely related; shortcomings of previous design paradigms motivate development of the theory, which in turn calls for new robust and efficient numerical algorithms.

Mainly two theoretical developments are studied in the thesis. Firstly, the spectral Kullback-Leibler approximation formulation is merged with simultaneous cepstral and covariance interpolation. For this formulation, both uniqueness of the solution, as well as smoothness with respect to data, is proven. Secondly, the theory is generalized to matrix-valued interpolation, but then only allowing for covariance-type interpolation conditions. Again, uniqueness and smoothness with respect to data is proven.

Three algorithms are presented. Firstly, a refinement of a previous algorithm allowing for multiple as well as matrix-valued interpolation in an optimization framework is presented. Secondly, an algorithm capable of solving the boundary case, that is, with spectral zeros on the unit circle, is given. This also yields an inherent numerical robustness. Thirdly, a new algorithm treating the problem with both cepstral and covariance conditions is presented.

Two design paradigms have sprung out of the complexity constrained analytical interpolation theory. Firstly, in robust control it enables low degree $\mathcal{H}_\infty$ controller design. This is illustrated by a low degree controller design for a benchmark problem in MIMO sensitivity shaping. Also, a user support for the tuning of controllers within the design paradigm for the SISO case is presented. Secondly, in ARMA estimation it provides unique model estimates, which depend smoothly on the data as well as enables frequency weighting. For AR estimation, a covariance extension approach to frequency weighting is discussed, and an example is given as an illustration. For ARMA estimation, simultaneous cepstral and covariance matching is generalized to include prefiltering. An example indicates that this might yield asymptotically efficient estimates.

Keywords: analytic interpolation, moment matching, Nevanlinna-Pick interpolation, spectral estimation, convex optimization, well-posedness, Kullback-Leibler discrepancy, continuation methods, ARMA modelling, robust control, $\mathcal{H}_\infty$ control

Mathematical Subject Classification (2000): 30E05, 93B36, 93E12, 62M10, 90C25, 93B29, 53C12, 93B40, 94A17

Sammanfattning

Analytisk interpoleringsteori har flera tillämpningar inom systemteori och reglerteknik. I synnerhet lösningar av låg grad, eller mer generellt av låg komplexitet, är särskilt intressanta då de möjliggör syntes av enklare system. Gradbegränsad interpoleringsteori har studerts sedan början av 80-talet och under det senaste decenniet har flertalet påtagliga framsteg gjorts.

Denna avhandling bidrar till analytisk interpolering med komplexitetsbivillkor i tre olika avseenden: teori, numeriska algoritmer och designparadigmer. Bidragen är nära sammanknutna; brister i tidigare designparadigmer motiverar utvidgningar av teorin, vilka in sin tur kräver nya robusta och effektiva numeriska algoritmer.

Huvudsakligen två teoretiska utvidgningar studeras i avhandlingen. För det första sammanlänkas formuleringen med avseende på den spektrala Kullback-Leiblerdiskrepansen med simultan kepstral- och kovariansinterpolering. För denna formulering bevisas både att en unik lösning finns och att den beror kontinuerligt på indata. För det andra utvecklas teorin till att omfatta matrisvärd interpolering, dock inskränkt till interpolering av kovarianstyp. Återigen bevisas entydighet och kontinuitet med avseende på data.

Tre algoritmer presenteras. Den första är en utveckling av en tidigare algoritm till att gälla för interpolering både med derivatavillkor och för matrisvärda funktioner i en optimeringsformulering. Den andra algoritmen kan lösa även randfallet med spektralnollställen på enhetscirkeln. Detta ger också en inneboende numerisk robusthet. Den tredje algoritmen behandlar fallet med både kepstral- och kovariansbivillkor.

Två designparadigmer har uppstått från den komplexitetsbegränsade analytiska interpoleringsteorin. Den första möjliggör design av $\mathcal{H}_\infty$-regulatorer av låg grad inom robust styrteori. Detta illustreras i ett benchmarkproblem för formning av MIMO-känslighetfunktionen. Ett användarstöd för att ställa in regulatordesignen inom paradigmen för SISO-fallet presenteras också. I den andra paradigmen skattas ARMA-modeller som är unika, beror kontinuerligt på data och kan frekvensviktas. I AR-fallet diskuteras frekvensviktning för skattningar inom kovariansutvidgningen med ett exempel som illustration. För ARMA-skattning generaliseras simultan kepstral- och kovariansmatchning till att omfatta förfiltrering. Ett exempel antyder att detta kan möjliggöra skattningar som är asymptotiskt effektiva.

Acknowledgments

Anders Lindquist, I am most grateful that you have shared your comprehensive experience, your commitment to mathematics, your wide-spread research network, and for the stimulating, self-organizing, research environment you have created at the division. Altogether this has been a scientific hotbed for me.

Ryozo Nagamune, you have been my scientific mentor, most frequent collaborator, and a dear friend all the way through my doctoral studies. Your solid work and nonnegotiable devotion to science have been great inspiration. I am honored and pleased that we have continued to collaborate after you left for Berkeley.

Next I wish to acknowledge my other collaborators and coauthors, whose work will be present in this thesis. Manfred Deistler, thanks for sharing your mathematical insights, developing my statistical mindset, and for twice hosting me in Vienna. Bo Wahlberg, I appreciate your clear thinking and your work free from prestige. The idea from our collaboration, however somewhat more elaborated upon, is to me the core part of this thesis. Vanna Fanizza, thanks for ambitiously and with solid mathematical thinking contributing to our joint work.

The division of Optimization and Systems Theory provides a great working environment thanks to a great faculty; in particular I want to thank a few. Ulf Jönsson, I am genuinely grateful to you for your unselfish work with courses and seminars as well as for giving all kind of advise and feedback. I sincerely hope you will be rewarded for this in your academic career. Anders Forsgren and Krister Svanberg, you are the teachers that made me start as a graduate student. At the end of the day, your work is the core part of the division to me. Claes Trygger, thank you for good collaboration with the basic course on Systems Theory and for all stimulating political discussions.

Petter Ögren, ever since you were my TA in the fall of 1995, you have been my professional guide and inspiration. Thanks for taking care of me when arriving at the division and thanks for becoming a great friend; your rapid sense of humor and high principles also made you an ideal officemate!

Henrik Rehbinder and Johan Karlsson, it has been very smooth sharing offices with you and I appreciate your company a lot. And Johan, if I had any positive influence in recruiting you to the division, that might be my most important contribution at the division; I wish you the best ever success with your research!

Per Enqvist, thank you being the expert on our theory who always allows time for explaining and for discussing with us less experienced.

I am also most thankful to all other past and present graduate students at the division that I have met. Sharing the goods and the bads of research and course work makes it easier and a whole lot more fun. The dynamic and open-minded atmosphere has made coffee breaks and lunches to a pleasure. Together with fascinating research this has made me look forward to go to work early every morning. To you in the most recent round of the "Krille course" – do enjoy the benefits of supporting each other and thank you for letting me be a nonaffiliated partner of your group.

I also wish to thank all my friends outside the division – not the least all my scouting friends – for providing a wonderful world despite the lack of Pick matrices. And thanks for each reminder of "närande" versus "tärande" – I believe they have encouraged me to try to keep the edge.

Finally, and the most, I wish to thank my immediate family: Elsa, Jan, Ingmar, Mats and more recently Malin and Carl. Thank you for caring, challenging, encouraging, and for giving the spirit to stand on the toes while not forgetting what really matters!

Stockholm, January 2005

Anders Blomqvist

Table of Contents

Chapter 1

Introduction

This thesis has its roots in fundamental systems theoretical problems concerning, preferably global, parameterizations of certain classes of linear systems. In [65] Kalman formulated the *rational covariance extension problem*, which amounts to parameterizing all rational Carathéodory functions of a certain degree that match a given finite covariance sequence. A related problem is parameterization of all internally stable sensitivity functions of bounded degree within $\mathcal{H}_\infty$ controller design.

In his thesis [50], Georgiou conjectured a complete parameterization but only proved the existence. It remained open until [32], where Byrnes, Lindquist, Gusev, and Matveev proved the uniqueness as well as a stronger assertion regarding the smoothness of the parameterization. Later, the theory has been developed significantly in various directions. This thesis can be seen as one such development.

The mathematical setting of this thesis is quite rich. It touches upon several large and developed mathematical areas. The main theorem of Section 3.1 deals with an approximation in the spectral domain, which combines, and thus generalizes, the results in [56, 25]. The approximation problem has a direct counterpart in analytic interpolation theory and can be interpreted as extensions of classical problems, such as the Nevanlinna-Pick interpolation problem. The analytic interpolation theory has also developed into operator theory starting with the seminal paper [94], and there are recent developments in that direction [28]. Our major tool for solving the approximation problem will be optimization theory and convexity. These will in fact imply uniqueness of the parameterization. In Section 3.2 the parameterization is shown to yield global smooth coordinates, utilizing tools from differential geometry and topology.

Beyond the purely mathematical background, each application needs its own tools; the applications in robust control require an $\mathcal{H}_\infty$ formulation, while applications in system identification and time series analysis require a statistical framework.

The theory is interesting *per se*, but it has also highly interesting applications. Two new significant paradigms for robust control [27, 81] and autoregressive moving

average (ARMA) estimation [25, 45] have evolved from the theory.

In robust control, the parameterization of bounded degree $\mathcal{H}_\infty$ transfer functions has accommodated design of controllers in a new fashion. In the conventional $\mathcal{H}_\infty$ implementation of Nevanlinna-Pick interpolation, the designer is referred to an intricate choice of weighting functions to satisfy the specifications. In this case, the degree of the weighting functions is added to the controller's degree. In several benchmark examples, controllers using the new paradigm have similar performance as conventional $\mathcal{H}_\infty$ controllers, but are of a much lower McMillan degree. Such examples will be given in this thesis. Constructing new methods for designing lower-order $\mathcal{H}_\infty$ controllers is an active research field, see for instance [105, p. 305].

For ARMA estimation, the paradigm give new global coordinates, which can be directly estimated from data. The transformation from the new coordinates to the ARMA parameters can readily be done by solving a convex optimization problem. This is in sharp contrast to Maximum Likelihood and Prediction Error estimators, which rely upon *nonconvex* optimization, yielding fundamental problems concerning smoothness with respect to data as well as uniqueness of the estimates. Also, failure modes need to be taken care of. In fact, constructing a statistically efficient ARMA estimator with guaranteed convergence and using reasonable computational effort, is an open problem. The relatively limited use of ARMA models in signal processing can probably be attributed to this fact, see for instance [100, p. 103]. Moreover, in ARMA estimation, the proposed framework accommodates the option of incorporating *a priori* information in the estimation procedure. By using a filter-bank with appropriate basis function, we can achieve high-resolution spectral estimation by emphasizing some parts of the spectrum. Another option is to use a common filter, a prefilter. Doing that for a Maximum Likelihood method requires a frequency domain formulation [73, 88] with weighting. A framework for prefiltering within the new paradigm, will be developed in Section 5.3.

There are many applications that fit into the paradigms above. For instance advanced design of hard disk drives is one suitable robust control application, and coding of speech one ARMA estimation application within signal processing. However, it is beyond the scope of the thesis to present any real-world applications in detail. Instead, the examples discussed will be of a more academic character, highlighting the features of the theory.

Below we illustrate with two simple examples – one from robust control and the other from ARMA estimation. Here the discussion is fairly shallow and the intention is to introduce some standard problems and the paradigms. These examples also illustrate some shortcomings of the paradigms, which we resolve in this thesis. In Chapter 5 we will study somewhat more realistic and challenging examples.

Example 1.0.1 (Sensitivity shaping). Consider the one-degree-of-freedom[1] feedback system depicted in Figure 1.1. Given a linear model of the plant P we wish to design a low-degree linear controller C, so that the closed loop system meet certain specifications. The lower case letters denotes the various signals in the system.

[1] A control design that does not allow for a prefilter.

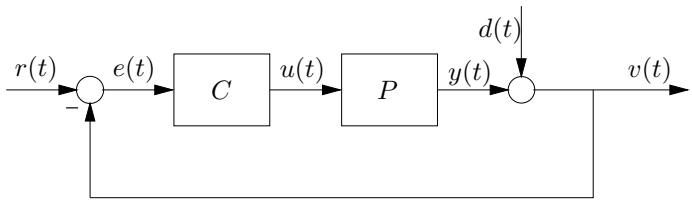

Figure 1.1: A one-degree-of-freedom feedback system.

Define the *sensitivity function* as

$$S := (I + PC)^{-1}. \tag{1.1}$$

It is well recognized in the robust control literature, that the sensitivity function has great influence on the closed-loop system's performance, such as robust stability, noise/disturbance attenuation, and tracking. In fact, sensible controller design can be accomplished by appropriate shaping of the sensitivity function. We will formulate our controller design problem in terms of the sensitivity function.

Consider the simple example, where a continuous time plant $P(s) = 1/(s-1)$ is given. It is well-known that to ensure that a feedback system is (internally) stable, we must have no unstable pole-zero cancellation in any transfer function of the closed-loop system. From the robust control literature we know that a necessary and sufficient condition for internal stability is that we have no unstable pole-zero cancellation between P and C in the sensitivity function and that the sensitivity function is stable. The given plant has an unstable pole at $s = 1$ and an unstable zero at infinity. To ensure internal stability we therefore need to fulfill the *interpolation conditions* $S(1) = 0$ and $S(\infty) = 1$.

Now assume that we are looking for simple solutions in the sense that we want the controller to be of a low (McMillan) degree. This typically leads to less computational effort in a digital implementation and a design that is easier to analyze. Not surprisingly, it turns out that bounding the degree of the sensitivity function, will also bound the degree of the controller, see [81]. In this trivial example we will look for sensitivity functions of degree one, and immediately we see that all solutions are of the form $S = (s-1)/(s+\delta)$.

We wish the closed-loop system to be robust against various noises and misspecifications of the plant. A successful way to achieve this is to bound the infinity norm of the sensitivity function, $\|S\|_\infty < \gamma$. This is known as $\mathcal{H}_\infty$ control design and we will work within such a framework. The lowest such bound, that is the infimum of $\|S\|_\infty$ over all stable S satisfying the interpolation conditions, will be denoted by γ_{opt}. There are optimal solutions achieving this bound, and their largest values are uniform over the spectrum. However, in general one would like to shape

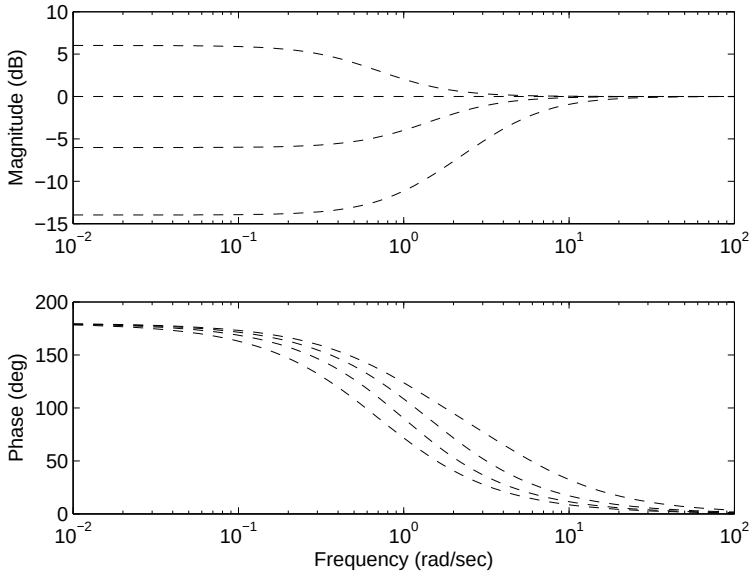

Figure 1.2: Frequency responses for the internally stabilizing solutions corresponding to $\delta = 0.5, 1, 2,$ and, 5.

the sensitivity function to obtain low sensitivity in a designated part of the spectrum, which, due to the water-bed effect [97], is done at the expense of higher sensitivity in some other part of the spectrum. To achieve this, it is customary to use weighting functions, which however could increase the degree of the sensitivity function considerably, and hence the controller. However, since we prefer sensitivity functions of low complexity, we would like to avoid weighting functions. To this end, and to allow for greater design flexibility, we consider suboptimal solutions, of which there are infinitely many. In our little example we require $\|S\|_\infty < \gamma := 2$.

In general it is highly non-trivial to parameterize all sensitivity functions fulfilling both the interpolation conditions and the norm condition. In our small example however, all solutions are parameterized by $\delta > 0.5$. We interpret δ as a *tuning parameter*. In Figure 1.2 Bode plots of the sensitivity function for $\delta = 0.5, 1, 2,$ and, 5 are drawn. We note that there is quite some difference between the different solutions. The corresponding controllers are given by $C = (1 - S)/PS$. In our case we get

$$C = \frac{a - \dfrac{s-1}{s+\delta}}{\dfrac{1}{s-1}\dfrac{s-1}{s+\delta}} = \delta + 1.$$

The cancellations in the computation are no coincidences. In fact, the following holds for any scalar plant.

Proposition 1.0.2. *[81] If the plant P is real-rational, strictly proper, and has n_p and n_z unstable poles and zeros, respectively, and the sensitivity function S is real-rational, internally stable, and fulfills* $\deg(S) \leq n_p + n_z - 1$*, then the controller $C = (1 - S)/PS$ is real-rational, proper and its degree satisfies*

$$\deg C \leq \deg P - 1. \tag{1.2}$$

This proposition justifies putting a degree constraint on the sensitivity function since it directly carries over to the controller. Moreover, it is often important to restrict the degree of the sensitivity function itself, since that corresponds to the dynamic behavior of the closed-loop system. However, note that this degree-bound in some applications with a high-order plant model might not be effective enough.

In the general case, in contrast to this simple example, it is unknown, and maybe impossible, to get a closed form set of solutions to the problem. Yet, it is in fact possible to parameterize all the solutions in this set. That is the key observation yielding a new design paradigm for robust control, see [27, 77, 83, 78, 81]. For each choice of the tuning parameters there is exactly one solution in the set. The tuning is performed so that the other design specifications are met; these can, for instance, be on the step response and the corresponding control signal.

The difficult part is the (complete) parameterization and that is the key problem in the complexity constrained analytic interpolation theory. However, there are also other issues that need to be resorted in order to enable a useful design method for practitioners. Some such issues that are studied in this thesis are listed below.

- In many applications there are also interpolation conditions on the derivative of the sensitivity function. For instance having a strictly proper continuous plant and requiring the controller to be strictly proper will lead to a derivative constraint at infinity. To deal with derivative conditions one can transform the problem to one without derivative conditions using bilinear transformations [101], which, however, is quite involved, see [58]. Instead, one can directly formulate the problem allowing for derivative constraints, see for instance [51]. In [9], software for such a formulation was presented. All theory and algorithms in this thesis allow for the derivative case. Examples 5.1.3 and 5.2.7 contain derivative conditions.

- In many applications, there are more than one control signal to be governed. Then the plant will be multi-input multi-output (MIMO) rather than single-input single-output (SISO), as in the example above. This thesis contains a generalization of the theory to the matrix-valued case which has been published in [8]. Example 5.1.3 is taken from there and illustrates that theory.

- To facilitate application of the theory to real-world control problems, development of accurate, reliable, robust, and numerically efficient software is of

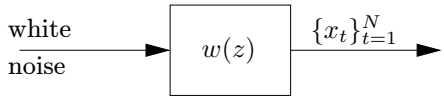

Figure 1.3: The shaping filter producing an ARMA process

key importance. Chapter 4 as well as Section 5.2, are devoted to development of algorithms to compute the solution corresponding to each choice of tuning parameters.

- A good design is achieved only if we tune the controller properly. Being a new paradigm, rules and support for tuning is most important. Also, user-friendly software to support the tuning is of vital importance. In Example 5.1.3, some basic ideas of how the tuning can be performed are included. A more comprehensive approach formulating the tuning problem itself as an optimization problem, is presented in Section 5.2 which is based on [82]. It is accompanied by the software [11].

Example 1.0.3 (ARMA estimation). Consider Figure 1.3. We assume, for the time being, that the measured, scalar data $\{x_t\}_{t=1}^N$ is generated by feeding white noise, say Gaussian, with variance λ^2 through a stable, causal, minimum-phase linear filter of some known degree. That implies that we should model the normalized transfer function of the shaping filter as:

$$w(z) = \frac{\sigma(z)}{a(z)},$$

where a and σ are monic stable polynomials of some degree. Given the measurement we want to determine the best possible model of the filter according to some criterion.

Let us consider the very simple example $a(z) \equiv 1$, $\sigma(z) = 1 - \sigma_1^{-1} z$ for $-1 < \sigma_1 < 1$, and $\lambda = 1$. This corresponds to a Moving Average process of order one, MA(1).

There is a large number of estimation schemes, estimators, that are applicable to the given problem. The estimator can be compared with respect to several criteria and under several different formulations. Criteria can be statistical, predictive ability, computation time, reliability, and smoothness of the estimated model with respect to data. Different estimators are best with respect to different criteria.

In statistics the Maximum Likelihood (ML) is probably the most widely used estimator, see for instance [20]. It is the best possible estimator with respect to the statistical criteria. The counterpart in the engineering literature is the Prediction Error Method (PEM), which minimizes the prediction error and is widely used for off-line estimation, [73]. The PEM and the ML methods are equivalent when the driving white noise is Gaussian. These estimators are based on nonconvex

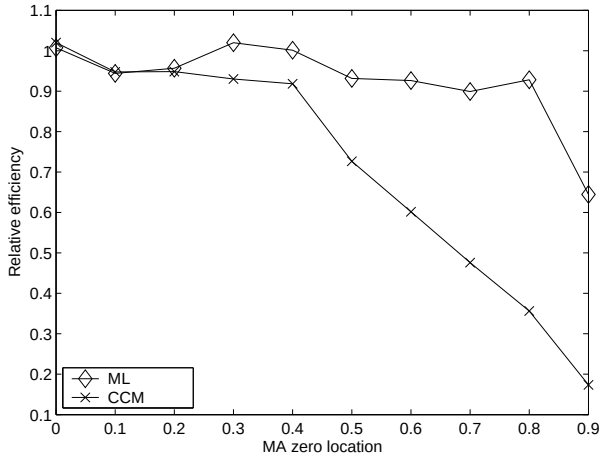

Figure 1.4: The estimated relative efficiency of the estimated zero of an MA(1) model using the ML and CCM estimators.

optimization and therefore typically computationally demanding. Also they need to treat failure modes caused by the nonconvexity.

In many signal processing applications the computational time and reliability are the limiting factors. Therefore, there is a large body of literature on algorithms which are fast and reliable, but typically, with larger variance of the estimates. This type of estimators are also used for initialization of ML estimators.

The theory and algorithms studied in this thesis constitute another estimator, which provides a new paradigm for ARMA estimation by introducing new, global coordinates, which can be directly estimated from the time series. The basic idea was first published in [25], where it was referred to as the Cepstral Covariance Matching (CCM) estimator since it exactly matches a window of covariances and cepstral coefficients. It was refined in [45]. There are different ways of estimating the covariances and cepstral coefficients yielding several versions of the estimator.

Now, for $0 \leq \sigma_1 \leq 0.9$ we generate a long data set, say $N = 1000$, and apply both the ML estimator `armax` in [71] and the CCM method of [45] with biased sample covariances and cepstral estimates based on a long AR model (length $L = 20$). The Cramér-Rao bound is known to be $N^{-1}(1-\sigma_1^2)$, see for instance [90, Chapter 5.2]. In Figure 1.4, the estimated *relative efficiency*, which is the ratio between the Cramér-Rao bound and the estimated variances, is plotted for each method based on a Monte Carlo simulation with 1000 realizations. The ML estimator is approximately efficient, that is, it has approximately relative efficiency 1, as expected from the theory. The CCM estimator seems to be efficient for a zero location close to origin but not otherwise, in that the relative efficiency is significantly less than one.

The MA(1) example illustrates both the idea and some of its present short-

comings of the new paradigm for ARMA estimation. Also note that the performance depends on several factors, for instance, what is the true process we study, what is the model class, the location in the parameter space, and the sample size. Next we will discuss some issues related to the ARMA identification problem.

- As seen from the example, the basic version of the CCM estimator does not seem to be an attractive alternative for high-quality estimation, not being asymptotically efficient – at least not with the currently known ways to estimate the cepstral coefficients. Later in the thesis, we will discuss an approach to improve the quality of the estimates, see Section 5.4 and Example 5.4.2 in particular. In fact, a modified CCM method presented there, seems to meet the Cramér-Rao bound. However, we will not study the statistical properties in detail.

- In the example the time series was generated by a model in the same class that we identify the model from. Generically, the true processes is generated by a more complex model than the one identified. This case is quite involved to analyze and less studied in the literature. However, in Example 5.3.2 such an example will be studied.

- In many applications we possess some *a priori* knowledge about the process. For instance one might know that there is a high frequency part generated by noise, in which we are not interested. Another case is when we are looking for a spectral component in a small frequency band. In the present setup, we can deal with such problems using a prefilter and/or a filter-bank. This enables high resolution spectral estimation within our framework and was first studied in [26]. The case of orthonormal basis functions was studied in [31, 5]. Prefiltering in the AR case was studied in [13], where it was shown, that under certain conditions, prefiltering in our framework is equivalent to weighting of the ML criterion in the frequency domain. This analysis will be included in conjunction with Example 5.3.2.

- In general the ARMA process might be vector-valued and then called a VARMA process. Estimation of VARMA models is more challenging than the scalar case since the nonconvexity of many formulations become a more significant obstacle. Extending the CCM method to the VARMA case would be most interesting since we rely on convexity. The matrix-valued extension in Section 3.3 is one somewhat restrictive possibility and we will not include any example of VARMA estimation.

Next we will outline the thesis and present which parts of the thesis that has been published elsewhere. The thesis is a part of a large research program at KTH lead by Prof. A. Lindquist in collaboration with Profs. C. Byrnes and T. Georgiou. A number of past and present doctoral students at KTH are also major contributors. The results presented in this theses are largely based on collaborate work with Dr.

Ryozo Nagamune, Prof. Anders Lindquist, Prof. Bo Wahlberg, Prof. Manfred Deistler, and Giovanna Fanizza in [4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 82]. That work has been approximately evenly distributed among the authors.

Chapter 2

This chapter gives some mathematical background as well as some background for $\mathcal{H}_\infty$ control and ARMA estimation. In doing that, almost all the notation used in the thesis is defined. Most of the material is classical and available in the literature and appropriate references are given. However, a couple of preliminary results related to the theory developed in this thesis are stated and proven; more precisely, Lemma 2.3.4 and Proposition 2.3.5 which are published, in slightly different form, in [82] and Lemma 2.4.1 which appeared in [8].

Chapter 3

This is the theory chapter containing the major theoretical developments in the thesis. The first section, Section 3.1, is a generalization/combination of results in [25] and [56], and is previously unpublished. In the following section, Section 3.2, the geometric results regarding the parameterization with cepstral and covariance coefficients in [25], are generalized to the situation in Section 3.1. Also this material is new. Finally, in Section 3.3, some results are generalized to the matrix-valued case, closely following [8].

Chapter 4

In this chapter, algorithms for computing the interpolants, and the spectral densities, corresponding to the theory in Chapter 3, are developed. The first algorithm, given in Section 4.1, treats the matrix-valued case but without cepstral-type constraints, and has been published in [8]. The second algorithm, in Section 4.2, takes another approach than the previous optimization-based algorithms and is thereby capable of treating the case with boundary spectral zeros. It was published in [6, 7]. The last algorithm, given in Section 4.3, treats the general formulation in the theory chapter, by considering a homotopy on a particular manifold. It is previously unpublished.

Chapter 5

This chapter contains four examples, two instances of ARMA estimation and two of robust control. The first example, in Section 5.1, applies our approach to a benchmark MIMO sensitivity shaping problem and was included in [8]. In Section 5.2 a framework for tuning the scalar sensitivity function is presented and a benchmark example is studied. The results are previously published in [82, 11]. In Section 5.3 a covariance extension approach to prefiltering is studied and illustrated with an example of AR estimation. The results are published in [13]. Finally, Section 5.4

contains an example of ARMA estimation using the full theory of Chapter 3 and the algorithm in Section 4.3. These results are new.

Chapter 6

This final chapter contains some conclusion and remarks of the thesis as well as a brief discussion regarding open problem related to the material of the thesis.

Chapter 2

Background and Notation

This chapter will briefly give the mathematical setting of the thesis. We will also formulate some mathematical problems as well as problems from applications and give some classical solutions. It is instructive to see the solutions presented in the thesis in the light of the classical solutions.

2.1 Spectral and Complex Analysis

In this section we will introduce the function spaces we use. We also mention the connection between spectral densities, spectral factors and positive-real functions. Finally we define the Kullback-Leibler discrepancy for spectral densities.

Let $\mathbb{T}$, $\mathbb{D}$, and $\mathbb{C}_+$ denote the unit circle, the open unit disc, and the open right half-plane in the complex plane, respectively. On the unit circle, we define all monotone, nondecreasing functions $\mu(\theta)$ to be the set of distribution functions. By the Lebesgue Decomposition and the Radon-Nikodym Theorem, see for instance [34, 93], any distribution function can uniquely be decomposed into an absolutely continuous component, a discrete (jump) component, and a singular component. Moreover, the absolutely continuous part can be written

$$d\mu_a(\theta) = \Phi(e^{i\theta})d\theta,$$

where Φ is the *spectral density*, which will be a key ingredient in this thesis. Any spectral density associated with an absolutely continuous distribution is in the set of positive, continuous, real-valued functions on the unit circle denoted by $\mathcal{C}_+$. Also, let $\mathcal{C}$ be the set of not necessarily positive, continuous, real-valued functions on the unit circle.

Consider a complex function f that is analytic in a disc with radius r centered

at the origin in the complex plane. Define the norms

$$M_p(r,f) = \begin{cases} \left(\frac{1}{2\pi}\int_0^{2\pi} |f(re^{i\theta})|^p d\theta\right)^{1/p}, & 0 < p < \infty, \\ \max\limits_{0\le\theta\le 2\pi} |f(re^{i\theta})|, & p = \infty. \end{cases}$$

Now we can define the complex[1] *Hardy spaces* $\mathcal{H}_p$ for $0 < p \le \infty$ as the set of complex functions f analytic in $\mathbb{D}$ and for which $M_p(r,f)$ remain bounded as $r \to 1$ from below. Thus, $\mathcal{H}_2$ is the space of complex functions analytic in the disc and with power series expansion $\sum a_n z^n$ such that $\sum |a_n|^2 < \infty$, while $\mathcal{H}_\infty$ is the space of bounded analytic functions in the disc.

The Hardy spaces can be identified with the subsets of the complex Lebesgue p-integrable functions on the unit circle, $\mathcal{L}_p(\mathbb{T})$, or simply $\mathcal{L}_p$, with vanishing negative Fourier coefficients. In particular, $\mathcal{L}_2$ will be a Hilbert space equipped with the inner-product

$$\langle f, g\rangle = \frac{1}{2\pi}\int_{-\pi}^{\pi} \overline{f(e^{-i\theta})} g(e^{i\theta}) d\theta,$$

which is inherited by $\mathcal{H}_2$. Note that spectral densities $\Psi \in \mathcal{L}_2$. We will use the inner-product notation extensively to simplify notation.

We will encounter the subspace of $\mathcal{H}_2$ consisting of the (strictly) *positive-real functions* (also called *Carathéodory* functions in mathematical literature). They have the property that they map the closed unit disc to the open right half-plane, that is, their real part is positive in the region of analyticity. Sometimes we will encounter the closure of the subspace, also containing the not necessarily strictly positive real functions, that is functions which map the closed unit disc to the *closed* right half-plane.

Next we will introduce some more subspaces that later will be most useful. Given a Hurwitz matrix $A \in \mathbb{C}^{(n+1)\times(n+1)}$, that is a matrix with all its eigenvalues in the closed open disc, we define the finite Blaschke product

$$B(z) := \frac{\det(zI - A^*)}{\det(I - Az)}.$$

The Blaschke product gives a natural orthogonal decomposition of $\mathcal{H}_2$ as

$$\mathcal{H}_2 = B\mathcal{H}_2 \oplus \mathcal{H}(B),$$

where $B\mathcal{H}_2$ is called the invariant subspace and $\mathcal{H}(B)$ the coinvariant subspace. The invariant subspace consists of all $\mathcal{H}_2$ functions that vanish on the spectrum of A^*, that is $f(A^*) = 0$, interpreted as a power series expansion evaluated with the matrix A^* as variable. It is called the invariant subspace since it is invariant under

[1] Sometimes, in particular regarding algorithms and applications, we will consider the subclass consisting of real functions, that is, with real coefficients in the Fourier expansion.

the shift, $zB\mathcal{H}_2 \subset B\mathcal{H}_2$. The coinvariant subspace is n-dimensional and can be expressed as

$$\mathcal{H}(B) = \left\{ \frac{a(z)}{\tau(z)} : a(z) = a_0 + a_1 z + \dots a_n z^n, \tau = \prod_{k=1}^{n} (1 - \overline{z_k}^{-1} z) \right\}.$$

It is called coinvariant since it is invariant under the adjoint to the shift operator. The space of constant functions will be given by $\mathcal{H}(B) \cap \mathcal{H}(B)^*$, whereas we define the union as

$$\mathcal{Q} := \mathcal{H}(B) \cup \mathcal{H}(B)^*.$$

Also, define the positive, symmetric subspace as

$$\mathcal{Q}_+ := \{Q \in \mathcal{Q} : Q(z) = Q^*(z), Q(z) > 0 \ \forall z \in \mathbb{T}\}.$$

Also, let $\overline{\mathcal{Q}_+}$ be the subset including nonnegative pseudo-polynomials, that is, $Q(z) \geq 0$ for all $z \in \mathbb{T}$.

Spectral factorization connects spectral densities on the circle and functions analytic in the unit disc, see for instance [34, p. 204f]. Given a spectral density $\Phi \in \mathcal{C}_+$ there exists a spectral factor $w(z) \in \mathcal{H}_2$ such that $w^{-1}(z) \in \mathcal{H}_2$ and

$$w(z)w^*(z) = \Phi(z), \text{ where } w^*(z) := \overline{w(\overline{z}^{-1})}.$$

In mathematical literature such a function is called *outer*. A rational outer function has all its poles and zeros outside the unit disc. Moreover, the factorization is unique up to orthogonal transformations. We shall denote the class of outer functions $\mathcal{P}$.

Another connection between spectral densities on the circle and functions analytic in the unit disc is manifested by the Riesz-Herglotz representation [1]:

$$f(z) = iv + \frac{1}{2\pi} \int_{-\pi}^{\pi} \frac{e^{i\theta} + z}{e^{i\theta} - z} \Phi d\theta, \tag{2.1}$$

where

$$v = \mathrm{Im} f(0).$$

Note that this f is positive-real. For a spectral densities $\Phi \in \mathcal{C}_+$ we have

$$\Phi(z) = 2\Re\{f(z)\} = f(z) + f^*(z) = w(z)w^*(z).$$

Thus, the spectral density is twice the real part of the analytic function f. Given a density, the analytic function f is uniquely determined up to an imaginary constant by (2.1).

Now, restrict the consideration to finite degree *real rational functions*:

$$w(z) = \lambda \frac{\sigma(z)}{a(z)} = \lambda \frac{z^m + \sigma_1 z^{m-1} + \dots \sigma_m}{z^n + a_1 z^{n-1} + \dots + a_n}. \tag{2.2}$$

Note that we here index the coefficients with decreasing powers of the variable. In particular we will be interested in rational functions with all poles and zeros outside the unit circle. Let the *Schur region* $\mathcal{S}_n$ be the n-dimensional smooth manifold of monic polynomials with all roots outside the unit disc. For simplicity of notation we will identify this function space with the space of coefficients:

$$\mathcal{S}_n = \left\{ a \in \mathbb{R}^n : z^n + a_1 z^{n-1} + \cdots + a_n \neq 0 \ \forall z \in \overline{\mathbb{D}} \right\}.$$

Normalized outer rational functions, that is with $\lambda = 1$, then belong to the direct product of two Schur regions

$$\mathcal{P}_{nm} := \mathcal{S}_n \times \mathcal{S}_m.$$

If the polynomials are of the same degree we simply write $\mathcal{P}_n$. Also, define the dense subset $\mathcal{P}^*_{nm}$ consisting of all coprime rational functions in $\mathcal{P}_{nm}$.

The topology of $\mathcal{P}^*_n$ is fairly complicated. Firstly, note that the Schur region $\mathcal{S}_n$ in general is nonconvex. Secondly, the coprimeness assumption divides the space into $n+1$ connected components, see [19, 96].

Also, introduce the function space $\mathcal{L}_n$ consisting of all not necessarily monic polynomials of degree at most n.

The spectral density corresponding to the rational spectral factor $w(z)$ can be written

$$\Phi(z) = w(z)w^*(z) = \lambda^2 \frac{\sigma(z)\sigma^*(z)}{a(z)a^*(z)} = \frac{P(z)}{Q(z)}, \tag{2.3}$$

where P and Q are *pseudo-polynomials* of degrees m and n defined as

$$\begin{aligned} P(z) &:= 1 + p_1/2(z + z^{-1}) + \cdots + p_m/2(z^m + z^{-m}), \\ Q(z) &:= q_0 + q_1/2(z + z^{-1}) + \cdots + q_n/2(z^n + z^{-n}). \end{aligned}$$

We can generalize the pseudo-polynomials to be represented in some other function space:

$$\frac{P(z)}{Q(z)} = \frac{P(z)}{\tau(z)\tau^*(z)} \Big/ \frac{Q(z)}{\tau(z)\tau^*(z)},$$

where we define

$$\tau(z) := \det(I - Az) = \tau_0 + \tau_1 z + \cdots + \tau_{n+1} z^{n+1}. \tag{2.4}$$

By taking $A = 0$ we recover the pseudo-polynomials. Also note, that if $\det A = 0$, $\tau_{n+1} = 0$ so $\tau(z)$ is of degree at most n. The generalized pseudo-polynomial can now be written

$$Q(z) = \sum_{k=0}^{n} \frac{q_k}{2} (G_k(z) + G_k^*(z)) = \frac{a(z)a^*(z)}{\tau(z)\tau^*(z)}, \tag{2.5}$$

for some basis functions $G_k(z)$ spanning $\mathcal{H}(B)$ and where $q(z)$ is a polynomial of degree n. Again we identify the space of pseudo-polynomials that are positive on the unit circle with the space of coefficients, given some basis functions:

$$\mathcal{Q}_+ = \left\{(q_0, q_1, \ldots, q_n) \in \mathbb{R}^{n+1} : Q(z) > 0, z \in \mathbb{T}\right\}.$$

We also define the subset for which the leading coefficient $q_0 = 1$ as $\mathcal{Q}_+^0$.

Remark 2.1.1. *Since Q is symmetric on the unit circle we can, when suitable, index the coefficients of $a(z)$ and $\sigma(z)$ in (2.2) in decreasing powers of the variable.*

Next consider the relation between rational spectral factors and rational positive-real functions:

$$\frac{b(z)}{a(z)} + \frac{b^*(z)}{a^*(z)} = \frac{b(z)a^*(z) + a(z)b^*(z)}{a(z)a^*(z)} = \frac{T(b(z))a(z)}{a(z)a^*(z)} = \frac{\sigma(z)\sigma^*(z)}{a(z)a^*(z)},$$

where we have defined the operator $T(b) : \mathcal{L}_n \to \mathcal{Q}_+$. It is well-known, see for instance [33, 30], that $T(b)$ is an invertible linear operator for all $b \in \mathcal{S}_n$. As for the functions, T will also be used as an operator between the coefficient spaces. In particular, considering the numerator we have the equation for the coefficients of the polynomials and pseudo-polynomials as

$$T(b)a = d, \tag{2.6}$$

that is, T is the Hankel + Toeplitz linear operator

$$T(b) := \begin{bmatrix} b_0 & \ldots & b_{n-1} & b_n \\ b_1 & & b_n & \\ \vdots & \iddots & & \\ b_n & & & \end{bmatrix} + \begin{bmatrix} b_0 & b_1 & \ldots & b_n \\ & b_0 & & b_{n-1} \\ & & \ddots & \vdots \\ & & & b_0 \end{bmatrix}.$$

The equation (2.6) will be studied in some detail in Section 4.2.

In this thesis we will be interested in comparing spectral densities. However, we will not define a true metric but rather a discrepancy. Since spectral densities can be interpreted as distribution functions in the spectral domain we will adopt a discrepancy from statistics called *spectral Kullback-Leibler discrepancy* [67]:

Definition 2.1.2 (Spectral Kullback-Leibler discrepancy). *Given two spectral densities $\Psi, \Phi \in \mathcal{C}_+$ with common zeroth moment, $\langle 1, \Psi \rangle = \langle 1, \Phi \rangle$, the spectral Kullback-Leibler discrepancy is given by*

$$\mathbb{S}(\Psi, \Phi) := \left\langle \Psi, \log \frac{\Psi}{\Phi} \right\rangle.$$

It is not symmetric in its arguments but jointly convex. It fulfills $\mathbb{S}(\Psi, \Phi) \geq 0$ with equality if and only if $\Psi = \Phi$, see for instance [56]. The spectral Kullback-Leibler discrepancy is a generalization of the entropy of a spectral density, which is recovered by taking $\Psi \equiv 1$. In [27] a similar generalization, namely $\langle P, \log \Phi \rangle$ where $P \in \mathcal{Q}_+$ was considered, and denoted *generalized entropy*.

2.2 Linear Systems

In this section we introduce some system-theoretical notions, which connect the spectral analysis and the applications in robust control, signal processing, and time series analysis. We will also state some spectral factorization results for rational functions of degree n.

A finite dimensional, discrete-time, linear, time-invariant system is a system of ordinary difference equations which can be represented as

$$\begin{aligned} x_{k+1} &= Ax_k + Bu_k, \\ y_k &= Cx_k + Du_k, \end{aligned} \tag{2.7}$$

where $k \in \mathbb{Z}$ is any integer denoting the time-indexing, $A \in \mathbb{C}^{n\times n}$, $B \in \mathbb{C}^{n\times q}$, $C \in \mathbb{C}^{p\times n}$, and $D \in \mathbb{C}^{p\times q}$. Here we call the processes u_k and y_k the input and the output, respectively. Applying the z-transform to the system equations (2.7) we get the frequency representation

$$\begin{aligned} z\hat{x}_k &= A\hat{x}_k + B\hat{u}_k, \\ \hat{y}_k &= C\hat{x}_k + D\hat{u}_k. \end{aligned} \tag{2.8}$$

The *transfer function* is the map between $\hat{u}_k$ and $\hat{y}_k$ and is given by $P(z) = D + C(zI - A)^{-1}B$. We say that P has the state-space representation (A, B, C, D) and that it is *real rational* whenever its state space matrices are real. The degree of the realization is $n = \dim(A)$. In general P will be a matrix-valued rational function. Obviously, each component of P will be a rational function of degree at most n. A state-space representation is said to be *minimal* if there is no other realization of the transfer function, that has a smaller degree. This minimal degree is called the *McMillan degree*.

We will be interested in particular classes of systems. An important property that we are interested in is *stability*, that is, the question whether the output remains bounded for certain classes of inputs. In general this is a property of a solution rather than of a system, but for our linear, time-invariant systems it is, in fact, a system property. We shall call a system *stable* if and only if the eigenvalue of A are in the open unit disc[2] and thus neglect the boundary case.

