Rational approximation techniques and frequency design: a Zolotarev problem and the Schur algorithm

Vincent Lunot

UNIVERSITÉ DE PROVENCE
U.F.R. M.I.M.
ÉCOLE DOCTORALE DE MATHÉMATIQUES ET INFORMATIQUE E.D. 184

THÈSE

présentée pour obtenir le grade de

DOCTEUR DE L'UNIVERSITÉ DE PROVENCE

Spécialité : Mathématiques Appliquées

par

Vincent Lunot

sous la direction du Dr. Laurent Baratchart

Titre :

Techniques d'approximation rationnelle en synthèse fréquentielle : problème de Zolotarev et algorithme de Schur

soutenue publiquement le 5 mai 2008

JURY

M. Alexander Borichev	Université de Provence	Président
M. Laurent Baratchart	INRIA Sophia Antipolis - Méditerranée	Directeur de thèse
M. Daniel Alpay	Ben-Gurion University of the Negev (Israel)	Rapporteur
M. Smain Amari	Royal Military College (Canada)	Rapporteur
M. Stéphane Bila	Institut de Recherche XLIM, Limoges	Examinateur
M. Stanislas Kupin	Université de Provence	Examinateur

INVITÉS

M. Damien Pacaud	Thales Alenia Space, Toulouse
M. Edward B. Saff	Vanderbilt University (USA)

Remerciements

Je remercie en premier lieu M. Alexander Borichev pour l'honneur qu'il me fait de présider ce jury.

Il m'est difficile d'exprimer en quelques mots ce que je dois à M. Laurent Baratchart, mon directeur de thèse. Sa perception globale de certains concepts mathématiques m'a été d'une grande aide à plusieurs reprises, lorsque je me retrouvais dans une impasse. Mais je lui dois surtout un enrichissement, tant sur le plan culturel qu'humain, acquis lors de ces quelques années passées au sein de son projet.

La première partie de cette thèse a été encadrée par M. Fabien Seyfert, qui m'a fait découvrir un univers passionnant, celui des méthodes numériques employées en optimisation et approximation rationnelle. Je lui suis tout particulièrement reconnaissant du temps qu'il a pu me consacrer à me fournir de nombreuses explications, le plus souvent d'une limpidité extrême.

Je remercie M. Daniel Alpay et M. Smain Amari qui m'obligent infiniment en acceptant de rapporter cette thèse.

Concernant l'approximation Schur, je remercie M. Stanislas Kupin et Mme Martine Olivi avec qui j'ai eu l'occasion de travailler. La capacité de M. Stanislas Kupin à reformuler de manière simplifiée des concepts complexes m'a permis une meilleure perception du problème. Le soutien régulier de Mme Martine Olivi m'a été d'une aide précieuse.

La partie applicative de cette thèse est certainement celle qui m'a apporté le plus sur le plan de la satisfaction personnelle.
L'intérêt prononcé de M. Smain Amari et M. Stéphane Bila pour mes résultats concernant le calcul des fontions de filtrage a été un facteur de motivation supplémentaire.
De plus, la collaboration avec M. Stéphane Bila, M. Philippe Lenoir et M. Abdallah Nasser, qui a permis de déboucher sur la fabrication de plusieurs filtres hyperfréquences multibandes a été réellement passionnante.

Mes trois années au sein du projet APICS seront inoubliables. Merci à tous ses membres.
Je remercie France Limouzis, Stéphanie Sorres et Christine Riehl pour leur aide concernant tous les problèmes organisationnels.
Un grand merci tout particulier à José Grimm, pour sa relecture attentive de certains chapitres de ma thèse, pour ses remarques pertinentes (bien que parfois formulées de manière surprenante), pour ses explications quant à certains problèmes informatiques, et plus généralement pour tout ce que j'ai appris grâce à lui.
Enfin, merci à Jean-Baptiste Pomet, Alban Quadrat, Juliette Leblond, Rania Bassila, Stéphane Rigat, mais aussi à Sapna Nundloll, du projet voisin, Comore.

Finalement, je tiens à remercier mes parents et mon frère, ainsi que mes amis, pour tous les moments de joie qu'ils ont pu m'apporter.

Contents

Chapter 0

Introduction

Mis à part l'introduction qui est en français, l'ensemble du manuscrit est rédigé en anglais. Étant moi-même fervent défenseur de la langue française, ce choix peut paraître surprenant. La justification est en grande partie pratique : la première partie étant basée sur une série d'articles en anglais ([Bila et al., 2006], [Lunot et al., 2007] et [Lunot et al., 2008]), il m'a semblé naturel de conserver cette langue. La seconde partie, quant à elle, devrait faire l'objet d'un futur article. La rédiger directement dans la langue internationale m'a donc paru approprié. De plus, certains éléments de ce travail pouvant intéresser d'autres scientifiques, il m'a semblé dommage d'en limiter l'accès aux seuls connaisseurs de la langue de Molière.
De surcroît, écrire tout un manuscrit dans une langue étrangère est un excellent exercice pour progresser dans sa pratique. En effet, cela permet d'assimiler du nouveau vocabulaire, mais aussi de se rendre compte, et ainsi de corriger, certaines grossières erreurs.
Enfin, l'anglais étant la langue scientifique internationale, son utilisation a aussi été choisie par respect pour la communauté scientifique.

Cette thèse traite deux problèmes : la résolution d'un problème de Zolotarev et l'approximation rationnelle sous contrainte Schur. Ces problèmes ont en fait deux points communs.
Le premier peut être perçu au niveau du domaine d'application. En effet, ces deux problèmes apparaissent lors de la fabrication de filtres hyperfréquences. La résolution du problème de Zolotarev permet de calculer des fonctions de filtrage optimales et trouve donc des applications en synthèse de filtres. L'approximation rationnelle Schur, quant à elle, permet l'identification de systèmes passifs, et donc en particulier de filtres.
Le deuxième point commun se situe au niveau théorique. Les deux problèmes sont de type $\max\min$, et les techniques employées dans leur étude font partie du domaine de l'approximation rationnelle.

La première partie traite d'un problème de Zolotarev. Le calcul de la solution d'un tel problème ayant déjà permis la réalisation de filtres hyperfréquences aux caractéristiques complexes, le sujet est abordé du point de vue applicatif.

Le premier chapitre présente très succinctement les filtres hyperfréquences. Il s'agit de filtres utilisés dans les satellites de télécommunications, et qui sont en fait une suc-

cession de cavités. Leur modèle théorique est une série de circuits résonnants, identifiée à un quadripôle. Celui-ci est représenté par une matrice 2×2 notée S, appelée matrice de transfert, qui permet de faire le lien entre les puissances entrantes et sortantes. Les termes S_{11} et S_{22} correspondent aux réflections de puissances, et les termes S_{12} et S_{21} aux transmissions.