Another class of systems that we will be particularly interested in, are those, which are *miniphase* (minimum-phase). Suppose we swap in- and output of a system, that is we run the system reversely. If the reverse system also is stable, then the system is miniphase. Note that for a system which is both stable and miniphase with transfer function $P(z)$, we have that $P^*(z)$ is an outer function in mathematical terminology.

Another important property is *passivity*. A system is passive if the energy of the input is larger or equal to the energy of the output for all time intervals. As

[2]Sometimes these systems are called asymptotically stable.

shown in [104] a linear time-invariant discrete time system is *passive* if and only if the transfer function is positive-real.

Stochastic realization theory is a topic in systems theory dealing with linear stochastic systems. One of the most celebrated results in stochastic realization theory is the Kalman-Yakubovich-Popov lemma, or the positive-real lemma, which gives a necessary and sufficient condition for positive-realness.

We follow the expositions in [70, 68]; see also [34]. Let $f(z)$ be a stable real rational function with a minimal state-space representation $(A, \bar{B}, C, \bar{D})$. Then

$$\Phi(z) = f(z) + f^*(z) = \begin{bmatrix} C(zI - A)^{-1} & I \end{bmatrix} M(E) \begin{bmatrix} (z^{-1}I - A^T)^{-1}C^T \\ I \end{bmatrix}, \tag{2.9}$$

where

$$M(E) = \begin{bmatrix} E - AEA^T & \bar{B}^T - AEC^T \\ \bar{B} - CEA^T & \bar{D} + \bar{D}^T - CEC^T \end{bmatrix}.$$

Therefore, if there exists an $E = E^T > 0$ fulfilling the LMI (Linear Matrix Inequality)

$$M(E) \geq 0, \tag{2.10}$$

we can factorize

$$M(E) = \begin{bmatrix} B \\ D \end{bmatrix} \begin{bmatrix} B^T & D^T \end{bmatrix}.$$

Consequently, (2.9) yields $\Phi(z) = w(z)w^*(z)$ where

$$w(z) = C(zI - A)^{-1}B + D,$$

is a spectral factor. A fundamental question is under what conditions there exists an E fulfilling (2.10), and the answer is given by the positive-real lemma.

Theorem 2.2.1 (The Positive-Real Lemma). *There exists a solution to* (2.10) *if and only if $f(z)$ is positive-real.*

We will also state a uniqueness result concerning the spectral factorization of rational functions of degree n, which also holds for the matrix-valued case.

Theorem 2.2.2. *[[68, Theorem 3.1] [34, Theorem 6.3]] Fix the minimal state-space realization $(A, \bar{B}, C, \bar{D})$ of $f(z)$. Then there is a one-to-one correspondence between the minimal-degree spectral factors $w(z)$ and the set of positive solutions E to* (2.10)*, modulo orthogonal transformations.*

To represent transfer functions and their corresponding spectral densities we will use basis functions. A suitable framework for this, which can be interpreted as filter-banks, is given in [52, 53]. Let $A \in \mathbb{C}^{n+1 \times n+1}$ and $B \in \mathbb{C}^{n+1}$. The pair (A, B) is called *reachable* if the reachability matrix

$$\Gamma := \begin{bmatrix} B & AB & \dots & A^nB \end{bmatrix},$$

has full rank. If, in addition, A have all its eigenvalues in $\mathbb{D}$, we define the basis functions

$$\begin{bmatrix} G_0 \\ G_1 \\ \vdots \\ G_n \end{bmatrix} := G(z) := (I - Az)^{-1}B = B + Az(Iz - A)^{-1}B,$$

In particular, if $\det A = 0$ one basis function will be a constant. Clearly, the basis functions G_k will be analytic in $\mathbb{D}$. Define a set of basis functions as

$$\mathcal{G} := \left\{ G(z) = (I - Az)^{-1}B : \begin{array}{l} A \in \mathbb{C}^{n+1\times n+1},\ B \in \mathbb{C}^{n+1}, \\ \mathrm{eig}(A) \subset \mathbb{D},\ (A,B) \text{ reachable},\ G_0(z) \equiv 1, \\ \langle G_0, G_k \rangle = \delta_{0k}, k = 1, \ldots, n+1, \end{array} \right\}. \tag{2.11}$$

For such basis function we define $\bar{G}$ by $G =: \begin{bmatrix} 1 & \bar{G}^T \end{bmatrix}^T$.

Some choices of (A, B) will be of particular interest. The all-pole standard basis is given by

$$A = \begin{bmatrix} 0 & & & \\ 1 & 0 & & \\ & \ddots & \ddots & \\ & & 1 & 0 \end{bmatrix}, \quad B = \begin{bmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix}, \quad \Rightarrow\ G(z) = \begin{bmatrix} 1 \\ z \\ \vdots \\ z^{n-1} \end{bmatrix}. \tag{2.12}$$

To get simple poles in z_k we can take

$$A = \begin{bmatrix} z_0 & & \\ & \ddots & \\ & & z_n \end{bmatrix}, \quad B = \begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix}, \quad \Rightarrow\ G(z) = \begin{bmatrix} (1 - z_0 z)^{-1} \\ \vdots \\ (1 - z_n z)^{-1} \end{bmatrix}. \tag{2.13}$$

Finally, taking

$$A = \begin{bmatrix} a & & & \\ (1-a^2) & a & & \\ \vdots & & \ddots & \\ (-a)^{n-1}(1-a^2) & (-a)^{n-2}(1-a^2) & \ldots & a \end{bmatrix}, \quad B = \begin{bmatrix} \sqrt{1-a^2} \\ -a\sqrt{1-a^2} \\ \vdots \\ (-a)^n\sqrt{1-a^2} \end{bmatrix},$$

$$\Rightarrow\ G(z) = \begin{bmatrix} \frac{\sqrt{1-a^2}}{1-az} \\ \frac{\sqrt{1-a^2}}{1-az}\frac{z-a}{1-az} \\ \vdots \\ \frac{\sqrt{1-a^2}}{1-az}\left(\frac{z-a}{1-az}\right)^n \end{bmatrix},$$

we get the so called Laguerre basis, see for instance [102].

In general, if (A, B) are balanced, that is, if $AA^* + BB^* = I$, then the basis functions are orthonormal, that is $\langle G_k, G_l \rangle = \delta_{kl}$. Among the above mentioned ones, the standard basis and the Laguerre basis are orthonormal. Orthonormal basis functions have been used in systems modeling and identification, see [15], and are also a natural choice in our framework, see [31, 5].

2.3 Some Parameterization Problems

In this section we will formulate the Nevanlinna-Pick interpolation problem with and without degree constraint. We state the solutions. Also using the covariance-type interpolation data, we reformulate the problem via some preliminary results. Finally we define the spaces of some interpolation data.

A classical problem in complex analysis is the existence and characterization of positive-real functions (Carathéodory functions) which interpolate some prescribed values. The problem can be formulated in several more or less equivalent ways, for instance:

Problem 2.3.1 (The Nevanlinna-Pick Interpolation Problem). *Given the interpolation data $\{(z_k, w_k)\}_{k=0}^n \subset \mathbb{D} \times \mathbb{C}_+$, does there exist a positive-real function f such that $f(z_k) = w_k,\ k = 0, 1, \dots n$? If so, characterize all solutions.*

The key result in the classical analytic interpolation theory is a necessary and sufficient condition on the interpolation data, $\{(z_k, w_k)\}_{k=0}^n$, for existence of a solution as well as a general characterization of all solutions. In fact, for the Nevanlinna-Pick formulation we have:

Theorem 2.3.2. *There exists a solution to the Nevanlinna-Pick interpolation problem 2.3.1 if and only if the* Pick matrix

$$\Sigma := \begin{bmatrix} \frac{w_0+\bar{w}_0}{1-z_0\bar{z}_0} & \frac{w_0+\bar{w}_1}{1-z_0\bar{z}_1} & \cdots & \frac{w_0+\bar{w}_n}{1-z_0\bar{z}_n} \\ \frac{w_1+\bar{w}_0}{1-z_1\bar{z}_0} & \frac{w_1+\bar{w}_1}{1-z_1\bar{z}_1} & \cdots & \frac{w_1+\bar{w}_n}{1-z_1\bar{z}_n} \\ \vdots & \vdots & & \vdots \\ \frac{w_n+\bar{w}_0}{1-z_n\bar{z}_0} & \frac{w_n+\bar{w}_1}{1-z_n\bar{z}_1} & \cdots & \frac{w_n+\bar{w}_n}{1-z_n\bar{z}_n} \end{bmatrix}, \tag{2.14}$$

is nonnegative definite.

If Σ is singular the solution is unique and otherwise all solutions can be represented by a linear-fractional transformation, see for instance [103, p. 286f]:

$$f(z) = \frac{t_1(z)g(z) + t_2(z)}{t_3(z)g(z) + t_4(z)}, \tag{2.15}$$

where $t_k(z)$ are certain rational functions of degree n and $g(z)$ is some arbitrary positive-real function.

The condition $\Sigma \geq 0$ is called the *Pick condition*. We will show this classical result in a much more general formulation in Section 2.4.

This problem was studied by Pick in [87] and later independently by Nevanlinna in [85]. Nevanlinna proved the theorem by constructing a solution with the so called *Nevanlinna algorithm*. The algorithm leads to the characterization in (2.15). Nevanlinna was inspired by Schur [95] who studied the related *Carathéodory problem*. This problem involves interpolation conditions on the function and its n first derivatives at the origin and the conditions for existence of a solution is positivity of a Toeplitz matrix. This result is called the Carathéodory-Toeplitz theorem. Also, note that taking $g(z) \equiv 1$ in (2.15) gives the so called *central solution* which is rational of degree at most n.

Problem 2.3.1 is stated in terms of positive-real functions. However, by means of the *Möbius*, or bilinear, transformation there are other equivalent formulations. Next to the positive-real functions the *bounded real* functions are the most widely used in control engineering. Bounded real functions are analytic in the unit disc and map the unit disc into itself (or possibly into a disc with radius γ). In mathematical literature they are called *Schur functions*. A well-written introduction to the classical analytic interpolation theory and the use of the Möbius transformation from a control viewpoint, can be found in [66].

Next we will formulate a variation of the classical Nevanlinna-Pick problem which will be the starting point of this thesis.

Problem 2.3.3 (The Nevanlinna-Pick Interpolation Problem with Degree Constraint). *Given* $\{(z_k, w_k)\}_{k=0}^{n} \subset \mathbb{T} \times \mathbb{C}_+$*, does there exist a positive-real function* f *of* degree at most n *such that* $f(z_k) = w_k,\ k = 0, 1, \ldots n$*? If so, characterize all solutions.*

The degree constraint is motivated by applications in engineering and time series analysis as mentioned in the introduction. Also, in the Schur version, where there are interpolation conditions on the function and its first n derivatives at the origin, the degree constraint turns the problem into the rational Carathéodory extension problem, first formulated by Kalman [65]. Therefore Problem 2.3.3 is also of pure systems theoretical interest.

Before stating the solution to Problem 2.3.3, we will formulate a more general class of analytic interpolation problem for which Problem 2.3.3 is a special case. In the sequel of papers [52, 53, 54, 55], the analytic interpolation problem, for both the scalar and matrix-valued case, is studied in this framework based on the idea of using basis functions in a filter-bank, or an input-to-state filter. The same idea was previously used in [26, 27]. In doing that, the state covariance matrix will exactly be the Pick matrix. This yields a general class of interpolation problems where the interpolation data is represented in terms of a set of basis functions $G \in \mathcal{G}$ and the Pick matrix Σ. The Pick matrix will have a certain structure as explained in the aforementioned papers. In fact,

$$\Sigma = WE + EW^*, \tag{2.16}$$

where E is the reachability Gramian and W is an interpolation data matrix commuting with A. That is, we have

$$W = w_0 I + w_1 A + \cdots + w_n A^n, \tag{2.17}$$

for some $w_k \in \mathbb{C}$. The reachability Gramian is defined by

$$E := \frac{1}{2\pi} \int_{-\pi}^{\pi} G(e^{i\theta}) G^*(e^{i\theta}) d\theta,$$

and can be computed by solving the Lyapunov equation $E - AEA^* = BB^*$. Also, W takes particular forms for different choices of basis functions. In the all pole case (2.12) we have

$$W = \begin{bmatrix} w_0 & 0 & \dots & 0 \\ w_1 & w_0 & & \\ \vdots & \ddots & \ddots & \\ w_n & \dots & w_1 & w_0 \end{bmatrix},$$

making Σ the standard Toeplitz matrix. In the simple pole case W is simply the diagonal matrix with the interpolation values:

$$W = \begin{bmatrix} w_0 & & \\ & \ddots & \\ & & w_n \end{bmatrix},$$

corresponding to the Pick matrix in (2.14). For a general problem W is a block-diagonal matrix corresponding to Jordan blocks in A. The interpolation conditions in terms of the positive-real function can then be expressed as

$$f(A^*) = W^*, \tag{2.18}$$

where the right-hand side is interpreted as evaluating the Fourier expansion with the matrix A^* as variable.

The positive-real functions in Problem 2.3.3 can be expressed in terms of their numerator and denominator polynomials $b(z)$ and $a(z)$. It is a trivial, but, as we shall see later, useful observation that the first degree constraint and the interpolation conditions lead to a linear, invertible relation between a and b. We have:

Lemma 2.3.4. *Let (A, B) be a given reachable pair and let the Pick matrix Σ be positive. Let $a(z)$ and $b(z)$ be polynomials of degree at most n such that $f(z) = b(z)/a(z)$ fulfills (2.18). Then there is a linear, invertible relation between the coefficients of $a(z)$ and $b(z)$:*

$$\overline{b} = \Gamma^{-1} W \Gamma \overline{a} =: K\overline{a}, \tag{2.19}$$

where Γ is the reachability matrix of (A, B).

Proof. By [52, Lemma 1] the interpolation conditions (2.18) is known to be equivalent to $\overline{f}(A)B = WB$. Substituting the rational representation we have

$$(\overline{a}(A))^{-1}\overline{b}(A)B = WB.$$

Since A and W commute this is equivalent to $\overline{b}(A)B = W\overline{a}(A)B$. Now the linear map can be constructed as:

$$\begin{aligned}
\overline{b}(A)B &= W\overline{a}(A)B, \\
(\overline{b}_0 I + \overline{b}_1 A + \cdots + \overline{b}_n A^n)B &= W(\overline{a}_0 I + \overline{a}_1 A + \cdots + \overline{a}_n A^n)B, \\
\begin{bmatrix} B & AB & \dots & A^n B \end{bmatrix} \begin{bmatrix} \overline{b}_0 \\ \overline{b}_1 \\ \vdots \\ \overline{b}_n \end{bmatrix} &= W \begin{bmatrix} B & AB & \dots & A^n B \end{bmatrix} \begin{bmatrix} \overline{a}_0 \\ \overline{a}_1 \\ \vdots \\ \overline{a}_n \end{bmatrix}, \\
\overline{b} &= \Gamma^{-1} W \Gamma \overline{a}, \\
\overline{b} &= K\overline{a}.
\end{aligned}$$

Here, Γ is invertible due to the reachability assumption. The matrix W can be brought to Jordan form using a similarity transformation U:

$$\hat{W} = UWU^{-1}. \tag{2.20}$$

The diagonal elements of $\hat{W}$ are the interpolation values (not derivative) and thus positive. Therefore $\hat{W}$ and thus W are positive definite. □

We have a direct result regarding the properties of K.

Proposition 2.3.5. *Let (A, B) and W be given as above. The spectrum of the matrix $K = \Gamma^{-1}W\Gamma$ is given by $\sigma(K) = \{w_{j0}\}_{j=1}^{n}$, where w_{j0} is the interpolation value in z_j for each j.*

Proof. Since (A, B) is a reachable pair, Γ is nonsingular. Then $\sigma(K) = \sigma(W)$ since K is a similarity transformation of W. Again bringing W to its Jordan form as in (2.20), $\hat{W}$ is triangular. The eigenvalues of $\hat{W}$ are simply the diagonal elements, since it is triangular. □

Therefore, for given interpolation data the polynomial $b(z)$ is a function of $a(z)$ (and vice versa). Also define the linear operator K on the set of *not necessarily monic polynomials*:

$$\begin{aligned}
K \quad : \quad \mathbb{R} \times \mathcal{S}_n &\rightarrow \mathbb{R} \times \mathcal{S}_n, \\
a(z) &\mapsto b(z) = K(a(z)).
\end{aligned}$$

As a direct consequence of Lemma 2.3.4, the set of interpolating positive-real function of degree at most n can be represented as

$$\mathcal{A}_n := \left\{ a \in \mathbb{R}^{n+1} : a_0 > 0, \frac{K(a(z))}{a(z)} \text{ positive-real} \right\}. \tag{2.21}$$

Then Problem 2.3.3 has a solution if and only if $\mathcal{A}_n$ is nonempty. If so, how can we characterize it?

Clearly, assuming the Pick matrix to be positive, there exists at least one solution to Problem 2.3.3, namely the central solution obtained by setting $g \equiv 1$ in Theorem 2.3.2. However, it was for a long time unclear how to characterize the set $\mathcal{A}_n$. In his thesis [50], Georgiou conjectured a complete parameterization in terms of the so called dissipation polynomial $d(z)$ in (2.6) or equivalently, by spectral factorization, the spectral polynomial $\lambda\sigma(z)$ in $\mathbb{R}_+ \times \mathcal{S}_n$. He also proved that, for each choice of dissipation polynomial, there exists an element in $\mathcal{A}_n$. Uniqueness of the parameterization remained open until [32].

Theorem 2.3.6. *[32, 30] Suppose the Pick matrix is positive. Then the map*

$$\begin{array}{rccc} H & : & \mathbb{R}_+ \times \mathcal{S}_n & \to & \mathcal{A}_n, \\ & & (\lambda, \sigma) & \mapsto & a, \end{array}$$

such that $a(z)K^*(a(z)) + a(z)^*K(a(z)) = \lambda^2\sigma(z)\sigma^*(z)$ *is a diffeomorphism.*

The statement of the theorem is in fact stronger than Georgiou's conjecture, which just claimed the map H to be bijective. The smoothness manifested by the diffeomorphism is most important in motivating numerical procedures as well as questions regarding well-posedness.

Next we study the case when a part of the density is pre-specified. Given some spectral density $\Psi \in \mathcal{C}_+$ and some basis functions $G \in \mathcal{G}$ we define

$$r_k := \frac{1}{2\pi}\int_{-\pi}^{\pi} G_k(e^{i\theta})\Psi(e^{i\theta})\Phi(e^{i\theta})d\theta = \langle G_k, \Psi\Phi\rangle\,, \tag{2.22}$$

for the spectral density Φ. Note that we can think of Ψ as the density of a prefilter that is applied to the signal. For the special case with $\Psi \equiv 1$ and G as in (2.12) the components r_k will be Fourier coefficients of the spectral density. In the setting of stochastic processes, these are exactly the covariances of the processes. For the special case of $\Psi \equiv 1$ and G as in (2.13) the components r_k are interpolation values on the positive-real part of the spectral density in the poles of the basis: $f(z_k) = r_k$. We will call (2.22) *generalized prefiltered covariances*. The corresponding Pick matrix is given by

$$\Sigma = \frac{1}{2\pi}\int_{-\pi}^{\pi} G(e^{i\theta})\Psi(e^{i\theta})\Phi(e^{i\theta})G^*(e^{i\theta})d\theta,$$

and is related to the interpolation data matrix W via (2.16). The covariance vector is then given by $r = WB$. Likewise, given the reachable pair (A, B) and the covariance vector r there is a unique Pick matrix $\Sigma(r)$. In fact, using the representation in (2.17) the coefficients w_j are given by $w = \Gamma^{-1}r$. Hence we can determine W and then compute Σ by (2.16). We define the set of feasible generalized prefiltered covariances as

$$\mathcal{R}_n := \left\{r \in \mathbb{C}^{n+1} : \Sigma(r) > 0\right\}.$$

In this thesis we will also study moments of the logarithm of the spectral density defined as

$$c_k := \frac{1}{2\pi}\int_{-\pi}^{\pi} G_k(e^{i\theta})\Psi(e^{i\theta})\log\Big(\Psi(e^{i\theta})\Phi(e^{i\theta})\Big)d\theta = \langle G_k, \Psi\log(\Psi\Phi)\rangle\,. \tag{2.23}$$

For the special case with $\Psi \equiv 1$ and G as in (2.12) the components c_k will be Fourier coefficients of the logarithm of the spectral density. In signal processing and speech processing, in particular, these are called *cepstral coefficients*, see for instance [14, 86]. In signal processing, the cepstral coefficients have traditionally been considered as an alternative to covariances for parameterizing AR models. However, in the innovative paper [25], it was shown that a combination of cepstral and covariance coefficients provide a *bona fide* coordinate systems for ARMA processes. In this thesis we will generalize that result. The basis G generalize the notion of cepstrum together with the prefiltering that Ψ represents. Therefore, we will call (2.23) the *generalized prefiltered cepstral coefficients*.

Next consider how to compute the cepstral coefficients, that is the case $\Psi \equiv 1$ and G as in (2.12), for a rational spectral density Φ. The integration in (2.23) can be done in a finite number of elementary operations if we proceed with some care as pointed out in [25, p. 46]. In fact, consider a Laurent expansion of $\log\Phi = \log w + \log w^*$ on a subset $\Omega \subset \mathbb{C}$, which is an intersection between an annulus of the unit circle containing no poles and zeros of Φ and a sector containing the real line. Note that we can not directly consider the whole unit circle since circling the origin adds 2π to the logarithm. Now a series expansion on the real line in Ω of $\log w$ and $\log w^*$ extends to the whole of Ω and in particular to the part of the unit circle contained in Ω. Due to the uniqueness of the Fourier transform, the Laurent expansion also holds for the rest of the unit circle. Now, the cepstral coefficients are quite easily computed, see for instance [86, 25], as

$$\begin{aligned} c_0 &= 2\log\lambda, \\ kc_k &= s_k(a) - s_k(\sigma), \end{aligned} \tag{2.24}$$

where $s_k(a) = z_1^k + z_2^k + \cdots + z_n^k$ for a Schur polynomial a with roots in z_k. Numerically, it is faster to compute s_k using the recursive Newton's Identities [25, 16].

As we are interested in the Nevanlinna-Pick problem with degree constraint, it is instrumental to define the set of covariances and cepstral coefficients that corresponds to a rational density of degree n. We will slightly generalize the definitions in [25, 69], which do this in an implicit fashion. Let $\Psi \in \mathcal{C}_+$ and $G, H \in \mathcal{G}$ be given. Define

$$\mathcal{X}_{nm} := \left\{ (r,c) \in \mathbb{C}^{n+1}\times\mathbb{C}^n : \begin{array}{l} r \in \mathcal{R},\ \lambda \in \mathbb{R}_+,\ \sigma \in \mathcal{S}_m,\ a \in \mathcal{S}_n, \\ r_k = \left\langle H_k, \Psi\lambda^2\dfrac{\sigma\sigma^*}{aa^*}\dfrac{\tau\tau^*}{tt^*}\right\rangle, k = 0,1,\ldots,n \\ c_k = \left\langle G_k, \Psi\log\Psi\lambda^2\dfrac{\sigma\sigma^*}{aa^*}\dfrac{\tau\tau^*}{tt^*}\right\rangle, k = 1,2,\ldots,m \end{array} \right\}, \tag{2.25}$$

where $\tau = \det(I - A_H z)$ and $t = \det(I - A_G z)$. In many situations we will have $m = n$ and $G = H$; then we will denote the set $\mathcal{X}_n$.

Remark 2.3.7. *The definition of $\mathcal{X}_n$ is implicit, making it as hard to check whether an element belongs to it, as to solve for the interpolating function. However, it will be of great theoretical value to define the set this way. For actual computation of an interpolant, there are ways to circumvent this difficulty as discussed and shown in the following chapters.*

2.4 Matrix-Valued Generalizations

In this section we will introduce some matrix-valued generalizations of the function classes *et cetera*. We will denote matrix-valued functions with capital letters to clarify what case we are treating. We also state and prove an intermediate result regarding the invertibility of the generalized linear Hankel + Toeplitz operator $T(\cdot)$.

An $\ell \times \ell$ matrix-valued function F that is analytic in the closed unit disc $\overline{\mathbb{D}}$ is called strictly positive-real if the spectral density function

$$\Phi(z) := \Re\{F(z)\} := \left[F(z) + F^*(z)\right], \text{ where } F^*(z) := \overline{F(\overline{z}^{-1})}^T,$$

is positive definite for all $z \in \mathbb{T}$. Here $\Re\{F(z)\}$ is the Hermitian generalization of the real part in the scalar case. If F is positive-real, so is F^{-1}. In particular, all the poles and zeros of F are located outside the unit circle, that is in $\overline{\mathbb{D}}^c$. The corresponding matrix-valued spectral density belongs to $\mathcal{C}_+^\ell$ and thus includes also nonrational densities. Given a density, the corresponding positive real function is given by the Riesz-Herglotz formula (2.1), now in its matrix form; see, for instance, [37]. Let $\mathcal{Q}_+^\ell$ denote the corresponding set of n^{th} order pseudo-polynomials. We also generalize the inner product to

$$\langle F(z), G(z) \rangle = \frac{1}{2\pi} \int_{-\pi}^{\pi} \operatorname{trace}\left[F^*(e^{i\theta}) G(e^{i\theta})\right] d\theta,$$

To each rational positive-real function F there corresponds an $\ell \times \ell$ matrix-valued stable, miniphase spectral factor $W(z)$ such that

$$W^*(z)W(z) = \Phi(z) = \Re\{F(z)\}, \tag{2.26}$$

which is unique modulo an orthogonal transformation. Determining W from F is a spectral factorization problem, which can be solved by determining the stabilizing solution of an algebraic Riccati equation, see for instance [34]. Conversely, if

$$W(z) = zC(I - zA)^{-1}B + D,$$

is any minimal realization of W, appealing to the equations of the Kalman-Yakubovich-Popov Lemma, there is a unique F satisfying (2.26), and it is given by

$$F(z) = z(B^*XA + D^*C)(I - zA)^{-1}B + B^*XB/2 + D^*D/2,$$

where X is the unique solution to the Lyapunov equation

$$X = A^*XA + C^*C.$$

Moreover, W is a proper rational function of the same McMillan degree as F, and so is the inverse W^{-1}.

Let the polynomial σ be the least common denominator of all entries in W^{-1}. Then there is a matrix polynomial $A(z)$ of the same degree as σ such that $W(z)^{-1} = A(z)/\sigma(z)$, and consequently

$$W(z) = \sigma(z)A(z)^{-1}.$$

Note that there is a slight risk of notational confusion since A also represents the basis functions. However, from the context it should be clear what is meant. In this representation, the degree $r := \deg \sigma$ is uniquely determined by F; to emphasize this we write $r(F)$. Now, define the class

$$\mathcal{F}_+(n) := \{F \text{ positive-real } : r(F) \leq n\}.$$

All functions $F \in \mathcal{F}_+(n)$ have McMillan degree at most ℓn, but, although this is a nongeneric situation, there are positive-real functions F of McMillan degree at most ℓn that do not belong to $\mathcal{F}_+(n)$. In fact, the standard observable and standard reachable realization of W^{-1} have dimension ℓr, see for instance [18, p. 106], and consequently W^{-1}, and hence F, has McMillan degree at most ℓr. Moreover, the standard observable realization may not be minimal, so there is a thin set of positive-real functions F of McMillan degree at most ℓn for which $r(F) > n$.

Generalize the Schur region to $\mathcal{S}_n^\ell$ consisting of all $\ell \times \ell$ outer real matrix polynomials of degree at most n with the first coefficient matrix upper triangular. Then $\mathcal{S}_n^\ell$ has real dimension $\ell^2 n + \frac{1}{2}\ell(\ell+1)$. Also, let $\mathcal{L}_n^\ell$ be the space of all $\ell \times \ell$ real matrix polynomials of degree at most n.

Let $Q(z)$ be of the form in (2.5) but take $Q_k \in \mathbb{C}^{\ell\times\ell}$. Again we identify the space of coefficients with the space of functions:

$$\mathcal{Q}_+^\ell = \left\{(Q_0, Q_1, \ldots, Q_n) : Q_k \in \mathbb{C}^{\ell\times\ell}, Q(z) > 0, z \in \mathbb{T}\right\}. \tag{2.27}$$

Note that the inequality means positive definite.

For $A \in \mathcal{S}_n^\ell$ we also generalize the linear operator $T(A)$:

$$\begin{array}{rccl} T(A) & : & \mathcal{L}_n^\ell & \to & \mathcal{L}_n^\ell, \\ & & V(z) & \mapsto & A(z)V^*(z) + V(z)A^*(z), \end{array}$$

Following [8] we will prove the following lemma regarding the invertibility of $T(A)$, generalizing the result in [33, 30] to the matrix case.

Lemma 2.4.1. *Let $A(z) \in \mathcal{S}_n^\ell$ and assume that $\det A(z)$ and $\det A^*(z)$ have no roots in common. Then, the linear map $T(A)$ is nonsingular.*

A proof of Lemma 2.4.1 restricted to the case when $\det A$ has all its roots in the complement of the closed unit disc can be found in [64]; see also [62, Theorem 3.1], which refers to [64]. Here we shall provide an independent proof of the more general statement of Lemma 2.4.1. Indeed, our proof is short and straight-forward. Moreover, the general statement given here was left as an open problem in [64, p. 28].

Proof. Since $T(A)$ is a linear map between Euclidean spaces of the same dimension, it suffices to prove that $T(A)$ is injective. Without restriction we may assume that $A(z)$ is upper triangular. In fact, let $U(z)$ be a unimodular matrix polynomial with $U(0)$ upper triangular such that $U(z)A(z)$ is upper triangular. Such a U indeed exists due to the procedure deriving the Smith form [48]. Then

$$UT(A)VU^* = (UA)(UV)^* + (UV)(UA)^* = 0,$$

if and only if $T(A)V = 0$. Moreover, the new V_0, that is $T(0)V(0)$, is still upper triangular. In this formulation

$$\det A(z) = a_{11}(z)a_{22}(z)\cdots a_{\ell\ell}(z),$$

where $a_{11}, a_{22}, \cdots, a_{\ell\ell}$ are the diagonal elements in $A(z)$. In particular, by assumption, no a_{ii} can have zeros in common with any a^*_{jj}. It then remains to prove that

$$AV^* + VA^* = 0, \tag{2.28}$$

implies $V = 0$. The proof is by induction. The statement clearly holds for $\ell = 1$. In fact, if $A(z_j) = 0$, then, by assumption, $A^*(z_j) \neq 0$, and hence, by (2.28), $V(z_j) = 0$. Consequently, we must have $V(z) = \lambda(z)A(z)$ for some real polynomial λ, which inserted into (2.28) yields

$$(\lambda + \lambda^*)AA^* = 0.$$

This implies that $\lambda = 0$ and hence that $V = 0$, as claimed. Now, suppose that (2.28) implies $V = 0$ for $\ell = k - 1$. Then, for $\ell = k$, (2.28) can be written

$$\left[\begin{array}{c|ccc} a_{11} & a_{12} & \cdots & a_{1k} \\ \hline 0 & & & \\ \vdots & & \hat{A} & \\ 0 & & & \end{array}\right] \left[\begin{array}{c|ccc} v^*_{11} & v^*_{21} & \cdots & v^*_{k1} \\ \hline v^*_{12} & & & \\ \vdots & & \hat{V}^* & \\ v^*_{1k} & & & \end{array}\right] + \left[\begin{array}{c|ccc} v_{11} & v_{12} & \cdots & v_{1k} \\ \hline v_{21} & & & \\ \vdots & & \hat{V} & \\ v_{k1} & & & \end{array}\right] \left[\begin{array}{c|ccc} a^*_{11} & 0 & \cdots & 0 \\ \hline a^*_{12} & & & \\ \vdots & & \hat{A}^* & \\ a^*_{1k} & & & \end{array}\right] = 0, \tag{2.29}$$

which, in particular, contains the $(k-1) \times (k-1)$ matrix relation $\hat{A}\hat{V}^* + \hat{V}\hat{A}^* = 0$ of type (2.28). Consequently, by the induction assumption, $\hat{V} = 0$, so, to prove that

$V = 0$, it just remains to show that the border elements $v_{11}, v_{12}, \ldots, v_{1k}, v_{21}, \ldots, v_{k1}$ are all zero. To this end, let us begin with the corner elements v_{1k} and v_{k1}. From the $(1,k)$ and $(k,1)$ elements in (2.29), we have

$$\begin{aligned} a_{11}v_{k1}^* + v_{1k}a_{kk}^* &= 0, & (2.30)\\ v_{1k}^* a_{kk} + v_{k1}a_{11}^* &= 0. & (2.31) \end{aligned}$$

In the same way as in the case $\ell = 1$, (2.30) implies that $v_{1k} = \lambda_{1k}a_{11}$ for some real polynomial λ_{1k}, and (2.31) implies that $v_{k1} = \lambda_{k1}a_{kk}$ for some real polynomial λ_{k1}, which inserted into (2.30) yields

$$(\lambda_{k1} + \lambda_{1k}^*)a_{11}a_{kk}^* = 0.$$

This implies that λ_{k1} and λ_{1k} are real numbers such that $\lambda_{1k} = -\lambda_{k1}$. However, by assumption, $V(0)$ is upper triangular, and $A(0)$ is upper triangular and nonsingular. Hence $v_{k1}(0) = 0$ and $a_{kk}(0) \neq 0$, implying that $\lambda_{k1} = v_{k1}(0)/a_{kk}(0) = 0$ and, consequently, $\lambda_{1k} = 0$. Since, therefore, $v_{1k} = 0$ and $v_{k1} = 0$, (2.29) now takes the form

$$\left[\begin{array}{c|c} \tilde{A} & * \\ \hline 0 & * \end{array}\right]\left[\begin{array}{c|c} \tilde{V}^* & 0 \\ \hline 0 & 0 \end{array}\right] + \left[\begin{array}{c|c} \tilde{V} & 0 \\ \hline 0 & 0 \end{array}\right]\left[\begin{array}{c|c} \tilde{A}^* & 0 \\ \hline * & * \end{array}\right] = 0,$$

which only yields the $(k-1)\times(k-1)$ matrix relation $\tilde{A}\tilde{V}^* + \tilde{V}\tilde{A}^* = 0$ of type (2.28). However, by the induction assumption, $\tilde{V} = 0$. Therefore, $V = 0$ in the case $\ell = k$ also, so, by mathematical induction, $V = 0$ for all k. □

For the matrix-valued case, we will only treat covariance-type interpolation conditions in this thesis:

$$R_k := \frac{1}{2\pi}\int_{-\pi}^{\pi} G_k(e^{i\theta})\Psi(e^{i\theta})\Phi(e^{i\theta})d\theta \quad k = 0, \ldots, n. \tag{2.32}$$

The space of interpolation values corresponding to a positive Pick matrix, we will denote $\mathcal{R}^\ell$.

The input-to-state framework for the basis functions $\mathcal{G}$ in (2.11), nicely generalizes to the matrix-valued case, as described in [54, 55] by letting $B \in \mathbb{C}^{(n+1)\times m}$.

Next we shall derive the Pick condition in this general setting, namely that $\Sigma > 0$ is a necessary and sufficient condition for existence of solutions to the Nevanlinna-Pick interpolation problem. Let $G \in \mathcal{G}$ and $\Psi(z) \in \mathcal{C}_+$ be given. If there exists a solution $\Phi(z) \in \mathcal{C}_+^\ell$, it needs to fulfill (2.32). Now, let $Q_k \in \mathbb{C}^{\ell\times\ell}$ and consider the sum

$$\begin{aligned} \mathrm{Re}\left\{\sum_{k=0}^{n}\mathrm{trace}(Q_k R_k)\right\} &= \mathrm{Re}\sum_{k=0}^{n}\mathrm{trace}\left(Q_k\frac{1}{2\pi}\int_{-\pi}^{\pi}G_k(e^{i\theta})\Psi(e^{i\theta})\Phi(e^{i\theta})d\theta\right),\\ &= \frac{1}{2\pi}\int_{-\pi}^{\pi}\mathrm{trace}(Q(e^{i\theta})\Psi(e^{i\theta})\Phi(e^{i\theta}))d\theta. \end{aligned} \tag{2.33}$$

The right-hand side will be positive for all $Q(z) \in \mathcal{Q}_+^\ell$, so a necessary conditions is that the sum on the left-hand side also is positive for all $Q(z) \in \mathcal{Q}_+^\ell$. Therefore we have

$$\operatorname{Re}\left\{\sum_{k=0}^{n} \operatorname{trace}(Q_k R_k)\right\} > 0, \forall\ Q(z) \in \mathcal{Q}_+^\ell. \tag{2.34}$$

Note that this condition is only in terms of the interpolation data R_k for $k = 0, \dots n$ and the basis functions $G(z)$. Therefore, again via the computation in (2.33), the positivity condition (2.34) holds whenever

$$\frac{1}{2\pi}\int_{-\pi}^{\pi} \operatorname{trace}(Q(e^{i\theta})R(e^{i\theta}))d\theta > 0, \forall\ Q(z) \in \mathcal{Q}_+^\ell, \tag{2.35}$$

where $R(z) \in \mathcal{C}^\ell$ fulfills the interpolation conditions in (2.32). Therefore, we have eliminated the positivity condition on the interpolating density, and moved it over to $Q(z)$. Being positive on the unit circle, $Q(z)$ admits a spectral factorization as

$$Q(z) = \Gamma(z)\Gamma^*(z), \tag{2.36}$$

where the miniphase, stable spectral factor has a representation as $\Gamma(z) = G(z)\Gamma$ for some matrix $\Gamma \in \mathbb{C}^{\ell(n+1)\times\ell}$. Plugging this into (2.35) we have

$$\begin{aligned}
&\frac{1}{2\pi}\int_{-\pi}^{\pi} \operatorname{trace}(Q(e^{i\theta})R(e^{i\theta}))d\theta \\
&\quad= \frac{1}{2\pi}\int_{-\pi}^{\pi} \operatorname{trace}(\Gamma(e^{i\theta})\Gamma^*(e^{i\theta})R(e^{i\theta}))d\theta, \\
&\quad= \frac{1}{2\pi}\int_{-\pi}^{\pi} \operatorname{trace}(\Gamma^*(e^{i\theta})R(e^{i\theta})\Gamma(e^{i\theta}))d\theta, \\
&\quad= \frac{1}{2\pi}\int_{-\pi}^{\pi} \operatorname{trace}(\Gamma\frac{1}{2\pi}\int_{-\pi}^{\pi} G^*(e^{i\theta})R(e^{i\theta})G(e^{i\theta})d\theta\Gamma), \\
&\quad= \operatorname{trace}(\Gamma\Sigma\Gamma),
\end{aligned} \tag{2.37}$$

where Σ is the Pick matrix. Therefore, the positivity condition in (2.34) holds if and only if the Pick matrix is positive. Thus we have proven the Pick condition in a quite general setting.

We note that specializing to the case when all interpolation points have multiplicity one, we obtain the classical Pick matrix

$$\Sigma = \left[\frac{R_i + R_j^*}{1 - z_i\overline{z}_j}\right]_{i,j=0}^{n}.$$

On the other hand, when there is only one interpolation point with multiplicity $n+1$ located at the origin, as in the classical Carathéodory extension problem, the

Pick matrix is the block Toeplitz matrix

$$\Sigma = \begin{bmatrix} R_0 + R_0^* & R_1^* & \cdots & R_n^* \\ R_1 & R_0 + R_0^* & \cdots & R_{n-1}^* \\ \vdots & \vdots & \ddots & \vdots \\ R_n & R_{n-1} & \cdots & R_0 + R_0^* \end{bmatrix}.$$

See for instance [38, 39] for the classical results.

One possible matricial generalization of the spectral Kullback-Leibler discrepancy is given below.

Definition 2.4.2 (Matricial Spectral Kullback-Leibler discrepancy). *Given one scalar spectral density $\Psi \in \mathcal{C}_+$ and one matrix-valued spectral density $\Phi \in \mathcal{C}_+^\ell$ with common zeroth moment, $\langle 1, \Psi \rangle = \langle 1, \log \det \Phi \rangle$, the matricial spectral Kullback-Leibler discrepancy is given by*

$$\mathbb{S}(\Psi, \Phi) := \left\langle \Psi, \log \frac{\Psi}{\det \Phi} \right\rangle.$$

This generalization is by no means the most general, but will be suitable for our framework.

2.5 Robust Control Theory

In this section we introduce the robust control theory. We also formulate the sensitivity shaping problem, for which we study some examples in Sections 5.1 and 5.2.

Control theory is the field studying the mathematical theory behind automatic control and it contains both analysis and synthesis. It bloomed fully in conjunction with the space race in the 1960's. During that period the linear-quadratic, LQ, control synthesis was developed. The LQ control minimizes an energy-type criterion, but does not put any emphasis on the worst case behavior. For case with no direct state measurements, this inevitably leads to problems concerning robustness various model errors and noises, see [40]. Starting in the 1980's the robust control theory, mainly within the $\mathcal{H}_\infty$ framework, has evolved. There is a large, and growing body, of literature on robust control, see for instance [105, 42, 41, 47].

In control applications, the mathematical problems are naturally stated in terms of bounded real rather than positive-real functions. For the continuous time case we consider the Hardy space $\mathcal{H}_\infty(i\mathbb{R})$ consisting of all complex-valued functions $f(s)$ which are analytic and *bounded* in the open right half-plane $\mathbb{C}_+$. In the discrete time case the function $f(z)$ has the unit disc $\mathbb{D}$ as the domain of analyticity and boundedness instead. However, as discussed in Section 2.3, by means of a Möbius transformation we can translate the problem into one for positive-real functions and thus remove the norm constraint. We will be particularly interested in real-rational

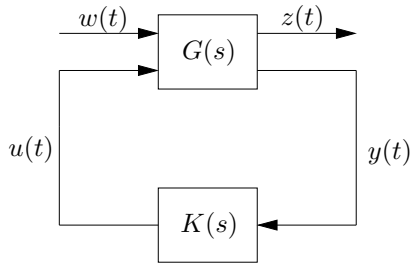

Figure 2.1: The generalized plant.

transfer functions $f(s)$, that is rational functions with real coefficients. These will be denoted $R\mathcal{H}_\infty$.

The standard setup in the robust control literature is using the generalized plant depicted in Figure 2.1. Here w is the exogenous input, y the sensor output, u the control input and z the regulated output. The transfer function $G(s)$ is called the generalized plant and $K(s)$ the controller. Here we have indicated the whole system to be in continuous time, but in general this might not be the case. In particular, in sample data control theory the generalized plant is typically taken in continuous time and the controller, being digitally implemented, is in discrete time. We will however, not touch upon these issues in this thesis.

We assume that all transfer functions are real-rational and that $G_{22}(s)$ is proper; that is, $D_{22} = 0$ for a state-space realization of G. Now introduce two additional input signals, v_1 and v_2, which are added to the outputs of G and K summing up to new y and u, respectively. Then the strict properness implies that all nine transfer functions from w, v_1, and v_2 to z, u, and y are stable, see [47]. If they are stable, that is, if they belong to $R\mathcal{H}_\infty$, K stabilizes the generalized plant. The closed-loop system is then called *internally stable*.