Le deuxième chapitre définit la notion de fonction de filtrage optimale, et introduit les divers résultats sous forme simplifiée. On montre tout d'abord que le carré du module de la transmission s'écrit sous la forme

$$|S_{12}|^2 = \frac{1}{1 + \left|\frac{p}{q}\right|^2}$$

où p et q sont deux polynômes tels que le degré de p est supérieur au degré de q. La fonction de filtrage F d'un filtre est définie par $F = p/q$. On dit qu'elle est optimale si pour un niveau de transmission donné dans les bandes passantes (notées I), la réflection est maximale dans les bandes stoppées (notées J). Comme le système est conservatif, la transmission S_{12} et la réflection S_{11} sont liées par la relation $|S_{11}|^2 + |S_{12}|^2 = 1$. Une fonction est donc optimale si elle est solution du problème normalisé suivant

$$\text{Trouver } (p,q) \text{ solution de : } \max_{(p,q)\in\mathcal{R}_m^n} \min_{\omega\in J} \left|\frac{p}{q}(\omega)\right|$$

où

$$\mathcal{R}_m^n = \left\{(p,q) \in \mathcal{P}_n(\mathbb{R}) \times \mathcal{P}_m^*(\mathbb{R}), \sup_{\omega\in I} \left|\frac{p}{q}(\omega)\right| \leq 1\right\}.$$

Si p/q est optimale, alors le signe de p est constant sur les bandes stoppées J et le signe de q est constant sur les bandes passantes I. On découpe donc le problème en sous-problèmes où le signe de p (resp. q) est imposé sur chaque bande stoppée (resp. passante). Ce sous-problème signé admet une unique solution, qui est caractérisée par une propriété d'alternation. On s'intéresse alors au calcul de cette solution. Pour cela, on adapte des algorithmes classiques d'approximation rationnelle. Un algorithme de type Remes (voir [Remes, 1934] ou [Powell, 1981]) est obtenu pour le cas polynomial. Le cas général (rationnel) utilise un algorithme de type correction différentielle (voir [Cheney and Loeb, 1961] ou [Braess, 1986]).

Le troisième chapitre traite un problème généralisé. Sur les bandes passantes, la fonction n'est plus supposée comprise entre -1 et 1 mais entre deux fonctions continues. Sur les bandes stoppées, le critère maximisé n'est plus la valeur absolue, mais l'écart à une fonction continue. De plus, les bandes passantes et stoppées ne sont plus des intervalles de longueur finie, mais des compacts, voir même pour le cas polynomial des compacts du compactifié d'Alexandroff, c'est-à-dire que la possibilité d'intervalles de longueur infinie est considérée. Enfin, un poids est ajouté.

Le quatrième chapitre présente la mise en pratique de la théorie. Les fonctions théoriques et les mesures obtenues après fabrication par les laboratoires de l'institut XLIM (Limoges) de deux filtres bi-bandes et un tri-bandes sont données.

La deuxième partie traite l'approximation rationnelle sous contrainte Schur. On appelle fonction Schur une fonction analytique et bornée en module par 1 dans le disque

unité. L'approximation d'une fonction Schur f par une fonction rationnelle elle-même Schur a d'importantes applications dans l'identification de systèmes passifs. Les techniques habituelles d'approximation rationnelle non-contrainte L^2 ne permettent pas de traiter un tel cas. En effet, lorsque la fonction f prend des valeurs proches (en module) de 1, l'approximant L^2 tournant autour de la fonction, celui-ci peut alors prendre des valeurs plus grandes que 1, et ainsi ne pas être Schur. L'idée est alors d'utiliser un algorithme de Schur multipoints ([Jones, 1988]) qui permet d'obtenir une fonction rationnelle garantie être Schur.

Le premier chapitre présente un tel algorithme. À partir d'une fonction Schur et d'une suite de points (α_k) dans le disque, celui-ci fournit une suite de fonctions Schur (f_k) et une suite de points du disque γ_k, appelés paramètres de Schur. L'algorithme est identifié à une fraction continue, dont les convergents d'ordre pair sont appelés fonctions rationnelles de Wall. Ces fonctions de Wall sont des fonctions Schur. Nous verrons par la suite qu'il s'agit de candidats intéressants pour l'approximation.

Le deuxième chapitre introduit les fonctions rationnelles orthogonales. La présentation est basée sur le livre [Bultheel et al., 1999].

Le troisième chapitre fournit un lien entre l'algorithme de Schur et les fonctions rationnelles orthogonales. À cette fin, nous associons par la transformée de Herglotz une mesure à la fonction f. Un théorème de type Geronimus (voir [Geronimus, 1944] pour la version traitant le cas des polynômes orthogonaux, ou [Langer and Lasarow, 2004] pour une version étendue aux fonctions rationnelles orthogonales) est ensuite présenté. Celui-ci montre que les paramètres de Schur sont liés aux valeurs des fonctions rationnelles orthogonales aux points α_k.

Le quatrième chapitre est une étude de différentes convergences. On y présente tout d'abord un résultat de type Szegő qui relie asymptotiquement les valeurs prises par les fonctions rationnelles orthogonales aux points α_k (tendant possiblement vers le cercle unité) aux valeurs prises par la fonction de Szegő de la mesure en ces mêmes points. On généralise ensuite des résultats de convergence obtenus pour l'algorithme de Schur classique dans [Khrushchev, 2001]. Lorsque les points α_k ne sont pas tous pris en 0 comme dans le cas classique, des poids de type noyau de Poisson en α_k apparaissent. La difficulté supplémentaire vient du fait qu'ici, les points (α_k) peuvent tendre vers le cercle. On obtient d'abord une convergence L^2 avec poids des fonctions de Schur f_n, puis pour les fonctions rationnelles de Wall, une convergence sur les compacts, une convergence par rapport à la distance pseudo-hyperbolique et à la métrique de Poincaré et une convergence de type L^2, toujours avec poids de type noyau de Poisson. En plus de ces extensions de [Khrushchev, 2001], nous construisons aussi une suite de points (α_n) pour laquelle nous obtenons une convergence de type "BMO asymptotique" des fonctions de Schur f_n.

Le cinquième chapitre est une étude numérique ayant pour objectif le calcul d'un approximant rationnel Schur de degré fixé. Pour cela, on constate tout d'abord que l'algorithme Schur multipoints fournit un paramétrage des fonctions rationnelles strictement Schur. Un processus d'optimisation est alors mis au point. Plusieurs exemples sont traités, et la comparaison est faite avec l'approximation non contrainte L^2.