The transfer function from w to z for some controller is given by a lower linear fractional transformation

$$z = \left[G_{11} + G_{12}K(I - G_{22})^{-1}G_{21}\right] w. \tag{2.38}$$

We will consider the general $\mathcal{H}_\infty$ control problem, which amounts to finding a real-rational $K(s)$ that minimizes the infinity norm of the transfer matrix in (2.38) such that $G(s)$ is internally stabilized. Associated, there is what we might call the general suboptimal $\mathcal{H}_\infty$ parameterization problem, where we want to parameterize all controllers such that the norm is bounded by some real γ. There are several control problems that fit into the framework: sensitivity shaping, gain-margin maximization, and robust stabilization to mention a few.

Bounding the infinity norm is of vital importance in applications. By Plancherel's theorem the $\mathcal{H}_\infty$ norm provides a bound on the the system gain:

$$\sup\{\|\hat{z}\|_2 : \hat{x} \in R\mathcal{H}_2, \|\hat{x}\|_2 \leq 1\},$$

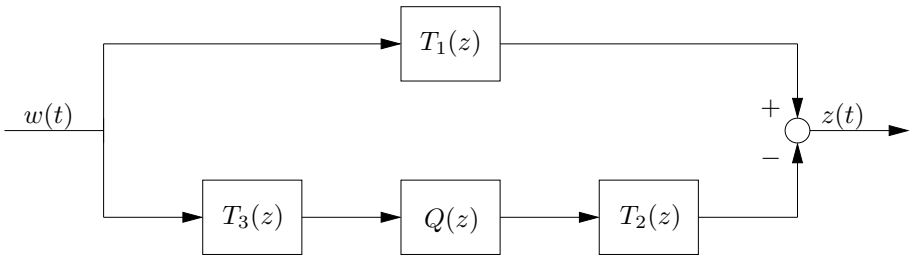

Figure 2.2: The model-matching setup.

where the Laplace or z-transformed signals belong to the real Hardy space $R\mathcal{H}_2$ of signals of bounded energy. This will enforce *robustness* of the closed loop system, a featured that was not attended to in the classical control theory. Since the $\mathcal{H}_\infty$ norm is the supremum over all cases it gives a *worst case* bound.

Consider Figure 2.2 illustrating the so called *model-matching problem*. The transfer functions $T_i(s)$ are assumed to be in $R\mathcal{H}_\infty$. We want to find all $Q(s) \in R\mathcal{H}_\infty$ such that the $\|T_1 - T_2QT_3\|_\infty < \gamma$. We can interpret this as given the model T_1, we want to match T_2QT_3 by choosing Q suitably. Clearly, this can be stated in a general $\mathcal{H}_\infty$ problem by taking

$$G = \begin{bmatrix} T_1 & T_2 \\ T_3 & 0 \end{bmatrix}, \text{ and } K = -Q.$$

As noted in [47, p. 19], K stabilizing G is equivalent to having $Q \in R\mathcal{H}_\infty$. A celebrated and more interesting result is the converse. Now assume that G is proper and that G_{22} is strictly proper; that is, that $D_{22} = 0$. Then there is an affine parameterization of all stabilizing K in terms of $Q \in R\mathcal{H}_\infty$. This is called the *Youla parameterization*, see, for instance, [42, 47]. Therefore, we can instead study this model matching problem which is easier to analyze. Once we have the solution, we can simply bring it back to K.

Given the controller parameterization a solution to the general $\mathcal{H}_\infty$ control problem can be computed either by solving a set of Riccati equations or by solving a linear matrix inequality (LMI). These are connected by the Kalman-Yacubovich-Popov lemma. The corresponding solution will then be of the same McMillan degree as the generalized plant unless an extra nonconvex rank condition is added. When tuning in the classical approach, design weighting functions are included in the generalized plant, see, for instance, [105, Chapter 6.2]. The additional degree of the weighting functions will then add to the controller degree.

The model matching problem can be brought into a Nevanlinna-Pick interpolation problem, see for instance [41, 61]. Here we will treat the scalar case whereas one generalization to the matrix-valued case is considered in Section 5.1. In the scalar case we can without loss of generality take $T_3 \equiv 1$. Define the combined

transfer function

$$T := T_1 - T_2 Q \Leftrightarrow Q = \frac{T_1 - T}{T_2}.$$

Clearly, $T \in R\mathcal{H}_\infty$ whenever $Q \in R\mathcal{H}_\infty$. However, the opposite does not hold since T_2 can have unstable zeros. To guarantee $Q \in R\mathcal{H}_\infty$ we must make sure that all unstable zeros of T_2 are cancelled. Suppose the unstable zeros of T_2 are simple and in z_j. Then we have

$$Q \in R\mathcal{H}_\infty \Leftrightarrow T \in R\mathcal{H}_\infty, T(z_j) = T_1(z_j), \forall j.$$

Thus, the general $\mathcal{H}_\infty$ parameterization problem is equivalent to the interpolation problem of parameterizing

$$\{T \in R\mathcal{H}_\infty : T(z_j) = T_1(z_j),\ \forall j, \|T\|_\infty < \gamma\}.$$

Using the Möbius transform this can be brought to the Nevanlinna-Pick problem 2.3.1. Adding a degree bound will bound the degree of T, which in turn typically will bound the degree of some devise in the control system.

One particular control problem that can be stated as a model matching problem is *sensitivity shaping*, which is a very important design problem within robust control. In this thesis we will study the sensitivity shaping problem in order to illustrate the design paradigm we are working on.

Consider the system setup in Figure 1.1. It is well-known that the sensitivity function, $S = (I + PC)^{-1}$ plays a major role in determining the behavior of the closed-loop system. In fact, a sensible controller design can be accomplished by a suitable design of the sensitivity function. In the paradigm that we study, we first design an internally stable sensitivity function and then compute the corresponding controller. This is the common approach in interpolation based design, see for instance [41, 61].

In general we want $|S|$ to be small for low frequencies to attain tracking and disturbance rejection. Pushing down the sensitivity in one frequency region unavoidably leads to a higher sensitivity in another region, by the so called *waterbed effect*, see for instance [41]. However, by imposing a uniform bound as well, we can also achieve a certain robustness, that is a gain-phase margin in the Nyquist plot. Other criteria for roll-off, bandwidth *et cetera* can be found [61, Chapter 3.5].

Substituting the Youla parameterization in the sensitivity function brings it to the model matching form. From there, interpolation conditions ensuring internal stability can be derived. As shown in [105, Theorem 5.5] the internal stability conditions can equivalently be stated as having $S \in R\mathcal{H}_\infty$ and that there is no unstable pole-zero cancellation in PK. See the examples in Chapter 5 for example on how to derive the interpolation conditions. Also, additional interpolation conditions can be a tool for shaping the sensitivity function, see [81].

2.6 Stationary Stochastic Processes

In this section, we introduce stationary stochastic processes and state the ARMA identification problem. We also treat model estimates and estimates of covariances as well as cepstra in a statistical framework, introducing the concept of unbiasness, consistency, and efficiency.

There is a huge literature on stochastic processes and the particular class we will consider. The literature is both in the statistical literature, see for instance [20, 17], and in the signal processing literature, [100, 90], and in systems science, [34, 98]. The book by Hannan and Deistler [60] provides a bridge between the statistical approach and the systems theoretical approach.

A stochastic process is a mathematical model of some uncertain evolution with time. We will consider regular stochastic processes, that is processes which are exclusively indeterministic. That means that the deterministic component in a Wold decomposition is zero. Usually it is defined as a family of stochastic variables indexed with some time $\{x_t, t \in T\}$ defined on a probability space. This is a far too general framework for us so we will introduce some restrictions.

We will restrict ourselves to *discrete time processes*, that is $T = \mathbb{Z}$. Also, we shall only consider *zero mean processes*, that is $\mathrm{E}x_t = 0$, where E denotes the expected value on the probability space. Mainly we will focus on *scalar real processes*, that is when the realizations of random variables are scalar and real. A realization of a stochastic process will be denoted a *time series*. A stationary process is *Gaussian* if and only if the distribution functions are normal.

Next, assume that the random variables x_t have finite variances for all t. Define the autocorrelation function as

$$\gamma_x(r, s) := \mathrm{Cov}(x_r, x_s) := \mathrm{E}(x_r x_s), \quad r, s \in \mathbb{Z}.$$

We will only consider (wide-sense) *stationary processes*. Such processes fulfill $\mathrm{E}|x_t|^2 < \infty$ for all $t \in \mathbb{Z}$ and $\gamma_x(r, s) = \gamma_x(r + t, s + t)$ for all $r, s, t \in \mathbb{Z}$. For stationary processes we define the covariances as $r_k = \gamma_x(0, k)$. It is well-known that a stationary stochastic processes defines a Hilbert space. This will be the link to the spectral analysis discussed previously in this chapter. Then the covariances just defined will agree with those defined in (2.22) without prefilter and with the standard basis.

A particular class of these processes is *white noise*. We will say that a process is white noise if x_t are independent identically distributed. If in addition the distribution is normal we shall call the process Gaussian white noise.

Next we shall introduce the autoregressive moving average (ARMA) processes. The class of ARMA processes is easy to analyze but also versatile enough to be useful in many applications. In particular the autoregressive (AR) processes and the moving average (MA) processes are subclasses of the class of ARMA processes. Also, the ARMA model class is the predominant class in parametric spectral estimation.

Consider the following difference equation

$$x_t + a_1 x_{t-1} + \cdots + a_n x_{t-n} = e_t + \sigma_1 e_{t-1} + \cdots + \sigma_m e_{t-m}, \tag{2.39}$$

where e_t is white noise with variance λ^2. Then x_t is an ARMA(n,m) process. We denote the ARMA parameters $\Theta := \begin{bmatrix} a_1 & a_2 & \dots & a_n & b_1 & b_2 & \dots & b_m \end{bmatrix}^T$. Letting q^{-1} denote the backward shift operator: $q^{-1}e_t = e_{t-1}$ the difference equation (2.39) can be written $a(q^{-1})x_t = \sigma(q^{-1})e_t$, where a and σ are polynomials of degrees n and m, respectively. We can now interpret the ARMA process as a linear *shaping filter* or *noise filter* as in Figure 1.3. The transfer function is then given by

$$w(q^{-1}) = \lambda \frac{\sigma(q^{-1})}{a(q^{-1})} = \lambda \frac{1 + \sigma_1 q^{-1} + \cdots + \sigma_m q^{-m}}{1 + a_1 q^{-1} + \cdots + a_n q^{-n}}.$$

Letting z denote the forward shift we can write the transfer function as

$$w(z) = \lambda \frac{(z^m + \sigma_1 z^{m-1} + \cdots + \sigma_m) z^{n-m}}{z^n + a_1 z^{n-1} + \cdots + a_n}.$$

The corresponding spectral density is given by $\lambda^2 |w(z)|^2$. In the spectral domain the factor z^{n-m} will cancel and we shall disregard it.

Of particular interest, and what will be studied in this thesis, are causal and invertible ARMA processes. Vaguely stated, an ARMA process is causal if the current value only depends on the past and invertible if it is stable as time evolves. More precisely, an ARMA process is causal if it can be expressed as

$$x_t = \sum_{j=0}^{\infty} \psi_j w_{t-j},$$

where ψ_j are some constants and invertible if, with π_j constant,

$$w_t = \sum_{j=0}^{\infty} \pi_j w_{t-j}.$$

Note that these are not properties of one stochastic process but rather as a relationship between two stochastic processes. Causality and invertibility correspond exactly to the shaping filter being miniphase and stable, as define earlier. Also note that an ARMA(n,n) model has a noise filter transfer function which is real rational of degree n.

A finite set of measurements of a stochastic process $\{x_t\}_{t=1}^N$ is a *time series*. The key problem of time series analysis and system identification is to determine a "good" model of the stochastic process that generated the times series. Here we will study how to estimate a model when the set of feasible models is restricted to ARMA(n,m) for some n and m. How to choose n and m, so called model order selection, is a topic in itself and will not be treated in this thesis.

The meaning of "a good model" needs to be clarified. It depends on the application. One common criterion is how well the model predicts future measurements. Another is how well the it describes the frequency characteristics of the stochastic process – for instance, how well a peak in the spectrum can be identified.

Even given a criterion there are several choices of how to formulate a mathematical problem to determine the estimate. Since we only have finite data we can only evaluate, for instance, the prediction criteria within some confidence. Also, depending on application, computation time to determine the estimate and reliability are important issues. As a result there is a vast amount of procedures, *estimators* for determining a model. In statistical literature the Maximum Likelihood (ML) estimator is very common. It maximizes the likelihood that the model explains the data. Under Gaussian assumption it coincides with the Prediction Error Method (PEM), see [73]. A major disadvantage with the ML estimator is that it relies on nonconvex optimization and may therefore not always be determined accurately. Therefore the estimate might not be unique or depend smoothly on data – see for instance [99]. Another class of estimators that is widely used is subspace methods, which are based on singular value decomposition of the Hankel matrix of Markov coefficients. However, as shown in [36] the subspace algorithms may fail. Yet, in practice they have proven very useful, not the least in the vector-valued case. When computation time is critical there are a number of estimators available – see, for instance, [90, 21].

In comparing the quality of the estimates from different estimators one can use a statistical approach. In fact, due to the probabilistic setting, the model parameters, Θ, will themselves be random variables. Now, make the quite restrictive assumption that the data is really generated by an ARMA(n,m) model with model parameters Θ_0. An estimator is said to be *unbiased* if

$$\mathrm{E}\Theta = \Theta_0,$$

where E denotes the expected value. Being unbiased is of course desirable, but clearly not enough. Depending on the distribution, an unbiased estimator might be good or not whereas a biased estimator might still be good. An estimator for which

$$\Theta \to \Theta_0, N \to \infty,$$

is said to be *consistent*. There are several notions of consistency – see, for instance, [90, p. 71]. Also the variance of the estimates is clearly of great interest. Due to the probabilistic setting there is a lower limit of the covariance matrix of Θ for an unbiased estimator, namely the *Cramér-Rao* lower bound; see for instance [98] and references therein. An unbiased estimator meeting the Cramér-Rao lower bound is said to be *efficient*. Also, the distribution can be of interest. If $\Theta \in N(\Theta_0, \Sigma)$ for some covariance matrix Σ the estimator is said to be normally distributed. An estimator being efficient as the sample length tends to infinity is asymptotically efficient. Likewise if the asymptotical distribution is normal the estimator is called asymptotically normally distributed and denoted by $\Theta \in AN(\Theta_0, \Sigma)$.

In the paradigm for ARMA estimation discussed in this thesis the statistical properties of the estimates of the generalized filtered covariances and cepstral coefficients will be as important as those of the ARMA parameters themselves. In fact, since we will define a smooth map between them and the ARMA parameters they will share statistical properties.

Covariances have been studied for a long time in engineering and statistics. The biased estimates are given by

$$\hat{r}_k := \frac{1}{N} \sum_{j=1}^{N-k} y_j y_{t+k}. \tag{2.40}$$

Dividing by $N - k$ yields an unbiased estimate. However, often the biased estimates are preferable since they guarantee the corresponding Pick matrix $\Sigma(r)$ to be positive definite. To estimate the generalized covariances, we use the previously discussed filter-bank approach to estimate the state covariance and then the least-squares formulation of [53, p. 34] to estimate a feasible r. To compute the prefiltered covariances, we simply feed the time series through a prefilter with density $\Psi(z)$. The statistical properties of standard covariances are well-known. The asymptotical variance is given by Bartlett's formula, see for instance [20].

A classical approach to ARMA modelling is to first estimate a finite set of covariances, $\{\hat{r}_k\}_{k=0}^n$ and then estimate a model. This is called rational covariance extension since it is the problem of parameterizing all ARMA models which are consistent with the given covariances and therefore determines an extension of the covariances sequence. The most celebrated solution is the linear predictive coding solution (LPC), which produces an AR model of order n. It is obtained by solving the Yule-Walker equations. The Levinson algorithm does this in a very fast fashion. In [22] it was shown that the LPC solution is equivalent to the solution maximizing the spectral entropy. The solutions in [27, 25] and in this thesis can be seen as extensions of the entropy-based method to ARMA models. For the case of MA modelling, a finite set of covariances does not provide a sufficient statistics, see for instance [90, p. 187]. However, in [99] a couple of convex schemes based upon using a window of sample covariances whose length increases with the times series length is proven to be asymptotically efficient. That approach can be interpreted as covariance fitting rather than extension, since the first covariances are not necessarily matched.

Cepstral coefficients are less studied than covariances. They are used in signal processing and speech processing in particular, see for instance [86]. We will mention three ways to estimate the cepstrum. In the first approach the discrete Fourier transform, DFT, of the times series is computed, then taken the logarithm of and finally the inverse discrete Fourier transform is applied. This is implemented in the MATLAB command `rceps`. In the second approach a windowed estimate of the spectrum is computed as

$$\hat{\Phi}(z) = \sum_{k=-L}^{L} r_k z^k,$$

The cepstrum is given by the inverse DFT of $\log \hat{\Phi}(z)$. The last approach is to estimate the spectrum by computing a high-order AR model. To compute the generalized filtered coefficients we again filter the data. The statistical properties of the cepstral coefficient estimates are less studied than for covariances. In [46] the DFT estimates are analyzed and in [75] the second and third approach are studied under quite restrictive assumptions. In [69] it is shown that the high-order AR-based estimates are consistent. These are also the estimates that seem preferable in applications.

For us the joint asymptotical distribution of the covariances and cepstral coefficients are of great interest. However, we are not aware of any results in the literature. The working paper [4] contains some results for windowed estimates.

In general the likelihood functional is nonconvex. However, in the classical paper [3] a very interesting theoretical result is proven: namely that the asymptotical likelihood functional is convex if the data is generated by a model in the class, and that the asymptotical ML estimate then is unique. In fact, as discussed in [23, p.126f] this has been an inspiration to the approach taken in the present research program and hence in this thesis as well.

Chapter 3

Complexity Constrained Analytic Interpolation Theory

In this chapter we will formulate a spectral estimation problem in terms of the spectral Kullback-Leibler discrepancy with respect to a given spectral density and with cepstral- and covariance-type interpolation conditions. As described in the previous chapter this has a direct counterpart in the analytic interpolation theory. We will prove that the solution fulfills a complexity constraint and that it is essentially unique. For the real case, we will show that the interpolation data thereby constitutes a global coordinatization of stable miniphase rational functions of degree n for each choice of the given density. Correspondingly, this yields a parameterization of the positive real functions. Finally, we provide a generalization to the matrix-valued interpolation problem with exclusively covariance-type conditions.

3.1 Spectral Kullback-Leibler Approximation with Cepstral- and Covariance-Type Constraints

Here we are interested in the problem of finding spectral densities that fulfill conditions on their cepstra and second order moments. By assumptions on the interpolation data, we will ensure that there exists infinitely many solutions. Out of those, we will be interested in the particular solution that has the smallest spectral Kullback-Leibler discrepancy, with respect to a given density. Moreover, we will show that this spectral density is essentially unique.

Consider the following infinite dimensional approximation problem:

Problem 3.1.1 (Kullback-Leibler Approximation). *Let $\Psi \in \mathcal{C}_+$ and $G, H \in \mathcal{G}$ be given. Assume that $(r, c) \in \mathcal{X}_{nm}$. Find any spectral density $\Phi \in \mathcal{C}_+$ that minimizes the spectral Kullback-Leibler discrepancy $\mathbb{S}(\Psi, \Phi)$ subject to the interpolation*

conditions

$$r_k = \langle H_k, \Phi \rangle \qquad k = 0, \ldots, n, \tag{3.1}$$

$$c_l = \langle G_l, \Psi \log \Phi \rangle \qquad l = 1, \ldots, m. \tag{3.2}$$

This Kullback-Leibler approximation problem is a generalization of the primal problem studied in [25, 45] in the style of [56]. In fact, in proving the theorem we shall follow these key references closely. Note that we let Ψ act as a frequency weighting of the log-spectrum of Φ.

The following theorem give the solution to the problem, its functional form, and conditions for a unique solution.

Theorem 3.1.2. *The solution to Problem 3.1.1 is of the form $\Phi = \Psi \hat{P}\hat{Q}^{-1}$ where $\hat{P} \in \mathcal{Q}_+^0$ and $\hat{Q} \in \mathcal{Q}_+$. Moreover, if $(\hat{P}, \hat{Q})$ are coprime they are unique.*

A key feature of the theorem is the functional form of the solution, which can be interpreted as a complexity constraint. For instance, taking $m = 0$, G is in (2.13), and $\Psi = \sigma\sigma^*/(\tau\tau^*)$ where $\sigma \in \mathcal{S}_n$ yields a complete parameterization of the Nevanlinna-Pick interpolation problem with degree constraint.

We will prove Theorem 3.1.2 using Lagrangian techniques. In fact, we will show that the dual, in mathematical programming sense, is

$$(\mathcal{D}) \quad \min_{(P,Q) \in \overline{\mathcal{Q}_+^0} \times \overline{\mathcal{Q}_+}} \underbrace{\langle Q, R \rangle - \langle P, \log R \rangle - \langle 1, P\Psi \rangle + \left\langle P\Psi, \log \frac{P\Psi}{Q} \right\rangle}_{=: \mathbb{J}(P,Q)}, \tag{3.3}$$

where $R(z) \in \mathcal{C}$ is any continuous function defined on $\mathbb{T}$, not necessarily positive, which fulfills the interpolation conditions (3.1) and (3.2).

As for the dual we will show the following also very central theorem, which will be the key in proving Theorem 3.1.2.

Theorem 3.1.3. *The dual problem $(\mathcal{D})$ is a convex optimization problem and has a solution $(\hat{P}, \hat{Q})$ where $\hat{Q}$ is an interior point, that is $\hat{Q} \in \mathcal{Q}_+$. Any corresponding spectral density of the form $\Psi\hat{P}\hat{Q}^{-1}$ fulfills the interpolation conditions* (3.1)*. If in addition $\hat{P} \in \mathcal{Q}_+^0$ also the interpolation conditions (3.2) are satisfied. Moreover, if $(\hat{P}, \hat{Q})$ are coprime, they are unique.*

Next we shall prove the main theorems.

Proof of Theorem 3.1.2. First we form the Lagrangian

$$\begin{aligned}
&L(P, Q, \Phi) \\
&\quad := \langle \Psi, \log \Psi - \log \Phi \rangle - \sum_{k=0}^{n} q_k (r_k - \langle H_k, \Phi \rangle) + \sum_{l=1}^{m} p_l (c_l - \langle G_l \Psi, \log \Phi \rangle), \\
&\quad = -q^T r + p^T c + \left\langle \sum_{k=0}^{n} q_k H_k, \Phi \right\rangle - \left\langle \sum_{l=1}^{m} p_l G_l + 1, \Psi \log \Phi \right\rangle, \\
&\quad = -\langle Q, R \rangle + \langle P, \log R \rangle + \langle Q, \Phi \rangle - \langle P\Psi, \log \Phi \rangle,
\end{aligned}$$

where we have defined

$$\begin{aligned} P &:= 1 + \frac{p_1}{2}(G_1 + G_1^*) + \cdots + \frac{p_m}{2}(G_m + G_m^*), \\ Q &:= q_0 + \frac{q_1}{2}(H_1 + H_1^*) + \cdots + \frac{q_n}{2}(H_n + H_n^*), \end{aligned}$$

and, R as *any* function, not necessarily positive, on the circle, which fulfills the interpolation conditions (3.1) and (3.2). Here p_k and q_k are complex numbers except q_0 which is real.

The dual optimization problem then is

$$(\mathcal{D}) \quad \min_{(P,Q)\in\overline{\mathcal{Q}_+^0}\times\overline{\mathcal{Q}_+}} \; - \inf_{\Phi\in\mathcal{C}_+} L(P,Q,\Phi). \tag{3.4}$$

We get additional conditions on P and Q, by noting where the dual functional attains an infinite value. Firstly, $Q(z) \geq 0$ for all $z \in \mathbb{T}$ since otherwise the term $\langle Q, \Phi\rangle$ can be arbitrary large. Secondly, also $P(z) \geq 0$ for $z \in \mathbb{T}$ since otherwise $-\langle P\Psi, \log\Phi\rangle$ can be made arbitrarily large. These are all the requirements[1].

Next we will show that any stationary point of the map $\Phi \mapsto L(P,Q,\Phi)$ fulfills the complexity constraint $\Phi = \Psi\hat{P}\hat{Q}^{-1}$. Consider any feasible change of Φ:

$$\begin{aligned} \delta L(P,Q,\Phi;\delta\Phi) &:= \lim_{\varepsilon\to 0}\frac{1}{\varepsilon}\Big(L(P,Q,\Phi+\varepsilon\delta\Phi) - L(P,Q,\Phi)\Big), \\ &= \langle Q, \delta\Phi\rangle - \lim_{\varepsilon\to 0}\frac{1}{\varepsilon}\left\langle P\Psi, \underbrace{\log\frac{\Phi+\varepsilon\delta\Phi}{\Phi}}_{=\varepsilon\frac{\delta\Phi}{\Phi}+\text{h.o.t.}}\right\rangle, \\ &= \left\langle \delta\Phi, Q - \frac{P\Psi}{\Phi}\right\rangle. \end{aligned}$$

Since we allow for all possible changes, any stationary point must satisfy the complexity constrain $\Phi = \Psi\hat{P}\hat{Q}^{-1}$. Evaluating the Lagrangian in the stationary point we have

$$\begin{aligned} L(P,Q,\frac{P\Psi}{Q}) &= -\langle Q, R\rangle + \langle P, \log R\rangle + \left\langle Q, \frac{P\Psi}{Q}\right\rangle - \left\langle P\Psi, \log\frac{P\Phi}{Q}\right\rangle, \\ &= -\mathbb{J}(P,Q), \end{aligned}$$

meaning that the dual problem in the Lagrangian relaxation is $(\mathcal{D})$ in (3.4).

Now, due to the definition of $\mathcal{X}_{nm}$ in (2.25), there exists at least one solution of the form $\Psi\hat{P}\hat{Q}^{-1}$ with $\hat{P} \in \mathcal{Q}_+^0$ and $\hat{Q} \in \mathcal{Q}_+$. Since the spectral Kullback-Leibler discrepancy is jointly convex we have that

$$L(\hat{P},\hat{Q},\hat{\Phi}) \leq L(\hat{P},\hat{Q},\Phi), \quad \forall\ \Phi \in \mathcal{C}_+. \tag{3.5}$$

[1] One might believe that as for Q we also need to require that $P \leq 0$ since $\log\Phi$ can be made arbitrarily large. However, since for any fix $P > 0$ and $Q > 0$ the linear term $\langle Q, \Phi\rangle$ will dominate the logarithmic term $-\log\Phi$ for large $|\Phi|$, this is in fact not the case.

However, for all Φ that fulfill the interpolation conditions (3.1) and (3.2), we have that

$$L(\cdot,\cdot,\Phi) = \mathbb{S}(\Psi,\Phi). \tag{3.6}$$

In particular $\hat{\Phi}$, again due to Theorem 3.1.3, fulfills the interpolation conditions. Therefore, combining (3.5) and (3.6) we have that

$$\mathbb{S}(\Psi,\hat{\Phi}) \leq \mathbb{S}(\Psi,\Phi), \quad \forall\ \Phi \in \mathcal{C}_+ \text{ satisfying (3.1) and (3.2)},$$

verifying the optimality of $\hat{\Phi}$. Appealing to Theorem 3.1.3 the solution is unique whenever $\hat{P}$ and $\hat{Q}$ are coprime. This concludes the proof of Theorem 3.1.2. □

Plugging that solution into the dual problem we get the dual $\mathcal{D}$ in (3.4). Next we will turn to the quite involved proof of Theorem 3.1.3. The proof is a fairly straightforward generalization of the corresponding proofs in [24, 29, 45].

Proof of Theorem 3.1.3. First we prove that the functional $\mathbb{J}(P,Q)$ is proper and bounded from below, that is, that inverse images of compact sets are compact in $\overline{\mathcal{Q}}_+^0 \times \overline{\mathcal{Q}}_+$. To this end, suppose that $(p^{(k)}, q^{(k)})$ is a sequence in $\mathbb{J}^{-1}((-\infty,\mu])$. To show that $\mathbb{J}^{-1}((-\infty,\mu])$ is compact it suffices to show that $(p^{(k)}, q^{(k)})$ has a subsequence that converges to a point in $\mathbb{J}^{-1}((-\infty,\mu])$.

First we show that $\overline{\mathcal{Q}}_+^0$ is compact. Clearly it is a closed subset of $\mathbb{R}^N$. We can factorize $P(z) = \lambda\sigma(z)\sigma^*(z)$ where $\sigma(z) \in \mathcal{S}_n$ and $p_0 = \lambda(1+\sigma_1^2+\cdots+\sigma_n^2)$. Clearly, the coefficients of $\sigma(z)$ are bounded and since $p_0 = 1$ also λ is bounded. Thus, also p_k for $k = 1,2,\ldots,n$ are bounded which implies that $\overline{\mathcal{Q}}_+^0$ is bounded and hence compact. The compactness of $\overline{\mathcal{Q}}_+^0$ implies that $p^{(k)}$ has a convergent subsequence.

As for $q^{(k)}$ we can factor out the constant, $Q^{(k)}(z) = q_0^{(k)}\tilde{Q}^{(k)}(z)$ where $\tilde{q}^{(k)} \in \overline{\mathcal{Q}}_+^0$. Since $\overline{\mathcal{Q}}_+^0$ is compact it suffices to show that $q_0^{(k)}$ has a convergent subsequence. Now we can write the dual functional in (3.3) as

$$\mathbb{J}(P^{(k)},Q^{(k)}) =: c_1^{(k)}q_0^{(k)} - c_2^{(k)}\log q_0^{(k)} - c_3^{(k)}, \tag{3.7}$$

where

$$\begin{aligned}
c_1^{(k)} &= \left\langle \tilde{Q}^{(k)}, R \right\rangle, \\
c_2^{(k)} &= \left\langle P^{(k)}\Psi, 1 \right\rangle, \\
c_3^{(k)} &= \left\langle P^{(k)}, \log R \right\rangle + \left\langle 1, P^{(k)}\Psi \right\rangle - \left\langle P^{(k)}\Psi, \log \frac{P^{(k)}\Psi}{\tilde{Q}^{(k)}} \right\rangle.
\end{aligned}$$

Clearly, since $P^{(k)}$ and $\tilde{Q}^{(k)}$ belong to $\overline{\mathcal{Q}}_+^0$ which is compact, $c_1^{(k)}$ and $c_2^{(k)}$ are positive and bounded away from positive infinity. Moreover, $c_3^{(k)}$ is bounded away from plus and minus infinity. Now, if $q_0^{(k)}$ would tend to 0 that second term in (3.7) would tend to infinity and not stay inside $\mathbb{J}^{-1}((-\infty,\mu])$. Likewise, if $q_0^{(k)}$ would tend

to positive infinity, the first term of (3.7) would tend to infinity. Thus we conclude that $q_0^{(k)}$ has a convergent subsequence and that $\mathbb{J}^{-1}((-\infty, \mu])$ is compact.

Since $\mathbb{J}$ is proper and defined on a closed, convex domain it attains a minimal point $(\hat{P}, \hat{Q})$ there. Next we will show that $\hat{Q}$ is an interior point. We shall proceed as in [29]. First consider the directional derivative of $\mathbb{J}(P, Q)$ in any feasible direction $\left\{\delta P : P + \delta P \in \overline{\mathcal{Q}}_+^0\right\}$ and $\left\{\delta Q : Q + \delta Q \in \overline{\mathcal{Q}}_+\right\}$:

$$\begin{aligned}
&\delta\mathbb{J}(P, Q; \delta P, \delta Q) \\
&\quad := \lim_{\varepsilon \to 0} \frac{1}{\varepsilon}\Big(\mathbb{J}(P + \varepsilon\delta P, Q + \varepsilon\delta Q) - \mathbb{J}(P, Q)\Big), \\
&\quad = \lim_{\varepsilon \to 0} \frac{1}{\varepsilon}\Big(\langle \varepsilon\delta Q, R\rangle - \langle \varepsilon\delta P, \log R\rangle - \langle 1, \varepsilon\delta P\Psi\rangle \\
&\qquad + \left\langle \varepsilon\delta P\Psi, \log \frac{(P + \varepsilon\delta P)\Psi}{Q + \varepsilon\delta Q} \right\rangle + \left\langle P\Psi, \log \frac{P + \varepsilon\delta P}{P} \right\rangle \\
&\qquad - \left\langle P\Psi, \log \frac{Q + \varepsilon\delta Q}{Q} \right\rangle \Big), \\
&\quad = \langle \delta Q, R\rangle - \langle \delta P, \log R\rangle - \langle \delta P, \Psi\rangle + \left\langle \delta P, \Psi \log \frac{P\Psi}{Q} \right\rangle \\
&\qquad + \left\langle P\Psi, \frac{\delta P}{P} \right\rangle - \left\langle P\Psi, \frac{\delta Q}{Q} \right\rangle, \\
&\quad = \left\langle \delta Q, R - \frac{P\Psi}{Q} \right\rangle - \left\langle \delta P, \log R - \Psi \log \frac{P\Psi}{Q} \right\rangle. \qquad (3.8)
\end{aligned}$$

For the moment, we will only study variations in $Q(z)$. Let $q \in \mathcal{Q}_+$ and $\overline{q} \in \partial\mathcal{Q}_+$ be arbitrary. Then $Q(z)$ is positive on the unit circle while $\overline{Q}(z)$ is nonnegative and equal to 0 for at least one $\theta_0 \in [-\pi, \pi]$. Define $q_\lambda := \overline{q} + \lambda(q - \overline{q})$ for $\lambda \in (0, 1]$ where $\overline{q}$ corresponds to $\overline{Q}(z)$. Then also $Q_\lambda(z)$ is positive on the unit circle. Consider the directional derivative in (P, Q_λ) in the direction $\delta Q = \overline{Q} - Q$ and keeping P constant:

$$\delta\mathbb{J}(P, Q_\lambda; 0, \overline{Q} - Q) = \left\langle \overline{Q} - Q, R - \frac{P\Psi}{Q_\lambda} \right\rangle = w^T(\overline{q} - q) - \left\langle P\Psi, \frac{\overline{Q} - Q}{Q_\lambda} \right\rangle. \quad (3.9)$$

Now, note that

$$\frac{d}{d\lambda}\frac{\overline{Q} - Q}{Q_\lambda} = -\frac{\overline{Q} - Q}{Q_\lambda^2}\frac{dQ_\lambda}{d\lambda} = \left(\frac{\overline{Q} - Q}{Q_\lambda}\right)^2 \geq 0,$$

and hence the integrand of the second term of (3.9) is a monotonically nondecreasing function of λ for all $z \in \mathbb{T}$. Thus the integrand tends pointwise on the unit circle to $(\overline{Q} - Q)/Q$ as $\lambda \to 0$. Since the $\{(\overline{Q} - Q)/Q_\lambda\}_\lambda$ is a Cauchy sequence in $\mathcal{L}_1(\mathbb{T})$ it converges almost everywhere to $(\overline{Q} - Q)/\overline{Q}$. However, since $(\overline{Q} - Q)/\overline{Q}$ has at

least one pole on the unit circle it is not summable and

$$-\left\langle P\Psi, \frac{\overline{Q}-Q}{Q_\lambda}\right\rangle \to \infty, \quad \lambda \to 0.$$

Consequently, $\delta\mathbb{J}(P, Q_\lambda; 0, \overline{Q}-Q) \to \infty$ as $\lambda \to 0$ for all $q \in \mathcal{Q}_+$ and $\overline{q} \in \partial\mathcal{Q}_+$. Hence, by [91, Lemma 26.2] $\mathbb{J}$ is an essentially smooth functional of Q and by [91, Theorem 26.3] it is essentially strictly convex with respect to Q. Thus we have proven that there exists a minimizer $(\hat{P}, \hat{Q}) \in \overline{\mathcal{Q}}_+^0 \times \mathcal{Q}_+$.

Since $\hat{Q}$ is an interior point the stationarity condition must be satisfied there. Taking $\delta Q = H_k + H_k^*$ and $\delta P = 0$ in (3.8) yields the stationarity condition

$$r_k = \langle H_k, R\rangle = \langle H_k, \Phi\rangle \quad k = 0, \ldots, n.$$

If in addition $\hat{P}$ is an interior point, and thus a stationary point, (3.8) also yields

$$c_l = \langle G_l, \log R\rangle = \langle G_l, \Psi \log \Phi\rangle \quad l = 1, \ldots, m.$$

We need to show that the optimal point is unique whenever $\hat{P}$ and $\hat{Q}$ are coprime. Consider the second variation:

$$\begin{aligned}
&\delta^2\mathbb{J}(P, Q; \delta P, \delta Q)\\
&\quad := \lim_{\varepsilon\to 0} \frac{1}{\varepsilon}(\delta\mathbb{J}(P+\varepsilon\delta P, Q+\varepsilon\delta Q; \delta P, \delta Q) - \mathbb{J}(P, Q; \delta P, \delta Q)),\\
&\quad = \lim_{\varepsilon\to 0} \frac{1}{\varepsilon}\left(\left\langle \delta Q, R - \frac{(P+\varepsilon\delta P)\Psi}{Q+\varepsilon\delta Q}\right\rangle - \left\langle \delta P, \log R - \Psi\log\frac{(P+\varepsilon\delta P)\Psi}{Q+\varepsilon\delta Q}\right\rangle\right.\\
&\qquad \left. - \left\langle \delta Q, R - \frac{P\Psi}{Q}\right\rangle + \left\langle \delta P, \log R - \Psi\log\frac{P\Psi}{Q}\right\rangle\right),\\
&\quad = \lim_{\varepsilon\to 0} \frac{1}{\varepsilon}\left(\left\langle \delta Q, \frac{Q(P+\varepsilon\delta P)\Psi - (Q+\varepsilon\delta Q)P\Psi}{(Q+\varepsilon\delta Q)Q}\right\rangle\right.\\
&\qquad \left. + \left\langle \delta P, \Psi(\log\frac{P+\varepsilon\delta P}{P} - \log\frac{Q+\varepsilon\delta Q}{Q})\right\rangle\right),\\
&\quad = \left\langle \delta Q, \frac{P\delta Q\Psi - \delta P Q\Psi}{Q^2}\right\rangle + \left\langle \delta P, \Psi\left(\frac{\delta P}{P} - \frac{\delta Q}{Q}\right)\right\rangle,\\
&\quad = \left\langle \frac{\Psi}{PQ^2}, (\delta P Q - P\delta Q)^2\right\rangle \geq 0.
\end{aligned}$$

Therefore, the dual functional $\mathbb{J}$ is convex. The second variation is zero only when $P\delta Q - \delta P Q$, that is,

$$\frac{P}{Q} = \frac{\delta P}{\delta Q}.$$

However, this is impossible if $\hat{P}$ and $\hat{Q}$ are coprime, since $p_0 \equiv 1$ implies that $\delta p_0 = 0$. Thus, $\mathbb{J}$ is strictly convex at $(\hat{P}, \hat{Q})$ if they are coprime, and the optimal point is indeed unique. That concludes the proof of Theorem 3.1.3. □

The statement of Problem 3.1.1 might appear intractable since, as stated in Remark 2.3.7, there is no available test for checking whether a point (r, c) belongs to $\mathcal{X}_{nm}$. The benefit of this formulation is the direct parameterization of ARMA models, on which we will elaborate more in the next section. Also, seen as an ARMA estimator and considering the asymptotical statistical properties of the parameters, the case when $(\hat{r}, \hat{c}) \notin \mathcal{X}_{nm}$ can easily be taken care of. In fact, we can take $(\hat{r}^N, \hat{c}^N) = (\hat{r}^{N-1}, \hat{c}^{N-1})$ whenever the N^{th} estimate falls outside $\mathcal{X}_{nm}$ and the initial estimate $(\hat{r}^0, \hat{c}^0)$ arbitrary in $\mathcal{X}_{nm}$. This will not affect the asymptotic behavior.

Yet, as a practical procedure in ARMA estimation and robust control, this is indeed an issue. Following [44, 45] we will study a *regularized* dual optimization problem, where we introduce a barrier-like term which will force the optimal point into the interior of the feasible region, that is, also with respect to the numerator pseudo-polynomial. More precisely, consider the problem

$$(\mathcal{D}_\lambda) \qquad \min_{(P,Q)\in\overline{\mathcal{Q}_+^0}\times\overline{\mathcal{Q}_+}} \mathbb{J}(P, Q) - \lambda \langle 1, \log P \rangle , \tag{3.10}$$

where $\lambda > 0$. Repeating the arguments in the proof of Theorem 3.1.3 one can readily show that the additional term, $-\lambda \langle 1, \log P \rangle$ is functional, which is proper and bounded from above, and whose derivative tends to negative infinity when the P tend to the boundary of $\mathcal{Q}_+^0$. Therefore the functional will still be proper and bounded from above so there exist a solution. Also, a parallel discussion with respect to P rules out the possibility to have a boundary solution. The first order variation (3.8) now becomes

$$\delta\mathbb{J}(P, Q; \delta P, \delta Q) = \left\langle \delta Q, R - \frac{P\Psi}{Q} \right\rangle + \left\langle \delta P, \log R - \Psi \log \frac{P\Psi}{Q} - \frac{\lambda}{P} \right\rangle .$$

At the stationary point we will therefore not quite match the cepstral estimate, but rather the modified estimate:

$$c_l = \langle G_l, \Psi \log \Phi \rangle - \lambda \left\langle 1, \frac{1}{P} \right\rangle \quad l = 1, \ldots, m. \tag{3.11}$$

In fact, we have the following result.

Theorem 3.1.4. *Let $(r, c) \in \mathcal{R}_n \times \mathbb{C}^n$. Then the regularized dual problem $(\mathcal{D}_\lambda)$ is a convex optimization problem and has an interior point solution $(\hat{P}, \hat{Q}) \in \mathcal{Q}_+^0 \times \mathcal{Q}_+$. Any corresponding spectral density of the form $\Psi\hat{P}\hat{Q}^{-1}$ fulfills the interpolation conditions* (3.1) *and* (3.11).

As for the special case studied in [45], we recover the original problem with $\lambda = 0$. When $\lambda \to \infty$ the regularization term tend to infinity unless $P \to 1$. Therefore, as argued in [45], the maximum entropy solution is recovered when $\lambda = \infty$. These properties make λ a natural choice for deformation parameter in a numerical continuation method, see [2], and the algorithm of [45] is based on this observation.

Remark 3.1.5. *The theorems of this section are generalizations of the results in [25, 56]. In fact, taking $\Psi = 1$ and G as the standard basis in (2.12) yields Theorem 5.1 of [25] while taking $m = 0$, that is no cepstral interpolation, yields Theorem 5 of [56].*

3.2 A Family of Global Coordinatizations of $\mathcal{P}_n^*$

In this section we shall show that the normalized generalized prefiltered covariances and generalized prefiltered cepstral coefficients provide a coordinatization of stable miniphase real rational functions of fixed degree for each choice of prefilter and each choice of basis functions. Hence we get a family of coordinatizations of $\mathcal{P}_n^*$ with the standard covariances and correlation coefficients as one member. By spectral factorization it is also a coordinatization of positive real functions of bounded degree. Note that in this section we only treat the real case rather than the complex case in Section 3.1. Thereby, all functions in $\mathcal{C}_+$ are real, the matrices (A, B) are real making $G \in \mathcal{G}$ real, and all interpolation data (r, c) is real.