Part I

A generalized Zolotarev problem with application to the synthesis of multi-band microwave filters

Every day, people use filters without even noticing it. Indeed, in our society where communications are omnipresent, filters are needed in order to select the relevant information. They can be found in many different systems such as mobile phones, radios, televisions, satellites, ... Therefore, it is not surprising that filters have been widely studied by the engineering community (see for example the books of [Kurokawa, 1969], [Hong and Lancaster, 2001] and [Cameron et al., 2007]).
However, since more and more performing filters are needed, new problems arise. In particular, being able to compute advanced filtering characteristics has become a major way of improving and simplifying the architecture of systems.
Some recent studies (e.g. [Cameron et al., 2005b], [Macchiarella and Tamiazzo, 2005] and [Lee and Sarabandi, 2008]) exposed methods using frequency transformations to design multiband microwave filters. However, these lack generality. Indeed, the response is limited to symmetric specifications or by the position of the transmission zeros.
For general specifications, some optimization methods are known (e.g. [Amari, 2000] and [Mokhtaari et al., 2006]). However, they do not guarantee the optimality of the response. Our purpose throughout this part is to give efficient ways to compute multiband filtering functions, that is giving algorithms which are proven to converge to the optimal solution. In this way, an automatic tool for computing the filtering functions can be implemented. We will adapt to that purpose some classical techniques of rational approximation.

This part is divided in four chapters. The first one introduces briefly microwave filters: a description is given and the theoretical model is presented. The second chapter gives methods for computing multiband filtering functions. The problem to solve is in fact a Zolotarev problem ([Todd, 1988]), that is finding a rational function bounded in modulus by one on some intervals whose infimum in modulus on some other intervals is maximal. We tried in this chapter to give the results in a simplified way. The Zolotarev problem, in the specific case where only two intervals are given (i.e. the rational functions are bounded by one over an interval, and we want to maximize the infimum on another interval), is studied in [Le Bailly and Thiran, 1998], and an algorithm is given. However, this algorithm uses the specific structure of the solution in this particular case, and does not extend to the general case with more than two intervals. Therefore, we present here two algorithms, which are adaptations of the Remes algorithm ([Remes, 1934]) and the differential-correction algorithm ([Cheney and Loeb, 1961]). The purpose of the third chapter is to give full proofs of the previous results. In fact, problems with more general constraints are studied. Finally, in the fourth chapter, multiband filters designed using the previous theory are presented.

Chapter 1

A short introduction to microwave filters

In this chapter, we will give a brief description of microwave filters. More details can be found in [Kurokawa, 1969], [Hong and Lancaster, 2001], [Baratchart et al., 1998] and [Sombrin, 2002]. We first present microwave filters and their different components. Next, the theoretical model is introduced.

1.1 Structure of a microwave filter

The purpose of a microwave filter is to select frequency ranges, i.e. to let the signal pass for some frequency ranges called the pass-bands and to stop the signal at some other frequencies, the stop-bands. Microwave filters work in frequency domains around the GHz, and their passbands are only a few MHz. A microwave filter is a passive system, only composed of a sequence of cavities (i.e. finite volumes delimited by metallic walls), electromagnetically coupled by irises (i.e. small apertures in the cavity). Fig. 1.1 shows a dual-mode microwave filter with six cavities. In Fig. 1.2, the reader can see the different components of a filter: the cavities, the irises, and some screws.

The design of a filter is complex, and is usually divided in two main steps: the synthesis

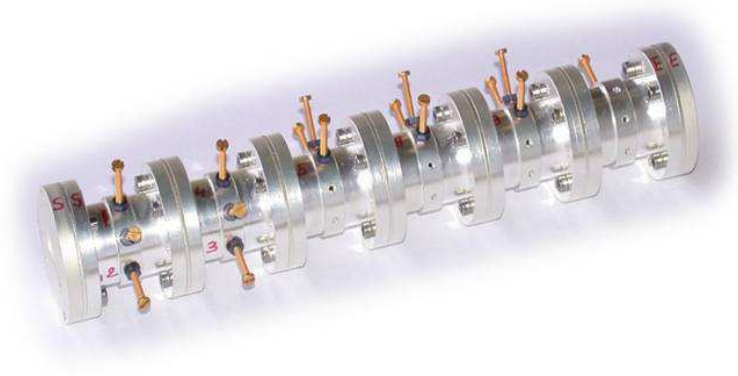

Figure 1.1: A dual-mode microwave filter with six cavities.

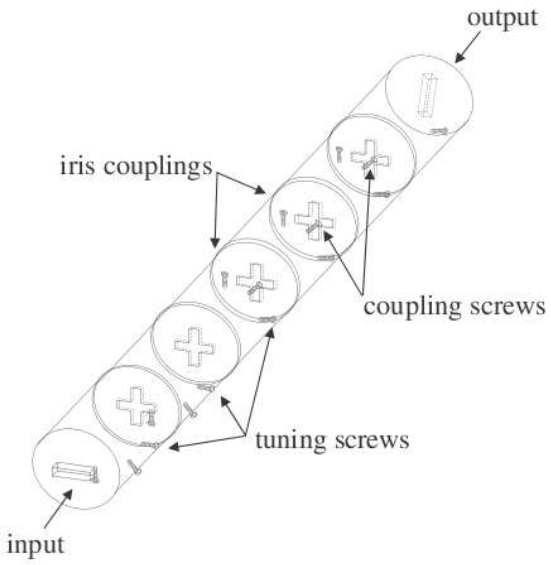

Figure 1.2: Irises and screws of a dual-mode microwave filter.

and the identification.
The first one consists in determining the physical parameters which meet the specifications, that is determining the topology, the number, the type and the size of the cavities, and the type and size of each iris. For example, the topology of the monomode filter in Fig. 1.3 is totally different from the one of the dual-mode filter in Fig. 1.2.
Once the main structure is determined, the filter is manufactured. Since the manufactured filter can not be perfect, tuning must be done. Tuning is realized by adjusting the screws embedded in each cavity. In this process, in order to determine which screw should be tuned, the actual parameters of the filter have to be identified and compared with the theoretical ones. This process is called the identification.

1.2 The scattering matrix

As stated before, the elementary component of a filter is a cavity. When fed through a waveguide, the effect of a cavity on the electric and magnetic fields in the waveguide section can be modeled as a RLC circuit ([Kurokawa, 1969], [Collin, 1991], [Matthaei, 1965]). More precisely, in a narrow band around the resonance frequency of the cavity, the amplitudes of the electric and magnetic fields of the feeding mode propagating in the waveguide behave like voltages and currents of a RLC circuit. In a similar way, several cavities connected one to the other by small apertures can be modeled as a sequence of circuits coupled electromagnetically (see Fig. 1.4). R_i, L_i and C_i denote respectively a resistor, an inductor and a capacitor. M_{ij} and r_{ij} are an inductive coupling and a resistor, which represent the interaction between the i-th and j-th resonator circuits. Z_1 and Z_2 are related to the electromagnetic couplings realized between the feeding mode and the resonating modes of the input and output cavities. The latter couplings are usually realized by the input and output irises. Depending of how many modes are excited in the cavity, the latter is modelled

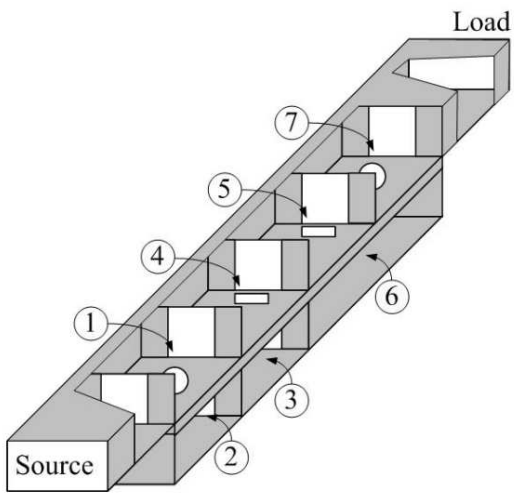

Figure 1.3: A monomode microwave filter with seven cavities.

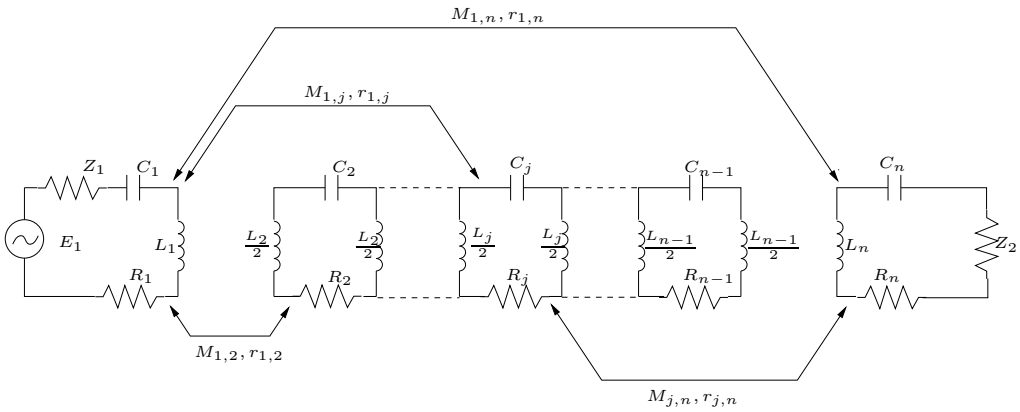

Figure 1.4: The equivalent electrical model.

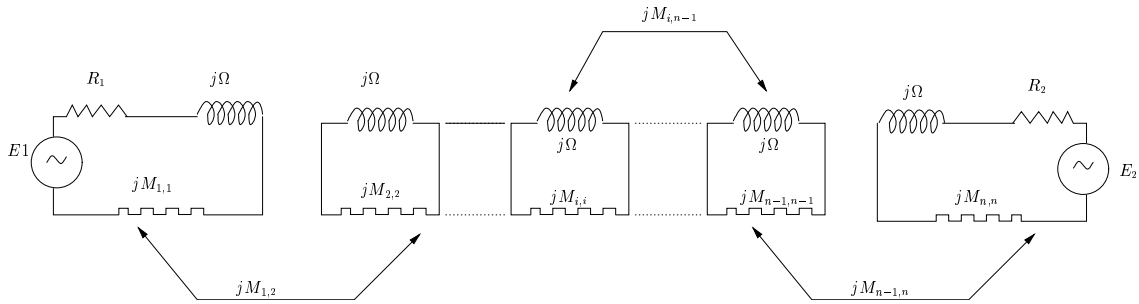

Figure 1.5: The low-pass prototype.

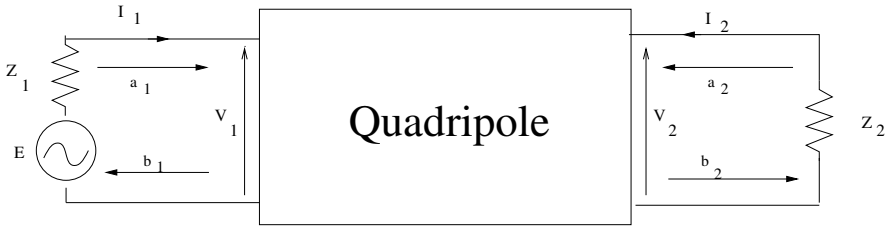

Figure 1.6: The quadripole model.

by one or two resonant circuits (one per mode). The resonance frequency of the resonating modes are entirely determined by the dimensions of the cavity. For circular cylindrical cavities and rectangular cavities, simple formulas are known ([Conciauro et al., 2000]). In the circuit representation, $\frac{1}{\sqrt{L_i C_i}}$ represents the frequency of the mode, and R_i represents the dissipation loss of the cavity.

In an ultimate approximation and normalization step, and when working in a very narrow band around the resonance frequencies of the cavities, the response of the RLC circuit (Fig. 1.4) is close to the response of the so-called low-pass prototype (Fig. 1.5) around the zero frequency (for details, see [Cameron et al., 2007] or [Sombrin, 2002]). In this transformation, the central frequency of the filter is cast to the zero frequency. In the low-pass circuit, magnetic couplings are replaced by constant admittance inverters ($jM_{i,l}$) and the LC elements are replaced by unity inductors and frequency-invariant reactances ($jM_{i,i}$).

When considering only the input and output, the previous circuits are in fact quadripoles, or two-port networks (see Fig. 1.6). When the first entry is powered, using Kirchhoff's law, we obtain a linear relation between the Laplace transforms of the currents I_1, I_2 (Fig. 1.6) and the voltages V_1, V_2 modelled by a 2×2 matrix Z :

$$\begin{pmatrix} V_1 \\ V_2 \end{pmatrix} = Z \begin{pmatrix} I_1 \\ I_2 \end{pmatrix}.$$

The matrix Z is called the impedance matrix. The entries of Z are rational functions of the variable $i\omega$, where ω denotes the frequency. Note that, due to the narrow-band approximation, the polynomials of the rational function do not necessarily have real valued coefficients.