We shall treat the normalized case, that is when $a, b, \sigma \in \mathcal{S}_n$ and where we normalize the covariance-type interpolation conditions to $r = r/r_0$. This will reduce the dimension of the problem by one and simplify the overall analysis somewhat. It can be perceived as a counterpart of analytically reducing the innovation variance in Maximum Likelihood ARMA estimation. Also see the discussion in [25, p. 29].

Since all functions are scalar in this section we write

$$\langle G, \Phi \rangle = \begin{bmatrix} \langle G_0, \Phi \rangle \\ \vdots \\ \langle G_n, \Phi \rangle \end{bmatrix},$$

which is a slight abuse of notation. However, it simplifies the presentation considerably.

Theorem 3.2.1. *Let $\Psi \in \mathcal{C}_+$ and $G \in \mathcal{G}$ be given. The corresponding generalized prefiltered normalized covariances $r_1, r_2, \ldots, r_n$ and the generalized prefiltered cepstral coefficients $c_1, c_2, \ldots, c_n$ provide a smooth coordinatization of $\mathcal{P}_n^*$.*

The theorem states that the map

$$\begin{array}{rccc} F & : & \mathcal{P}_n^* & \to & \mathcal{X}_n, \\ & & (a, \sigma) & \mapsto & (r, c), \end{array} \tag{3.12}$$

where r and c are the generalized filtered normalized covariances and cepstral coefficients in (2.22) and (2.23), respectively, is a diffeomorphism. The normalization means that $r_0 = 1$ and we have taken $r = (r_1, \ldots, r_n)$. A direct consequence of the theorem is

Corollary 3.2.2. *The map F is a homeomorphism and $\mathcal{X}_n$ has the same topological properties as $\mathcal{P}_n^*$.*

Remark 3.2.3. *Theorem 3.2.1 is a generalization of [25, Theorem 3.1] which is recovered by taking $\Psi \equiv 1$ and G as the standard basis in (2.12). Our proof has the same structure as that of [25] but introducing Ψ render some technical difficulties.*

The rest of this section is devoted to the proof of Theorem 3.2.1. One might believe that F is a diffeomorphism as a direct consequence of some global inverse function theorem, such as Hadamard's global inverse function theorem [59]. However, the rather complicated topology of $\mathcal{P}_n^*$ and $\mathcal{X}_n$, see Chapter 2, make such global theorems not applicable. Instead, we will perform a global analysis of two foliations of the manifold $\mathcal{P}_n$ in order to prove that F is a *local* diffeomorphism at each point of $\mathcal{P}_n^*$. In fact we will prove:

Theorem 3.2.4. *The map F if a local diffeomorphism on $\mathcal{P}_n^*$.*

In order to prove Theorem 3.2.4, we will study two sets of submanifolds of $\mathcal{P}_n$. In fact, they both form n-dimensional foliations of $\mathcal{P}_n$. For $k = 0, \ldots, n$, define the maps

$$\begin{array}{rccl} \xi_k & : & \mathcal{P}_n & \to & \mathbb{R}, \\ & & (a, \sigma) & \mapsto & \left\langle G_k, \Psi \dfrac{\sigma\sigma^*}{aa^*} \right\rangle. \end{array} \tag{3.13}$$

Normalization with the zeroth generalized prefiltered covariance gives $\eta : \mathcal{P}_n \to \mathbb{R}^n$ with components $\eta_k = \xi_k/\xi_0$, $k = 1 \ldots n$. The normalization makes η a map to the *generalized prefiltered correlation coefficients*. We have that $\mathcal{R}_n = \eta(\mathcal{P}_n)$, where $\mathcal{R}_n \subset \mathbb{R}^n$ with the previously described normalization. Given $r \in \mathcal{R}_n$ define the first set of submanifolds as the subsets of $\mathcal{P}_n$ matching r, that is,

$$\mathcal{P}_n(r) := \eta^{-1}(r).$$

As for the second foliation, we define the map $\zeta : \mathcal{P}_n \to \mathbb{R}^n$ to the cepstral coefficients:

$$\begin{array}{rccl} \zeta_k & : & \mathcal{P}_n & \to & \mathbb{R}, \\ & & (a, \sigma) & \mapsto & \left\langle G_k, \Psi \log \dfrac{\sigma\sigma^*}{aa^*} \right\rangle, \end{array} \tag{3.14}$$

for $k = 1 \ldots n$. The set of feasible cepstra is given by $\mathcal{C}_n = \zeta(\mathcal{P}_n)$. Now, given $c \in \mathcal{C}_n$, define the second set of submanifolds as the subsets of $\mathcal{P}_n$ with cepstra c, that is,

$$\mathcal{P}_n(c) := \zeta^{-1}(c).$$

First we will state and prove a preliminary result is regarding a linear map. Let $\phi \in \mathcal{S}_n$. Consider the linear map from the vector space of polynomials of degree at most $n-1$:

$$\begin{array}{rccl} \vartheta_\phi & : & \mathcal{L}_{n-1} & \to & U \subset \mathbb{R}^n, \\ & & u & \mapsto & \left\langle \bar{G}, \dfrac{T(\phi)u}{\phi\phi^*} \right\rangle. \end{array}$$

The map is invertible, generalizing [25, Lemma 4.1]. However, we can not directly generalize the proof since $\langle \Psi, T(\phi)u/(\phi\phi^*)\rangle$ is, in general, nonzero for nonconstant Ψ.

Lemma 3.2.5. *The linear map ϑ_ϕ is a bijection.*

Proof. Start with injectivity by supposing that $\vartheta_\phi u = 0$. Then

$$\left\langle G_k, \Psi \frac{T(\phi)u}{\phi\phi^*} \right\rangle = 0, \quad k = 1, 2, \dots n.$$

By symmetry this also holds for $k = -1, -2, \cdots - n$. Therefore

$$\left\langle G_k \frac{\tau\tau^*}{\phi\phi^*}, \frac{\phi\phi^*}{\tau\tau^*} \Psi \frac{T(\phi)u}{\phi\phi^*} \right\rangle = 0, \quad k = \pm 1, \pm 2, \cdots \pm n.$$

Now let $\hat{G} \in \mathcal{G}$ be a set of basis functions corresponding to $(\hat{A}, \hat{B})$ such that $\phi = \det(I - \hat{A}z)$. Since $\langle 1, G_k \rangle =< 1, \hat{G}_k >= 0$ we then have that

$$\left\langle \hat{G}_k, \frac{\phi\phi^*}{\tau\tau^*} \Psi \frac{T(\phi)u}{\phi\phi^*} \right\rangle = 0, \quad k = \pm 1, \pm 2, \cdots \pm n.$$

Now, since

$$\frac{T(\phi)u}{\phi\phi^*} = \frac{u}{\phi} + \frac{u^*}{\phi^*},$$

with u/ϕ strictly proper, taking an appropriate linear combination we have

$$\left\langle \frac{T(\phi)u}{\phi\phi^*}, \frac{\phi\phi^*}{\tau\tau^*} \Psi \frac{T(\phi)u}{\phi\phi^*} \right\rangle = \left\| \frac{T(\phi)u}{\phi\tau} w \right\|^2 = 0,$$

where w is the spectral factor of Ψ. Hence $T(\phi)u = 0$ and by the invertibility of T, see for instance [30, Lemma 2.1], we also have $u = 0$. Hence ϑ_ϕ is injective. Being a linear map between vector spaces of the same real dimension, it is also surjective, and hence bijective. □

Now, we can prove the first major result regarding the submanifolds of $\mathcal{P}_n$.

Proposition 3.2.6. *The manifolds $\mathcal{P}_n(c)$ are smooth n-manifolds and their tangent space $T_{(a,\sigma)}\mathcal{P}_n(c)$ consists of those $(u, v) \in \mathcal{L}_{n-1} \times \mathcal{L}_{n-1}$ for which*

$$\left\langle G_k, \Psi \frac{T(\sigma)v}{\sigma\sigma^*} \right\rangle = \left\langle G_k, \Psi \frac{T(a)u}{aa^*} \right\rangle, \tag{3.15}$$

for $k = 1 \dots n$. Moreover the connected components of the n-manifolds $\{\mathcal{P}_n(c) : c \in \mathcal{C}_n\}$ form the leaves of a foliation of $\mathcal{P}_n$.

Proof. The tangent vector of $\mathcal{P}_n(c)$ at (a,σ), $T_{(a,\sigma)}\mathcal{P}_n(c)$ are the vectors in the kernel of the Jacobian of ζ at (a,σ). For $u,v \in \mathcal{L}_{n-1}$:

$$D_{(u,v)}\zeta(a,\sigma) = \lim_{\varepsilon\to 0}\frac{1}{\varepsilon}\left(\zeta(a+\varepsilon u,\sigma+\varepsilon v)-\zeta(a,\sigma)\right).$$

Applying the calculation

$$\lim_{\varepsilon\to 0}\frac{1}{\varepsilon}\left(\log(\sigma+\varepsilon v)-\log\sigma\right) = \lim_{\varepsilon\to 0}\frac{1}{\varepsilon}\log\left(1+\epsilon\frac{v}{\sigma}\right) = \frac{v}{\sigma},$$

we get that

$$\begin{aligned} D_{(u,v)}\zeta(a,\sigma) &= \left\langle \bar{G}, \Psi\left(\frac{v}{\sigma}+\frac{v^*}{\sigma^*}-\frac{u}{a}-\frac{u^*}{a^*}\right)\right\rangle, \\ &= \left\langle \bar{G}, \Psi\left(\frac{T(\sigma)v}{\sigma\sigma^*}-\frac{T(a)u}{aa^*}\right)\right\rangle. \end{aligned} \tag{3.16}$$

Thus we have proven that the tangent space is given by (3.15). Since both the maps ϑ_a and ϑ_σ are bijective linear maps by Lemma 3.2.5, the tangent space is of dimension n. Hence the rank of $\mathrm{Jac}(\zeta)|_{(a,\sigma)}$ is then full for all feasible (a,σ). By the implicit function theorem $\mathcal{P}_n(c)$ are then smooth n-manifolds.

Since $\mathrm{Jac}(\zeta)|_{(a,\sigma)}$ is full rank, ζ is a submersion, and hence the connected components of the n-manifolds $\{\mathcal{P}_n(c) : c \in \mathcal{C}_n\}$ form the leaves of a foliation of $\mathcal{P}_n$. □

Now we have a parallel statement for $\mathcal{P}_n(r)$, which we will prove in a similar fashion. Here, the normalization makes the analysis somewhat more involved but the generalization from [25] is more direct.

Proposition 3.2.7. *The manifolds $\mathcal{P}_n(r)$ are smooth n-manifolds and their tangent space $T_{(a,\sigma)}\mathcal{P}_n(r)$ consists of those $(u,v) \in \mathcal{L}_{n-1}\times\mathcal{L}_{n-1}$ for which*

$$\left\langle G_k, \Psi\frac{T(\sigma)v}{aa^*}\right\rangle = \left\langle G_k, \Psi\frac{\sigma\sigma^*}{aa^*}\frac{T(a)u}{aa^*}\right\rangle + \varphi(a,\sigma,u,v)\left\langle G_k, \Psi\frac{\sigma\sigma^*}{aa^*}\right\rangle, \tag{3.17}$$

for $k = 0\ldots n$ and where

$$\varphi(a,\sigma,u,v) := D_{(u,v)}\log\xi_0(a,\sigma) = \frac{D_{(u,v)}\xi_0(a,\sigma)}{\xi_0(a,\sigma)}.$$

Moreover the connected components of the n-manifolds $\{\mathcal{P}_n(r) : r \in \mathcal{R}_n\}$ form the leaves of a foliation of $\mathcal{P}_n$.

Proof. Again we compute the directional derivative of η at $(a,\sigma) \in \mathcal{P}_n$ in the direction $(u,v) \in \mathcal{L}_{n-1}\times\mathcal{L}_{n-1}$. We have

$$D_{(u,v)}\eta_k(a,\sigma) = \frac{1}{\xi_0(a,\sigma)}D_{(u,v)}\xi_k(a,\sigma) - \frac{\xi_k(a,\sigma)}{\xi_0(a,\sigma)^2}D_{(u,v)}\xi_0(a,\sigma), \tag{3.18}$$

where

$$D_{(u,v)}\xi_k(a,\sigma) = \left\langle G_k, \Psi\left(\frac{T(\sigma)v}{aa^*} - \frac{\sigma\sigma^*}{aa^*}\frac{T(a)u}{aa^*}\right)\right\rangle. \tag{3.19}$$

Multiplying (3.18) with $\xi_0(a,\sigma) = r_0 > 0$ we get the the kernel of $\mathrm{Jac}(\eta)|_{(a,\sigma)}$ to consist of all $(u,v) \in \mathcal{L}_n \times \mathcal{L}_n$ such that

$$\left\langle G_k, \Psi\frac{T(\sigma)v}{aa^*}\right\rangle = \left\langle G_k, \Psi\frac{\sigma\sigma^*}{aa^*}\frac{T(\sigma)v}{aa^*}\right\rangle + \varphi(a,\sigma,u,v)\left\langle G_k, \Psi\frac{\sigma\sigma^*}{aa^*}\right\rangle, \tag{3.20}$$

for $k = 1 \dots n$. Since $\eta_0 = \xi_0/\xi_0 = 1$ (3.20) trivially also holds for $k = 0$. This establishes (3.17). Next we will prove that the tangent space is n-dimensional for all $(a,\sigma) \in \mathcal{P}_n$. Let p be a polynomial of degree n defined by $p(z) := v(z) + \varphi a(z)$. Then the tangent equations can be written as

$$\Pi p = \Upsilon u,$$

where the linear operators $\Pi : \mathcal{L}_n \to \mathbb{R}^{n+1}$ and $\Upsilon : \mathcal{L}_{n-1} \to \mathbb{R}^{n+1}$ are given by

$$\Pi p := \left\langle G, \Psi\frac{\sigma\sigma^*}{aa^*}\frac{T(\sigma)p}{aa^*}\right\rangle \text{ and } \Upsilon u := \left\langle G, \Psi\frac{T(\sigma)u}{aa^*}\right\rangle.$$

To see this, note that $T(a)a/(aa^*) = 2$. Now, Π is in fact injective. Assume that $\Pi p = 0$. By changing basis functions from G to some $\tilde{G} \in \mathcal{G}$ associated with $(\tilde{A}, \tilde{B})$ such that $\det(I - \tilde{A}z) = a(z)$ we have that, for some nonsingular U, that

$$\begin{aligned} \Pi p = U\left\langle \tilde{G}, \frac{aa^*}{\tau\tau^*}\Psi\frac{\sigma\sigma^*}{aa^*}\frac{T(\sigma)p}{aa^*}\right\rangle = 0, \\ \Rightarrow \quad \left\langle \tilde{G}_k, \Psi\frac{\sigma\sigma^*}{\tau\tau^*}\frac{T(\sigma)p}{aa^*}\right\rangle = 0,\ k = 0, \pm 1, \cdots \pm n. \end{aligned}$$

Now, taking appropriate linear combinations we have that

$$0 = \left\langle \frac{T(\sigma)p}{aa^*}, \Psi\frac{\sigma\sigma^*}{\tau\tau^*}\frac{T(\sigma)p}{aa^*}\right\rangle = \left\| w\frac{\sigma T(\sigma)p}{a\tau}\right\|^2,$$

where w is the spectral factor of Ψ. Hence $T(\sigma)p = 0$, implying that $p = 0$. Thus Π is injective. Then we have $p = \Pi^{-1}\Upsilon v$.

Since the leading coefficient of p is $\varphi/2$, this defines an affine map $L : \mathcal{L}_{n-1} \to \mathcal{L}_{n-1}$ sending u to $v := \Pi^{-1}\Upsilon u - \varphi a/2$. Then $T_{(a,\sigma)}\mathcal{P}_n(r)$ consists of those $(u,v) \in \mathcal{L}_{n-1} \times \mathcal{L}_{n-1}$ such that $v = Lu$ which hence is n dimensional. Therefore the rank of $\mathrm{Jac}(\eta)|_{(a,\sigma)}$ is full for all $(a,\sigma) \in \mathcal{P}_n$ so that $\mathcal{P}_n(r)$ are smooth n-manifolds by the implicit function theorem.

As in the proof of Proposition 3.2.6, η is a submersion and the claim follows. □

Next we shall study the intersection of the tangent spaces $T_{(a,\sigma)}\mathcal{P}_n(c)$ and $T_{(a,\sigma)}\mathcal{P}_n(r)$. Whenever the intersection is a unique point, the submanifolds $\mathcal{P}_n(c)$ and $\mathcal{P}_n(r)$ are complementary and provide a coordinatization.

Theorem 3.2.8. *The tangent spaces $T_{(a,\sigma)}\mathcal{P}_n(r)$ and $T_{(a,\sigma)}\mathcal{P}_n(c)$ are complementary in $\mathcal{P}_n^*$. The dimension of $\Theta := T_{(a,\sigma)}\mathcal{P}_n(r) \cap T_{(a,\sigma)}\mathcal{P}_n(c)$ is the degree of the greatest common divisor.*

Proof. First consider the equations for $T_{(a,\sigma)}\mathcal{P}_n(c)$:

$$\left\langle G_k, \Psi \frac{T(\sigma)v}{\sigma\sigma^*} \right\rangle = \left\langle G_k, \Psi \frac{T(a)u}{aa^*} \right\rangle, \ k = \pm 1, \cdots \pm n, \tag{3.21}$$

which can be written

$$\left\langle G_k \frac{\tau\tau^*}{\sigma\sigma^*}, \frac{\sigma\sigma^*}{\tau\tau^*}\Psi \frac{T(\sigma)v}{\sigma\sigma^*} \right\rangle = \left\langle G_k \frac{\tau\tau^*}{\sigma\sigma^*}, \frac{\sigma\sigma^*}{\tau\tau^*}\Psi \frac{T(a)u}{aa^*} \right\rangle, \ k = \pm 1, \cdots \pm n.$$

Now let $\hat{G} \in \mathcal{G}$ be a set of basis functions corresponding to $(\hat{A}, \hat{B})$ such that $\sigma = \det(I - \hat{A}z)$. Since $\langle 1, G_k \rangle =< 1, \hat{G}_k >= 0$ we then have that

$$\left\langle \hat{G}_k, \frac{\sigma\sigma^*}{\tau\tau^*}\Psi \frac{T(\sigma)v}{\sigma\sigma^*} \right\rangle = \left\langle \hat{G}_k, \frac{\sigma\sigma^*}{\tau\tau^*}\Psi \frac{T(a)u}{aa^*} \right\rangle, \ k = \pm 1, \cdots \pm n.$$

Now, since

$$\frac{T(\sigma)v}{\sigma\sigma^*} = \frac{v}{\sigma} + \frac{v^*}{\sigma^*},$$

with v/σ strictly proper, taking an appropriate linear combination we have

$$\left\langle \frac{T(\sigma)v}{\sigma\sigma^*}, \frac{\sigma\sigma^*}{\tau\tau^*}\Psi \frac{T(\sigma)v}{\sigma\sigma^*} \right\rangle = \left\langle \frac{T(\sigma)v}{\sigma\sigma^*}, \frac{\sigma\sigma^*}{\tau\tau^*}\Psi \frac{T(a)u}{aa^*} \right\rangle,$$

that is

$$\left\langle \frac{T(\sigma)v}{\tau\tau^*}, \Psi \frac{T(\sigma)v}{\sigma\sigma^*} \right\rangle = \left\langle \frac{T(\sigma)v}{\tau\tau^*}, \Psi \frac{T(a)u}{aa^*} \right\rangle.$$

Taking linear combinations of (3.21) corresponding to $T(\sigma)v/(\tau\tau^*)$ yields

$$\left\langle \frac{T(\sigma)v}{\tau\tau^*}, \Psi \frac{T(\sigma)v}{\sigma\sigma^*} \right\rangle = \left\langle \frac{T(\sigma)v}{\tau\tau^*}, \Psi \frac{T(a)u}{aa^*} \right\rangle + \left\langle 1, \frac{T(\sigma)v}{\tau\tau^*} \right\rangle \left\langle \Psi, \frac{T(\sigma)v}{\sigma\sigma^*} - \frac{T(a)u}{aa^*} \right\rangle.$$

Since $T(\sigma)v/(\tau\tau^*)$ is a density for nonzero v, combining the expressions we have

$$\left\langle \Psi, \frac{T(\sigma)v}{\sigma\sigma^*} \right\rangle = \left\langle \Psi, \frac{T(a)u}{aa^*} \right\rangle,$$

on $T_{(a,\sigma)}\mathcal{P}_n(c)$. Hence, we have that

$$\left\langle G_k, \Psi \frac{T(\sigma)v}{\sigma\sigma^*} \right\rangle = \left\langle G_k, \Psi \frac{T(a)u}{aa^*} \right\rangle,$$

for $k = 0, \pm 1, \cdots \pm n$ on $T_{(a,\sigma)}\mathcal{P}_n(c)$. The equations describing $T_{(a,\sigma)}\mathcal{P}_n(r)$ are

$$\left\langle G_k, \Psi \frac{T(\sigma)v}{aa^*} \right\rangle = \left\langle G_k, \Psi \frac{\sigma\sigma^*}{aa^*} \frac{T(\sigma)v}{aa^*} \right\rangle + \varphi(a,\sigma,u,v) \left\langle G_k, \Psi \frac{\sigma\sigma^*}{aa^*} \right\rangle,$$

for $k = 0, \pm 1, \cdots \pm n$. Now, taking appropriate linear combinations we have that

$$\begin{aligned} \left\langle \Psi, \frac{T(\sigma)v}{\tau\tau^*} \right\rangle &= \left\langle \Psi, \frac{\sigma\sigma^*}{\tau\tau^*} \frac{T(a)u}{aa^*} \right\rangle, \\ \left\langle \Psi, \frac{T(\sigma)v}{\tau\tau^*} \right\rangle &= \left\langle \Psi, \frac{\sigma\sigma^*}{\tau\tau^*} \frac{T(a)u}{aa^*} \right\rangle + \varphi \left\langle \Psi, \frac{\sigma\sigma^*}{\tau\tau^*} \right\rangle. \end{aligned}$$

We conclude that $\varphi = 0$ on Θ.

Thus on Θ we have

$$\begin{aligned} \left\langle G_k, \Psi \frac{T(\sigma)v}{\sigma\sigma^*} \right\rangle &= \left\langle G_k, \Psi \frac{T(a)u}{aa^*} \right\rangle, \qquad k = 0, \pm 1, \cdots \pm n, \\ \left\langle G_k, \Psi \frac{T(\sigma)v}{aa^*} \right\rangle &= \left\langle G_k, \Psi \frac{\sigma\sigma^*}{aa^*} \frac{T(\sigma)v}{aa^*} \right\rangle, \ k = 0, \pm 1, \cdots \pm n. \end{aligned}$$

Again taking appropriate linear combinations we have that

$$\begin{aligned} \left\langle \frac{T(\sigma)v}{\tau\tau^*}, \Psi \frac{T(\sigma)v}{\sigma\sigma^*} \right\rangle &= \left\langle \frac{T(\sigma)v}{\tau\tau^*}, \Psi \frac{T(a)u}{aa^*} \right\rangle, \\ \left\langle \frac{T(a)u}{\tau\tau^*}, \Psi \frac{T(\sigma)v}{aa^*} \right\rangle &= \left\langle \frac{T(a)u}{\tau\tau^*}, \Psi \frac{\sigma\sigma^*}{aa^*} \frac{T(a)u}{aa^*} \right\rangle. \end{aligned}$$

Again letting w be the spectral factor of Ψ and defining

$$f_1 = \frac{T(\sigma)v}{\tau\sigma^*} w \text{ and } f_2 = \frac{T(a)u\,\sigma}{\tau aa^*} w,$$

the equations can be written $\|f_1\|^2 = \langle f_1, f_2 \rangle$ and $\langle f_1, f_2 \rangle = \|f_2\|^2$. By the parallelogram law we then have

$$\|f_1 - f_2\| = \|f_1\|^2 + \|f_2\|^2 - 2\langle f_1, f_2 \rangle = 0.$$

Hence $f_1 = f_2$ on the unit circle implying that

$$\frac{v}{\sigma} + \frac{v^*}{\sigma^*} = \frac{T(\sigma)v}{\sigma\sigma^*} = \frac{T(a)u}{aa^*} = \frac{u}{a} + \frac{u^*}{a^*},$$

on the unit circle. Being real polynomials this need to hold also for the positive real part. Moreover, it clearly holds for $u = v = 0$ so now consider the nontrivial case. We can write the equation as

$$\frac{v}{u} = \frac{\sigma}{a}.$$

This has no solution if (a, σ) are coprime, which establishes that $T_{(a,\sigma)}\mathcal{P}_n(r)$ and $T_{(a,\sigma)}\mathcal{P}_n(c)$ are complementary in $\mathcal{P}_n^*$. On the other hand, if they have a common factor of degree d, then u and v can be any polynomials of degree $n-1$ with a common factor of degree at least $d-1$, hence defining a vector space of dimension d establishing the rest of the claim. □

We summarize with the proofs of the main theorems.

Proof of Theorem 3.2.4. Since $T_{(a,\sigma)}\mathcal{P}_n(r)$ and $T_{(a,\sigma)}\mathcal{P}_n(c)$ are complementary in $\mathcal{P}_n^*$ by Theorem 3.2.8, the kernels of $\mathrm{Jac}(\eta)|_{(a,\sigma)}$ and $\mathrm{Jac}(\zeta)|_{(a,\sigma)}$ are complementary at any $(a, \sigma) \in \mathcal{P}_n^*$. Hence the Jacobian of the joint map F is full rank. By the implicit function theorem F is a local diffeomorphism on $\mathcal{P}_n^*$. □

Proof of Theorem 3.2.1. First we prove that the map F is a bijection as a consequence of Theorem 3.1.2 in the previous section. Take $H = G$. Let $Q(z) = \lambda_1 a(z)a^*(z)$ and $P(z) = \lambda_2 \sigma(z)\sigma^*(z)$ where λ_1 and λ_2 are taken so that $p_0 = 1$ and $r_0 = 1$. Since P and Q are coprime F is a bijection by Theorem 3.1.2. Together with Theorem 3.2.4 this implies that F is a diffeomorphism. □

3.3 Matrix-Valued Spectral Estimation with Complexity Constraint

This section will make a generalization of the theory to matrix-valued functions and densities, which is of interest both from a theoretical and an applicational viewpoint. There are several ways to generalize the scalar analytic interpolation theory to both matrix values and tangential interpolation. The approach taken here follows [8] and facilitates a direct generalization of the convex optimization approach, but it is somewhat restrictively in the class of interpolants. In fact, we will only provide a parameterization of $\mathcal{F}_+(n)$ defined in Section 2.4. We will only treat the case of covariance type interpolation conditions, corresponding to setting $m = 0$ in (3.2). First we formulate a matricial generalization of the Kullback-Leibler approximation problem. Then we state a theorem giving the uniqueness of the solution and finally a theorem concerning the smoothness of the parameterization.

The approximation problem that we will solve here is stated below.

Problem 3.3.1 (Matrix-Valued Kullback-Leibler Approximation). *Let $\Psi \in \mathcal{C}_+$ and $G \in \mathcal{G}$ be given. Assume that $R \in \mathcal{R}^\ell$. Find any spectral density $\Phi \in \mathcal{C}_+^\ell$ that minimizes the matricial spectral Kullback-Leibler discrepancy $\mathbb{S}(\Psi, \Phi)$ subject to the interpolation conditions*

$$R_k \quad = \quad \frac{1}{2\pi}\int_{-\pi}^{\pi} G_k(e^{i\theta})\Phi(e^{i\theta})d\theta \quad k = 0, \ldots, n. \tag{3.22}$$

We can now state a theorem giving the solution to this minimization problem, which we shall call the primal optimization problem.

Theorem 3.3.2. *The solution to Problem 3.3.1 is of the form* $\Phi = \Psi\hat{Q}^{-1}$ *where* $\hat{Q} \in \mathcal{Q}_+^\ell$. *Via the Riesz-Herglotz representation* (2.1) *this established a one-one correspondence between interpolants* $F \in \mathcal{F}_+(n)$ *and* $\Psi \in \mathcal{Q}_+$.

The primal problem is a constrained optimization problem over the infinite-dimensional space $\mathcal{C}_+^\ell$, which is hard to solve directly. We observe that the optimization problem has only finitely many constraints and thus a finite-dimensional dual, in mathematical programming terms. In fact, as for the scalar case in the previous section, we shall demonstrate that $\hat{Q}$ in $\Phi = \Psi\hat{Q}^{-1}$ can be determined by solving the dual optimization problem, namely the problem to find a $Q \in \mathcal{Q}_+^\ell$ that minimizes the functional

$$\mathbb{J}_\Psi(Q) := \langle Q, R\rangle - \frac{1}{2\pi}\int_{-\pi}^{\pi} \Psi(e^{i\theta}) \log\det Q(e^{i\theta}) d\theta. \tag{3.23}$$

This will be formalized in the next theorem.

Theorem 3.3.3. *Given* $R \in \mathcal{R}^\ell$ *and any* $\Psi \in \mathcal{C}_+$, *the minimization problem*

$$\min_{Q\in\mathcal{Q}_+^\ell} \mathbb{J}_\Psi(Q), \tag{3.24}$$

has a unique optimal solution. If $\Psi \in \mathcal{Q}_+$, *the unique optimal solution* $\hat{Q}$ *give a unique interpolant* $F \in \mathcal{F}_+(n)$ *by* (2.1). *The optimal solution* $\hat{Q}$ *depends smoothly on the interpolation data. In particular, the map* $\mathcal{I} : \mathcal{Q}_+^\ell \to \mathcal{R}^\ell$ *with components*

$$\mathcal{I}_k(Q) := \frac{1}{2\pi}\int_{-\pi}^{\pi} \alpha_k(e^{i\theta})\Psi(e^{i\theta})Q(e^{i\theta})^{-1} d\theta, \quad k = 0, 1, \dots, n, \tag{3.25}$$

is a diffeomorphism.

In order to solve the primal problem we form the Lagrangian

$$L(\Phi, Q) := \mathbb{S}(\Psi, \Phi) - \mathrm{Re}\left\{\sum_{k=0}^{n}\sum_{i=1}^{\ell}\sum_{j=1}^{\ell} q_k^{ji}\left[r_k^{ij} - \frac{1}{2\pi}\int_{-\pi}^{\pi} G_k(e^{i\theta})\Phi_{ij}(e^{i\theta})d\theta\right]\right\},$$

where r_k^{ij} and Φ_{ij} are the matrix components of R_k and Φ respectively, and then solve the dual problem to minimize

$$-\inf_{\Phi\in\mathcal{C}_+^\ell} L(\Phi, Q),$$

with respect to the Lagrange multipliers q_k^{ij}, which are complex numbers except when $k = 0$ when they are real and $q_0^{ji} = q_0^{ij}$. Here, Q is the generalized pseudo-polynomial (2.27) formed by taking Q_k to be the $\ell \times \ell$ matrix $[q_k^{ij}]_{i,j=1}^\ell$ for $k =$

$0, 1, \ldots, n$. Then the Lagrangian can be written by

$$\begin{aligned} L(\Phi, Q) &= \frac{1}{2\pi}\int_{-\pi}^{\pi} \Psi(e^{i\theta})\left(\log \Psi(e^{i\theta}) - \log\det \Phi(e^{i\theta})\right) d\theta, \\ &\quad - \langle Q, R\rangle + \frac{1}{2\pi}\int_{-\pi}^{\pi} \operatorname{trace}\{Q(e^{i\theta})\Phi(e^{i\theta})\} d\theta. \end{aligned} \tag{3.26}$$

Here $R \in \mathcal{R}^{\ell}$ is any function defined on the unit circle, which fulfill the interpolation conditions in (3.22). Clearly, the Lagrangian will be unbounded if Q is allowed to have negative eigenvalues on the unit circle. Hence, we determine the infimum for each $Q \in \mathcal{Q}_+^{\ell}$. To this end, we want to determine a Φ such that the directional derivative

$$\begin{aligned} &\delta L(\Phi, Q; \delta\Phi) \\ &\quad := \lim_{\varepsilon\to 0} \frac{L(\Phi + \varepsilon\delta\Phi, Q) - L(\Phi, Q)}{\varepsilon}, \\ &\quad = -\frac{1}{2\pi}\int_{-\pi}^{\pi} \Psi \lim_{\varepsilon\to 0}\frac{1}{\varepsilon}\log\left[\frac{\det(\Phi+\varepsilon\delta\Phi)}{\det\Phi}\right] d\theta + \frac{1}{2\pi}\int_{-\pi}^{\pi}\operatorname{trace}\{Q\delta\Phi\} d\theta, \end{aligned}$$

equals zero in all directions $\delta\Phi$ such that $\Phi + \varepsilon\delta\Phi \in \mathcal{C}_+^{\ell}$ for some $\varepsilon > 0$. However, since

$$\log\left[\frac{\det(\Phi+\varepsilon\delta\Phi)}{\det\Phi}\right] = \log\det(I + \varepsilon\Phi^{-1}\delta\Phi) = \log\prod_{j=1}^{\ell}(1+\varepsilon\lambda_j) = \sum_{j=1}^{\ell}\log(1+\varepsilon\lambda_j),$$

where $\lambda_1(e^{i\theta}), \lambda_2(e^{i\theta}), \ldots, \lambda_\ell(e^{i\theta})$ are the eigenvalues of $\Phi(e^{i\theta})^{-1}\delta\Phi(e^{i\theta})$, and $\log(1+\varepsilon\lambda_j) = \varepsilon\lambda_j + O(\varepsilon^2)$, we have

$$\lim_{\varepsilon\to 0}\frac{1}{\varepsilon}\log\left[\frac{\det(\Phi+\varepsilon\delta\Phi)}{\det\Phi}\right] = \sum_{j=1}^{\ell}\lambda_j = \operatorname{trace}(\Phi^{-1}\delta\Phi). \tag{3.27}$$

Consequently, in terms of the inner product the directional derivative can be written as

$$\delta L(\Phi, Q; \delta\Phi) = \left\langle \delta\Phi, \Psi\Phi^{-1} - Q\right\rangle, \tag{3.28}$$

which equals zero for all $\delta\Phi$ if and only if $\Phi = \Psi Q^{-1}$. Inserting this into (3.26) we obtain

$$L(\Psi Q^{-1}, Q) =: -\mathbb{J}_{\Psi}(Q) + \frac{1}{2\pi}\int_{-\pi}^{\pi}\Psi(e^{i\theta})d\theta,$$

where

$$\mathbb{J}_{\Psi}(Q) = \langle Q, R\rangle - \frac{1}{2\pi}\int_{-\pi}^{\pi}\Psi(e^{i\theta})\log\det Q(e^{i\theta})d\theta. \tag{3.29}$$

Hence, modulo an additive constant, $L(\Psi Q^{-1}, Q)$ is precisely the dual function. We want to show that this functional is strictly convex and that it has a unique minimum in $\mathcal{Q}_+^\ell$. To this end, we form the directional derivative

$$\begin{aligned}\delta\mathbb{J}_\Psi(Q;\delta Q) &:= \lim_{\varepsilon\to 0}\frac{\mathbb{J}_\Psi(Q+\varepsilon\delta Q)-\mathbb{J}_\Psi(Q)}{\varepsilon},\\ &= \langle \delta Q, R\rangle - \frac{1}{2\pi}\int_{-\pi}^{\pi}\Psi(e^{i\theta})\lim_{\varepsilon\to 0}\frac{1}{\varepsilon}\log\left[\frac{\det(Q+\varepsilon\delta Q)}{\det Q}\right]d\theta,\\ &= \left\langle \delta Q, R-\Psi Q^{-1}\right\rangle, \end{aligned} \tag{3.30}$$

where we have performed the same calculation as in (3.27). We need to determine a $Q\in\mathcal{Q}_+^\ell$ such that $\delta\mathbb{J}_\Psi(Q;\delta Q)=0$ for all δQ of the form

$$\delta Q(e^{i\theta}) = \Re\left\{\sum_{k=0}^{n}\delta Q_k G_k(e^{i\theta})\right\}, \tag{3.31}$$

where δQ_k, $k=0,1,\dots,n$, are arbitrary complex $\ell\times\ell$ matrices, except for δQ_0 that is real and symmetric. Inserting (3.31) into (3.30), we obtain

$$\begin{aligned}&\delta\mathbb{J}_\Psi(Q;\delta Q)\\ &\quad= \mathrm{Re}\left\{\sum_{k=0}^{n}\mathrm{trace}\left(\delta Q_k\frac{1}{2\pi}\int_{-\pi}^{\pi}G_k(e^{i\theta})\left[R(e^{i\theta})-\Psi(e^{i\theta})Q(e^{i\theta})^{-1}\right]d\theta\right)\right\},\\ &\quad= \mathrm{Re}\left\{\sum_{k=0}^{n}\mathrm{trace}\Big(\delta Q_k\big[R_k-\mathcal{I}_k(Q)\big]\Big)\right\},\end{aligned}$$

where $\mathcal{I}_0(Q),\mathcal{I}_1(Q),\dots,\mathcal{I}_n(Q)$ are defined as in (3.25).

Lemma 3.3.4. *The stationarity condition $\delta\mathbb{J}_\Psi(Q;\delta Q)=0$ holds for all δQ of the form* (3.31) *if and only if $\mathcal{I}_k(Q)=R_k$, $k=0,1,\dots,n$.*

Proof. For an arbitrary (k,i,j) with $k\neq 0$, take all components of $\delta Q_0,\delta Q_1,\dots,\delta Q_n$ equal to zero except δq_k^{ji}, which we take to be $\lambda+i\mu$ with λ and μ arbitrary. Then, letting u_k^{ij} be the real part and v_k^{ij} the imaginary part of $\mathcal{I}_k^{ij}(Q)-r_k^{ij}$, we obtain

$$\delta\mathbb{J}_\Psi(Q;\delta Q) = \mathrm{Re}\{(\lambda+i\mu)(u_k^{ij}+iv_k^{ij})\} = \lambda u_k^{ij}-\mu v_k^{ij},$$

and hence $r_k^{ij}=\mathcal{I}_k^{ij}(Q)$, as claimed. If $k=0$, μ and v_k^{ij} equal to zero, so the same conclusion follows. The reverse statement is trivial. □

It remains to show that there is a $Q\in\mathcal{Q}_+^\ell$ such that the stationary condition $\delta\mathbb{J}_\Psi(Q;\delta Q)=0$ holds.

Theorem 3.3.5. *Let $\Psi \in \mathcal{Q}_+$, and suppose that $R \in \mathcal{R}^\ell$. The dual functional $\mathbb{J}_\Psi : \mathcal{Q}_+^\ell \to \mathbb{R}$ is strictly convex and has a unique minimum $\hat{Q}$. Moreover,*

$$\frac{1}{2\pi}\int_{-\pi}^{\pi} G_k(e^{i\theta})\Psi(e^{i\theta})\hat{Q}(e^{i\theta})^{-1}d\theta = R_k, \quad k = 0, 1, \ldots, n. \tag{3.32}$$

Proof. To prove that $\mathbb{J}_\Psi$ is strictly convex we form

$$\begin{aligned}
\delta^2\mathbb{J}_\Psi(Q;\delta Q) &:= \lim_{\varepsilon\to 0}\frac{\delta\mathbb{J}_\Psi(Q+\varepsilon\delta Q;\delta Q) - \delta\mathbb{J}_\Psi(Q;\delta Q)}{\varepsilon},\\
&= \lim_{\varepsilon\to 0}\frac{1}{\varepsilon}\left\langle \delta Q, \Psi\left[Q^{-1} - (Q+\varepsilon\delta Q)^{-1}\right]\right\rangle,\\
&= \lim_{\varepsilon\to 0}\frac{1}{\varepsilon}\left\langle \delta Q, \Psi\left[I - (I+\varepsilon Q^{-1}\delta Q)^{-1}\right]Q^{-1}\right\rangle.
\end{aligned}$$

However, $(I+\varepsilon Q^{-1}\delta Q)^{-1} = I - \varepsilon Q^{-1}\delta Q + O(\varepsilon^2)$ for sufficiently small $\varepsilon > 0$, and hence

$$\delta^2\mathbb{J}_\Psi(Q;\delta Q) = \left\langle \delta Q, \Psi Q^{-1}\delta Q Q^{-1}\right\rangle.$$

Now, since $Q \in \mathcal{Q}_+^\ell$ is positive definite on the unit circle, there is a nonsingular matrix function S such that $Q^{-1} = SS^*$. Then, using the commuting property of the trace, we have

$$\operatorname{trace}(\delta Q Q^{-1}\delta Q Q^{-1}) = \operatorname{trace}(S^*\delta Q S S^*\delta Q S),$$

and hence

$$\delta^2\mathbb{J}_\Psi(Q;\delta Q) = \langle S^*\delta Q S, \Psi(S^*\delta Q S)\rangle \geq 0,$$

taking the value zero if and only if $S^*\delta Q S = 0$ or, equivalently, $\delta Q = 0$. Consequently, the Hessian of $\mathbb{J}_\Psi(Q)$ is positive definite for all $Q \in \mathcal{Q}_+^\ell$, implying that $\mathbb{J}_\Psi$ is strictly convex, as claimed. The rest of the proof is the same *mutatis mutandis* as the one in [29] as well as the proof of Theorem 3.1.3. Since the linear term $\langle Q, R\rangle$ is positive and linear growth is faster than logarithmic, the function $\mathbb{J}_\Psi$ is proper, that is, the inverse images of compact sets are compact. In particular, if we extend the function $\mathbb{J}_\Psi$ to the boundary of $\mathcal{Q}_+^\ell$, it has compact sublevel sets. Consequently, $\mathbb{J}_\Psi$ has a minimum, $\hat{Q}$, which is unique by strict convexity. We need to rule out that $\hat{Q}$ lies on the boundary. To this end, note that the boundary of $\mathcal{Q}_+^\ell$ consists of the Q for which $\det Q$ has a zero on the unit circle, and for which the directional derivative $\delta\mathbb{J}_\Psi(Q;\delta Q) = -\infty$ for all δQ pointing into $\mathcal{Q}_+^\ell$. Therefore, since $\mathcal{Q}_+^\ell$ is an open set, $\delta\mathbb{J}_\Psi(\hat{Q};\delta Q) = 0$ for all δQ of the form (3.31), and therefore (3.32) follows from Lemma 3.3.4. □

Proof of Theorem 3.3.2. First we show that the matricial Kullback-Leibler discrepancy is convex with respect to its second argument. Proceeding as in the calculation leading to (3.28) yields

$$\delta\mathbb{S}(\Psi,\Phi;\delta\Phi) = \left\langle \delta\Phi, \Psi\Phi^{-1}\right\rangle,$$

and, following the lines of the corresponding proof in Theorem 3.3.5,

$$\delta^2\mathbb{S}(\Psi,\phi;\delta\Phi) \geq 0,$$

with equality if and only if $\delta\Phi = 0$. Hence $\mathbb{S}$ is strictly convex. Let $\hat{Q}$ be the optimal solution of the dual problem. Then, since $\mathbb{S}$ is strictly convex, so is $\Phi \mapsto L(\Phi,\hat{Q})$. Clearly $\hat{\Phi} := \Psi\hat{Q}^{-1}$ belongs to $\mathcal{C}_+^\ell$, and, by (3.28), it is a stationary point of the map $\Phi \mapsto L(\Phi,\hat{Q})$. Hence

$$L(\hat{\Phi},\hat{Q}) \geq L(\Phi,\hat{Q}), \quad \text{for all } \Phi \in \mathcal{C}_+^\ell. \tag{3.33}$$

However, by Theorem 3.3.5, $\hat{\Phi}$ satisfies the interpolation condition (3.22), and consequently

$$L(\hat{\Phi},\hat{Q}) = \mathbb{S}(\Psi,\hat{\Phi}).$$

Therefore, it follows from (3.33) that

$$\mathbb{S}(\Psi,\Phi) \geq \mathbb{S}(\Psi,\hat{\Phi}),$$

for all $\Phi \in \mathcal{C}_+^\ell$ that satisfies the interpolation condition (3.22), establishing optimality of $\hat{\Phi}$. □

To finish the proof of Theorem 3.3.3 it remains to establish that the map $\mathcal{I} : \mathcal{Q}_+^\ell \to \mathcal{R}^\ell$ is a diffeomorphism. To this end, first note that $\mathcal{Q}_+^\ell$ and $\mathcal{R}^\ell$ are both convex, open sets in $\mathbb{R}^{2n\ell^2+\frac{1}{2}\ell(\ell+1)}$ and hence diffeomorphic to $\mathbb{R}^{2n\ell^2+\frac{1}{2}\ell(\ell+1)}$. Moreover, the Jacobian of $\mathcal{I}$ is the Hessian of $\mathbb{J}_\Psi$, which is negative definite on $\mathcal{Q}_+^\ell$, as shown in the proof of Theorem 3.3.5. Hence, by Hadamard's global inverse function theorem [59], $\mathcal{I}$ is a diffeomorphism.