In practice, we can only measure the amplitude and phase of the incident and reflected waves. These waves are denoted by a_1, a_2 (incident waves) and b_1,b_2 (see Fig.1.6) and are

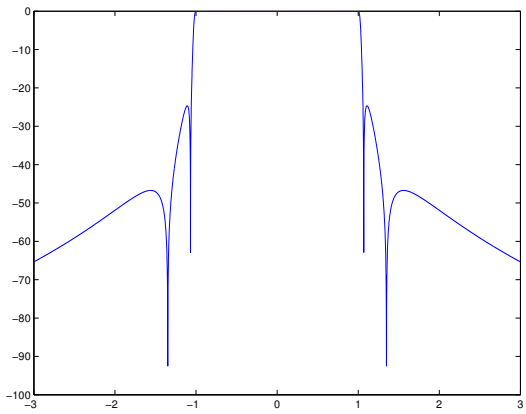

Figure 1.7: Transmission $|S_{12}|^2$ versus normalized frequency, y in dB (i.e. the plotted function is $20\log_{10}|S_{12}|$)

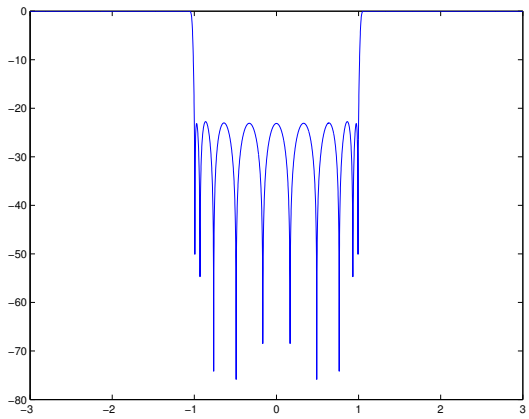

Figure 1.8: Reflection $|S_{22}|^2$ versus normalized frequency, y in dB (i.e the plotted function is $20\log_{10}|S_{22}|$)

defined by

$$\begin{aligned} a_1 &= \frac{1}{2}\left(\frac{V_1}{\sqrt{Z_1}} + \sqrt{Z_1}I_1\right), \\ b_1 &= \frac{1}{2}\left(\frac{V_1}{\sqrt{Z_1}} - \sqrt{Z_1}I_1\right), \\ a_2 &= \frac{1}{2}\left(\frac{V_2}{\sqrt{Z_2}} + \sqrt{Z_2}I_2\right), \\ b_2 &= \frac{1}{2}\left(\frac{V_2}{\sqrt{Z_2}} - \sqrt{Z_2}I_2\right). \end{aligned}$$

The square modulus of these quantities can be seen as the transmitted and reflected powers at the input and output of the filter. The relation between the input and the output is given by a 2×2 matrix S whose entries are denoted by S_{ij}, $1 \leq i, j \leq 2$:

$$\begin{pmatrix} b_1 \\ b_2 \end{pmatrix} = \begin{pmatrix} S_{11} & S_{12} \\ S_{21} & S_{22} \end{pmatrix} \begin{pmatrix} a_1 \\ a_2 \end{pmatrix}$$

Definition 1.2.1 *The matrix S is called the scattering matrix of the filter.*

The terms S_{11} and S_{22} represent the reflection, and S_{12} and S_{21} the transmission. In Fig. 1.7 and 1.8, the transmission and the reflection of an ideal monoband filter are plotted (the passband is $I = [-1, 1]$ and the stopbands are $J_1 = [-3, -1.1]$ and $J_2 = [1.1, 3]$).

Definition 1.2.2 *We call attenuation level in a stopband the value (in dB) of the minimum of the absolute value of the transmission $-20\log_{10}|S_{12}|$ in this band, and we call return loss in a passband the value (in dB) of the minimum of the absolute value of the reflection $-20\log_{10}|S_{22}|$ in this band.*

In Fig. 1.7 and 1.8, the attenuation level in the stopband $[1.1, 3]$ is equal to 30 dB, and the return loss in the passband $[-1, 1]$ is equal to 22 dB.

The matrices S and Z are related by

$$S = Z_0^{-1/2}(Z - Z_0)(Z + Z_0)^{-1} Z_0^{1/2}$$

where

$$Z_0 = \begin{pmatrix} Z_1 & 0 \\ 0 & Z_2 \end{pmatrix}.$$

We assume that the microwave filter is a stable causal linear system without loss, i.e. we assume that R_i and r_{ij} are small and can be approximated by 0. As the filter is modelled by a finite sequence of resonant circuits, it is a finite dimension system. Therefore, the entries S_{ij} of the scattering matrix are *rational functions, analytic in the right half-plane.* Furthermore, the reciprocity law implies the equality $S_{12} = S_{21}$ and the conservativity of the system implies that S *is an inner matrix*, i.e. $S(i\omega)^t \overline{S(i\omega)} = Id$ for all $\omega \in \mathbb{R}$. As the filter is supposed to be a perfect reflector without phase shift at infinite frequencies, we impose $\lim_{z\to\infty} S(z) = Id$.

Chapter 2

Computation of optimal multiband filtering functions

This chapter is essentially a compilation of the following articles: [Bila et al., 2006], [Lunot et al., 2007] and [Lunot et al., 2008]. The purpose is to give efficient ways to compute multiband filtering functions. As stated in the introduction, no existing method is totally satisfactory. We first define the optimal filtering function as the solution of a Zolotarev problem. We therefore study such a solution, and next, give two algorithms to compute it. In this chapter, the results are just given. The proofs will be given in the next chapter.

2.1 Statement of the synthesis problem

Starting from the scattering matrix, we state our problem as a max min problem. We next show that this problem can be divided into easier sub-problems. The characterization of the solution of such a sub-problem is given.

2.1.1 Polynomial structure of the S matrix

We have seen in the previous chapter that the scattering matrix S of a filter has the following properties:

- The entries of S are rational functions analytic in the right half-plane, i.e. analytic in $\{z \in \mathbb{C},\ Re(z) \geq 0\}$,
- S is an inner matrix (i.e. $S(i\omega)^t\overline{S(i\omega)} = Id$ for all $\omega \in \mathbb{R}$),
- $S_{12} = S_{21}$,
- $\lim_{z\to\infty} S(z) = Id$.

For a polynomial p, we denote by $\tilde{p}$ the polynomial given by $\tilde{p}(z) = \overline{p(-\bar{z})}$. Note that we have $\overline{p(i\omega)} = \tilde{p}(i\omega)$ for all $\omega \in \mathbb{R}$. We now give the polynomial structure of the scattering matrix.

Proposition 2.1.1 *If a 2×2 matrix S satisfies the above properties, then there exist polynomials p, q and d and an integer n such that*

$$S = \frac{1}{d} \begin{bmatrix} p & q \\ q & (-1)^n \widetilde{p} \end{bmatrix}.$$

with d and p monic of degree n.
Furthermore :

1. *the roots of d are in the left half-plane $\{z \in \mathbb{C}, \quad Re(z) < 0\}$,*
2. *the degree of q satisfies $d^\circ q \leq n-1$,*
3. *$q = (-1)^{n+1}\widetilde{q}$, and*
4. *$d\widetilde{d} = p\widetilde{p} - (-1)^n q^2$.*

Proof Since the entries of S are rational functions, $\det(S)$ is a rational function. We define the polynomials r and d by

$$\frac{r}{d} = \det(S)$$

with r and d relatively prime and d monic. We denote by n the degree of d. Since S is inner, writing $s = i\omega$ with $\omega \in \mathbb{R}$, we have $S(s)^t\overline{S(s)} = Id$ so

$$\det(S(s)^t)\det(\overline{S(s)}) = 1.$$

Note that $\det(S)$ is not the zero function. We get

$$1 \quad = \quad \det(S(s))\overline{\det(S(s))} = |\det(S(s))|^2 = \frac{r(s)\overline{r(s)}}{d(s)\overline{d(s)}} = \frac{r(s)\widetilde{r}(s)}{d(s)\widetilde{d}(s)}.$$

Thus,

$$r(i\omega)\widetilde{r}(i\omega) = d(i\omega)\widetilde{d}(i\omega) \text{ for all } \omega \in \mathbb{R}$$

and, since a non-zero polynomial has a finite number of roots, we obtain

$$r\widetilde{r} = d\widetilde{d}.$$

Since S is stable, all the roots of d are in the left half-plane. As r and d are relatively prime, their roots are distinct. We therefore deduce that the roots of r are exactly the roots of $\widetilde{d}$. Consequently, there is a complex number γ such that $r = \gamma\widetilde{d}$. Thus $\det(S) = \gamma\frac{\widetilde{d}}{d}$. Furthermore, since $|\det(S)| = 1$ on the imaginary axis, $|\gamma| = 1$.
We denote by C the matrix

$$C = \begin{pmatrix} S_{22} & -S_{12} \\ -S_{21} & S_{11} \end{pmatrix}$$

Since $\det(S) \neq 0$, S is invertible and $S^{-1} = C/\det(S)$ so

$$\gamma\widetilde{d}S^{-1} = dC. \tag{2.1}$$

As $S(s)^{-1} = \widetilde{S}(s)^t$ on the imaginary axis, the entries of S^{-1} and $\widetilde{S}^t$ are equal on an infinity of points. Since these entries are rational functions, they are equal everywhere : $S^{-1} = \widetilde{S}^t$.

S being stable, $\widetilde{S}$ has all its poles in the right half-plane. But $S^{-1} = \widetilde{S}^t$, so $\gamma\widetilde{d}S^{-1}$ also has its poles in the right half-plane. Since the entries of C are, up to a sign, the entries of S, C is stable. Thus, dC is also stable. Therefore, using the equality (2.1), we deduce that dC is a rational matrix with no poles in $\mathbb{C}$. Consequently, dC is a polynomial matrix. Then, we obtain that dS is also a polynomial matrix. We therefore get the existence of polynomials p, q, u and v such that

$$S = \frac{1}{d}\begin{bmatrix} p & q \\ u & v \end{bmatrix}.$$

Furthermore, S is symmetric, so $u = q$. We have $S^{-1} = \widetilde{S}^t$ and $S^{-1} = C/\det(S)$ therefore

$$\frac{\gamma}{\widetilde{d}}\begin{bmatrix} v & -q \\ -q & p \end{bmatrix} = \frac{1}{\widetilde{d}}\begin{bmatrix} \widetilde{p} & \widetilde{q} \\ \widetilde{q} & \widetilde{v} \end{bmatrix}.$$

Thus, we deduce that $v = \overline{\gamma}\widetilde{p}$ and

$$q = -\overline{\gamma}\widetilde{q}. \tag{2.2}$$

We get

$$S = \frac{1}{d}\begin{bmatrix} p & q \\ -\overline{\gamma}\widetilde{q} & \overline{\gamma}\widetilde{p} \end{bmatrix}.$$

Since $S^t\widetilde{S} = Id$, we have

$$\frac{1}{d\widetilde{d}}\begin{bmatrix} p & q \\ -\overline{\gamma}\widetilde{q} & \overline{\gamma}\widetilde{p} \end{bmatrix}\begin{bmatrix} \widetilde{p} & -\gamma q \\ \widetilde{q} & \gamma p \end{bmatrix} = Id.$$

Looking at the first entry of the previous matrix, we obtain $\frac{p\widetilde{p}+q\widetilde{q}}{d\widetilde{d}} = 1$, that is, using (2.2),

$$p\widetilde{p} + q\widetilde{q} = p\widetilde{p} - \gamma q^2 = d\widetilde{d}.$$

Since $\lim_{s\to\infty} S(s) = Id$, p is monic of degree n and the degree of q is at most $n-1$. Furthermore, $\lim_{s\to\infty} S_{22}(s) = 1$, so $\overline{\gamma}\widetilde{p}$ is monic of degree n. But the leading coefficient of $\overline{\gamma}\widetilde{p}$ is $\overline{\gamma}(-1)^n z^n$. Therefore, $\gamma = (-1)^n$.

■

In fact, in the previous representation of S, n is the number of resonators (e.g. [Cameron, 1999]). Note that, as $q = (-1)^{n+1}\tilde{q}$, the roots of q are symmetric with respect to the imaginary axis. Therefore, q is, up to a rotation, a polynomial with real coefficients. More precisely $z \mapsto i^{n+1}q(iz)$ is a polynomial with real coefficients.

Using the previous proposition, the squared modulus of the transmission parameter is expressed as

$$\begin{aligned} |S_{21}(i\omega)|^2 &= \left|\frac{q}{d}(i\omega)\right|^2 = \frac{q\tilde{q}}{d\tilde{d}}(i\omega) = \frac{q\tilde{q}}{p\tilde{p}+q\tilde{q}}(i\omega) \\ &= \frac{1}{1+\frac{p\tilde{p}}{q\tilde{q}}(i\omega)} = \frac{1}{1+\left|\frac{p(i\omega)}{q(i\omega)}\right|^2} \\ &= \frac{1}{1+|F(i\omega)|^2} \end{aligned} \tag{2.3}$$

where $F = \frac{p}{q}$ is known as the filtering or characteristic function. In practice, the measurements give values of the filtering function F.