Remark 3.3.6. *The results in this section are in fact a slight generalization of those in [8] in that Ψ is taken arbitrary in $\mathcal{C}_+$ rather than in $\mathcal{Q}_+$.*

Chapter 4

Numerical Algorithms

In order to exploit the theoretical results in the previous chapter, we will need accurate, reliable, robust, and numerically efficient algorithms for computing the interpolants and densities. A large part of the research presented in this thesis are studies of numerical algorithms for this purpose. In fact, developing software is a key issue in control and systems engineering, see for instance the survey [63]. In this chapter we will present three different numerical implementations solving certain special cases of the problem.

Depending on application, different aspects are critical. In the $\mathcal{H}_\infty$ applications, the capability in terms of generality (MIMO, derivative etc) as well as accuracy, are critical whereas computation time is secondary since the design is done offline[1]. Meanwhile, for ARMA estimation the reliability and computation time might be critical in online applications.

Generally we will focus more on accuracy, generality and robustness than on computation time and reliability. This is manifested by the absence of computation time comparison and discussions of algorithmic complexity. These are however still important and will be left as open problems in Chapter 6.

4.1 An Optimization-Based Algorithm without Cepstral-Type Conditions

In this section we will restrict ourselves to the case when we only have covariance-type interpolation conditions. Meanwhile we will allow for matrix-valued functions as in Section 3.3. This section is based on [9, 8, 10].

Recall that, by Theorem 3.3.3, for each choice of $\Psi \in \mathcal{C}_+$, there is a unique solution to the interpolation problem and this solution is obtained by determining

[1] For large-scale MIMO problems and in an iterative design procedure, the computation time can be an issue.

the unique maximizer over $\mathcal{Q}_+^\ell$ of the dual functional

$$\mathbb{J}_\Psi(Q) := \langle Q, R\rangle - \langle \Psi, \log\det Q\rangle\,. \tag{4.1}$$

This functional has the property that its gradient is infinite on the boundary of $\mathcal{Q}_+^\ell$. This is precisely the property that buys us properness of the dual functional (4.1), and therefore it is essential in the proof of Theorem 3.3.3. However, from a computational point of view, this property is undesirable, especially if the maximum is close to the boundary. In fact, it adversely affects the convergence properties of any Newton-type algorithm. For this reason, following [43, 80], we first reformulate the optimization problem to eliminate this property. This is done at the expense of global convexity, but the new functional is still locally strictly convex in a neighborhood of a unique minimizing point. Thus, if we were able to choose the initial point in the convexity region, a Newton method would work well. However, finding such an initial point is a highly nontrivial matter. Therefore, again following [43, 80], we want to design a homotopy continuation method that determines a sequence of points converging to the minimizing point.

We replace $Q(z)$ with its spectral factor $\Gamma(z)$ as in (2.36). Following the calculation in (2.37), the right hand side of (4.1) can also be written as

$$\operatorname{trace}(\Gamma^*\Sigma\Gamma) - \langle \Psi, \log\det \Gamma\Gamma^*\rangle\,,$$

where Σ is the Pick matrix. Let us now assume that the interpolation data (A, B, R) is real; we call this the self-conjugate case. Then the space $\mathcal{Q}_+^\ell$ has dimension $\ell^2 n + \frac{1}{2}\ell(\ell+1)$ and the matrix coefficients $A_0, A_1, \ldots, A_n$ in

$$A(z) := \tau(z)\Gamma(z) = A_0 + A_1 z + \cdots + A_n z^n, \tag{4.2}$$

are real. We also assume that A_0 is upper triangular. The block matrix of coefficient matrices A then belong to $\mathcal{S}_n^\ell$. In terms of the spectral factor $\Gamma(z)$ of $Q(z)$ we have that $A(z) = \tau(z)G(z)\Gamma$ which defines a nonsingular linear real transformation T such that $\Gamma = TA$. Under this change of coordinates, the Pick matrix becomes $T^T\Sigma T$, and, since $\arg\det A(e^{-i\theta}) = -\arg\det A(e^{i\theta})$, the functional in (4.1) can be written

$$\mathbb{J}_\Psi(Q) = J_\Psi(A) + 2\,\langle \Psi, \log\tau\rangle\,,$$

where the new cost functional

$$J_\Psi(A) = \operatorname{trace}\left(A^T T^T \Sigma T A\right) - 2\,\langle \Psi, \log\det A\rangle\,, \tag{4.3}$$

is defined on the space $\mathcal{S}_n^\ell$. We will now study the optimization problem for the functional $J_\Psi(A)$.

Proposition 4.1.1. *The functional $J_\Psi : \mathcal{S}_n^\ell \to \mathbb{R}$ has a unique stationary point and is locally strictly concave about this point.*

Proof. Since $\Gamma(z) := A(z)/\tau(z)$ is a uniquely defined (outer) spectral factor of $Q(z)$, the map $U : \mathcal{S}_n^\ell \to \mathcal{Q}_+^\ell$ sending A to $Q(z) = \Theta(z)AA^*\Theta^*(z)$, where

$$\Theta(z) := \frac{1}{\tau(z)} \begin{bmatrix} I_\ell & zI_\ell & \cdots & z^n I_\ell \end{bmatrix},$$

is a bijection with first and second directional derivatives

$$\begin{aligned} \delta U(A;\delta A) &= \Theta(z)\big(A(\delta A)^* + (\delta A)A^*\big)\Theta^*(z), \\ \delta^2 U(A;\delta A) &= 2\Theta(z)\big((\delta A)(\delta A)^*\big)\Theta^*(z). \end{aligned}$$

Now, $\delta A \mapsto \delta U(A;\delta A)$ is an injective linear map between Euclidean spaces of the same dimension, and hence it is bijective. In fact, since $\det A(z)$ has all its roots in the complement of the closed unit disc, the homogeneous equation

$$A(z)\Delta^*(z) + \Delta(z)A^*(z) \equiv 0, \quad \Delta(z) := \Theta(z)\delta A,$$

has the unique solution $\Delta(z) \equiv 0$ by Lemma 2.4.1. Therefore, since

$$J_\Psi(A) = \mathbb{J}_\Psi(U(A)) + 2\left\langle \Psi, \log \tau \right\rangle,$$

the directional derivative

$$\delta J_\Psi(A;\delta A) = \delta \mathbb{J}_\Psi(U(A);\delta U(A;\delta A)),$$

is zero for all δA if and only if $\delta \mathbb{J}_\Psi(Q;\delta Q) = \left\langle \delta Q, R - \Psi Q^{-1} \right\rangle$ is zero for all δQ. Consequently, J_Ψ has a stationary point at $\hat{A}$ if and only if $\mathbb{J}_\Psi$ has a stationary point at $U(\hat{A})$. However, $\mathbb{J}_\Psi$ has exactly one such point, and hence the same holds for J_Ψ. Moreover, since $\delta^2 \mathbb{J}_\Psi(Q;\delta Q) = \left\langle \delta Q, \Psi Q^{-1}\delta Q Q^{-1} \right\rangle > 0$ for all $\delta Q \neq 0$ and $\delta \mathbb{J}_\Psi(\hat{Q};\delta Q) = 0$ at the minimizer $\hat{Q}$, the second directional derivative

$$\delta^2 J_\Psi(\hat{A};\delta A) = \delta^2\ \mathbb{J}_\Psi(U(\hat{A});\delta U(\hat{A};\ \delta A)) + \delta \mathbb{J}_\Psi(U(\hat{A});\delta^2 U(\hat{A};\delta A)),$$

is positive for sufficiently small $\delta A \neq 0$. Therefore, J_Ψ is strictly convex in some neighborhood of $\hat{A}$. □

The Gradient and the Hessian of the New Functional

In order to use Newton's method to solve the new optimization problem, we need to determine the gradient and the Hessian of J_Ψ. We begin with the gradient.

Proposition 4.1.2. *Let the basis function $G \in \mathcal{G}$, the density $\Psi \in \mathcal{C}_+$, the real $\ell \times \ell$ matrix-valued Fourier coefficients*

$$W_k = \frac{1}{2\pi} \int_{-\pi}^{\pi} e^{ik\theta} \Psi(e^{i\theta}) \big(A^*(e^{i\theta})A(e^{i\theta})\big)^{-1} d\theta, \quad k = 0, 1, \ldots, n, \tag{4.4}$$

and the Pick matrix be Σ *be given. The gradient of* J_Ψ *is given by*

$$\frac{\partial J_\Psi}{\partial A}(A) = 2\big(T^T \Sigma T - W(A)\big)A, \tag{4.5}$$

where the $(n+1)\ell \times (n+1)\ell$ *matrix* $W(A)$ *is the block Toeplitz matrix*

$$W(A) := \begin{bmatrix} W_0 & W_1 & \cdots & W_n \\ W_1^T & W_0 & \ddots & \vdots \\ \vdots & \ddots & \ddots & W_1 \\ W_n^T & \cdots & W_1^T & W_0 \end{bmatrix}. \tag{4.6}$$

Proof. To establish the expression (4.5) in Proposition 4.1.2 for the gradient

$$\frac{\partial J_\Psi}{\partial A}(A) = 2\left(T^T \Sigma T A - \frac{\partial}{\partial A}\left\langle \log\det A, \Psi \right\rangle\right), \tag{4.7}$$

of (4.3), we need to determine

$$\begin{aligned}
&\frac{\partial}{\partial A_k}\left\langle \log\det A, \Psi\right\rangle \\
&\quad = \frac{1}{2\pi}\int_{-\pi}^{\pi} \frac{\partial}{\partial A_k} \log\det A^*(e^{i\theta})\Psi(e^{i\theta}) d\theta, \\
&\quad = \frac{1}{2\pi}\int_{-\pi}^{\pi} e^{-ik\theta} A^*(e^{i\theta})^{-T}\Psi(e^{i\theta}) d\theta, \\
&\quad = \frac{1}{2\pi}\int_{-\pi}^{\pi} \Psi(e^{i\theta})\big(A^*(e^{i\theta})A(e^{i\theta})\big)^{-T} A^T(e^{i\theta}) e^{-ik\theta} d\theta, \\
&\quad = \frac{1}{2\pi}\int_{-\pi}^{\pi} \Psi(e^{i\theta})\big(A^*(e^{i\theta})A(e^{i\theta})\big)^{-T} \begin{bmatrix} e^{-ik\theta} I_\ell & \cdots & e^{-i(n-k)\theta} I_\ell \end{bmatrix} d\theta \cdot A, \\
&\quad = \begin{bmatrix} W_k^T & \cdots & W_0 & \cdots & W_{n-k} \end{bmatrix} A,
\end{aligned}$$

where W_k is defined by (4.4). Putting these together we obtain (4.5). □

Next, we explain how to actually compute $W_0, W_1, \ldots, W_n$. Now, restrict to the case when $\Psi \in \mathcal{Q}_+$ so that it can be represented as $\Psi = \rho\rho^*/(\tau\tau^*)$. This case is particularly interesting in applications. We have

$$\Psi\big(A^*A\big)^{-1} = \rho\rho^*\Big[(\tau A)^*(\tau A)\Big]^{-1}.$$

We can determine $\hat{W}_0, \hat{W}_1, \ldots, \hat{W}_{2n}$ in the expansion

$$\Big[(\tau A)^*(\tau A)\Big]^{-1} = \hat{W}_0 + \hat{W}_1 z + \hat{W}_1^T z^{-1} + \cdots + \hat{W}_{2n} z^{2n} + \hat{W}_{2n}^T z^{-2n} + \cdots,$$

by solving a system of linear equations. Now, defining

$$\mu(z) := \mu_0 + \sum_{s=1}^{n} \mu_\ell (z^s + z^{-s}) = \rho(z)\rho^*(z),$$

we can identify matrix coefficients of equal powers in z in

$$\mu\Big[(\tau A)^*(\tau A)\Big]^{-1} = W_0 + W_1 z + W_1^T z^{-1} + \cdots + W_n z^n + W_n^T z^{-n} + \cdots,$$

to obtain

$$\begin{bmatrix} W_0 \\ W_1 \\ \vdots \\ W_n \end{bmatrix} = \left(\begin{bmatrix} \hat{W}_0 & \hat{W}_1^T & \cdots & \hat{W}_n^T \\ \hat{W}_1 & \hat{W}_0 & \cdots & \hat{W}_{n-1}^T \\ \vdots & \vdots & \ddots & \vdots \\ \hat{W}_n & \hat{W}_{n-1} & \cdots & \hat{W}_0 \end{bmatrix} + \begin{bmatrix} \hat{W}_0 & \hat{W}_1 & \cdots & \hat{W}_n \\ \hat{W}_1 & \hat{W}_2 & \cdots & \hat{W}_{n+1} \\ \vdots & \vdots & \iddots & \vdots \\ \hat{W}_n & \hat{W}_{n+1} & \cdots & \hat{W}_{2n} \end{bmatrix} \right) \begin{bmatrix} \mu_0 I/2 \\ \mu_1 I \\ \vdots \\ \mu_n I \end{bmatrix}.$$

Next we turn to the Hessian computation. Let $\otimes$ denote the Kronecker product, see for instance [57].

Proposition 4.1.3. *The Hessian of J_Ψ is given by*

$$\frac{\partial^2}{(\partial vecA)^2} J_\Psi(A) = 2(I_\ell \otimes T^T \Sigma T) - 2\frac{\partial^2}{(\partial vecA)^2} \langle \log\det A, \Psi \rangle .$$

Here the component of the second term are obtained by rearranging the elements in

$$\left(\frac{\partial}{\partial A_j} \otimes \frac{\partial}{\partial A_k} \right) \langle \log\det A, \Psi \rangle = -S_{j+k}^T, \quad j,k = 0,1,\ldots,n, \tag{4.8}$$

where $S_0, S_1, \ldots, S_{2n}$ are defined via the expansion

$$\Psi(z)\Big(vecA(z)^{-1}\Big)\Big(vecA(z)^{-T}\Big)^T = \sum_{-\infty}^{\infty} S_k z^{-k}. \tag{4.9}$$

Proof. Since

$$\frac{\partial^2 \big(\mathrm{trace} A^T T^T \Sigma T A\big)}{\partial \mathrm{vec} A^2} = 2(I_\ell \otimes T^T \Sigma T),$$

it remains to establish (4.8). Since

$$\frac{\partial}{\partial A_j} \otimes \frac{\partial}{\partial A_k} \langle \log\det A, \Psi \rangle = \frac{1}{2\pi} \int_{-\pi}^{\pi} \frac{\partial}{\partial A_j} \otimes A^*(e^{i\theta})^{-T} \Psi(e^{i\theta}) e^{-ik\theta} d\theta,$$

(4.8) would follow if we could show that

$$\frac{\partial}{\partial A_j} \otimes A^*(z)^{-T} = -z^{-j} \mathrm{vec}(A^*)^{-T} (\mathrm{vec} A^{-*})^T. \tag{4.10}$$

Since $A^{-*}(z)^T A^*(z)^T \equiv I$, denoting the (s,t) element of A_j by A_j^{st} we obtain

$$\frac{\partial}{\partial A_j^{st}} A^*(z)^{-T} = -(A^*)^{-T} \frac{\partial (A^*)^T}{\partial A_j^{st}} (A^*)^{-T} = -z^{-j} (A^*)^{-T} e_s e_t^T (A^*)^{-T},$$

and therefore

$$\begin{aligned}
\frac{\partial}{\partial A_j} \otimes A^*(z)^{-T} &:= \begin{bmatrix} \frac{\partial}{\partial A_j^{11}} A^*(z)^{-T} & \cdots & \frac{\partial}{\partial A_j^{1\ell}} A^*(z)^{-T} \\ \vdots & \ddots & \vdots \\ \frac{\partial}{\partial A_j^{\ell 1}} A^*(z)^{-T} & \cdots & \frac{\partial}{\partial A_j^{\ell\ell}} A^*(z)^{-T} \end{bmatrix}, \\
&= -z^{-j} \begin{bmatrix} (A^*)^{-T} e_1 \\ \vdots \\ (A^*)^{-T} e_m \end{bmatrix} \begin{bmatrix} A^{-*} e_1 \\ \vdots \\ A^{-*} e_m \end{bmatrix}^T, \\
&= -z^{-j} \mathrm{vec}(A^*)^{-T} (\mathrm{vec} A^{-*})^T,
\end{aligned}$$

establishing (4.10). □

Again consider the case when $\Psi \in \mathcal{Q}_+$. Still, since the left hand side of (4.9) is the product of three factors, two of which have Laurent expansions with infinitely many terms, one might wonder how to determine the coefficients $S_0, S_1, \ldots, S_{2n}$ in a finite number of operations. This can be achieved by observing that $\Psi(z)\big(A(z)^T \otimes A(z)^{-1}\big)$ has the same elements as (4.9), appropriately rearranged, and can be factored as the product of two finite and one infinite Laurent expansion.

First expand

$$\Psi(z)(A^{-T}(z) \otimes A^{-1}(z)) = \cdots + \tilde{S}_{2n} z^{-2n} + \cdots + \tilde{S}_1 z^{-1} + \tilde{S}_0 + \cdots,$$

and transform $\tilde{S}_k$ to the coefficient matrices of $\Psi(\mathrm{vec}A^{-1})(\mathrm{vec}A^{-T})^T$ by comparing the elements of $A^{-T} \otimes A^{-1}$ with those of $(\mathrm{vec}A^{-1})(\mathrm{vec}A^{-T})^T$. The computation of $\tilde{S}_k$ can be done by first observing that

$$\begin{aligned}
A^{-T} \otimes A^{-1} &= (A^*)^T (A^{-*})^T A^{-T} \otimes (A^* A^{-*} A^{-1}), \\
&= (A^*)^T (A^* A)^{-T} \otimes (A^* (A A^*)^{-1}), \\
&= ((A^*)^T \otimes A^*)((A^* A)^{-T} \otimes (A A^*)^{-1}), \\
&= ((A^*)^T \otimes A^*)((A^* A)^T \otimes (A A^*))^{-1}, \\
&= ((A^*)^T \otimes A^*)((A^T \otimes A)((A^*)^T \otimes A^*))^{-1}, \\
&= ((A^*)^T \otimes A^*)((A^T \otimes A)(A^T \otimes A)^*)^{-1},
\end{aligned}$$

where we have used properties of the Kronecker product that may be found in, for

instance [57]. Multiplying this by Ψ then yields

$$\begin{aligned}\Psi(A^{-T}\otimes A^{-1}) &= \mu\big((A^*)^T\otimes A^*\big)\big((\tau A^T\otimes A)(\tau A^T\otimes A)^*\big)^{-1},\\ &= \underbrace{(\mu_0+\mu_1(z+z^{-1})+\cdots+\mu_n(z^n+z^{-n}))}_{\mu}\\ &\quad\times\underbrace{(U_0+U_1z^{-1}+\cdots+U_{2n}z^{-2n})}_{(A^*)^T\otimes A^*}\underbrace{(T_0+T_1z+T_1^Tz^{-1}+\cdots)}_{\big((\tau A^T\otimes A)(\tau A^T\otimes A)^*\big)^{-1}},\\ &= \tilde{S}_0+\tilde{S}_{-1}z^{-1}+\cdots+\tilde{S}_{-2n}z^{-2n}+(\text{other terms}),\end{aligned}$$

from which we can compute $\tilde{S}_k$.

The Maximum Entropy Solution

The optimization problem to minimize J_Ψ is particularly simple if $\Psi\equiv 1$. In this case, and only in this case, the problem can be reduced to one of solving a system of linear equations. This solution is generally called the *maximum entropy solution*. In fact, since $\det A(z)$ has no zeros in $\overline{\mathbb{D}}$, by the mean-value theorem of harmonic functions,

$$\frac{1}{2\pi}\int_{-\pi}^{\pi}\log|\det A(e^{i\theta})|d\theta=\log|\det A(0)|.$$

Consequently, since $\arg\det A(e^{-i\theta})=-\arg\det A(e^{i\theta})$,

$$J_1(A)=\mathrm{trace}A^TT^T\Sigma TA-2\log\det A_0.$$

Since $\det A(z)$ has no zeros in the unit disc, A_0 is nonsingular. Therefore, setting the gradient of $J_1(A)$ equal to zero, we obtain

$$T^T\Sigma TA=\hat{I}A_0^{-T},\quad \hat{I}=\begin{bmatrix}I_\ell & 0 & \cdots & 0\end{bmatrix}^T, \tag{4.11}$$

and therefore $A_0=\hat{I}^TA=\hat{I}^TT^T\Sigma T^{-1}\hat{I}A_0^{-T}$, which yields

$$A_0A_0^T=\hat{I}^TT^T\Sigma T^{-1}\hat{I}. \tag{4.12}$$

First solving (4.12) for the unique Cholesky factor and inserting into (4.11), the equation (4.11) reduces to a linear system of equations that has a unique solution A since the Pick matrix Σ is positive definite and T nonsingular. We will denote the solution A^{ME}.

A Continuation Method

Now, we would like to find the maximizer of J_Ψ for an arbitrary $\Psi\in\mathcal{Q}_+$. To this end, we construct a homotopy between the gradient of J_1 and the gradient of J_Ψ

along the lines of [43, 80], allowing us to pass from the maximum entropy solution to the solution of interest. Hence the interpolant is not affected. Now, for any $\nu \in [0,1]$, define

$$\Psi_\nu(z) := 1 + \nu\big(\Psi(z) - 1\big).$$

Then, since $\mathcal{Q}_+(1,n)$ is convex, $\Psi_\nu \in \mathcal{Q}_+(1,n)$ for $0 \leq \nu \leq 1$. By Proposition 4.1.1, the functional

$$J_{\Psi_\nu}(A) = \mathrm{trace} A^T T^T \Sigma T A - 2\left\langle \log\det A, \Psi_\nu \right\rangle,$$

has a unique minimum at $\hat{A}(\nu)$ and is locally strictly convex in some neighborhood of $\hat{A}(\nu)$. This point is the unique solution in $\mathcal{S}_n^\ell$ of the nonlinear equations

$$h(A,\nu) := \frac{\partial J_{\Psi_\nu}(A)}{\partial \mathrm{vec} A} = 0.$$

Then the function $h:\ \mathcal{S}_n^\ell \times [0,1] \to \mathbb{R}^{(n+1)\ell^2}$ is a homotopy from the gradient of J_1 to the gradient of J_Ψ. In particular, $\hat{A}(0)$ is the central solution. In view of the strict local convexity of J_{Ψ_ν} in a neighborhood of $\hat{A}(\nu)$, the Jacobian of $h(A,\nu)$ is negative definite at $\hat{A}(\nu)$. Consequently, by the implicit function theorem, the function $\nu \mapsto \hat{A}(\nu)$ is continuously differentiable on the interval $[0,1]$, and

$$v(\hat{A}(\nu)) := \frac{d}{d\nu}\mathrm{vec}\hat{A}(\nu) = -\left(\frac{\partial h}{\partial \mathrm{vec} A}(A,\nu)\right)^{-1}\left(\frac{\partial h}{\partial \nu}(A,\nu)\right)\Bigg|_{A=\hat{A}(\nu)}, \tag{4.13}$$

where the inverted matrix is the Hessian of J_{Ψ_ν} that can be determined as in Proposition 4.1.3. This is called the *Euler direction*. We want to follow the trajectory $\hat{A}(\nu)$ defined by the solution to this differential equation with the maximum entropy solution as the initial condition. To this end, we construct an increasing sequence of numbers $\nu_0, \nu_1, \ldots, \nu_N$ on the interval $[0,1]$ with $\nu_0 = 0$ and $\nu_N = 1$. Then, for $k = 1, 2, \ldots, N$, we solve the nonlinear equation $h(A, \nu_k) = 0$ for $\mathrm{vec}\hat{A}(\nu_k)$ by Newton's method. The Newton steps are given by

$$\mathrm{vec}A(\nu_{k+1}) - \mathrm{vec}A(\nu_k) = -\frac{\partial^2 h}{\partial \mathrm{vec}A(\nu_k)^2}(A(\nu_k),\nu_k)h(A(\nu_k),\nu_k). \tag{4.14}$$

The numbers $\nu_0, \nu_1, \ldots, \nu_N$ have to be chosen close enough so that, for each $k = 1, 2, \ldots, N$, $A_0(\nu_k)$ lies in the local convexity region of $J_{\Psi_{\nu_k}}$, guaranteeing that Newton's method converges to $\hat{A}(\nu_k)$. Strategies for choosing $\nu_0, \nu_1, \ldots, \nu_N$ are given in [43, 80]. This choice is a trade-off between convergence and staying in the locally convex region. We summarize in Algorithm 1.

Here the error bound $c_1(\nu)$ is chosen so that we iterate closely enough to the trajectory and an acceptable final gradient error is achieved. One can also add a constraint in the absolute interpolation error – at least for the final iteration.

Note that the algorithms in [43, 80, 9] are special cases of the algorithm proposed here. We will not provide a convergence proof, but for the scalar case with standard basis functions a convergence proof is available in [44].

Algorithm 1. *Optimization-Based Algorithm*

Compute A^{ME} by solving (4.11) and (4.12).
Set $A \leftarrow A^{ME}$ and $\nu \leftarrow 0$
while $\nu < 1$
  begin Predictor step
    Determine Euler direction $v(A(\nu))$ by (4.13)
    Determine step length $\delta\nu_*$ by the rules in [80]
    Set $\delta\nu \leftarrow \min\{\delta\nu_*, 1 - \nu\}$
    while $A + \delta\nu v(A(\nu)) \notin \mathcal{S}_n^\ell$
      Set $\delta\nu \leftarrow \delta\nu/2$
    end while
    Set $\nu \leftarrow \nu + \delta\nu$
    Set $A \leftarrow A + \delta\nu v(A(\nu))$
  end Predictor step
  begin Corrector step
    while $\max\{h_k(A, \nu)\} > c_1(\nu)$
      Determine Newton step δA by (4.14)
      Set $\delta\nu \leftarrow 1$
      while $A + \delta\nu\delta A \notin \mathcal{S}_n^\ell$
        Set $\delta\nu \leftarrow \delta\nu/2$
      end while
      Set $A \leftarrow A + \delta\nu\delta A$
    end while
  end Corrector step
end while

4.2 Solving the Equation $T(a)Ka = d$

In this section we will again treat the case without cepstral-type conditions and we will restrict even more, namely to the scalar case and the case where $\Psi \in \overline{\mathcal{Q}}_+$. The latter is a restriction in the sense that the $\overline{\mathcal{Q}}_+$ is finite dimensional, while $\mathcal{C}_+$ is not, but a generalization in the sense that we allow for nonnegative rather than positive densities on the unit circle. Due to continuity this will also imply that the parameterization is better scaled close to the boundary of $\overline{\mathcal{Q}}_+$ which has proven to be advantageous in applications. Moreover, the algorithm proposed in this section will be used also for the case with cepstral conditions in the next chapter. This section is based on [6, 7].

More precisely, we will design a numerical method to solve any problem of the following type: let a set of basis functions $G \in \mathcal{G}$ and interpolation data, in the form of a positive definite Pick matrix Σ, be given. For *any* $\Psi \in \overline{\mathcal{Q}}_+$, compute the solution to the Kullback-Leibler Approximation Problem 3.1.1, extended to allow for the present choice of Ψ.

Some Topological Results Regarding the Parameterization

The theory developed in Chapter 3 treats the positive case, that is when densities and pseudo-polynomials are positive on the circle, the shaping filter is strictly miniphase and strictly stable, and the positive-real function is strictly so. In [51] it was proven, for the scalar case without cepstral matching, that the parameterization extends to the boundary case. More precisely, consider the extended map

$$\begin{array}{rccl} M & : & \overline{\mathcal{A}}_n & \rightarrow & \overline{\mathcal{Q}}_+, \\ & & a(z) & \mapsto & a(z)K^*(a(z)) + a(z)^*K(a(z)). \end{array} \tag{4.15}$$

Theorem 4.2.1. *[51] Consider the standard basis* (2.12)*. Then the map M is bijective.*

We can state the result in terms of the solvability of (2.6).

Corollary 4.2.2. *For each vector $d \in \overline{\mathcal{Q}}_+$ the system of nonlinear equations*

$$m(a) := T(a)Ka - d = 0,$$

has a unique solution in $\overline{\mathcal{A}}_n$.

Next, we will discuss important properties of the map M defined by (4.15). As was stated in Theorem 4.2.1, it is a bijection. However, we can actually show that the map M is a *homeomorphism*. In the region of *strictly* positive real interpolants, it is even a *diffeomorphism*, as shown by Byrnes and Lindquist in [30]. Both of these properties will turn out to be vital in justifying the numerical continuation method proposed here. First we have the following theorem.

Theorem 4.2.3. *The map M defined by* (4.15) *is a homeomorphism.*

In order to prove this theorem, we introduce normalized versions of the sets, namely

$$\overline{\mathcal{A}}_n^0 := \left\{ a \in \overline{\mathcal{A}}_n : \ \langle a(z), K(a(z)) \rangle = 1 \right\}, \tag{4.16}$$

$$\overline{\mathcal{Q}}_+^0 := \left\{ d \in \overline{\mathcal{Q}}_+ : \ d_0 = 1 \right\}. \tag{4.17}$$

We can state a normalized version of the theorem as

Lemma 4.2.4. *The map $M^0 : \overline{\mathcal{A}}_n^0 \rightarrow \overline{\mathcal{Q}}_+^0$ defined by*

$$M^0(a(z)) := a(z)K^*(a(z)) + a(z)^*K^*(a(z)), \tag{4.18}$$

is a homeomorphism.

Proof. First, note that the normalization $\langle a(z), K(a(z))\rangle = 1$ is equivalent to setting

$$\left\langle M^0(a(z)), 1\right\rangle = 2. \tag{4.19}$$

The map M^0 is a bijection by Theorem 4.2.1. Hence, it suffices to show that it is continuous and that the domain $\overline{\mathcal{A}}_n^0$ is closed and bounded.

First we address the boundedness of $\overline{\mathcal{A}}_n^0$. Let us take $a \in \mathcal{A}_n^0$ and form the corresponding positive real function

$$f(z) := \frac{K(a(z))}{a(z)}.$$

Note that the coefficient vectors a and Ka can be written respectively as $a_0\tilde{a}$ and $b_0\tilde{b}$ where $\tilde{a}, \tilde{b} \in \overline{\mathcal{S}}_n$ with constant terms equal one. Also note that due to the assumption on the first basis function G_0 in (2.11) we will get one interpolation condition at origin: $f(0) = w_0$. This yields the relation $b_0 = w_0 a_0$. Hence, $f(z) = w_0\tilde{b}(z)/\tilde{a}(z)$.

Since $f(z)$ is positive real, the roots of $\tilde{a}(z)$ are in the closed unit disc. Thus, and since a_0 is taken positive, it suffices to show that a_0 is bounded from above. In order to prove the boundedness of a_0, we need to utilize the positivity of the Pick matrix.

$$\Sigma = \frac{1}{2\pi}\int_{-\pi}^{\pi} G(f+f^*)G^* d\theta = \frac{w_0}{2\pi}\int_{-\pi}^{\pi} G\left(\frac{\tilde{a}}{\tau}\right)^{-1} \frac{\tilde{a}\tilde{b}^* + \tilde{b}\tilde{a}^*}{\tau\tau^*}\left(\frac{\tilde{a}}{\tau}\right)^{-*} G^* d\theta.$$

Since $\tilde{a}/\tau \in \mathrm{Span}\{G_k\}$, there is a vector $t \in \mathbb{C}^{n+1}$ such that $\tilde{a}/\tau = t^H G$. Therefore, and since the Pick matrix is positive definite, there is an $\varepsilon > 0$ such that

$$a_0^2\varepsilon \le a_0^2 t^H \Sigma t = \frac{1}{2\pi}\int_{-\pi}^{\pi} \frac{M^0(a(e^{i\theta}))}{\left|\tau(e^{i\theta})\right|^2} d\theta \le \frac{N}{2\pi}\int_{-\pi}^{\pi} M^0(a(e^{i\theta}))d\theta = 2N,$$

for some finite number N satisfying $N \ge \max_{\theta\in[-\pi,\pi]} 1/\left|\tau(e^{i\theta})\right|^2$. Here, the finiteness of N follows from the assumption that the basis function matrix A is Hurwitz, and the last equality follows from (4.19). Since ε and N are independent of a_0, we have shown that a_0 is bounded from above, and thus $\overline{\mathcal{A}}_n^0$ is bounded.

Secondly, we deal with the continuity of M^0. For two arbitrary elements a_1 and a_2 in $\overline{\mathcal{A}}_n^0$, we have

$$\begin{aligned} M^0(a_1) - M^0(a_2) &= a_1^* K(a_1) + K(a_1)^* a_1 - \left(a_2^* K(a_2) + K(a_2)^* a_2\right), \\ &= (a_1 - a_2)^* K(a_1) + a_2^* K(a_1 - a_2), \\ &\quad + K(a_1)^*(a_1 - a_2) + K(a_1 - a_2)^* a_2. \end{aligned}$$

Therefore, since the linear map K is bounded and the set $\overline{\mathcal{A}}_n^0$ is bounded, it is easy to show that there exists a constant $C > 0$ satisfying

$$\|M^0(a_1) - M^0(a_2)\| \le C\|a_1 - a_2\|.$$

Since a_1 and a_2 are arbitrary in $\overline{\mathcal{A}}_n^0$, we have proven the continuity of M^0 in $\overline{\mathcal{A}}_n^0$.

Finally the closedness of $\overline{\mathcal{A}}_n^0$ follows immediately by using the property of continuity and bijectivity of M^0 and the closedness of $\tilde{\mathcal{D}}$, see for instance [92]. This completes the proof. □

Proof of Theorem 4.2.3. The map G can be written as a composite of three maps:

$$M = M^1 \circ M^2 \circ M^3,$$

where

$$\begin{aligned}
&M^1 : \mathcal{A} \mapsto \overline{\mathcal{A}}_n^0 \times \mathbb{R}_+ : && M^1(a) := \begin{bmatrix} \dfrac{a}{(\langle a, K(a)\rangle)^{1/2}} \\ \langle a, K(a)\rangle \end{bmatrix}, \\
&M^2 : \overline{\mathcal{A}}_n^0 \times \mathbb{R}_+ \mapsto \overline{\mathcal{Q}}_+^0 \times \mathbb{R}^+ : && M^2(a, r) := \begin{bmatrix} M^0(a) \\ r \end{bmatrix}, \\
&M^3 : \overline{\mathcal{Q}}_+^0 \times \mathbb{R}_+ \mapsto \overline{\mathcal{Q}}_+ : && M^3(d, r) := rd.
\end{aligned}$$

The maps M^1 and M^3 play the roles of normalization and inverse scaling, respectively. Due to Lemma 4.2.4, the map M^2 is homeomorphic. Since the maps M^1 and M^3 are also homeomorphic, so is the composite map M. □

An Algorithm Based on a Homotopy

To construct a homotopy, we notice that if d is chosen as $\tau := [2\tau_0, \tau_1, \cdots, \tau_n]^T \in \overline{\mathcal{Q}}_+^0$ which consists of coefficients of the trigonometric polynomial

$$\sum_{k=0}^{n} \tau_k (z^k + z^{-k}) := \frac{1}{N} \prod_{k=0}^{n} (1 - z_k z)(1 - z_k z)^*,$$

where N is a scaling to make $\tau_0 = 1$, then the corresponding system of nonlinear equations

$$g_\tau(a) := T(a)Ka - \tau = 0, \tag{4.20}$$

is easy to solve. In fact, as shown in the previous section, the maximum entropy solution can be determined by solving (4.11) and (4.12), that is a matrix factorization and a system of linear equations, respectively.

We design a convex homotopy $h : \mathbb{R}^{n+1} \times [0, 1] \to \mathbb{R}^{n+1}$ as

$$\begin{aligned}
h(a, \nu) &:= (1 - \nu) g_\tau(a) + \nu g(a), \\
&= T(a)Ka - \tau + \nu(\tau - d), \quad \nu \in [0, 1].
\end{aligned}$$

Note that $h(a, 0) = g_\tau(a)$ and $h(a, 1) = g(a)$. Therefore, $h(a, 0) = 0$ is easy to solve, while $h(a, 1) = 0$ is our problem at hand.

For each $\nu \in [0, 1]$, the system $h(a, \nu) = 0$ has a unique solution in $\hat{\mathcal{A}}$, due to Corollary 4.2.2 and the convexity of $\overline{\mathcal{Q}}_+$. Let us denote the unique solution

of the system $h(a, \nu) = 0$ as $\hat{a}(\nu)$, and we call the class $\{\hat{a}(\nu)\}_{\nu=0}^{1}$ the *trajectory*. Our objective is to trace this implicitly defined trajectory numerically from $\nu = 0$ to $\nu = 1$, and to obtain $\hat{a}(1)$. For this purpose, we use a numerical continuation method with predictor-corrector steps. Before proceeding the exposition of the predictor-corrector steps, we shall analyze the properties of the trajectory.

To apply a continuation method to trace a trajectory, the trajectory should enjoy some favorable properties, since otherwise, the method is likely to break down or end up with a incorrect solution. For example, the trajectory should have sufficient smoothness, such as continuity and differentiability, but neither a bifurcation nor a turning point, see for instance [2]. We shall next study these properties for the trajectory $\{\hat{a}(\nu)\}_{\nu=0}^{1}$.

First, due to Theorem 4.2.3 and Theorem 2.3.6, the property of $\{\hat{a}(\nu)\}_{\nu=0}^{1}$ concerning the smoothness follows immediately.

Proposition 4.2.5. *The trajectory $\{\hat{a}(\nu)\}_{\nu=0}^{1}$ is continuously differentiable in the interval $[0, 1)$. In addition, it is continuous at $\nu = 1$.*

A direct consequence of Proposition 4.2.5 is the following.

Corollary 4.2.6. *The trajectory $\{\hat{a}(\nu)\}_{\nu=0}^{1}$ does not have any turning point in $[0, 1)$.*

Because of the uniqueness of $\hat{a}(\nu)$ for each $\nu \in [0, 1]$, we also have that

Corollary 4.2.7. *The trajectory $\{\hat{a}(\nu)\}_{\nu=0}^{1}$ does not have any bifurcation.*

The three properties above justify the use of a continuation method to trace the trajectory numerically, see, for instance, [2].

Owing to the continuous differentiability of the trajectory, we can express the trajectory as a solution of an ordinary differential equation with an initial value. When we take a derivative of $h(\hat{a}(\nu), \nu) = 0$ with respect to ν, we have

$$\frac{\partial h}{\partial a}(\hat{a}(\nu), \nu) \cdot \frac{d\hat{a}}{d\nu}(\nu) + \frac{\partial h}{\partial \nu}(\hat{a}(\nu), \nu) = 0,$$

or equivalently,

$$\frac{d\hat{a}}{d\nu}(\nu) = v(\hat{a}(\nu)), \quad v(\hat{a}(\nu)) := -\left[\frac{\partial h}{\partial a}(a, \nu)\right]^{-1} \frac{\partial h}{\partial \nu}(a, \nu)\bigg|_{a=\hat{a}(\nu)}. \tag{4.21}$$

The invertibility of the Jacobian $\partial h/\partial a$ is guaranteed on the trajectory in the interval $[0, 1)$ because of the differentiability of $\hat{a}$. Since we can easily compute $\hat{a}(0)$, the problem to solve is the ordinary differential equation with an initial value:

$$\begin{cases} \dfrac{d\hat{a}}{d\nu}(\nu) & = & v(\hat{a}(\nu)), \\ \hat{a}(0) & : & \text{given.} \end{cases}$$

To solve this initial value problem numerically, we use predictor-corrector steps.

In the predictor step, given a point $\hat{a}(\nu)$ on the trajectory, we move the point to a new point as

$$a(\nu + \delta\nu) := \hat{a}(\nu) + \delta\nu \cdot v(\hat{a}(\nu)).$$

This point may not be on the trajectory, and hence we use the notation $a(\nu + \delta\nu)$ instead of $\hat{a}(\nu + \delta\nu)$. To determine the new point $a(\nu + \delta\nu)$, we need to compute $v(\hat{a}(\nu))$ and to provide $\delta\nu$.

The directional vector $v(\hat{a}(\nu))$ consists of two factors. One is the inverse of $\partial h/\partial a$, which is the Jacobian of g and can be written explicitly as

$$\nabla g(a) = T(a)K + T(Ka).$$

Note that, due to the differentiability of $g(a)$, the Jacobian is nonsingular for all ν in the interval $[0, 1)$. The other factor is

$$\frac{\partial h}{\partial \nu} = \tau - d,$$

which is independent of both a and ν, and can thus be computed offline.

The step size must be chosen in a careful way. This is because a small step size will cause a long time to arrive at $\nu = 1$ and to obtain $\hat{a}(1)$, while a large step size might ruin the convergence rate of corrector step that follows the predictor step. Next, we propose a reasonable way to make this trade-off.

On the trajectory, we have $h(\hat{a}(\nu), \nu) = 0$, and in particular,

$$e_1^T h(\hat{a}(\nu), \nu) = e_1^T T(\hat{a}(\nu))K\hat{a}(\nu) - e_1^T \tau = 2\hat{a}(\nu)^T K\hat{a}(\nu) - 2 = 0.$$

One of the criterion for not deviating too far from the trajectory is to require

$$1 - \mu \le a(\nu + \delta\nu)^T K a(\nu + \delta\nu) \le 1 + \mu,$$

for some small number $\mu > 0$. For a given direction $v(\hat{a}(\nu))$, we can compute the maximal step length $\delta\nu_*$ as

$$\begin{aligned}
(a + \delta\nu_* v(\hat{a}))^T K (a + \delta\nu_* v(\hat{a})) &= 1 \pm \mu, \\
\delta\nu_*^2 v(\hat{a})^T K v(\hat{a}) + \delta\nu_* (v(\hat{a})^T K a + a^T K v(\hat{a})) \pm \mu &= 0, \\
\delta\nu_*^2 + \delta\nu_* \underbrace{\frac{v(\hat{a})^T K a + a^T K v(\hat{a})}{\delta a^T K v(\hat{a})}}_{=:p} \pm \underbrace{\frac{\mu}{\delta a^T K v(\hat{a})}}_{=:q} &= 0.
\end{aligned}$$

We pick the smallest positive solution:

$$\delta\nu_* = \begin{cases} -\frac{p}{2} - \sqrt{\frac{p^2}{4} - |q|} & \text{if } p < 0 \ \&\ \frac{p^2}{4} > |q|, \\ -\frac{p}{2} + \sqrt{\frac{p^2}{4} + |q|} & \text{otherwise.} \end{cases} \tag{4.22}$$

Algorithm 2. *Nonlinear Positive Real Equations*

Compute a^{ME}, for instance by (4.11) and (4.12)
Set $a \leftarrow a^{ME}$ and $\nu \rightarrow 0$
while $\nu < 1$
  begin Predictor step
    Determine Euler direction $v(a(\nu))$ by (4.21)
    Determine step length $\delta\nu_*$ in (4.22)
    Set $\delta\nu \leftarrow \min\{\delta\nu_*, 1-\nu\}$
    while $a + \delta\nu v(a(\nu)) \notin \mathcal{A}_n$
      Set $\delta\nu \leftarrow \delta\nu/2$
    end while
    Set $\nu \leftarrow \nu + \delta\nu$
    Set $a \leftarrow a + \delta\nu v(a(\nu))$
  end Predictor step
  begin Corrector step
    while $\max\{h_k(a,\nu)\} > c_2(\nu)$
      Determine Newton step δa by (4.23)
      Set $\delta\lambda \leftarrow 1$
      while $a + \delta\lambda\delta a \notin \overline{\mathcal{A}}_n$
        Set $\delta\lambda \leftarrow \delta\lambda/2$
      end while
      Set $a \leftarrow a + \delta\lambda\delta a$
    end while
  end Corrector step
end while

In the corrector step, given a point $a(\nu + \delta\nu)$ which is obtained in the predictor step, we pull the point back to the trajectory by fixing ν, and obtain $\hat{a}(\nu + \delta\nu)$. This is equivalent to solving the system of nonlinear equations

$$\tilde{h}(a) := h(a, \nu + \delta\nu) = 0.$$

We use Newton's method with an initial point $a(\nu + \delta\nu)$ to solve this system numerically. Newton's method uses iterations

$$\delta a_{k+1} = a_{k+1} - a_k = -\nabla\tilde{h}(a_k)^{-1} \cdot \tilde{h}(a_k), \quad k = 0, 1, 2, \ldots. \tag{4.23}$$

The Jacobian $\nabla\tilde{h}(a_k)$ is nonsingular if a_k is close enough to the trajectory, due to the continuity of the Jacobian.

The complete algorithm is given in Algorithm 2. Here the error bound $c_2(\nu)$ is chosen so that $c_2(1)$ is the desired accuracy at the end, and so that the we iterate closely enough to the trajectory.

We will not provide any convergence proof for the algorithm. However, it is our numerical experience that the algorithm is both fast, robust, and accurate

and preferable to Algorithm 1 whenever it is applicable. For instance the embedded solver in the implementation described in [11, 82] is an implementation of Algorithm 2.

4.3 Simultaneously Solving the Cepstral and Covariance Equations

In this section we will study the scalar problem also allowing for the cepstral-type conditions. Since we have generalized the theory of [25], the numerical algorithms of [44, 45] are no longer applicable. Here we will construct a homotopy with respect to the covariances and cepstral interpolation data from some initial values to some desired values. By choosing the initial values so that the covariance-type conditions are satisfied, we can construct a homotopy which lies in the connected sub-manifold $\mathcal{P}_n(r)$. Thereby, we can solve the inverse problem of going from the interpolation data to a model by solving a set of ordinary differential equations. The well-posedness shown in Section 3.2 is critical in motivating the here proposed algorithm.

We shall consider the case of $m = n$ being the default. Also, the smoothness properties in Section 3.2 were derived for this case. However, a generalization to the arbitrary (n, m) seems within reach. Moreover we will assume that the prefilter density is rational of degree n. In this section we shall index the polynomial coefficients in decreasing powers of the variable.

Assume that the prefilter density Ψ is given by

$$\Psi = \frac{\hat{a}\hat{a}^*}{\hat{\sigma}\hat{\sigma}^*},$$

where $\hat{a}$ and $\hat{\sigma}$ are of order n. Typically, for ARMA estimation, we will take $(\hat{a}, \hat{\sigma})$ as a preliminary estimate of the ARMA model. We also choose some basis function $\tilde{G} \in \mathcal{G}$. The map ξ, defined by (3.13), then has the components

$$\begin{aligned} \xi_k \quad : \quad \mathcal{P}_n \quad &\to \quad \mathbb{R}, \\ (a, \sigma) \quad &\mapsto \quad \left\langle \tilde{G}_k, \frac{\hat{a}\hat{a}^*}{\hat{\sigma}\hat{\sigma}^*} \frac{\sigma\sigma^*}{aa^*} \right\rangle, \end{aligned}$$

for $k = 0 \dots n$. The normalized coefficients are as before given by $\eta_k = \xi_k / \xi_0$ for $k = 0 \dots n$. Likewise, for the cepstral-type equations, we have the map ζ defined in (3.14) with components

$$\begin{aligned} \zeta_k \quad : \quad \mathcal{P}_n \quad &\to \quad \mathbb{R}, \\ (a, \sigma) \quad &\mapsto \quad \left\langle \tilde{G}_k, \frac{\hat{a}\hat{a}^*}{\hat{\sigma}\hat{\sigma}^*} \log \frac{\sigma\sigma^*}{aa^*} \right\rangle, \end{aligned}$$

for $k = 1 \dots n$. The map F is given by (3.12). Let r and c be some given data normalized so that $r_0 = 1$. As noted in Section 3.1, generic data might not belong

to $\mathcal{X}_n$. To circumvent this issue we study some regularization of the problem. Here we consider the regularization terms

$$s_k(\sigma) := \left\langle \tilde{G}_k, \frac{\hat{a}\hat{a}^*}{\hat{\sigma}\hat{\sigma}^*}\frac{1}{\sigma\sigma^*} \right\rangle, \ k = 1 \dots m,$$

for the corresponding cepstral equations.

First we will state and prove an immediate algebraic result formalizing the transformation from one set of basis functions to another set. This proposition has been published in [12].

Proposition 4.3.1. *Let (A, B) and $(\tilde{A}, \tilde{B})$ corresponding to $G, \tilde{G} \in \mathcal{G}$ be given. Then, for any $q, \tilde{q} \in \mathbb{C}^{n+1}$ such that $Q(z) = q^T G(z) + q^*(G^*)^T(z)$ and $\tilde{Q}(z) = \tilde{q}^T\tilde{G}(z) + \tilde{q}^*(\tilde{G}^*)^T(z)$ with common numerator $Q(z)\tau(z)\tau^*(z) = \tilde{Q}(z)\tilde{\tau}(z)\tilde{\tau}^*(z)$,*

$$\tilde{q} = Vq,$$

where V is the invertible matrix

$$\begin{aligned} V &:= \tilde{\Gamma}^{-T} L_{\tilde{\tau}}^{-1} T(\tilde{\tau})^{-1} T(\tau) L_\tau^T \Gamma^T, \\ L_r &:= \begin{bmatrix} r_0 & & & \\ r_1 & r_0 & & \\ \vdots & & \ddots & \\ r_n & r_{n-1} & \dots & r_0 \end{bmatrix}, \end{aligned}$$

$T(r)$ is given by (2.6), and Γ is the reachability matrix.

Proof. For each q we have

$$Q(z) = q^T G(z) + q^*(G^*)^T(z) = \frac{b(z)}{\tau(z)} + \frac{b^*(z)}{\tau^*(z)} = \frac{b(z)\tau^*(z) + b^*(z)\tau(z)}{\tau(z)\tau^*(z)} = \frac{d(z)}{\tau(z)\tau^*(z)},$$

Note that both $\tau(z)$ and $\tilde{\tau}(z)$ are polynomials of degree at most n due to the assumption that $\det A = \det \tilde{A} = 0$. The coefficients of the pseudo-polynomial $d(z)$ are given by $T(\tau)b$. Noting that

$$a_0 + a_1 z + \cdots = \frac{b(z)}{\tau(z)} = \frac{q^T}{2} G(z) = \frac{q^T}{2}(B + ABz + \cdots),$$

we have that $a = \Gamma^T q/2$. Comparing coefficients we also have that $b = L_\tau a$. Combining the expressions, the coefficients of $d(z)$ is given by $T(\tau)L_\tau\Gamma^T q/2$. Now, since $\tilde{Q}(z) = d(z)/(\tilde{\tau}(z)\tilde{\tau}(z)^*)$ we have that

$$T(\tau)L_\tau\Gamma^T q = T(\tilde{\tau})L_{\tilde{\tau}}\tilde{\Gamma}^T\tilde{q}.$$

All the included matrices are invertible: T by [33], L since $\tau_0 = \tilde{\tau}_0 = 1$, and Γ due to the reachability assumption. Thus, $\tilde{q} = Vq$ with V invertible. □

We will consider a particular choice of basis function, namely orthonormal basis functions $G = (I - Az)^{-1}B$ and so that $\det(I - Az) = \hat{\sigma}(z)$ and another set of orthonormal basis function $\tilde{G} = (I - \tilde{A}z)^{-1}\tilde{B}$ such that $\det(I - \tilde{A}z) = \hat{a}(z)$. Then we apply Proposition 4.3.1 and instead consider the simplified maps with components

$$\begin{aligned} \xi_k(a,\sigma) &= \left\langle G_k, \frac{\sigma\sigma^*}{aa^*} \right\rangle, \\ \zeta_k(a,\sigma) &= \left\langle G_k, \log\frac{\sigma\sigma^*}{aa^*} \right\rangle, \\ s_k(\sigma) &= \left\langle G_k, \frac{1}{\sigma\sigma^*} \right\rangle, \end{aligned}$$

for $k = 0, 1, \ldots m$.

Remark 4.3.2. *Another obvious choice of basis function is to take $\tilde{G}$ as the standard basis defined in* (2.12)*. This typically enable fast evaluation of the functions ξ and ζ as well as their derivatives. However, it is our experience that the method presented here yields better numerical scaling.*

Since the manifold $\mathcal{P}_n^*$, or equivalently $\mathcal{X}_n$, is known to have $n+1$ connected components, see, for instance, [19, 96], we need to be somewhat careful in designing a homotopy from a known solution to the desired solution. We will use the following result proven by Byrnes and Lindquist:

Corollary 4.3.3. *[30, Corollary 5.5] The submanifolds $\mathcal{P}_n(r)$ are connected.*

Therefore, we wish to start in a point (a_0, σ_0) which fulfills the covariances type conditions. Supposing that the prefilter density corresponds to a model, which is close to the desired, a natural initial point is to take $\sigma_0 = \hat{\sigma}$. Then, to find an a_0 such that the covariance-type conditions hold, we can use Algorithm 2. Note that in general this $a_0 \neq \hat{a}$. Then we can construct a homotopy from $F(a_0, \sigma_0) = (r, c_0)$ to the desired values $F(a, \sigma) = (r, c)$ which stays in one component of $\mathcal{P}_n^*$ as

$$\begin{aligned} h \quad : \quad [0,1] \times \mathcal{P}_n &\rightarrow U \subset \mathbb{R}^{2n}, \\ (\mu, a, \sigma) &\mapsto F(a,\sigma) - \varepsilon \begin{bmatrix} 0 \\ s(\sigma) \end{bmatrix} - \mu \begin{bmatrix} r_0 \\ c \end{bmatrix} - (1-\mu) \begin{bmatrix} r_0 \\ c_0 \end{bmatrix}. \end{aligned} \tag{4.24}$$

The initial point is given by $h(0, a_0, \sigma_0) = 0$ and the desired solution is given by the nonlinear equation $h(1, a, \sigma) = 0$. We have a *trajectory* defined by

$$\{(a,\sigma) \in \mathcal{P}_n(r) : h(\mu, a, \sigma) = 0, \mu \in [0,1]\}.$$

Since $\mathcal{P}_n(r)$ is a subset of $\mathcal{P}_n^*$, we have by Theorem 3.2.1 that F restricted to $\mathcal{P}_n(r)$ is a diffeomorphism onto its image. Hence it has a full rank Jacobian there, and 0 is a regular value of the homotopy h. Thus we will get a smooth curve from the

initial point to the solution. In particular we have no turning point, bifurcations, and the curve is of finite length.

We define the initial value problem as in [2], by differentiating h with respect to μ. Let $x := (a, \sigma)$.

$$\frac{\partial h}{\partial x}\frac{\partial x}{\partial \mu} + \frac{\partial h}{\partial \mu} = 0.$$

Here

$$\frac{\partial h}{\partial x} = \frac{dF}{dx} - \varepsilon \begin{bmatrix} 0 & 0 \\ 0 & \dfrac{ds}{d\sigma} \end{bmatrix},$$

so F being diffeomorphic implies that $\partial h/\partial x$ is full rank for all x for some $\varepsilon > 0$. Hence we get the initial value problem as

$$\left\{ \begin{array}{rcl} \dfrac{dx}{d\mu} & = & -\left(\dfrac{\partial h}{\partial x}\right)^{-1} \dfrac{\partial h}{\partial \mu}, \\ x(0) & = & 0. \end{array} \right. \tag{4.25}$$

To solve the initial value problem one can apply some predictor-corrector method, see [2], or some other ordinary differential equation solver. For the predictor-corrector method we have the Euler step as

$$v(\mu, x) := -\left(\frac{\partial h}{\partial x}\right)^{-1} \frac{dh}{d\mu}, \tag{4.26}$$

and Newton steps as

$$x_{k+1} - x_k = -\left(\frac{\partial h}{\partial x}\right)^{-1} h(\mu, x). \tag{4.27}$$

In any case we will need to evaluate the right-hand side of the (4.25). We have

$$\frac{\partial h}{\partial \mu} = -\begin{bmatrix} r \\ c \end{bmatrix} + \begin{bmatrix} r_0 \\ c_0 \end{bmatrix} \text{ and } \frac{\partial h}{\partial x} = \begin{bmatrix} \dfrac{d\eta}{da} & \dfrac{d\eta}{d\sigma} \\ \dfrac{d\zeta}{da} & \dfrac{d\zeta}{d\sigma} \end{bmatrix} - \varepsilon \begin{bmatrix} 0 & 0 \\ 0 & \dfrac{ds}{d\sigma} \end{bmatrix}.$$

First, consider the covariance part of the equation. We will need to evaluate the covariances ξ_k themselves. Note that

$$\xi_k = \left\langle G_k, \frac{\sigma\sigma^*}{aa^*} \right\rangle = \left\langle G_k, \frac{b}{a} + \frac{b^*}{a^*} \right\rangle,$$

where $b = T_a^{-1} T_\sigma \sigma/2$. Hence, defining $\Gamma_k = \begin{bmatrix} B & AB & \dots & A^k B \end{bmatrix}$, we can evaluate the integrals as

$$\xi = (I + e_1 e_1^T) a (A)^{-1} \Gamma_n b,$$

where e_1 is the first unit vector of length n. From (3.19) we have that

$$\begin{aligned}\frac{d\xi_k}{da_l} &= -\left\langle G_k, \frac{\sigma\sigma^*}{aa^*}\frac{T(a)z^l}{aa^*}\right\rangle = -\left\langle G_k, \frac{p_l}{a^2} + \frac{p_l^*}{a^{2*}}\right\rangle, \\ \frac{d\xi_k}{d\sigma_l} &= \left\langle G_k, \frac{T(\sigma)z^l}{aa^*}\right\rangle = \left\langle G_k, \frac{q_l}{a} + \frac{q_l^*}{a^*}\right\rangle,\end{aligned}$$

for some polynomials $p_l(z)$ and $q_l(z)$ of degrees $2n$ and n, respectively. In fact, the polynomial coefficients are immediately computable as

$$\begin{aligned}p_l &= T_{a^2}^{-1} M_{T_\sigma \sigma/2} T_a e_l, \\ q_l &= T_a^{-1} T_\sigma e_l,\end{aligned}$$

where

$$\begin{aligned}M_{d_1} \;:\; \mathcal{Q}_n &\to \mathcal{Q}_{2n}, \\ d_2 &\mapsto d_1 d_2,\end{aligned}$$

is the linear operator corresponding to multiplication of two pseudo-polynomials. Then the derivatives can be evaluated as

$$\begin{aligned}\frac{d\xi}{da_l} &= -(I + e_1 e_1^T) a(A)^{-2} \Gamma_{2n} p_l, \\ \frac{d\xi}{d\sigma_l} &= (I + e_1 e_1^T) a(A)^{-1} \Gamma_n q_l.\end{aligned}$$

The derivative of η are then readily computable using (3.18). Next, consider the cepstral part of the equation. From (3.16) we have that

$$\begin{aligned}\frac{d\zeta_k}{da_l} &= -\left\langle G_k, \frac{z^l}{a}\right\rangle, \\ \frac{d\zeta_k}{d\sigma_l} &= \left\langle G_k, \frac{z^l}{\sigma}\right\rangle.\end{aligned}$$

Gathering the equations, we have that

$$\begin{aligned}\frac{d\zeta}{da} &= -\begin{bmatrix}0 & I\end{bmatrix} a(A)^{-1}\Gamma_n \begin{bmatrix}0 & I\end{bmatrix}^T, \\ \frac{d\zeta}{d\sigma} &= \begin{bmatrix}0 & I\end{bmatrix} \sigma(A)^{-1}\Gamma_n \begin{bmatrix}0 & I\end{bmatrix}^T.\end{aligned}$$

For the regularization term, the derivative is determined as for the covariance-type conditions:

$$\frac{ds_k}{d\sigma_l} = -\left\langle G_k, \frac{T(\sigma)z^l}{\sigma^2\sigma^{2*}}\right\rangle = -\left\langle G_k, \frac{\pi_l}{\sigma^2} + \frac{\pi_l^*}{\sigma^{2*}}\right\rangle,$$

where $\pi_l = T_{\sigma^2}^{-1} T_\sigma e_l$. We can evaluate the integral as

$$\frac{ds}{d\sigma_l} = -\begin{bmatrix}0 & I\end{bmatrix} \sigma(A)^{-2}\Gamma_{2n}\pi_l.$$

Algorithm 3. *Cepstral- and Covariance Matching*

Set $a \leftarrow a^{ME}$, $\sigma \leftarrow \hat{\sigma}$, and $\mu \leftarrow 0$
while $\mu < 1$
  begin Predictor step
    Determine Euler direction v by (4.26)
    Set $d\mu \leftarrow c_3(\mu) \leq 1 - \mu$.
    while $x + d\mu v \notin \mathcal{P}_n^*$
      Set $d\mu \leftarrow d\mu/2$
    end while
    Set $x \leftarrow x + d\mu v$
  end Predictor step
  begin Corrector step
    while $\max\{h_k(\mu, x)\} > c_4(\mu)$
      Determine Newton step v by (4.27)
      Set $d\nu \leftarrow 1$
      while $x + d\nu v \notin \mathcal{P}_n^*$
        Set $d\nu \leftarrow d\nu/2$
      end while
      Set $x \leftarrow x + d\nu v$
    end while
  end Corrector step
end while

Algorithm 3 is a predictor-corrector algorithm for solving the cepstral and covariance equations simultaneously. Here $c_3(\mu)$ is some function of μ determining the step size in the predictor step. With small increments we follow the trajectory closely, but that increases the number of steps. How to choose $c_3(\mu)$ is therefore a trade-off. The function $c_4(\mu)$ affects the accuracy in the corrector step, that is how close to the trajectory we need to be before taking a new predictor step. The value $c_4(1)$ gives the accuracy of the final solution. To test whether $x \in \mathcal{P}_n^*$, we compute the Schur parameters of a and σ and check whether they are less than one in modulus.

Chapter 5

Applications

In this chapter we will illustrate the theory of Chapter 3 and the numerical algorithms of Chapter 4 by considering four examples in the two categories sensitivity shaping and ARMA estimation. In neither example we provide any physical background describing the complexity and challenges which each real-world problem represents. Yet they should illustrate the applicability of the theory within some engineering disciplines. Both robust control examples deal with sensitivity shaping, but also other problem classes can be treated, see [79, 78].

5.1 A MIMO Sensitivity Shaping Benchmark Problem

Our first example is a sensitivity shaping problem for a plant with vector-valued in- and output. Being able to treat the MIMO case is an important development for the robust control applications. Since the high order of conventional $\mathcal{H}_\infty$ controllers often become a serious problem in the MIMO case, it is in the MIMO case our approach has the chance to outperform the conventional methods.

First we will discuss continuous-time MIMO sensitivity shaping in general. Now P in Figure 1.1 is a linear control system with a vector-valued input $u(t)$ and a vector-valued output $y(t)$, having a rational transfer function $P(s)$ with unstable poles and non-miniphase zeros. The sensitivity function in (1.1) is also a rational matrix-valued function. Substituting the Youla-parameterization yields a model matching form:

$$S(s) = T_1(s) - T_2(s)Q(s)T_3(s), \tag{5.1}$$

where T_j, $j = 1, 2, 3$ and Q are stable rational matrices with Q arbitrary. To avoid some technical complications and simplify notation, let us assume that the plant P is square and full rank, that is $\det P(s) \not\equiv 0$. Then both T_2 and T_3 are square and full rank. Now, the (transmission) zeros of T_2 and T_3 are located at the zeros respectively the poles of the plant P. By inner-outer factorizations $T_2 = \Theta_2\tilde{T}_2$ and $T_3 = \tilde{T}_3\Theta_3$, respectively, the non-miniphase zeros of the plant are transferred to the inner function Θ_2 and the unstable poles to the inner function Θ_3. Moreover,

the outer factor $\tilde{T}_2$ contains the relevant information about "relative degree" of P. In particular, $\tilde{T}_2(\infty)$ has the same rank as $P(\infty)$. Then, following the procedure in [35], we define $\tilde{S} := \phi\Theta_2^* S\Theta_3^*$ and $\tilde{T}_1 := \phi\Theta_2^* T_1\Theta_3^*$, where $\phi := \det\Theta_2 \det\Theta_3$. Hence (5.1) can be transformed into

$$\tilde{S}(s) = \tilde{T}_1(s) - \phi(s)\tilde{T}_2(s)Q(s)\tilde{T}_3(s), \quad \|S\|_\infty = \|\tilde{S}\|_\infty, \tag{5.2}$$

where ϕ is a scalar inner function having zeros at the unstable poles and zeros of P. If these poles and zeros, denoted by $s_0, s_1, \ldots, s_n$, are distinct and $P(\infty)$ has full rank, the interpolation conditions required for internal stability become

$$\tilde{S}(s_k) = \tilde{T}_1(s_k), \quad k = 0, 1, \ldots, n, \tag{5.3}$$

whereas any multiple point has to be handled in a separate way. If s_k is an interpolation point of multiplicity ν so that $s_k = s_{k+1} = \cdots = s_{k+\nu-1}$, then the equations in (5.3) corresponding to $s_{k+1} = \cdots = s_{k+\nu-1}$ are replaced by

$$\tilde{S}^{(j)}(s_k) = \tilde{T}_1^{(j)}(s_k), \ j = 1, \ldots, \nu - 1. \tag{5.4}$$

If $P(\infty)$ is rank deficient, we also need to add interpolation conditions at infinity to ensure that the controller is proper. To see this, recall that $P(\infty)$ has the same rank as $\tilde{T}_2(\infty)$. Therefore, if $P(\infty)$ is rank deficient, then $v^T\tilde{T}_2(\infty) = 0$ for some v, and hence, in view of (5.2), we have the interpolation condition

$$v^T\tilde{S}(\infty) = v^T\tilde{T}_1(\infty). \tag{5.5}$$

If $\tilde{T}_2$, and thus P, is strictly proper, this interpolation condition becomes

$$\tilde{S}(\infty) = U_0 := \tilde{T}_1(\infty). \tag{5.6}$$

More generally, if in addition the first $k-1$ Markov parameters are zero, that is $A_1 = \cdots = A_{k-1} = 0$ in the expansion

$$\tilde{T}_2(s^{-1}) = A_1 s + A_2 s^2 + A_3 s^3 + \cdots,$$

and A_{k+1} is full rank, a similar argument shows that

$$\frac{d^j}{ds^j}\tilde{S}(s^{-1})\Big|_{s=0} = U_j := \frac{d^j}{ds^j}\tilde{T}_1(s^{-1})\Big|_{s=0}, \quad j = 1, \ldots, k-1, \tag{5.7}$$

and

$$v^T\frac{d^k}{ds^k}\tilde{S}(s^{-1})\Big|_{s=0} = v^T\frac{d^k}{ds^k}\tilde{T}_1(s^{-1})\Big|_{s=0} \tag{5.8}$$

for any v such that $v^T A_k = 0$. We would like to express all these conditions as interpolation conditions involving some analytic function and its derivatives. To this end, introduce the modified sensitivity function

$$Z(s) := \tilde{S}(s^{-1}), \tag{5.9}$$

which has the same analyticity properties as S (and as $\tilde{S}$), that is Z is analytic in the right half of the complex plane. Then, to avoid tangential conditions, we replace conditions (5.6),(5.7), and (5.8) by

$$Z^{(j)}(0) = U_j, \quad j = 0, 1, \ldots, k. \tag{5.10}$$

Likewise, (5.3) becomes

$$Z(s_k^{-1}) = \tilde{T}_1(s_k),$$

whereas (5.4) corresponds to easily computed but somewhat more complicated expressions for $Z^{(j)}(s_k^{-1})$,$j = 1, \ldots, \nu - 1$.

Remark 5.1.1. *These interpolation conditions in terms of Z are sufficient but may not be necessary. In fact, the tangential conditions* (5.5) *and* (5.8) *have been allowed to hold in all directions v. The reason for this is that tangential interpolation is not covered by the theory developed in this thesis.*

Remark 5.1.2. *In our problem formulation, we do not allow for interpolation points on the boundary of the analyticity region. Therefore we shall move the interpolation point $s = 0$ in* (5.10) *slightly into the open right half plane.*

Given some $\gamma > \gamma_{opt}$, we consider the whole class of stable Z satisfying the required interpolation conditions and some complexity constraint. To bring this problem in conformity with the problem formulation in Chapter 3, we transform first the interpolation points in the right half plane to $z_0, z_1, \ldots, z_n$ in the unit circle, via the linear fractional transformation $z = (s-1)(s+1)^{-1}$, and then the

$$F(z) := \left[\gamma I - Z\left(\frac{1+z}{1-z}\right)\right]\left[\gamma I + Z\left(\frac{1+z}{1-z}\right)\right]^{-1}.$$

For each Z satisfying $\|Z\|_\infty = \|S\|_\infty < \gamma$, the new function F is analytic in the unit disc and has the property that $F(z) + F^*(z) > 0$ for all $z \in \mathbb{T}$, that is, it is a matrix-valued positive-real function. The interpolation conditions become

$$F(z_k) = R_k, \tag{5.11}$$

for each k such that z_k has multiplicity one and

$$\frac{1}{j!}F^{(j)}(z_k) = R_{k+j}, \quad j = 0, 1, \ldots, \nu - 1, \tag{5.12}$$

whenever z_k has multiplicity ν and $z_k = z_{k+1} = \cdots = z_{k+\nu-1}$. It is straightforward, but tedious in the multiple-pointcase, to determine the interpolation values $R_0, R_1, \ldots, R_n$. Using the filter-bank, that is the input-to-state, framework for representing the interpolation points, enable fairly direct translation of the interpolation data.

Next we will consider a design problem which illustrates the design paradigm in the MIMO case. To compute each interpolant we will use Algorithm 1 in Section 4.1. This sensitivity shaping for a MIMO plant is taken from a popular textbook on multivariable control by Maciejowski [74] and the design has been published in [8].

Example 5.1.3 (Flight Control). The control system in [74] describes the vertical-plane dynamics of an airplane and can be linearized to yield a linear system P with three inputs, three outputs and five states, namely

$$\begin{aligned} \dot{x} &= Ax + Bu, \\ y &= Cx + Du, \end{aligned}$$

where

$$\begin{aligned} A &= \begin{bmatrix} 0 & 0 & 1.1320 & 0 & -1.000 \\ 0 & -0.0538 & -0.1712 & 0 & 0.0705 \\ 0 & 0 & 0 & 1.0000 & 0 \\ 0 & 0.0485 & 0 & -0.8556 & -1.013 \\ 0 & -0.2909 & 0 & 1.0532 & -0.6859 \end{bmatrix}, \\ B &= \begin{bmatrix} 0 & 0 & 0 \\ -0.12 & 1.0000 & 0 \\ 0 & 0 & 0 \\ 4.4190 & 0 & -1.665 \\ 1.5750 & 0 & -0.0732 \end{bmatrix}, \\ C &= \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \end{bmatrix}, \\ D &= \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}. \end{aligned}$$

This system is not asymptotically stable due to the pole at the origin. It is strictly proper ($D = 0$) and the first Markov coefficient

$$CB = \begin{bmatrix} 0 & 0 & 0 \\ -0.12 & 1 & 0 \\ 0 & 0 & 0 \end{bmatrix},$$

is rank deficient. To compare our result with that of [74], we want to design a one-degree-of-freedom controller C as in Figure 1.1 that renders the closed-loop system robust against various disturbances. More precisely, the specifications are

- Bandwidth about 10 rad/s
- Zero sensitivity at zero frequency; $S(0) = 0$
- Well-damped step responses

By exploiting the design freedom offered by choosing the *design parameters*, namely an upper limit γ of the gain, the spectral zeros, and additional interpolation constraints, we shape the sensitivity function to meet the specifications, while limiting the degree of the controller. The spectral zeros are the roots of $\sigma(z)$ in

$\Psi = \sigma(z)\sigma^*(z)/(\tau(z)\tau^*(z)) \in \mathcal{Q}_+$. Note that the class of solutions increases when we increase γ. For each choice of spectral zeros, we get a particular solution in the solution set.

First we deal with the pole at origin. By perturbing the A matrix we move the pole into the open right half-plane, generating an interpolation point as described in Remark 5.1.2. More precisely, we move the pole to 10^{-6} by increasing the upper left element of A to 10^{-6}. This will ensure the sensitivity S to be zero near zero frequency. In terms of the modified sensitivity function Z in (5.9), this yields the interpolation condition $Z(10^6) = 0$. Since the plant is strictly proper and the first Markov parameter CB is rank deficient, we need to add interpolation conditions for Z and Z' at zero in accordance with (5.10). Moving the interpolation condition slightly into the open right half-plane (Remark 5.1.2), these conditions become

$$Z(10^{-8}) = I, \quad Z'(10^{-8}) = U_1 := \begin{bmatrix} 0 & 0 & 0 \\ 0 & -2 \cdot 10^{-6} & 0 \\ 0 & 0 & -2 \cdot 10^{-6} \end{bmatrix}.$$

To force the controller to be strictly proper and create a steep "roll-off" of the complementary sensitivity function, we also add the condition $Z''(10^{-8}) = 0$. Then the class of bounded interpolants becomes

$$\left\{ Z \in R\mathcal{H}_\infty : \begin{array}{c} Z(10^{-8}) = I,\ Z'(10^{-8}) = U_1, Z''(10^{-8}) = 0, \\ Z(10^6) = 0,\ \|Z\|_\infty < \gamma \end{array} \right\},$$

where γ is a bound to be selected in the design. By means of a linear fractional transformation and an appropriate scaling, we transform the problem to the form considered in Section 3.3, yielding the family

$$\left\{ F \in \mathcal{F}_+(3) :\ F(0) = 1.9250I,\ F'(0) = F''(0) = 0,\ F(0.9997) = I \ \right\},$$

for the particular choice of γ described next. We now tune the design parameters to meet the design specifications. First we pick the upper bound $\gamma = 3.16$ (10 dB). However, the actual infimum norm of the sensitivity will be smaller. Furthermore, we want to peak the sensitivity function somewhat above 10 rad/s. We can achieve this by choosing spectral zeros close to the imaginary axis in the corresponding region. Here, we first pick the points $\{60, \pm 40i\}$ and transform them to the unit disc by the same linear fractional transformation as for the interpolation points. By rescaling each resulting root to have absolute value less than 0.95, if necessary, we avoid numerical difficulties and prevent the peak of $|S|$ from becoming too high. In this way, we obtain the spectral zeros $\{0.3969, 0.4936 \pm 0.4998i\}$, which we use in Algorithm 1 to determine the corresponding unique interpolant F. Then we transform back to S and calculate $C(s) = P(s)^{-1}\big(S(s)^{-1} - I\big)$. In Table 5.1 we compare our control design with the $\mathcal{H}_\infty$ design using the weighting functions of [74, pp. 306-315]. In Figure 5.1 the (singular-value) frequency responses of the sensitivity and the complementary sensitivity of both designs are plotted, and in

Table 5.1: Comparison between the proposed and conventional design.

	Proposed design	Conventional design
Controller degree	8	17
Peak $\|S\|_\infty$ (dB)	1.3419	1.3582
Peak $\|T\|_\infty$ (dB)	0.9984	1.2328
Bandwidth S (rad/s)	7.3938	4.6202
Bandwidth T (rad/s)	16.1141	16.4140

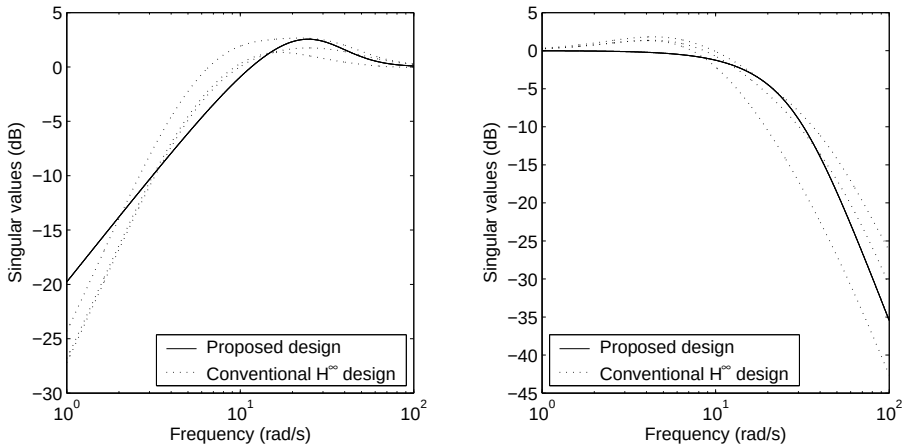

Figure 5.1: Singular value plots of the sensitivity and the complementary sensitivity.

Figure 5.2 the step responses are depicted. Note that in our design the three curves coincide. Clearly, both designs meet the design specifications. We emphasize that although our design meets the specifications at least as well as does the conventional $\mathcal{H}_\infty$ design, the McMillan degree of our controller is only half of that of the conventional $\mathcal{H}_\infty$ controller.

Since $\mathcal{H}_\infty$ control design often leads to controllers of high degree, it is therefore customary to apply some method of model reduction. This is typically done by balanced truncation [76], where states that correspond to relatively small entries on the diagonals of the balanced observability/controllability Gramian are removed. Although such procedures are quite *ad hoc*, a certain reduction in degree can often be done without unacceptable degradation in performance. An interesting question is now whether the conventional $\mathcal{H}_\infty$ design in the present example can be reduced to the same degree as our design, namely eight, without unacceptable degradation. The answer is "No". To see this we have used the DC gain matching function in MATLAB's Control Toolbox. Successively removing states in the conventional $\mathcal{H}_\infty$ design, we found that the controller can be reduced to degree eleven without

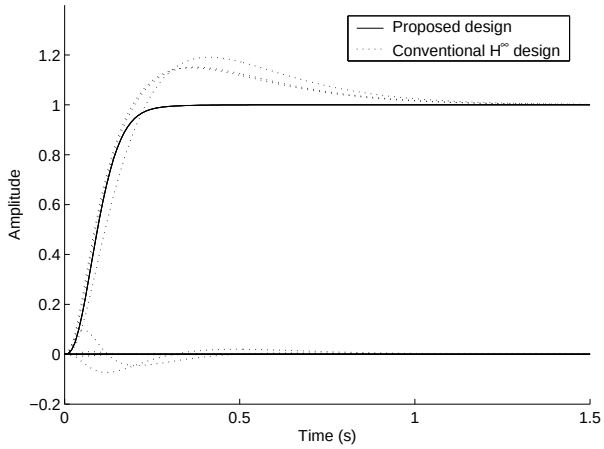

Figure 5.2: The step responses.

Table 5.2: Comparison between model-reduced controllers.

	Proposed design	Conventional design
Controller degree	6	11
Peak $\|S\|_\infty$ (dB)	1.3419	1.3593
Peak $\|T\|_\infty$ (dB)	0.9984	1.2327
Bandwidth S (rad/s)	7.3938	4.6202
Bandwidth T (rad/s)	16.1141	16.4140

loss of internal stability and without undue degradation in performance, whereas reduction to ten leads to an unacceptable design. Of course, model reduction could also be applied to the proposed design. In fact, the degree of proposed controller can be reduced to six without unacceptable degradation in performance, restoring the ratio in the control degree between the two methods. The results are displayed in Table 5.2.

The corresponding (singular-value) frequency responses of the sensitivity and the complementary sensitivity are displayed in Figure 5.3, and the step responses are depicted in Figure 5.4. Our design still is of considerably smaller McMillan degree while meeting the design specifications at least as well as the $\mathcal{H}_\infty$ design.

Remark 5.1.4. *In interpreting these model-reduction results we need to observe that the interpolation conditions used in our procedure to ensure internal stability are in fact only sufficient, see Remark 5.1.1. Modifying our procedure to handle tangential interpolation would allow us to use necessary and sufficient interpolation conditions. This would reduce the total number of interpolation conditions imposed*

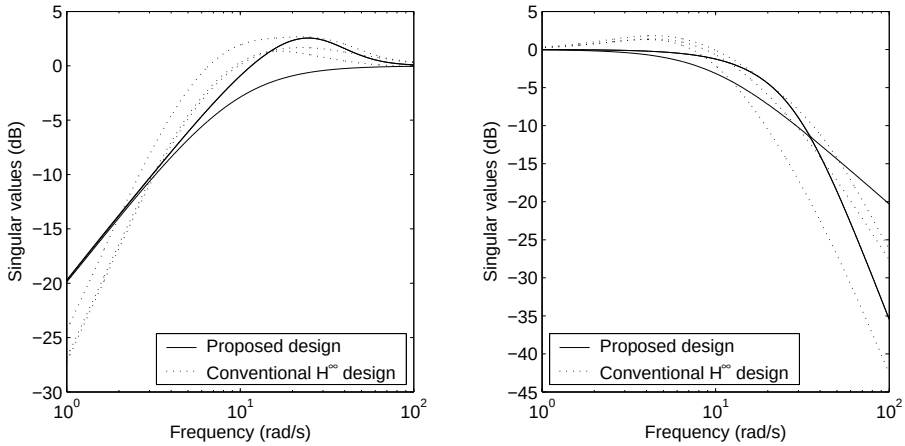

Figure 5.3: Singular value plots of the sensitivity and the complementary sensitivity corresponding to the model reduced controllers.

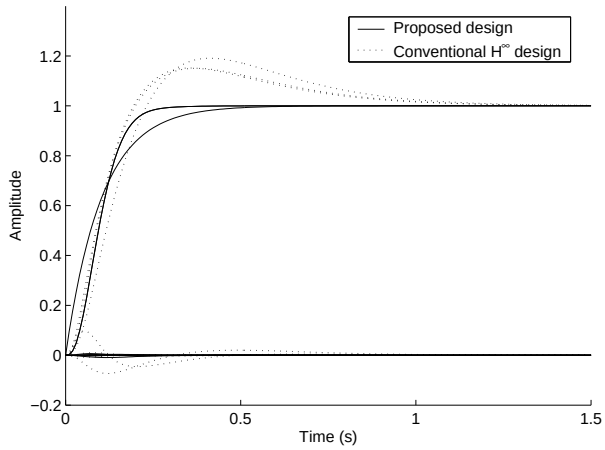

Figure 5.4: The step responses for the reduced-order designs.

on the sensitivity function, thus very likely leading to a lower degree controller. In fact, modifying our approach in this way, it is quite possible that the sixth-degree controller obtained above after model reduction of our design could then be obtained directly (without model reduction) using appropriate tuning.

5.2 A Scalar Sensitivity Shaping Benchmark Problem

As seen from the previous example the design unavoidably amounts to some tuning of the controllers. There, the spectral zeros and the $\mathcal{H}_\infty$ upper bound γ were directly used as tuning parameters. The effect of a certain choice of spectral zeros is sometimes hard to predict, not always making them suitable as design parameters. In this section we will discuss an alternative approach which formulates an approximation problem as a tool for tuning the controllers. We will also illustrate the method on a benchmark example. This section is entirely based on [82, 11].

A Sensitivity Shaping Problem

Consider Figure 1.1 and the sensitivity function S in (1.1) for the discrete time case. Assume that, at a given finite number N of frequencies $\theta := \{\theta_k\}_{k=1}^N \subset [0, \pi]$, a "desired" frequency response $s := \{s_k\}_{k=1}^N \subset \mathbb{C}$ of S is given, and we try to find a "best-approximate" sensitivity function S from a class of "allowable" sensitivity functions (see Figure 5.5[1]). Next, what we mean by "best-approximate" and "allowable" will be explained.

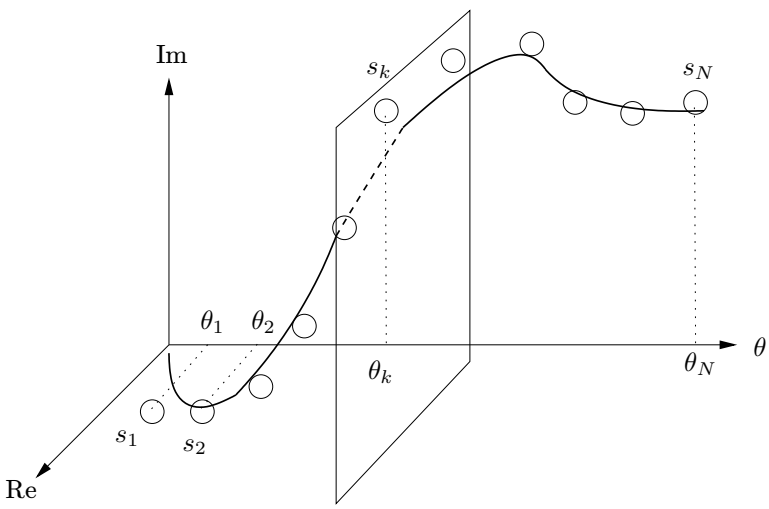

Figure 5.5: The frequency response of a "best-approximate" sensitivity function S (solid curve) to data s_k (circles) at frequencies θ_k on θ-axis.

To clarify the meaning of "best-approximation", we need to introduce a discrepancy between the desired frequency response data (θ, s) and a sensitivity function

[1]The 3-D plot in Figure 5.5 can be interpreted as a combination of the gain plot and the phase plot in the Bode diagram.

S. We use the weighted squares sum[2]:

$$d_w((\theta, s), S) := \frac{1}{2} \sum_{k=1}^{N} \frac{w_k}{|s_k|^2} \left| S(e^{i\theta_k}) - s_k \right|^2, \tag{5.13}$$

where the weights $w := \{w_k\}_{k=1}^N$ are positive scalars to be chosen by the designer; if one wants a better approximation at the frequency θ_k, one can choose a large w_k relative to weights at other frequencies. In (5.13), the term $\left| S(e^{i\theta_k}) - s_k \right|$ is the distance of two complex numbers $S(e^{i\theta_k})$ and s_k in the complex plane; see the dashed arrow in Figure 5.5. A "best-approximate" sensitivity function S is the one which minimizes this discrepancy for given (w, θ, s).