In the case of a single passband, one can show that all the roots of p (respectively q) are real numbers and are in the passband (respectively in the stopband), e.g. see [Le Bailly and Thiran, 1998]. Furthermore, the optimal function is equiripple in the bands, i.e. there are $d^\circ p + 1$ points in the passband where the maximum is reached, and $d^\circ q + 1$ points in the stopband where the minimum is reached.
For given transmission zeros (i.e. q is fixed), a formula using the arccosh function allows the computation of a polynomial p that yields an equiripple filtering characteristic ([Cameron, 1999]). The latter formula in fact gives the solution to the so-called third Zolotarev optimization problem that, roughly speaking, specifies in mathematical terms the notion of a "best" filtering function for a bandpass filter. Whereas in the multi-band situation explicit formulas no longer exist for F, we show in the following that the original Zolotarev problem adapted to a single passband can easily be extended to take into account several passbands and stopbands.

2.1.2 Zolotarev problem

Let $I_1, \ldots, I_r$ and $J_1, \ldots, J_s$ be a collection of $r + s$ finite closed intervals on the real axis, non reduced to a point. The intervals $(I_i)_{1 \le i \le r}$ represent the pass-bands whereas $(J_i)_{1 \le i \le r}$ represent the stop-bands. Therefore, they are disconnected two by two. We note I the union of all the pass-bands and J the union of the stop-bands:

$$I = \bigcup_{i=1}^{r} I_i \text{ and } J = \bigcup_{i=1}^{s} J_i.$$

The "best" multi-band response is such that the transmission and the reflection are as big as possible respectively on the pass-bands I and on the stop-bands J. Since the system is conservative ($|S_{11}|^2 + |S_{12}|^2 = 1$), this is equivalent to saying that the modulus of the transmission is as big as possible in the pass-bands I and as small as possible in the stop-bands J. Using the expression of the transmission (see equation (2.3)), the correct way to formulate the previous problem is to maximize the following ratio:

$$\max_{(p,q) \in \mathcal{P}_n(\mathbb{C}) \times \mathcal{P}_m(\mathbb{R})} \frac{\min\limits_{\omega \in J} \left| \frac{p}{q}(\omega) \right|}{\max\limits_{\omega \in I} \left| \frac{p}{q}(\omega) \right|}$$

where $\mathcal{P}_k(\mathbb{K})$ is the set of polynomials of degree at most k with coefficients in $\mathbb{K}$. If a pair (p, q) which maximizes the above ratio is found, a multiple of this pair also maximizes it. Therefore, we choose to normalize the ratio by assuming that $\max_{\omega \in I} |p/q(\omega)| = 1$. Thus, we obtain the following normalized optimization problem specifying what the best filtering function is

$$\text{find } (p, q) \text{ solution of:} \quad \max_{(p,q) \in R_m^n} \min_{\omega \in J} \left| \frac{p}{q}(\omega) \right| \tag{2.4}$$

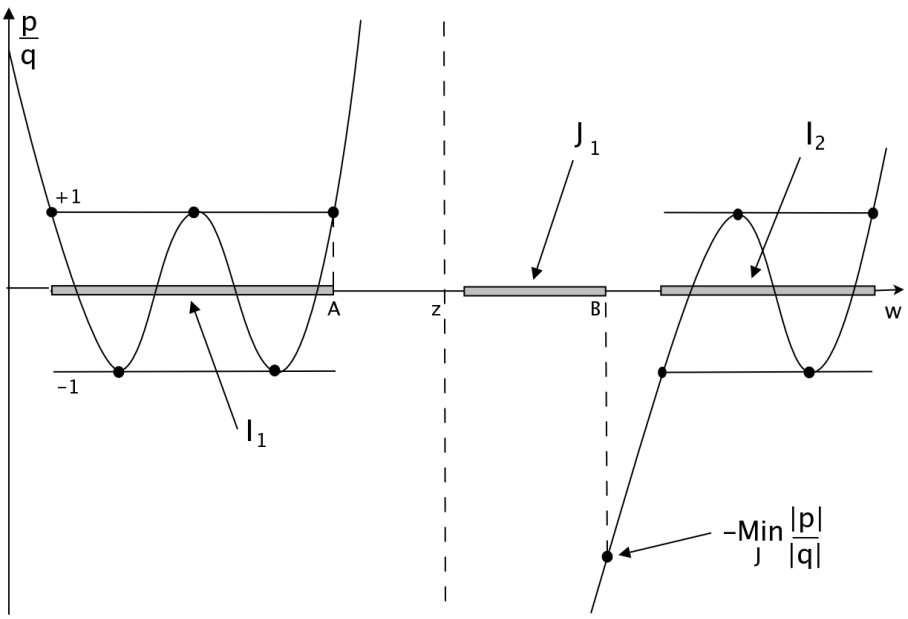

Figure 2.1: Graph of a function p/q in R_m^n for the case of two passbands I_1, I_2 and one stopband J_1.

where R_m^n is the set of the rational functions of numerator (resp. denominator) degree at most n (resp. m) bounded by 1 in the pass-bands:

$$R_m^n = \left\{ (p,q) \in \mathcal{P}_n(\mathbb{C}) \times \mathcal{P}_m^*(\mathbb{R}), \left\| \frac{p}{q} \right\|_I \leq 1 \right\}$$

and $\|.\|_I$ is the sup norm over the set I. Fig. 2.1 gives an example where $I = I_1 \cup I_2$, $J = J_1$, $n = 7$ and $m = 1$.

Since the constant polynomial 1 is in R_m^n and has a minimum equal to 1 in J, an optimal solution P/Q of the problem (2.4) has a criterion $\min_{\omega \in J} |p/q(\omega)|$ at least equal to 1. Therefore, $P \neq 0$, and we can assume that P is monic. Then setting $p(s) = i^n P(-is)$ and $q(s) = \frac{i}{\epsilon} Q(-is)$ yields a scattering matrix with the lowest possible transmission in all the stopbands J_i, provided $|S_{21}|^2 \geq \frac{1}{1+\epsilon^2}$ in the passbands I_i.

2.1.3 Real Zolotarev problem

In this work, we consider solving problem (2.4) under the additional condition that p is a polynomial with real coefficients. Therefore, the problem in which we are interested is

$$\text{find } (p,q) \text{ solution of:} \quad \max_{(p,q) \in \mathcal{R}_m^n} \min_{\omega \in J} \left| \frac{p}{q}(\omega) \right| \tag{2.5}$$

with

$$\mathcal{R}_m^n = \left\{ (p,q) \in \mathcal{P}_n(\mathbb{R}) \times \mathcal{P}_m^*(\mathbb{R}), \left\| \frac{p}{q} \right\|_I \leq 1 \right\}.$$

and we give the following definition of an optimal filtering function:

Definition 2.1.2 *A filtering function $\frac{p}{q}$ is said to be optimal if it is a solution of the real Zolotarev problem 2.5.*

In particular, considering the real problem implies that the synthesized scattering matrix satisfies $S_{1,1} = S_{2,2}$, which is clearly an extra condition. On the one hand the latter guarantees, for example, that the response can be synthesized in a cul-de-sac topology, but on the other hand the solution to the "complex" Zolotarev problem can achieve better results (because less restricted).