We will call a sensitivity function S "allowable" if it satisfies the following four conditions:

(C1) the internal stability condition,

(C2) n_e interpolation conditions $S(\lambda_j) = \eta_j$, $j = 1, \ldots, n_e$, which are specified at points $\lambda_j \in \mathbb{C}$, not only outside the unit disc but also different from unstable poles and zeros of the plant,

(C3) the $\mathcal{H}_\infty$ norm bound condition $\|S\|_\infty < \gamma$ and γ is chosen to be large enough so that there exists an S which satisfies (C1)–(C3), and

(C4) rationality and a degree condition, i.e., S must be real rational and $\deg S \leq n := n_p + n_z + n_e - 1$, where n_p and n_z are the number of unstable poles and zeros, including infinite zeros and counting multiplicities, of the plant P, respectively.

The motivations for these conditions are as follows. (C1) is a standard requirement for any practical feedback system. (C2)–(C4) are motivated by the work in [27, 81]. (C2) increases the flexibility of the shaping design, see [81], where these conditions are called *additional interpolation constraints.* We may not need this condition for achieving required performance, in which case, we just set $n_e = 0$. As for (C3), there are motivations from both control viewpoint and optimization viewpoint. From control viewpoint, the constraint (C3) is called the *gain-phase margin constraint*, see for instance [61, p. 20], and (C3) is important to avoid a large peak gain of S ensuring a large stability margin. From an optimization viewpoint, (C3) is useful to avoid choosing an initial point far from the solution in nonconvex optimization that we need to solve; see [82]. (C4) restricts the class to a degree constrained one, which eventually leads to a restriction on the controller degree, as stated in Proposition 1.0.2. Note that the number of additional conditions are just added to the degree of the controller and sensitivity function.

With definitions of the discrepancy d_w in (5.13) and the class of allowable sensitivity functions

$$\mathcal{T} := \{S : \ S \text{ satisfies (C1)–(C4)}\}, \tag{5.14}$$

[2]Division by $|s_k|^2$ is for normalization. We assume $s_k \neq 0$.

the *sensitivity shaping problem* to be considered in this section is, for given weights w and data (θ, s), to solve an optimization problem:

$$\inf_{S \in \mathcal{T}} d_w((\theta, s), S). \tag{5.15}$$

The infimum may not be achieved by any S in $\mathcal{T}$. In the next subsection, we will reduce the problem (5.15) to a finite dimensional constrained nonlinear least-squares problem by expressing the class $\mathcal{T}$ in terms of a finite dimensional parameter vector.

Remark 5.2.1. The set $\mathcal{T}$ is actually the real rational solution set to the *Nevanlinna-Pick interpolation problem with degree constraint* and directly corresponds to the polynomial set $\mathcal{A}_n$ in (2.21) as shown in the following subsection. The degree bound in (C4) is chosen as n to guarantee the nonemptyness of the set $\mathcal{T}$.

Reduction to a Nonlinear Least-Squares Problem

Next we will reduce the optimization problem (5.15) to a finite dimensional nonlinear least-squares problem. Suppose that S is a feasible point of the optimization problem in (5.15), that is $S \in \mathcal{T}$. Then, since S satisfies (C4), it can be factored as

$$S(z) = \frac{b(z)}{a(z)}, \tag{5.16}$$

where $a(z) := \bar{z}^T a$, $b(z) := \bar{z}^T b$, $a \in \mathbb{R}^{n+1}$, $b \in \mathbb{R}^{n+1}$ and $\bar{z} := [z^n, \cdots, z, 1]^T$. In addition, since S satisfies (C1) and (C2), S needs to fulfill $n_p + n_z + n_e (= n + 1)$ interpolation/derivative conditions at unstable poles and zeros (including infinite zeros) of the plant, as well as at points specified by (C2). Due to these $n + 1$ conditions, we can derive a linear relation between b and a as

$$b = Ka, \tag{5.17}$$

for a uniquely determined real matrix K as in (2.19). In this context we have the following corollary to Proposition 2.3.5.

Corollary 5.2.2. *Suppose that a sensitivity function S satisfies the conditions (C1)–(C4). Then, the matrices $\gamma I + K$ and $\gamma I - K$ are invertible.*

Proof. Due to the norm condition (C3), the interpolation conditions on the function value fulfill $f(z_j) = w_j < \gamma$. Then $\gamma I + K$ and $\gamma I - K$ are invertible by Proposition 2.3.5. □

Since S satisfies (C3), S must be stable and meet the norm condition $\|S\|_\infty < \gamma$. The stability condition can be stated that the denominator vector a needs to be in the Schur stability region $\mathcal{S}_n$ with $a_0 > 0$. The norm condition can be expressed as

$$\gamma^2 \left|a(e^{i\theta})\right|^2 - \left|b(e^{i\theta})\right|^2 > 0, \quad \forall \theta,$$

which leads to spectral factorization

$$\gamma^2 a(z)a(z^{-1}) - b(z)b(z^{-1}) = \sigma(z)\sigma(z^{-1}), \tag{5.18}$$

for a unique[3] spectral factor $\sigma(z) := \bar{z}^T\sigma$ with $\sigma \in \mathcal{S}_n$.

So far, we have explained that, for each $S \in \mathcal{T}$, there corresponds some $a \in \mathcal{A}_n$, where $\mathcal{A}_n$ is an open set in $\mathbb{R}^{n+1}$ defined by

$$\mathcal{A}_n := \left\{ a \in \mathcal{S}_n : \gamma^2 \left|e(\theta)^T a\right|^2 - \left|e(\theta)^T K a\right|^2 > 0,\ \forall\theta \right\},$$

with $e(\theta) := [e^{in\theta}, e^{i(n-1)\theta}, \cdots, 1]^T$. The converse is trivial; for each $a \in \mathcal{A}_n$, the function $S := (\bar{z}^T K a)/(\bar{z}^T a)$ is in $\mathcal{T}$. We have also explained that, for each $a \in \mathcal{A}_n$, there corresponds to a unique $\sigma \in \mathcal{S}_n$. Actually, a much stronger assertion holds for the map between $\mathcal{A}_n$ and $\mathcal{S}_n$, as stated in Theorem 2.3.6.

Due to this parameterization of $\mathcal{T}$, we can reduce the sensitivity shaping problem (5.15) to the following finite dimensional constrained NLS problem:

$$\inf_{\sigma\in\mathcal{S}_n} \frac{1}{2}\sum_{k=1}^{N} \frac{w_k}{|s_k|^2}\left|\frac{e_k^T K h(\sigma)}{e_k^T h(\sigma)} - s_k\right|^2, \tag{5.19}$$

where $e_k := e(\theta_k)$, $k = 1, \ldots, N$.

Remark 5.2.3. *The problem* (5.15) *can also be reduced to a finite dimensional constrained NLS problem with respect to a:*

$$\inf_{a\in\mathcal{A}_n} \frac{1}{2}\sum_{k=1}^{N} \frac{w_k}{|s_k|^2}\left|\frac{e_k^T K a}{e_k^T a} - s_k\right|^2. \tag{5.20}$$

However, it is our numerical experience that it is advantageous to solve (5.19) *rather than* (5.20).

Next, we will discuss the map h from $\mathcal{S}_n$ to $\mathcal{A}_n$, in Theorem 2.3.6. Let us express a nonlinear map h from $\mathcal{S}_n$ to $\mathcal{A}_n$ as a composition of three maps:

$$h := h_3 \circ h_2 \circ h_1. \tag{5.21}$$

We will explain next what these three maps are.

First, the map h_1 is defined in the domain $\mathcal{S}_n$ as

$$h_1(\sigma) := \frac{1}{2}T(\sigma)\sigma, \quad \sigma \in \mathcal{S}_n. \tag{5.22}$$

It was shown in [32] that the map h_3 is a diffeomorphism from $\mathcal{S}_n$ to the $\mathcal{Q}_+$.

[3]Without the positivity condition $\alpha_0 > 0$, the spectral factor σ would be determined uniquely up to sign.

Next, the map h_2 is defined in the domain $\mathcal{Q}_+$ as the inverse map of

$$g_2(\hat{a}) := T(\hat{a})\hat{K}\hat{a}, \quad \hat{a} \in \hat{\mathcal{A}}_n. \tag{5.23}$$

The domain of g_2 is an open set in $\mathbb{R}^{n+1}$:

$$\hat{\mathcal{A}}_n := \left\{ \hat{a} \in \mathcal{S}_n : \min_{\theta\in[-\pi,\pi]} \mathrm{Re}\left[\frac{\hat{e}(\theta)^T\hat{K}\hat{a}}{\hat{e}(\theta)^T\hat{a}}\right] > 0 \right\},$$

where $\hat{e}(\theta) := \left[1, e^{i\theta}, \cdots, e^{in\theta}\right]^T$ and

$$\hat{K} := (\gamma I - K)(\gamma I + K)^{-1}. \tag{5.24}$$

The set $\hat{\mathcal{A}}_n$ is a set of denominator coefficient vectors for strictly positive real functions. Since the map g_2 was proven to be a diffeomorphism in [32, 30], its inverse map $h_2 := g_2^{-1}$ is well-defined.

Finally, the linear map h_3 is defined in the domain $\hat{\mathcal{A}}_n$ by

$$h_3(\hat{a}) := (\gamma I + K)^{-1}\hat{a}, \quad \hat{a} \in \hat{\mathcal{A}}_n. \tag{5.25}$$

Now, we will state that the map h in (5.21) is actually a map appeared in Theorem 2.3.6, by analyzing the properties of the three maps h_k, $k = 1, 2, 3$.

Proposition 5.2.4. *The maps h_k, $k = 1, 2, 3$, are diffeomorphisms from $\mathcal{S}_n$ to $\mathcal{Q}_+$, from $\mathcal{Q}_+$ to $\hat{\mathcal{A}}_n$ and from $\hat{\mathcal{A}}_n$ to $\mathcal{A}_n$, respectively. Their Jacobians are given by*

$$\begin{aligned}
\frac{\partial h_1}{\partial \sigma}(\sigma) &= T(\sigma), \\
\frac{\partial h_2}{\partial d}(d) &= \left[T(h_2(d))\hat{K} + T(\hat{K}h_2(d))\right]^{-1}, \\
\frac{\partial h_3}{\partial \hat{a}}(\hat{a}) &= (\gamma I + K)^{-1}.
\end{aligned}$$

Proof. The derivatives are obtained via direct calculations from the definition of each map. The diffeomorphisms of h_1 and h_2 are the results in [32, 30]. Thus, we have only to prove that the map h_3 is onto the set $\mathcal{A}_n$, because if this is the case, the diffeomorphism of h_3 follows from the linearity and invertibility of the map.

To prove that h_3 is onto $\mathcal{A}_n$, suppose that a is in $\mathcal{A}_n$, that is,

$$a \in \mathcal{S}_n, \quad \gamma^2 \left|e(\theta)^T a\right|^2 - \left|e(\theta)^T K a\right|^2 > 0, \forall\theta. \tag{5.26}$$

We want to show that $\hat{a} := (\gamma I + K)a$ is in $\hat{\mathcal{A}}_n$, that is,

$$\hat{a} \in \mathcal{S}_n, \quad \min_{\theta\in[-\pi,\pi]} \mathrm{Re}\left[\frac{\hat{e}(\theta)^T\hat{K}\hat{a}}{\hat{e}(\theta)^T\hat{a}}\right] > 0. \tag{5.27}$$

If we define $S(z) := (\bar{z}^T K a)/(\bar{z}^T a)$, then S is analytic in $\mathbb{D}^c$ and takes the absolute value less than γ at each point on $\mathbb{T}$ due to (5.26). Using a bilinear transformation, define

$$F(z) := \frac{\gamma - S(z^{-1})}{\gamma + S(z^{-1})} = \frac{\hat{\bar{z}}^T(\gamma I - K)a}{\hat{\bar{z}}^T(\gamma I + K)a}, = \frac{\hat{\bar{z}}^T \hat{K}\hat{a}}{\hat{\bar{z}}^T \hat{a}},$$

where $\hat{\bar{z}} := [1, z, \cdots, z^n]^T$. Then, F is analytic for $|z| \leq 1$ and takes a positive real value at each point on $\mathbb{T}$. In addition, the sign of the first element in a and that in $\hat{a}$ are the same for the following reason. Because the function F is positive real, the first element of $\hat{a}$ and that of $\hat{K}\hat{a}$ must have the same sign. This means that the first element of $\hat{a}$ and that of $\hat{a} + \hat{K}\hat{a} = (\gamma I + K)a + (\gamma I - K)a = 2\gamma a$ have the same sign, and thus $\hat{a} \in \mathcal{S}_n$. Therefore, we have established $\hat{a} \in \hat{\mathcal{A}}_n$, and hence the surjectivity of h_3. □

Due to Proposition 5.2.4, as well as the chain rule, we have arrived at the following assertion.

Theorem 5.2.5. *The map h in (5.21) is a diffeomorphism from $\mathcal{A}$ to $\mathcal{S}_n$. Its derivative is given by*

$$\frac{\partial h}{\partial \sigma}(\sigma) = \frac{\partial h_3}{\partial \hat{a}}((h_2 \circ h_1)(\sigma)) \cdot \frac{\partial h_2}{\partial d}(h_1(\sigma)) \cdot \frac{\partial h_1}{\partial \sigma}(\sigma). \tag{5.28}$$

Remark 5.2.6. *One may think that it is beneficial to use d, instead of σ, as optimization variables, since the set $\mathcal{Q}_+$ is convex. However, our numerical expe-rience tells us that σ-parameterization often gives better solutions than d-parameterization.*

Solving the Nonlinear Least-Squares Problem

In order to solve the sensitivity shaping problem, we need a reliable and numerically robust algorithm to solve the optimization problem in (5.19). By "solving" we mean finding a local minimizer or an approximation of a local infimizer. The nonlinear least squares approach enables the use of the standard Gauss-Newton and Levenberg-Marquardt algorithms, see for instance [84], provided they are modified, for instance as described in [82]. There is also an implementation available in [11].

A Design Example

Now we are ready to deal with the benchmark design problem. The flowchart of our controller design procedure is depicted in Figure 5.6. In the flowchart, "NLSsolver" is the nonlinear least-squares optimization solver which realizes the method just described. The NLSsolver can be regarded as a *blackbox* whose inputs are a plant transfer function P and design parameters, and whose output is a sensitivity function S, and thus, a controller $C = (1-S)/PS$. As for inputs, the plant is given and fixed, whereas the design parameters are tunable for performance improvements. We have developed a user-friendly interface [11] that realizes this flowchart.

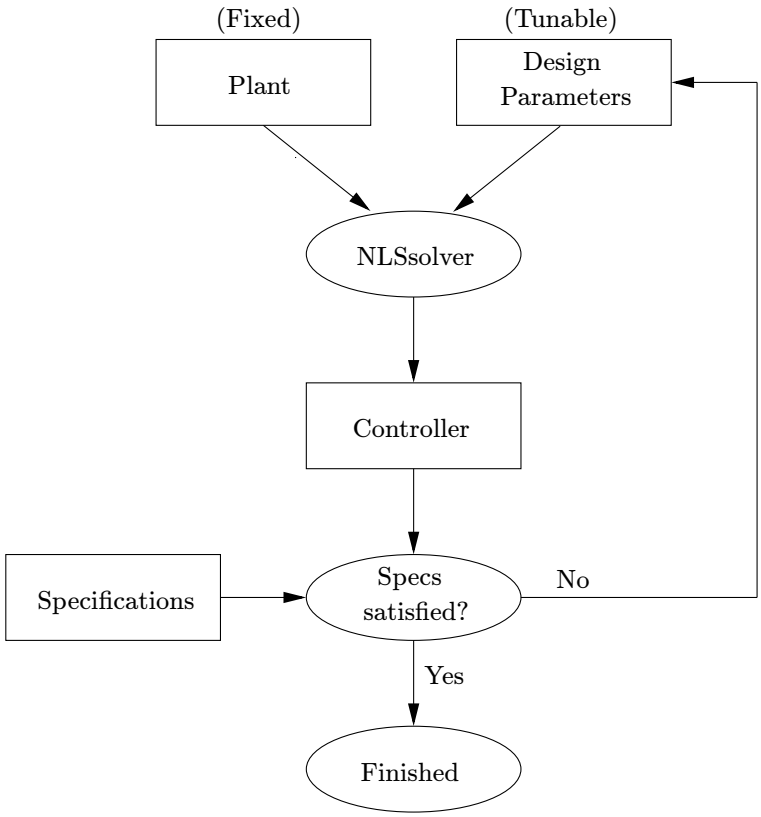

Figure 5.6: Flowchart of the design procedure

Next we shall consider a benchmark problem, the "Flexible Beam" design problem, which is taken from a standard textbook on conventional $\mathcal{H}_\infty$ control, namely *Feedback Control Theory* [41, Sections 10.3 & 12.4].

Example 5.2.7 (Flexible Beam). Consider the standard feedback system depicted in Figure 1.1. The plant P is given by

$$P(s) = \frac{-6.4750s^2 + 4.0302s + 175.7700}{s(5s^3 + 3.5682s^2 + 139.5021s + 0.0929)},$$

which has one unstable pole at $s = 0$, one non-minimum phase zero at $s = 5.5308$, and one zero at infinity of multiplicity two. Our goal in this problem is to design a strictly proper controller C which satisfies, for a step reference r,

- the settling time is less than 8 seconds,
- the overshoot is less than 10 %, and

- the control input fulfills $|u(t)| \leq 0.5$ for all t.

Remark 5.2.8. *In our controller design, the specifications must be stated in the frequency domain. Time domain specifications will be translated into approximate frequency domain ones, and after controller design based on the frequency domain specifications, we will check if the original time domain specifications are indeed satisfied.*

In [41], the first two requirements in the time domain have been translated into a requirement in the frequency domain as a desired sensitivity function:

$$S_d(s) := \frac{s(s+1.2)}{s^2+1.2s+1}.$$

We also aim at designing a sensitivity function similar to S_d, with extra consideration of control input constraint.

Remark 5.2.9. *A controller for this problem was designed in [9, 10] by a manual tuning of our optimization parameter σ. The design procedure here is more systematic than the design [9, 10]. In fact, the manual tuning of σ was quite heuristic when it comes to the input constraint.*

Using S_d, we extract our desired frequency response at a finite number of frequencies. We take 100 discrete points in the frequency $\left[10^{-3}, 10^{3}\right]$ (rad/sec), equally distanced in the logarithmic scale, as

$$\omega := \{\omega_k\}_{k=1}^{100}.$$

With these points, we set our desired frequency response (θ, s) in the discrete-time setting as

$$\begin{aligned} \theta &:= \left\{\theta_k :\ e^{i\theta_k} = \frac{1+i\omega_k}{1-i\omega_k},\ \omega_k \in \omega\right\}, \\ s &:= \{s_k := S_d(i\omega_k),\ \omega_k \in \omega\}. \end{aligned}$$

Since we have initially no information on the frequency emphasis, the weights are set as

$$w := \{w_k := 1,\ k = 1, \ldots, 100\},$$

and the uniform upper bound of the sensitivity gain is chosen as

$$\gamma := 1.5.$$

We do not use any additional interpolation condition in this problem. From the gain plot of S_d, we would like to have a peak gain around 1 rad/sec. Therefore, we always set the initial point for optimization to a σ in $\mathcal{S}_n$ that has its roots at $\pm 0.95i$, which corresponds to a solution having the peak of its frequency response close to 1 (rad/sec) in the continuous-time setting.

With the initial selection of design parameters, NLSsolver outputs a controller and a sensitivity function as

$$C_0(s) := \frac{75.66s^3 + 54s^2 + 2111s + 1.406}{s^4 + 10.1s^3 + 452.4s^2 + 2755s + 3238}, \tag{5.29}$$

$$S_0(s) := \frac{s^4 + 5.193s^3 + 426.9s^2 + 659.7s}{s^4 + 5.193s^3 + 426.9s^2 + 561.7s + 541.9}. \tag{5.30}$$

Several frequency and time responses are plotted in Figures 5.7 and 5.8.

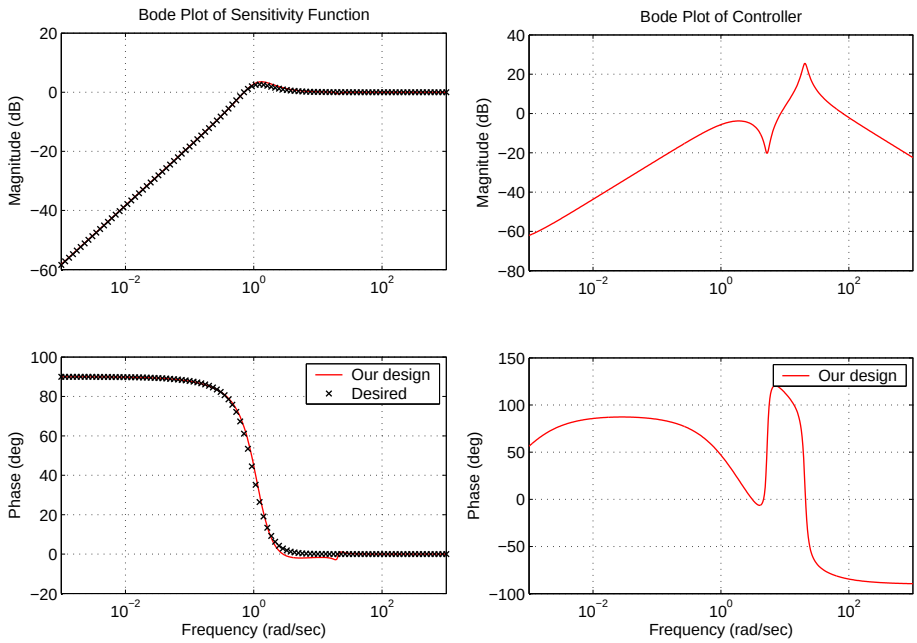

Figure 5.7: Bode plots of the sensitivity function and controller for the initial design.

The uppermost plot in Figure 5.7 shows the Bode plot of S_0 with the desired frequency response (θ, s). As can be seen, the NLSsolver indeed generates S_0 approximating the given data (θ, s).

Now, we check the original time domain specifications. Figure 5.8 shows the step response and the input signal. Although the step response meets the specification, the input signal is too large to fulfill the specification $|u(t)| \leq 0.5$. Therefore, we need to update some of our design parameters, and redesign a controller.

To see the cause of the large input signal, we consider the Bode plot of the controller C_0 in Figure 5.7. From the figure, we see that there is a sharp gain peak around 20 rad/sec. In fact, this frequency coincides with the frequency of the input

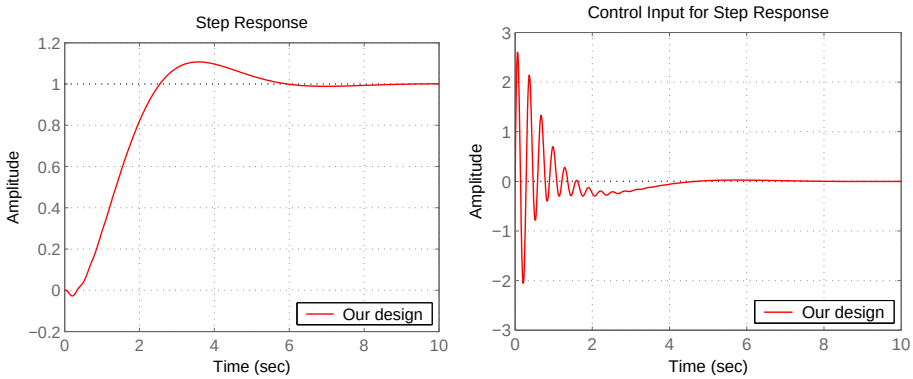

Figure 5.8: The step response and the corresponding control signal for the initial design.

oscillation. Therefore, one natural way to suppress the input is to lower the gain peak of C.

Now, we update the design parameters. Since $C = (1 - S)/PS$, we need to make S close to one to decrease the gain of C. Desired frequency response s_k is almost one around frequency 20 rad/sec, and thus, we increase the weight w_k around the frequency to fit S closer to s_k. Note that we do not change other design parameters in this example. After some trial and error, we have chosen weights w as in Figure 5.9, that results in the following controller and sensitivity function:

$$C(s) = \frac{2.706s^3 + 1.931s^2 + 75.51s + 0.05028}{s^4 + 7.698s^3 + 33.59s^2 + 126.8s + 143}, \tag{5.31}$$

$$S(s) = \frac{s^4 + 2.789s^3 + 19.9s^2 + 29.13s}{s^4 + 2.789s^3 + 19.9s^2 + 25.62s + 19.38}. \tag{5.32}$$

The resulting Bode plots and response signals are shown in Figures 5.10 and 5.11, together with response signals from the design of [41].

The figures show that the sharp peak disappeared in the gain of C, which has been done at the price of degradation of sensitivity fitting, and that the original time domain specifications are indeed satisfied. Also, one can see that we have obtained a similar performance to that in [41]. We stress that the controller (5.31) is half the degree of the one obtained in [41].

Remark 5.2.10. *The role of weights w is similar to that of weighting functions W in conventional $\mathcal{H}_\infty$ control, in that they emphasize suppression of gain at some frequency regions. However, the weights here have two advantages over weighting functions. One is that w do not assume rationality, and rather arbitrary, while the weighting functions W must be rational in most cases. This will increase the*

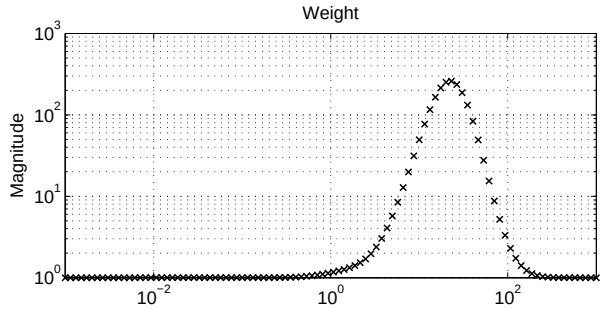

Figure 5.9: Weight w.

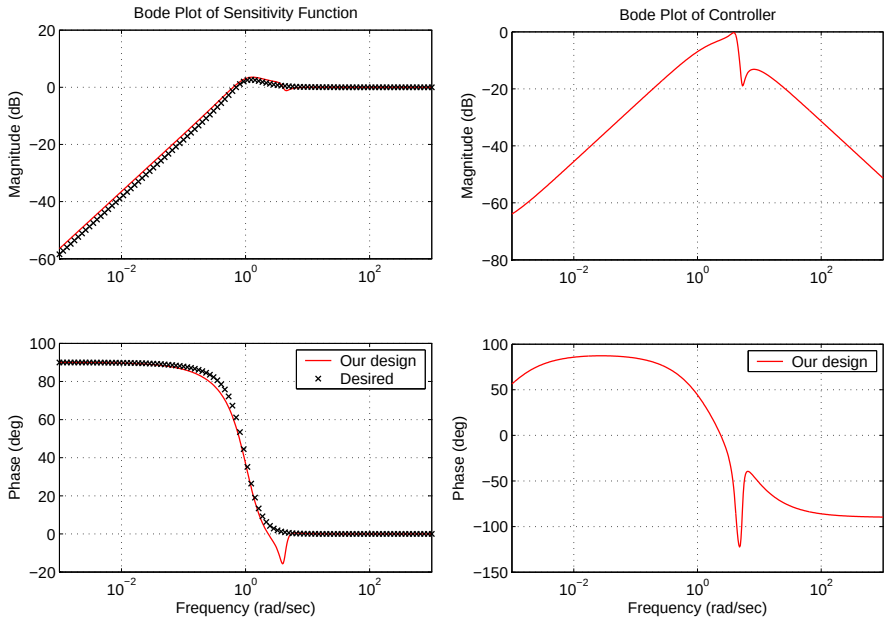

Figure 5.10: Bode plots of the sensitivity function and the controller for the final design.

flexibility of the design. The other is that w *do not increase* $\deg C$*; it just changes the cost function to be minimized. On the other hand,* W *typically increases* $\deg C$ *by* $\deg W$*.*

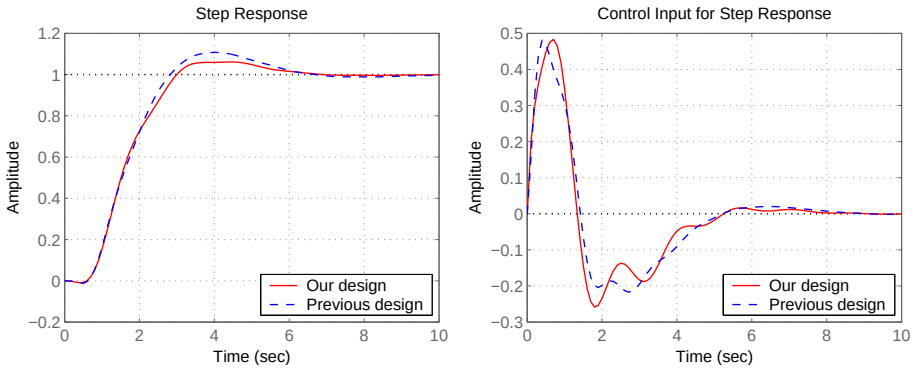

Figure 5.11: The step response and the corresponding control signal for the final design.

5.3 AR Estimation With Prefiltering

Here we will study how one can affect the model error distribution in the frequency domain in the case of AR estimation using prefiltering. It turns out that under certain conditions prefiltering in the studied paradigm is equivalent to weighting in the frequency domain maximum likelihood, FDML, estimation in the AR case. The idea in this section should also be applicable to the ARMA estimation case using the full agility of the theory in Chapter 3. This part is based on the results presented in [13].

Throughout this section, we shall consider the case when the true process is more complex than the model to be identified. In particular, attention is given to models which are good in a certain frequency region.

Consider the system setup in Figure 5.12. Here H_m represents all unmodeled characteristics of the true random process $y(t)$. Given a time series of measurements of $y(t)$ we want to estimate a good model of $H(z)$ provided that we know that most of its energy is in the lower frequency domain.

One standard approach to affect the bias distribution in prediction error methods is to prefilter the prediction errors,

$$\epsilon_f(t,\theta) := L_f(q)\epsilon(t,\theta) = [H(q,\theta)]^{-1} L_f(q)y(t),$$

where $L_f(q)$ is a fixed prefilter and $H(q,\theta)$ is the noise filter parameters set, see for instance [73, 72]. The corresponding estimate equals

$$\hat{\theta}_{PEM(f)} := \arg\min_{\theta} \frac{1}{N}\sum_{t=1}^{N} \epsilon_f^2(t,\theta).$$

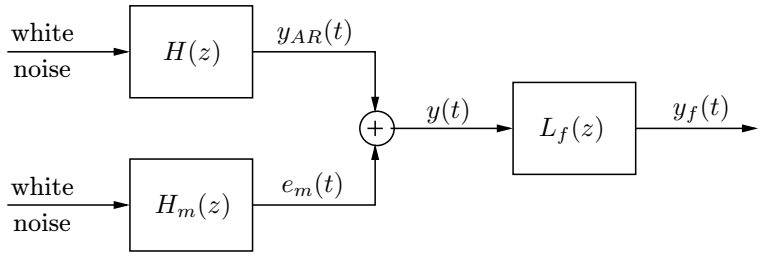

Figure 5.12: The system setup for prefiltering.

Now, $L_f(q)y(t) = L_f(q)H_0(q)e(t)$, and the effect of the prefilter is to change the true noise filter to $L_f(q)H_0(q)$. Therefore, without any prior knowledge of the process, such as the zero locations, prefiltering cannot improve a PEM estimate since the estimator simply recovers the prefilter. For the case of undermodeling, a prefilter can even disprove the solution if it recovers an arbitrary part of a prefilter rather than the process.

Filtering is often easier to view in the frequency domain. The ML method can also be formulated using frequency domain data, see for instance [73, 89]. Let

$$Y_N(\omega) := \frac{1}{\sqrt{N}} \sum_{t=1}^{N} y(t) e^{-i\omega t},$$

be the discrete Fourier transform of the sequence $\{y(t),\ t = 1 \dots N\}$. By also including the noise variance σ^2 as a parameter, the FDML cost function equals, see [73, p. 230],

$$V_{FDML}(\theta, \sigma^2) := \frac{1}{2\pi} \int_{-\pi}^{\pi} \frac{|Y_N(\omega)|^2}{\sigma^2 |H(e^{i\omega}, \theta)|^2} d\omega + \frac{1}{2\pi} \int_{-\pi}^{\pi} \log \left(\sigma^2 |H(e^{i\omega}, \theta)|^2 \right) d\omega. \tag{5.33}$$

This is the discrete time version of the classical Whittle likelihood and can be interpreted as the likelihood of the asymptotic distribution of the periodogram, see [20, p. 347]. Here we have used an integral formulation, instead of the more common summation over the frequencies $\omega_k = 2\pi k/N$, $k = 1 \dots N$:

$$\tilde{V}_{FDML}(\theta, \sigma^2) := \frac{1}{2\pi N} \sum_{k=1}^{N} \frac{|Y_N(\omega_k)|^2}{\sigma^2 |H(e^{i\omega_k}, \theta)|^2} + \frac{1}{2\pi N} \sum_{k=1}^{N} \log \left(\sigma^2 |H(e^{i\omega_k}, \theta)|^2 \right). \tag{5.34}$$

The difference between (5.33) and (5.34) is negligible for large N. In [88] this frequency domain approach is further refined by also taking the initial state of the noise filter into account to obtain a leakage free spectral representation.

To obtain a frequency weighted estimate, FDML(f), we can assign a non-negative weight, $W(\omega) \geq 0$, to each frequency in (5.33):

$$V_{FDML(f)}(\theta, \sigma^2) := \frac{1}{2\pi} \int_{-\pi}^{\pi} W(\omega) \frac{|Y_N(\omega)|^2}{\sigma^2 |H(e^{i\omega}, \theta)|^2} d\omega + \frac{1}{2\pi} \int_{-\pi}^{\pi} W(\omega) \log \left(\sigma^2 |H(e^{i\omega}, \theta)|^2 \right) d\omega.$$

One possible frequency weighting is to give a constant weight to all frequencies in a desired region and zero weight to all others. In the discrete case the cost function then is of the form

$$\tilde{V}_{FDML(f)}(\theta, \sigma^2) = \frac{1}{M} \sum_{\omega_k \in \Omega} \frac{|Y_N(\omega_k)|^2}{\sigma^2 |H(e^{i\omega_k}, \theta)|^2} + \frac{1}{M} \sum_{\omega_k \in \Omega} \log \left(\sigma^2 |H(e^{i\omega_k}, \theta)|^2 \right), \tag{5.35}$$

where Ω is a specified set of M important frequencies.

It is possible to perform analytic minimization of the cost-functions with respect to σ^2 as shown in [73, p. 230]. For $V_{FDML(f)}(\theta, \sigma^2)$ this leads to the estimates

$$\hat{\theta}_{FDML(f)} := \arg\min_{\theta} \left[\bar{W} \log \left(\frac{1}{2\pi} \int_{-\pi}^{\pi} W(\omega) \frac{|Y_N(\omega)|^2}{|H(e^{i\omega}, \theta)|^2} d\omega \right) + \frac{1}{2\pi} \int_{-\pi}^{\pi} W(\omega) \log |H(e^{i\omega}, \theta)|^2 d\omega \right], \tag{5.36}$$

and

$$\hat{\sigma}^2_{FDML(f)} := \frac{1}{\bar{W}} \frac{1}{2\pi} \int_{-\pi}^{\pi} W(\omega) \frac{|Y_N(\omega)|^2}{|H(e^{i\omega}, \hat{\theta}_{FDML})|^2} d\omega,$$

where $\bar{W} := \frac{1}{2\pi} \int_{-\pi}^{\pi} W(\omega) d\omega$. For any monic, stable and inversely stable transfer function $H(e^{i\omega}, \theta)$ we have

$$\frac{1}{2\pi} \int_{-\pi}^{\pi} \log |H(e^{i\omega}, \theta)|^2 d\omega = 0, \tag{5.37}$$

and hence the second term in (5.36) disappears if $W(\omega) = 1$ (no weighting). This term is, however, most important for the weighted case to regularize the optimization problem, and can be viewed as a kind of barrier function in constrained optimization.

Next we will introduce prefiltering in our framework. Given a prefilter $L_f(z)$ with the corresponding spectral density $\Psi = L_f(z) L_f^*(z)$, consider the filtered signal

$$y_f(t) := L_f(z) y(t).$$

Compute the generalized filtered covariances and generalized filtered cepstral coefficients. Now, solving the Kullback-Leibler Approximation Problem 3.1.1 with

$\Psi = L_f(z)L_f^*(z)$ and the filtered data (r, c) we will get a solution of the form $\Phi = \Psi\hat{P}\hat{Q}^{-1}$. Here Ψ comes from the prefilter and the spectral factor w of $\hat{P}\hat{Q}^{-1}$ will be a model of the underlying ARMA model.

In the AR case, this yields the method:

1. Assume that the system can be well approximated by an AR model in the frequency region of interest.
2. Choose a prefilter $L_f(z)$ to stress the important frequency range, that is $|L_f(e^{i\omega})|$ is large for frequencies of interest, and small otherwise. Typically, $L_f(z)$ is a low-pass filter to remove high frequency disturbances, or a band-pass filter which focuses the fit to a certain frequency band.
3. Pick a set of basis functions $G \in \mathcal{G}$. Prefilter the data and feed it through the filter-bank corresponding to the basis functions. Estimate the state covariance matrix Σ and compute a feasible estimate thereof, see [53, 5].
4. Solve the dual optimization problem $(\mathcal{D})$ using Algorithm 1 and compute the corresponding spectral factor $w(z)$.

We note that the proposed estimator, which we shall call the prefiltered covariance extension, CE(f), estimator, is very closely related to the FDML(f) estimator. In fact, we have the following proposition:

Proposition 5.3.1. *Consider the AR estimation problem with G as the standard basis in* (2.12)*. Suppose that $W(\omega) = |L_f(e^{i\omega})|^2$. Then the CE(f) and FDML(f) estimates are the same.*

Proof. In the case considered $m = 0$, $Q(z) = a(z)a^*(z) = [\sigma^2|H(z,\theta)|^2]^{-1}$, and $\Psi(z) = L_f(z)L_z^*(z)$ in the dual function (3.3). We then have, with $\mathbb{J}(Q)$ as a short-hand notation for $\mathbb{J}(1, Q)$ defined in (3.3),

$$\begin{aligned}\mathbb{J}(Q) &= \langle \Psi, \log Q\rangle - \langle Q, R\rangle + \text{const},\\ &= -\left\langle L_f L_f^*, \log \sigma^2|H(z,\theta)|^2\right\rangle \\ &\qquad - \left\langle \frac{1}{\sigma^2|H(z,\theta)|^2}, r_0 + \sum_{k=0}^{n} r_k(z^k + z^{-k}) + S(z)\right\rangle + \text{const},\end{aligned}$$

where $S(z)$ is *any* function of z^k for all $k > n$ and $k < -n$. Now take

$$\hat{\Phi}(\omega) := \frac{1}{|L_f(e^{i\omega})|^2}\left[r_0 + \sum_{k=0}^{N-1} r_k(e^{i\omega k} + e^{-i\omega k})\right].$$

Then $\hat{\Phi}(\omega) = |L_f(z)|^2|Y_N(\omega)|^2$ by [90, p. 106f] and hence, since we assume that $W(\omega) = |L_f(e^{i\omega})|^2$

$$\begin{aligned}\mathbb{J}(Q) &= \left\langle |L_f|^2, \log \sigma^2|H(z,\theta)|^2\right\rangle - \left\langle \frac{1}{\sigma^2|H(z,\theta)|^2}, |L_f|^2|Y_N(\omega)|^2\right\rangle + \text{const},\\ &= -V_{FDML(f)} + \text{const}.\end{aligned}$$

Therefore the FDML(f) and CE(f) estimates will be the same. □

Note that Proposition 5.3.1 only applies to the AR case. It can be seen as a generalization of the fact that the maximum entropy and the ML estimates agree for the AR case without prefiltering. Also, this only applies to the integral formulation. In fact, with the finite sum formulation of the FDML(f) functional, there is no guarantee that there exist an interior point minimizer. The following counterexample illustrates this.

Counterexample . *Consider the first order example where* $Q(z) = q_0 + \frac{q_1}{2}(z+z^{-1})$. *Let* Ω *consists of only two points* 0 *and* $\pi/2$ *and let* $|Y(0)|^2 = c$ *and* $|Y(\pi/2)|^2 = 1$. *The finite functional then is*

$$\tilde{W}_{FDML(f)}(Q) = \frac{1}{2}((c+1)q_0 + cq_1 - \log(q_0+q_1) - \log(q_0)).$$

The stationarity conditions are

$$\begin{aligned} c + 1 - \frac{1}{q_0+q_1} - \frac{1}{q_0} &= 0, \\ c - \frac{1}{q_0+q_1} &= 0, \end{aligned}$$

which yield

$$\begin{aligned} q_0 &= 1, \\ q_1 &= \frac{1}{c} - 1. \end{aligned}$$

For $q_0 = 1$ *the set of positive pseudo polynomials is given by*

$$\left\{ Q(z) = 1 + \frac{q_1}{2}(z+z^{-1}) : -1 < q_1 < 1 \right\}.$$

Therefore we have that only for $c > 1/2$ *there exist an interior minimizer.*

For the integral formulation this will never happen, see Theorem 3.1.3.

A key advantage of this type of prefiltering is the smoothness with respect to the choice of prefilter. Clearly, the filtered sample covariances depend smoothly on the choice of prefilter. Moreover, appealing to the well-posedness stated in Theorem 3.2.1, the smoothness carries over to the model estimate. Thus, at least a minor mismatch between L_f and H_m is acceptable.

Another, previously proposed, method for incorporating prior knowledge in the covariances extension framework is the THREE algorithm in [26]. There, a non-default choice of basis functions, put into a filter-bank, filters the data and enables high-resolution estimation. In fact, the method proposed in this section can be fully combined with the THREE approach since the filter-bank framework is already incorporated in our approach. For clarity of presentation we will not elaborate on the choice of basis functions in the filter-bank in the following example which illustrates the proposed procedure as well as compares it to the FDML(f) estimator as well as the PEM estimator. This example is taken from [13].

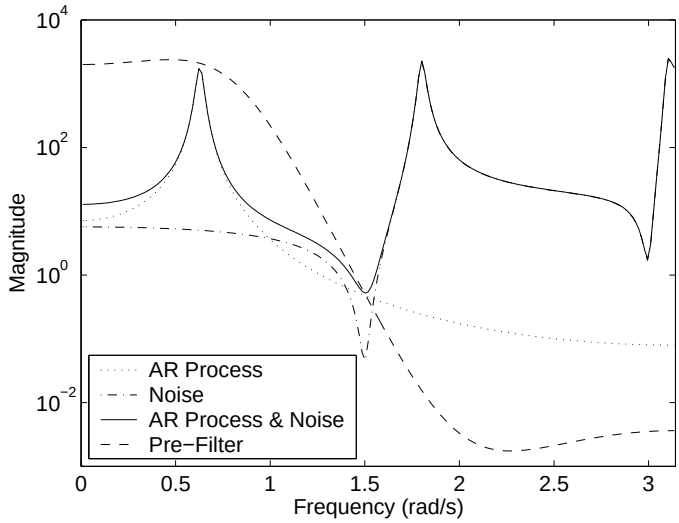

Figure 5.13: The spectrum of the AR(2) process and the noise together with the spectral density of the prefilter.

Example 5.3.2 (AR(2) in colored noise). This example amounts to identifying an AR(2) process in colored noise. The signal $y(t)$ is generated as in Figure 5.12. Let the roots of $a(z)$ be in $0.98e^{\pm 0.2\pi i}$ and let the driving noise be Gaussian white with unit variance. Let the added noise be colored by driving Gaussian white noise with unit variance through the shaping filter

$$H_m(z) = \frac{(z-0.98e^{1.5i})(z-0.98e^{3i})}{(z-0.98e^{1.8i})(z-0.98e^{3.1i})}\frac{(z-0.98e^{-1.5i})(z-0.98e^{-3i})}{(z-0.98e^{-1.8i})(z-0.98e^{-3.1i})}.$$

This noise has most of its power in the higher frequency region. The signal-to-noise ratio of the signal $y(t)$ will then be high around the peak located at the frequency 0.2π while it will be very small at high frequencies. Assuming that we possess this *a priori* knowledge of the process, we use, for instance, the low-pass prefilter

$$L_f(z) = \frac{(z-0.6e^{2i})^3(z-0.6e^{-2i})^3}{(z-0.6e^{0.8i})^3(z-0.6e^{-0.8i})^3}, \tag{5.38}$$

in the method. The spectral representation of $y_{AR}(t)$, $e_m(t)$, and the prefilter are plotted in Figure 5.13.

Now we apply the proposed CE(f) method. We compare our estimator to the PEM estimator of [73] and the FDML(f) estimator. In the latter we take

$$\Omega = \left\{ \omega_k = \frac{2\pi k}{N} : \omega_k \in [0.1\pi, 0.3\pi] \right\}.$$

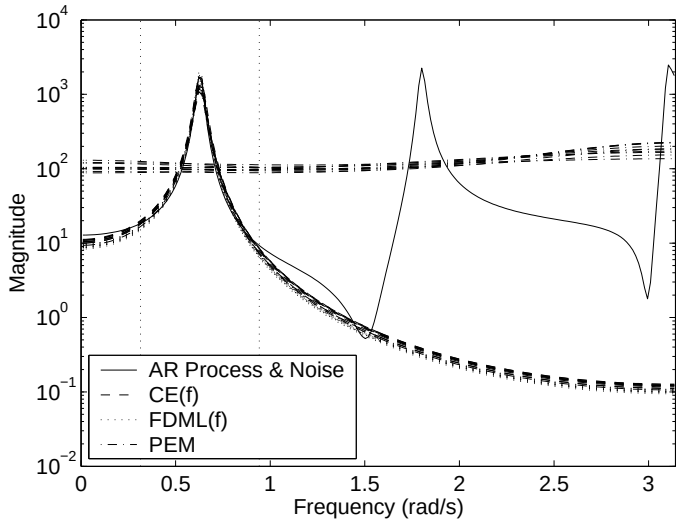

Figure 5.14: The spectral densities for ten estimations of Example 5.3.2. The proposed estimator CE(f) and the FDML(f) estimator seem to be robust against noise at high frequency whereas the PEM is not.

In a sense this corresponds to an ideal band-pass filter and thus differs significantly from $L_f(z)$ in (5.38).

In Figure 5.14 the frequency responses for ten different realizations of the noises $\{e(t)\}$ and $\{e_m(t)\}$ are given. We clearly see that the FDML(f) and CE(f) estimators are better in the lower frequency region than the PEM estimator. However, this is of course at the expense of a worse match for high frequencies. That the FDML(f) and CE(f) estimators give approximately the same result despite a large difference in prefiltering/weighting indicate a low sensitivity with respect to the choice of prefilter/weight.

To make a more precise comparison of the estimators we use some different error measures, both in the time and frequency domains. In the frequency domain we compare the magnitude of the frequency response. Due to the logarithmic behavior, the comparison is performed in decibels. We use two measures: both the absolute deviation over the whole spectrum and the absolute deviation in the frequency range $[0.1\pi,\ 0.3\pi]$ that we are particularly interested in (indicated by vertical dotted line in Figure 5.14). As for a time domain error measure, we use the variance of the prediction errors (PE). In order to estimate the PE for a particular model and the nominal system we generate 1000 data points and compute the corresponding PE. Also, in the time domain, we use an alternative measure that focuses on the dynamics close to the spectral peak of the AR(2) model. Here we have chosen to drive the sample prediction errors through the same filter as used as prefilter,

Table 5.3: The estimated errors in the time and frequency domains.

	Spectral Error(dB)	Zoomed Spectral Error(dB)	Prediction Error	Filtered Prediction Error
PEM*	18.83	11.33	347.1	9703
Std. dev.	0.38	0.38	21.8	1133
FDMLf	24.76	1.07	834.5	907
Std. dev.	0.38	0.11	80.1	44
CEf	23.67	1.71	830.3	920
Std. dev.	0.40	0.46	79.8	47

(5.38). This is fair since the prefilter is a design choice. The error estimates based on 100 Monte Carlo simulations are shown in Table 5.3.

We note that the PEM estimator is better when averaging over the whole frequency region while the CE(f) and the FDML(f) estimators are better in the desired frequency region. The latter two have comparable performance despite the fact that the FDML(f) uses an ideal band-pass filter whereas CE(f) uses H_m. This indicates the insensitivity to the choice of prefilter. This is an expected result of the smoothness incorporated in our approach as well as of the fact that we only use the prefilter as weighting of different frequency regions. The result also carries over to the prediction error (PE) and the filtered PE.

5.4 ARMA Estimation

Our final example will be estimation of an ARMA model from data generated from a model in the same class. As seen from Example 1.0.3 in the introduction chapter, simultaneously matching a direct estimate of the covariances and cepstral coefficients is not expected to yield a statistically efficient estimate. However, we observed that the estimator seemed to be efficient, or close to efficient, when the MA zero was close to origin. Then, the covariances and the cepstral coefficients were computed for data that was close to white noise. Here we will explore this by trying to prefilter the data to make it close to white noise. This is sometimes called *whitening*[4].

Consider Figure 5.15. Here $\{x_t\}_{t=1}^N$ is the measured data in the same way as in Example 1.0.3. Assume that we have a preliminary estimate of $w(z)$, say $\tilde{w}(z)$. By choosing $\psi(z) = \tilde{w}^{-1}(z)$ the filtered data $\{y_t\}_{t=1}^N$ will be closer to white noise if the preliminary estimate were decent. Now, estimate the biased covariances and cepstral coefficients of the filtered data. Given these, we formulate the Kullback-Leibler

[4]In particular one can interpret the Prediction Error Method, PEM, as whitening the data. By feeding the data reversely through the current model estimate the prediction errors are obtained. The PEM estimate is the minimizer of the sum of squared prediction errors.

approximation problem, see Problem 3.1.1, with $\Psi = \psi\psi^*$. From Theorem 3.1.2 we know that there is a unique solution of the form $\Phi = \Psi\hat{P}\hat{Q}^{-1}$. Since the prefilter contributes multiplicatively with Ψ, we can ignore that factor and our estimate of the ARMA model will be the stable, miniphase spectral factor of $\hat{P}\hat{Q}^{-1}$.

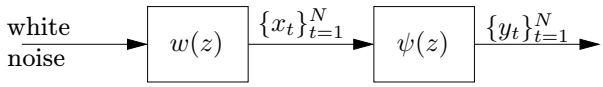

Figure 5.15: The prefiltering for ARMA estimation.

We can iteratively change the prefilter ψ by using the new estimates of the model. In Procedure 1 we propose one possible such scheme, which uses Algorithm 3. Being iterative, it is not clear whether the procedure converges. However, we will not study this issue in this thesis.

Procedure 1. *Prefiltered Cepstral-Covariance Matching, CCM(f)*

Set $\hat{a} \leftarrow 1$ and $\hat{\sigma} \leftarrow 1$
for $k = 1, \ldots, 5$
  Set $\varepsilon \leftarrow c_5(k)$
  Set $\Psi \leftarrow \hat{a}\hat{a}^*/(\hat{\sigma}\hat{\sigma}^*)$
  Estimate r and c from original data
  Determine a and σ using Algorithm 3
end for

No matter how interesting, it is beyond the scope of this thesis to include an exhaustive statistical analysis of proposed CCM(f) estimator. However, we shall prove one fairly immediate result, which ought to be a key result in any statistical analysis of the estimator.

Theorem 5.4.1. *Let $\Psi = (\tau\tau^*)/(\sigma\sigma^*) \in \mathcal{Q}_+$ and $G \in \mathcal{G}$ such that $\det(I - Az) = \tau(z)$ be given. Also let a time series of length N generated from an ARMA process with parameters Θ_0 be given. Assume that some estimation procedure estimates the generalized prefiltered covariances and cepstral coefficients Ξ such that*

$$\Xi \text{ is } AN\left(\Xi_0, N^{-1}U\right), \quad U = \begin{pmatrix} U_1 & U_2 \\ U_2^T & U_3 \end{pmatrix},$$

where Ξ_0 are the parameters corresponding to Θ_0. Then

$$\Theta \text{ is } AN(\Theta_0, N^{-1}W),$$

where W is the covariance matrix

$$W = \left[\frac{\partial F}{\partial \Theta}\right]^{-1} \begin{pmatrix} D & 0 \\ 0 & I \end{pmatrix} \begin{pmatrix} U_1 & U_2 \\ U_2^T & U_3 \end{pmatrix} \begin{pmatrix} D & 0 \\ 0 & I \end{pmatrix}^T \left[\frac{\partial F}{\partial \Theta}\right]^{-T} \Bigg|_{\Theta_0},$$

where F is the map defined in (3.12) *and D is given by*

$$D := r_0^{-1} \begin{bmatrix} -r_1/r_0 & 1 & 0 & \cdots & 0 \\ -r_2/r_0 & 0 & 1 & & 0 \\ \vdots & \vdots & & \ddots & \\ -r_n/r_0 & 0 & 0 & & 1 \end{bmatrix}. \tag{5.39}$$

Proof. The Jacobian of the map mapping the generalized prefiltered covariances to the normalized ditto, that is, $\begin{bmatrix} r_0 & r_1 & \dots & r_n \end{bmatrix}^T \mapsto \begin{bmatrix} r_1/r_0 & r_2/r_0 & \dots & r_n/r_0 \end{bmatrix}^T$, is given by (5.39). By Theorem 3.2.1 the map F has an everywhere invertible Jacobian. Since F is a diffeomorphism and the diagonal elements of U are nonzero, so are the diagonal elements of W. Hence the claim follows by Proposition 6.4.3 in [20]. □

The theorem tells us that the CCM(f) estimates inherit the statistical properties from the estimates of the covariances and cepstral coefficients. The underlying reason is of course the smoothness of the map F discussed in detail in Section 3.2. Thus, if we can estimate the covariances and the cepstral coefficients statistically efficiently, then we automatically have a statistically efficient estimate of the ARMA model. Also, consistently estimated cepstral coefficients and covariences yield consistent estimates of ARMA models.

As noted in Section 2.6, the joint distribution of generalized prefiltered covariances and cepstral coefficients is unknown, also in the case of unfiltered coefficients with the standard basis. Yet, Theorem 5.4.1 serves as a conceptual tool in understanding the following example, where we will study the idea presented above for ARMA(n,n) models. This is believed to be a generic example, in comparison to the overly simplified example in the introduction with a simple real zero. However, we acknowledge that higher order models are more useful in practice – not the least in term of numerical issues.

Example 5.4.2 (ARMA(n,n)). Consider Figure 5.15. First take the true model to have poles in $0.5e^{\pm 2i}$ and zeros in $0.98e^{\pm i}$. The corresponding density is plotted in Figure 5.16. Note that the zero close to the unit circle creates a frequency region with low magnitude of the spectrum. This makes the identification harder. Also compare to the case in Example 1.0.3 when the simple zero tended towards the circle.

Given a time series consisting of the measurements x_t with sample lengths $N = 200, 400, \dots, 12800$ we try to identify the filter. We compare three estimators. As reference we use the Maximum-Likelihood estimator implemented in `armax` in [71]. The estimator will be denote ML. The second estimator is the Cepstral-Covariance Matching estimator without prefiltering and with the standard basis, as presented in [25, 45], though computed by an implementation of Algorithm 3. More precisely, we use the standard biased sample covariances in (2.40). For estimation of the cepstral coefficients we first estimate long AR models of orders $L = 10, 15, \dots, 40$ for the different sample sizes, respectively. The corresponding

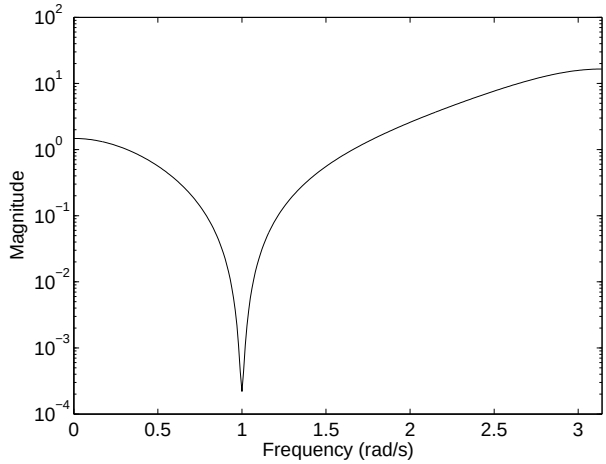

Figure 5.16: The spectrum of the AR(2,2) process.

ARMA models are computed using an implementation of Algorithm 3 with the default prefilter $\Psi \equiv 1$. We will call the estimator CCM. The last estimator is based on Procedure 1 where we recursively estimate new, generalized, covariances and cepstral coefficients. To estimate the generalized prefiltered covariances we apply the input-to-state framework and estimate the state-covariance. We use the least-squares approach suggested in [53, p. 34] to estimate a feasible estimate of the interpolation data matrix W; see also the discussion in [5]. For the generalized cepstrum, we again estimate long AR models of orders $L = 10, 15, \ldots, 40$. From the AR model we can directly compute the generalized prefiltered cepstral coefficients.

First we make a statistical comparison of the parameter estimate. Since the variance is asymptotical decoupled from estimating the other ARMA parameters, see for instance [90], we will only compare the parameters $\begin{bmatrix}\sigma_1 & \sigma_2 & a_1 & a_2\end{bmatrix}$. In Table 5.4 the estimated means and variances of the parameter estimates for the different methods using a Monte Carlo simulation with 500 runs is given. We note that all methods seem to be unbiased. Moreover, the variance for the CCM(f) estimator is approximately the same as for the ML estimator, which in turn is approximately efficient. Meanwhile, the unfiltered estimator, CCM, does not seem to be efficient. Thus the example indicates that the prefiltering seem to make the CCM method approximately asymptotically efficient.

Another way of comparing the estimators is to compute some error measure of each estimate and then by the Monte Carlo simulation estimate what the mean error is. In Figures 5.17 and 5.18 the estimated prediction error and Kullback-Leibler discrepancy relative to the true model are plotted as a function of the sample size. The prediction errors are computed by feeding white Gaussian noise with 100 samples through the filters and then estimating the prediction errors. The

Table 5.4: The mean and standard deviation in the estimation of the ARMA parameters for $N = 12800$.

	True	ML		CCM(f)		CCM	
		Mean	Std.dev.	Mean	Std.dev.	Mean	Std.dev.
a_1	0.000	−0.000	0.010	0.000	0.009	0.009	0.013
a_2	−0.250	−0.249	0.010	−0.250	0.010	−0.244	0.014
σ_1	−1.070	−1.067	0.003	−1.065	0.004	−1.008	0.007
σ_2	0.980	0.974	0.006	0.972	0.006	0.884	0.007

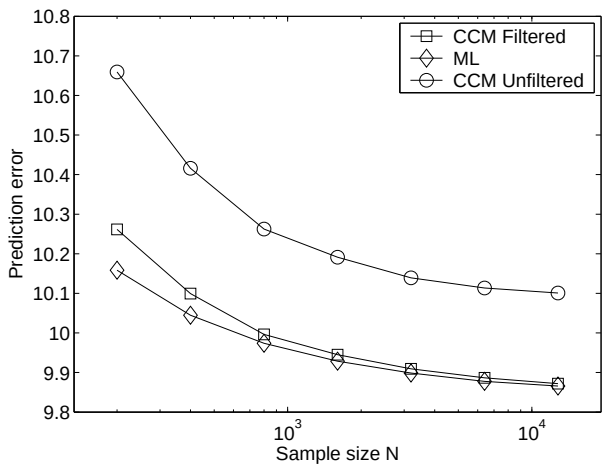

Figure 5.17: The prediction error as a function of the sample length for the ARMA(2,2) model.

Kullback-Leibler discrepancy used is $\mathbb{S}(\Phi_{true}, \Phi_{est})$ where Φ_{true} and Φ_{est} are the normalized densities of the true and estimated densities, respectively. We again note that, independent of comparison method, the CCM(f) seem to have asymptotical error similar to ML while CCM's error seems larger. This highlights the reason for introducing the prefiltering.

Finally we will consider randomly generated true systems in order to illustrate the robustness of the approach and its implementation. We will consider ARMA(n,n) model for $n = 3, 4$, and, 5 where the numerator and denominator polynomials are stabilized polynomials (with respect to the unit circle) with normally distributed coefficients of zero mean and unit variance. We use the same estimators as described above for the sample lengths $N = 200, 800$, and, 3200 and high-order AR models of orders $L = 10, 20$, and, 30, respectively. For each sample length we generate one model of each order. We estimate the prediction error as

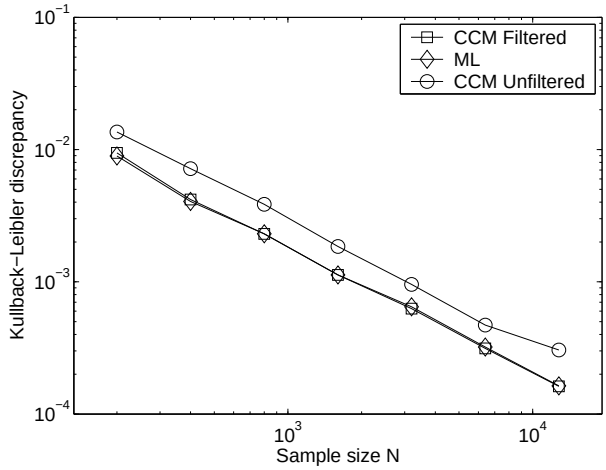

Figure 5.18: The Kullback-Leibler discrepancy as a function of the sample length for the ARMA(2,2) model.

above, normalize it with the ML prediction error for each example (in order to emphasize each realization as much) and run a Monte Carlo simulation with 500 runs. The result is displayed in Figure 5.19. We note a similar behavior as in the ARMA(2,2) case.

Remark 5.4.3. *An observation made in Example 1.0.3 is that the high order AR based cepstral estimate seems to be asymptotically efficient for white noise. This example indicates that this might generalize to other ARMA models. Therefore it can be in place to conjecture, that if the prefilter is the inverse of the true model, the covariance and cepstral estimates are asymptotically efficient and hence so the corresponding ARMA estimates by Theorem 5.4.1. Furthermore, the example indicates that using a recursive estimation scheme for determining the prefilter might also constitute an asymptotically efficient estimator.*

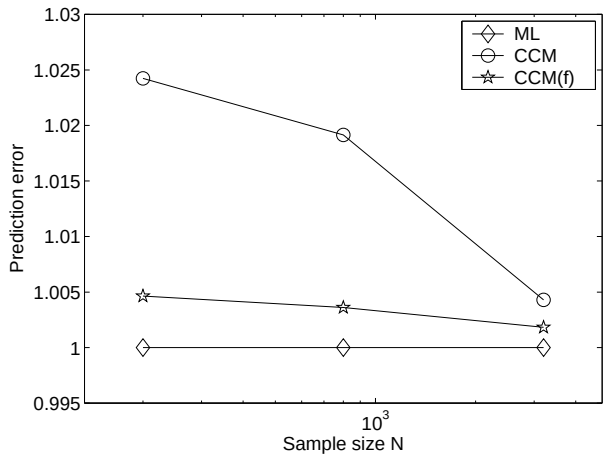

Figure 5.19: The normalized prediction error as a function of the sample length for the ARMA models of various order.

Chapter 6

Conclusions and Open Problems

In this thesis we have generalized the complexity constrained analytic interpolation theory. Together with new numerical algorithms for computation of the interpolants and the corresponding spectral densities this has enabled a significant development of two paradigms in robust control design and ARMA estimation. The theory *per se* as well as the examples motivate the study.

The results are encouraging. The problems are set in a very nice and general mathematical framework. Yet, controller designs in the studied benchmark problems outperform previous state-of-the-art designs and in ARMA estimation the prefiltering allow for better performance than time-domain ML estimation. Altogether, this should encourage implementation in real systems.

More fundamentally one should ask what methods to develop for these problems. The rapid development of accessible computational power as well as digital implementation make this question even more interesting. There seems to be a consensus that increased computational power alone will never be enough for a significant development. The reason is of course that the computational complexity obeys the curse of dimensionality. Then, two major options are to either develop widely-spread and used methods or to question the current methods and, if called upon, develop new methods eliminating the shortcomings of the previous ones. This thesis belong the the latter option. What will be most rewarding approach is of course a secret of the future.

In the engineering applications motivating this study, mathematics is simply a tool. Thereby, we are left with the freedom of choosing the mathematical formulation. In this thesis we advocate to state a mathematical problem that is well-posed. More precisely, we use well-posedness in the sense of Hadamard, namely that i) a solution exists, ii) the solution is unique, and iii) the solution depends continuously on the data in some reasonable topology. A problem that is not well-posed is said to be ill-posed. By stating a well-posed problem severe numerical sensitivity is avoided.

We formulate a problem for which a solution exist. Our key idea to obtain

uniqueness is strict convexity which is proven formulating a convex problem with a unique solution. Subsequently, we can also prove smoothness of the solution with respect to data, manifesting that we have stated a well-posed problem. The convex program also suggest that efficient numerical solutions are possible to develop.

Conventional $\mathcal{H}_\infty$ design relies on convex programming. The obstacle is that tuning of the design requires complexity-inflating weighting functions. Therefore our mathematical formulation is tempting since the degree is kept constant. Meanwhile, maximum likelihood ARMA estimation is an ill-posed problem. However, its ultimate statistical performance, that is meeting the Cramér-Rao bound, makes it attractive despite the ill-posedness. Meanwhile, the findings of this thesis indicate that our approach potentially can achieve ultimate asymptotical statistical performance while being a well-posed problem.

Our framework only allows for linear systems. A concern is whether this class of systems is wide enough. What model class that is appropriate of course heavily depends on the application in mind. Also, future developments of nonlinear systems might make them much more applicable than today. Yet, the experience is that simplicity is desirable.

The work presented in this thesis leave several related question open. Some of these are listen and commented below.

- A major topic discussed in this thesis is numerical algorithms. Clearly having robust, accurate and reliable algorithms is a prime concern. Meanwhile, for online implementation for instance within ARMA estimation but also higher order MIMO robust control problems, the numerical efficiency, that is the computational complexity, is critical. For ARMA estimation the preliminary implementation gives a computation time that is of the same magnitude as `armax` in [71]. However, the well-posedness suggests that faster algorithms might be within reach. One potential improvement within the homotopy approach taken in all the algorithms in Chapter 4, is to study higher order techniques for solving the corresponding ODEs. Recent work discussing the ODE in a more general framework can be found in [49]. Also, at least for the covariance type conditions, "fast" algorithms might be possible. Future work on the algorithms most likely also includes a thorough analysis of the algorithms' *computational complexity*. As for convergence results a major reason for studying this well-posed problem is of course guaranteed convergence. A proof manifesting this can be of theoretical interest.

- The matrix-valued case need much more attention in future work. Without doubt, it is the matrix-valued case that the degree constraint can be most useful in robust control applications. Likewise it is for vector-valued processes in the ARMA identification problem that the nonconvexity cause most severe problems. In robust control capability to treat tangential interpolation would also be useful. The matrix-valued generalization studied in this thesis is clearly interesting. Yet a more general approach would be valuable. A careful

choice of generalization of the Kullback-Leibler discrepancy as well as convex parameterization of the interpolant are the key issues to be studied. However, it is not obvious that a generalization most rewardingly will be studied in an optimization framework. A generalization from a geometric and moment matching viewpoint might be more suitable.

- The results presented in Section 5.2 simplifies the tuning of $\mathcal{H}_\infty$ controllers within the paradigm significantly. However, in contrast to the other results in the thesis, this problem is nonconvex. An alternative path is to study whether cepstral coefficients are useful as design parameters; either by estimating them from a pointwise given desired spectrum as in Section 5.2 or by directly giving the user a convenient way of choosing them.

- Example 5.4.2 indicates that the CCM(f) estimator might be asymptotically efficient. A thorough statistical analysis would be in place. By Theorem 5.4.1 this will in essence be a study of the joint distribution of the cepstral and covariance estimates. The prefilter complicates the analysis significantly – not the least because one need to study whether the prefilter converges to something in the iterative procedure.

- In the robust control applications we frequently encounter boundary interpolation values, that is interpolation points on the unit circle. Developing the theory to incorporate the boundary interpolation case would be interesting. As for computation Algorithm 2 seems to generalize immediately, whereas by construction, the optimization approach and the interpolation equation approach which contain barrier-like terms do not.

- We have only studied fixed degree interpolation problems. For both theoretical and applicational reasons characterization of minimal degree interpolants can be interesting. Admittedly, this is a difficult problem also within our framework since it requires characterization of pole-zero cancellation of the interpolant.

- In the presented work the basis functions for the covariances and cepstral coefficients need to be the same in order to parameterize solutions of degree n. An interesting topic would be to study whether there is some more freedom in choosing the basis functions. In a potential application to $\mathcal{H}_\infty$ controller design this might be an interesting option.

Bibliography

[1] N. I. Akhiezer. *The Classical Moment Problem*. Oliver and Boyd Ltd, 1965.

[2] E. L. Allgower and K. Georg. *Numerical Continuation Methods*. Springer-Verlag, 1990.

[3] K. J. Åström and T. Söderström. Uniqueness of the maimum likelihood estimates of the parameters of an ARMA model. *IEEE Trans. Automat. Control*, 19(6):769–773, 1974.

[4] A. Blomqvist and M. Deistler. Statistical properties of simultaneous ceostral and covariance matching. Working paper, 2003.

[5] A. Blomqvist and G. Fanizza. Identification of rational spectral densities using orthonormal basis functions. In *The proceedings of Symposium on System Identification*, Rotterdam, The Netherlands, 2003.

[6] A. Blomqvist, G. Fanizza, and R. Nagamune. Computation of bounded degree Nevanlinna-Pick interpolants by solving nonlinear equations. In *The proceedings of the 42nd IEEE Conference on Decision and Control*, pages 4511–4516, 2003.

[7] A. Blomqvist, G. Fanizza, and R. Nagamune. Computation of bounded degree Nevanlinna-Pick interpolants by solving nonlinear equations. Technical report, Mittag-Leffler Institute, 2003. Spring, no. 17.

[8] A. Blomqvist, A. Lindquist, and R. Nagamune. Matrix-valued Nevanlinna-Pick interpolation with complexity constraint: an optimization approach. *IEEE Trans. Automat. Control*, 48(12):2172–2190, 2003.

[9] A. Blomqvist and R. Nagamune. An extension of a Nevanlinna-Pick interpolation solver to cases including derivative constraints. In *The proceedings of the 41st IEEE Conference on Decision and Control*, pages 2552–2557, Orlando, Florida, 2002.

[10] A. Blomqvist and R. Nagamune. Optimization-based computation of analytic interpolants of bounded degree. Submitted to Systems and Control Letters, 2004.

[11] A. Blomqvist and R. Nagamune. *Sshaper User's Manual*, June 2004. Software available at `www.math.kth.se/~andersb/software.html`.

[12] A. Blomqvist and B. Wahlberg. A data driven orthonormal parameterization of the generalized entropy maximization problem. In *The proceedings of the Sixteenth International Symposium on Mathematical Theory of Networks and Systems*, 2004.

[13] A. Blomqvist and B. Wahlberg. On affecting the frequency domain error distribution in autoregressive spectral estimation using prefiltering. Submitted to IEEE Trans. Signal Processing, 2004.

[14] B. P. Bogert, J. R. Healy, and J. W. Tukey. The quefrency alanysis of time series for echoes: Cepstrum, pseudo-autocovariance, cross-cepstrum and saphe cracking. In M. Rosenblatt, editor, *Prooceedings of the Symposium on Time Series Analysis*, pages 209–243. John Wiley and Sons, 1963.

[15] J. Bokor, P. Heuberger, B. Ninness, T. Oliveira e Silva, P. Van den Hof, and B. Wahlberg. Modelling and identification with orthogonal basis functions. In *Workshop Notes, 14:th IFAC World Congress, Workshop nr 6*, Beijing, PRC, 7 1999.

[16] P. Borwein and T. Erdélyi. *Polynomials and Polynomial Inequalities*. Graduate Texts in Mathematics. Springer, 1995.

[17] G. Box and G. Jenkins. *Time Series Analysis: forcasting and control*. Holden-Day, 1976.

[18] R. W. Brockett. *Finite Dimensional Linear Systems*. John Wiley & Sons, New York, 1970.

[19] R. W. Brockett. Some geometric questions in the theory of linear systems. *IEEE Trans. Automat. Control*, 21(4):449–455, 1976.

[20] P. J. Brockwell and R. A. Davis. *Time Series: Theory and Methods, Second Edition*. Springer Series in Statistics. Springer, 1991.

[21] P. Broersen. Automatic spectral analysis with time series models. *IEEE Trans. Instrum. Meas.*, 51(2):211–216, 2002.

[22] J. Burg. *Maximum Entropy Spectral Analysis*. PhD thesis, Stanford University, 1975.

[23] C. I. Byrnes. On the global analysis of linear systems. In *Mathematical control theory*, pages 99–139. Springer, New York, 1999.

[24] C. I. Byrnes, P. Enqvist, and A. Lindquist. Cepstral coefficient, covariance lags and pole-zero models for finite data strings. *IEEE Trans. Signal Processing*, 49(4):677–693, April 2001.

[25] C. I. Byrnes, P. Enqvist, and A. Lindquist. Identifiability and well-posedness of shaping-filter parameterizations: A global analysis approach. *SIAM J. Contr. and Optimiz.*, 41(1):23–59, 2002.

[26] C. I. Byrnes, T. T. Georgiou, and A. Lindquist. A new approach to spectral estimation: A Tunable High-Resolution Spectral Estimator. *IEEE Trans. Signal Processing*, 48(11):3189–3205, November 2000.

[27] C. I. Byrnes, T. T. Georgiou, and A. Lindquist. A generalized entropy criterion for Nevanlinna-Pick interpolation with degree constraint. *IEEE Trans. Automat. Control*, 46(6):822–839, June 2001.

[28] C. I. Byrnes, T. T. Georgiou, A. Lindquist, and A. Megretski. Generalized interpolation in H^∞ with a complexity constraint. To appear in Trans. Amer. Math. Soc.

[29] C. I. Byrnes, S. V. Gusev, and A. Lindquist. A convex optimization approach to the rational covariance extension problem. *SIAM J. Contr. and Optimiz.*, 37(1):211–229, 1998.

[30] C. I. Byrnes and A. Lindquist. On the duality between filtering and Nevanlinna-Pick interpolation. *SIAM J. Contr. and Optimiz.*, 39(3):757–775, 2000.

[31] C. I. Byrnes and A. Lindquist. A convex optimization approach to generalized moment problems. In K. Hashimoto, Y. Oishi, and Y. Yamamoto, editors, *Cybernetics in the 21st Century: Festschrift in Honor of Hidenori Kimura on the Occasion of his 60th Birthday*, Control and Modeling of Complex Systems, pages 3–21. Birkhauser, 2003.

[32] C. I. Byrnes, A. Lindquist, S. V. Gusev, and A. S. Matveev. A complete parameterization of all positive rational extensions of a covariance sequence. *IEEE Trans. Automat. Control*, 40(11):1841–1857, November 1995.

[33] C. I. Byrnes, A. Lindquist, and Y. Zhou. On the nonlinear dynamics of fast filtering algorithms. *SIAM J. Contr. and Optimiz.*, 32(3):744–789, 1994.

[34] P. E. Caines. *Linear Stochastic Systems*. John Wiley & Sons, New York, 1988.

[35] B.-C. Chang and J. B. Pearson. Optimal Disturbance Reduction in Linear Multivariable Systems. *IEEE Trans. Automat. Control*, 29(10):880–887, October 1984.

[36] A. Dahlén, A. Lindquist, and J. Mari. Experimental evidence showing that stochastic subspace identification methods may fail. *Systems and Control Lett.*, 34(5):303–312, 1998.

[37] Ph. Delsarte, Y. Genin, and Y. Kamp. Orthogonal Polynomial Matrices on the Unit Circle. *IEEE Trans. Circuits and Systems*, 25(3):149–160, March 1978.

[38] Ph. Delsarte, Y. Genin, and Y. Kamp. The Nevanlinna-Pick Problem for Matrix-valued Functions. *SIAM J. Appl. and Math.*, 36:47–61, Feb 1979.

[39] Ph. Delsarte, Y. Genin, and Y. Kamp. Schur Parametrization of Positive Definite Block-Toeplitz Systems. *SIAM J. Appl. and Math.*, 36:34–46, Feb 1979.

[40] J. C. Doyle. Guaranteed magins for LQG regulators. *IEEE Trans. Automat. Control*, 23(4):756–757, 1978.

[41] J. C. Doyle, B. A. Francis, and A. R. Tannenbaum. *Feedback Control Theory*. MacMillan Publishing Company, New York, 1992.

[42] G. E. Dullerud and F. G. Paganini. *A Course in Robust Control Theory: A Convex Approach*. Springer, New York, 1999.

[43] P. Enqvist. A homotopy approach to rational covariance extension with degree constraint. *Int. J. Appl. Math. Comput. Sci.*, 11(5):1173–1201, 2001.

[44] P. Enqvist. *Spectral Estimation by Geometric, Topological and Optimization Methods*. PhD thesis, Royal Institute of Technology, Stockholm, Sweden, 2001.

[45] P. Enqvist. A convex optimization approach to ARMA(n,m) model design from covariance and cepstral data. 2004.

[46] Y. Ephraim and M. Rahim. On second-order statistics and linear estimation of cepstral coefficients. *IEEE Trans. Signal Processing*, 7(2):162–176, 1999.

[47] B. A. Francis. *A Course in H_∞ Control Theory.* Lecture Notes in Control and Information Sciences. Springer-Verlag, 1987.

[48] F. R. Gantmacher. *The Theory of Matrices.* Chelsea, New York, 1959.

[49] T. T. Georgiou. Solution of the general moment problem via a one-parameter imbedding. Submitted to IEEE Trans. Automat. Control.

[50] T. T. Georgiou. *Partial realization of covariance sequences.* PhD thesis, Univ. Florida, Gainesville, 1983.

[51] T. T. Georgiou. The interpolation problem with a degree constraint. *IEEE Trans. Automat. Control*, 44(3):631–635, March 1999.

[52] T. T. Georgiou. Signal estimation via selective harmonic amplification: MUSIC, Redux. *IEEE Trans. Signal Processing*, 48(3):780–790, March 2000.

[53] T. T. Georgiou. Spectral estimation via selective harmonic amplification. *IEEE Trans. Automat. Control*, 46(1):29–42, 2001.

[54] T. T. Georgiou. Spectral analysis based on the state covariance: the maximum entropy spectrum and linear fractional parametrization. *IEEE Trans. Automat. Control*, 47(11):1811–1823, 2002.

[55] T. T. Georgiou. The structure of state covariances and its relation to the power spectrum of the input. *IEEE Trans. Automat. Control*, 47(7):1056–1066, 2002.

[56] T. T. Georgiou and A. Lindquist. Kullback-Leibler approximation of spectral density functions. *IEEE Trans. Information Theory*, 49(11):2910–2917, November 2003.

[57] A. Graham. *Kronecker products and matrix calculus with applications.* John Wiley & Sons, 1981.

[58] M. Green and D. J. N. Limebeer. *Linear Robust Control.* Prentice Hall, 1995.

[59] J. Hadamard. Sur les correspondances ponctuelles. In *Oeuvres, Editions du Centre Nationale de la Researche Scientifique*, pages 383–384. Paris, 1968.

[60] E.J. Hannan and M. Deistler. *The Statistical Theory of Linear Systems.* John Wiley and Sons, 1988.

[61] J. W. Helton and O. Merino. *Clasical Control using H^∞ Methods.* SIAM, Philadelphia, 1998.

[62] D. Henrion and M. Šebek. Efficient numerical method for the discrete-time symmetric matrix polynomial equation. *IEE Proc. Control Theory Appl.*, 5(145):443–447, 1998.

[63] S. Van Huffel, V. Sime, A. Varga, S. Hammarling, and F. Delebecque. High-Performance numerical software for control. *IEEE Contr. Syst. Mag.*, 24(1):60–76, 2004.

[64] J. Ježek. Symmetric matrix polynomial equations. *Kyberneika*, 22(1):19–30, 1986.

[65] R. E. Kalman. Realization of covariance sequences. In I. Gohberg, editor, *Toeplitz Centennial*, volume 4 of *Operator Theory: Advances and Applications*, pages 331–342. 1981.

[66] H. Kimura. Application of classical interpolation theory. In N. Nagai, editor, *Linear Circuits, Systems and Signal Processing*, pages 61–85. Marcel Dekker Inc., New York, 1990.

[67] S. Kullback. *Information Theory and Statistics*. Wiley Publications in Statistics. John Wiley & Sons, 1959.

[68] A. Lindquist. Stochastic realization theory. Lectures in "Matematisk systemteori, fk".

[69] A. Lindquist. Notes on covariance lags and cepstral coefficients. 2002.

[70] A. Lindquist and G. Picci. *Identification, Adaptation, Learning: The Science of Learning Models from Data*, volume 153 of *Nato ASI Series, Series F*, chapter Geometric Methods for State Space Identification, pages 1–69. Springer, 1996.

[71] L. Ljung. Systems identification toolbox. Mathworks MATLAB toolbox.

[72] L. Ljung. Estimation focus in system identification: Prefiltering, noise models, and prediction. In *IEEE Conference on Decision and Control*, pages 2810–2815, Dec. 1999.

[73] L. Ljung. *System Identification, Theory for the User*. Prentice Hall, 1999.

[74] J. M. Maciejowski. *Multivariable Feedback Design*. Addison-Wesley, Wokingham U.K., 1989.

[75] N. Merhav and C.-H. Lee. On the asymptotic statistical behavior of emperical cepstral coefficients. *IEEE Trans. Signal Processing*, 41(5):1990–1993, May 1993.

[76] B. C. Moore. Principal component analysis in linear systems: controllability, observability, and model reduction. *IEEE Trans. Automat. Control*, 26(1):17–32, 1981.

[77] R. Nagamune. Sensitivity reduction for SISO systems using the nevanlinna-pick interpolation with degree constraint. In *The proceeding of MTNS*, Perpignan, France, June 2000.

[78] R. Nagamune. *Robust Control with Complexity Constraint: A Nevanlinna-Pick Interpolation Approach*. PhD thesis, Royal Institute of Technology, Stockholm, Sweden, 2002.

[79] R. Nagamune. Simultaneous robust regulation and robust stabilization with degree constraint. In *The proceeding of MTNS*, University of Notre Dame, Indiana, August 2002.

[80] R. Nagamune. A robust solver using a continuation method for Nevanlinna-Pick interpolation with degree constraint. *IEEE Trans. Automat. Control*, 48(1):113–117, 2003.

[81] R. Nagamune. Closed-loop shaping based on Nevanlinna-Pick interpolation with a degree bound. *IEEE Trans. Automat. Control*, 49(2):300–305, 2004.

[82] R. Nagamune and A. Blomqvist. Sensitivity shaping with degree constraint by nonlinear least-squares optimization. To appear in Automatica, 2005.

[83] R. Nagamune and A. Lindquist. Sensitivity shaping in feedback control and analytic interpolation theory. In J. L. Menaldi, E. Rofman, and A. Sulem, editors, *Optimal Control and Partial Differential Equations*. IOS Press, Ohmsha, 2001.

[84] S. G. Nash and A. Sofer. *Linear and Nonlinear Programming*. McGraw Hill, 1996.

[85] R. Nevanlinna. Über beschränkte Funktionen, die in gegebenen Punkten vorgeschriebene Werte annehmen. *Ann. Acad. Sci. Fenn. Ser. A*, 13(1), 1919.

[86] A. V. Oppenheim and R. W. Shafer. *Digital Signal Processing*. Prentice Hall, London, 1975.

[87] G. Pick. Über die Beschränkungen analytischer Funktionen, welche durch vorgegebene Funktionswerte bewirkt werden. *Math. Ann.*, 77:7–23, 1916.

[88] R. Pintelon and J. Schoukens. Time series analysis in the frequency domain. *IEEE Trans. Signal Processing*, 47(1):206–210, 1999.

[89] R. Pintelon and J. Schoukens. *System Identification, A Frequency Domain Approach*. IEEE Press, 2001.

[90] B. Porat. *Digital Processing of Random Signals, Theory & Methods*. Prentice Hall, 1994.

[91] R. T. Rockafellar. *Convex Analysis*. Princeton University Press, Princton, NJ, 1970.

[92] W. Rudin. *Principles of Mathematical Analysis, Third edition*. McGraw-Hill, New York, 1976.

[93] W. Rudin. *Real and Complex Analysis*. McGraw-Hill, New York, 1987.

[94] D. Sarason. Generalized interpolation in H^∞. *Trans. Amer. Math. Soc.*, 127:179–203, 1967.

[95] I. Schur. Über Potenzreihen, die im Innern des Einheitskreises beschränkt sind. *Journal für die reine und angewandte Mathematik*, 147:205–232, 1917.

[96] G. Segal. The topology of spaces of rational functions. *Acta Mathematica*, 143:39–72, 1979.

[97] M. M. Seron, J. H. Braslavsky, and G. C. Goodwin. *Fundamental Limitations in Filtering and Control*. Springer-Verlag, 1997.

[98] T. Söderström and P. Stoica. *System Identification*. Prentice Hall, 1989.

[99] P. Stoica, T. McKelvey, and J. Mari. MA estimation in polynomial time. *IEEE Trans. Signal Processing*, 48(7):1999–2012, 2000.

[100] P. Stoica and R. Moses. *Introduction to Spectral Analysis*. Prentice Hall, 1997.

[101] A. Tannenbaum. Feedback stabilization of linear dynamical plants with uncertainty in the gain factor. *Int. J. Control*, 32(1):1–16, 1980.

[102] B. Wahlberg. Orthonormal basis functions models: A transformation analysis. *SIAM Review*, pages 689 – 705, 11 2003.

[103] J. L. Walsh. *Interpolation and Approximation by Rational Functions in the Complex Domain*, volume 20. American Mathematical Society, Colloquium Publications, Providence, R. I., 1956.

[104] J. C. Willems. Realization of systems with internal passivity and symmetry constraints. *Journal of The Franklin Institute*, 301(6):605–621, 1976.

[105] K. Zhou. *Essentials of Robust Control*. Prentice-Hall, New Jersey, 1998.