2.1.4 Sign combinations and characterization of the solution

Sign combinations

Our goal is now to eliminate the absolute value in (2.5) to get a "linear" version of the problem. If $\frac{P}{Q}$ is an optimal solution of (2.5) and is irreducible (i.e. $\gcd(P,Q) = 1$) then, as the value of the max min in (2.5) is positive, P has no zero in J and, as the absolute value of $\frac{P}{Q}$ is bounded by one over I, Q has no zero in I. Therefore, P has constant sign in every interval J_j and Q has constant sign in every interval I_i. So there exists a sign function σ (such that $\sigma(\omega) = \pm 1$) that is constant in every interval I_i and J_j such that $\frac{P}{Q}$ has a representative in the convex set

$$\mathcal{A}_m^n = \left\{ (p,q) \in \mathcal{P}_n \times \mathcal{P}_m^*, \quad \forall \omega \in J : p(\omega)\sigma(\omega) \geq 0, \forall \omega \in I : q(\omega)\sigma(\omega) \geq 0, \left\| \frac{p}{q} \right\|_I \leq 1 \right\}.$$

Of course, we do not know the signs in advance, but there are only a finite number of possible combinations of them. For every combination of signs on the intervals, we therefore define a signed version of (2.5) by

$$\text{find } (p,q) \text{ solution of:} \qquad \max_{(p,q)\in\mathcal{A}_m^n} \min_{\omega\in J} \frac{\sigma(\omega)p(\omega)}{q(\omega)}. \tag{2.6}$$

Solving (2.6) for all possible sign combinations and retaining the overall best solution yields an optimal solution of (2.5).
If $m > 0$, the number of different possible choices of sign is $2^{\text{number of intervals}}$. However, as $\left|\frac{p}{q}\right| = \left|\frac{-p}{q}\right| = \left|\frac{p}{-q}\right| = \left|\frac{-p}{-q}\right|$, we can only consider $2^{\text{number of intervals}}/4 = 2^{\text{number of intervals}-2}$ choices of sign. We choose the convention that the signs on the first pass-band and on the first stop-band are positive.
If $m = 0$, only the signs over the intervals J_i have to be taken in account, and therefore, the number of different possible choices of sign is $2^{\text{number of stopbands } J_i - 1}$.

For example, suppose that we want to compute the "best" filtering function of a filter with three stop-bands J_1, J_2, J_3 and two pass-bands I_1, I_2. In this case, the number of different possible choices of sign is $2^{\text{number of intervals}-2} = 2^3 = 8$. Then, the eight possible

choices of sign are :

$$\begin{array}{ccccc} J_1 & I_1 & J_2 & I_2 & J_3 \\ + & + & + & + & + \\ + & + & + & + & - \\ + & + & + & - & - \\ + & + & + & - & + \\ + & + & - & - & + \\ + & + & - & - & - \\ + & + & - & + & - \\ + & + & - & + & + \end{array}.$$

Computing the solution of the problem for this eight choices of sign, and taking the overall "best" result yields the solution to the original problem.

Note that for a tri-band filter (three pass-bands and four stop-bands), the number of choices of sign to consider is equal to $2^5 = 32$. Suppose you want to compute the filtering function of a 10-band filter, then you have to consider $2^{19} = 524288$ choices of sign. We can see here the biggest drawback of this theory: we will never be able to compute a filter with numerous bands using this method. However, in practice, we are usually interested in dual-band filters, and from time to time in tri-band filters, for which the amount of signed problems to solve is quite low.

In the following, we will denote by J^+, J^-, I^+ and I^- the union of intervals J_i defined by

$$\begin{array}{ll} J^+ = \cup_{i=1}^r \{J_i, \sigma(J_i) = 1\}, & J^- = \cup_{i=1}^s \{J_i, \sigma(J_i) = -1\}, \\ I^+ = \cup_{i=1}^r \{I_i, \sigma(I_i) = 1\}, & I^- = \cup_{i=1}^s \{I_i, \sigma(I_i) = -1\}. \end{array}$$

In order to obtain the all pole case (i.e. $m = 0$), the polynomial q has to be taken equal to 1, and the signs are only considered in the intervals J_i.

Characterization of the solution

Imagine that we are trying a numerical method to compute the solution of the polynomial sub-problem defined by

- $n = 7$, $m = 0$
- $I = [-1, -0.3] \cup [0.5, 1]$,
- $J^+ = [-5, -1.1] \cup [-0.2, 0.4]$,
- $J^- = [1.1, 5]$,

and that we obtain the result in Fig. 2.2.

In an optimization process, numerical problems often happen, therefore checking the veracity of the result whenever it is possible is only good sense. However, looking at the previous result, it seems difficult to say whether it is good or not. In fact, intuitively, we could expect a better solution by improving the oscillation in the left pass-band. For this reason, being able to check whether a function is optimal or not seems to be useful.

For a given sign function σ, we now give a way of testing whether a rational function of "full rank" (where no simplification between numerator and denominator occurs) is a

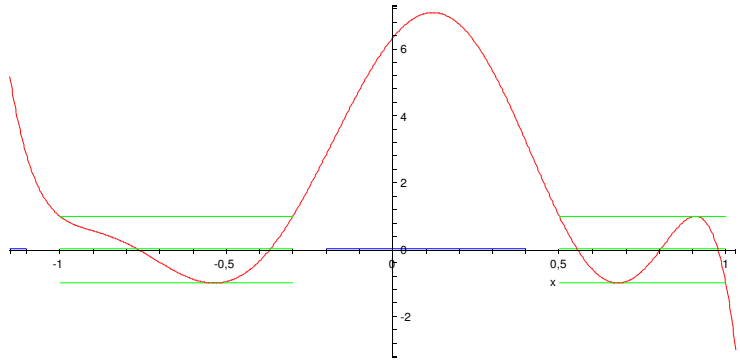

Figure 2.2: Best polynomial of degree at most seven?

solution of (2.6). The latter is based on an alternation property. Let λ be the value of the minimum of $\left|\frac{p}{q}\right|$ on J. We define the following sets of "extreme" points:

$$E^+(p,q) = \left\{\omega \in I, \frac{p}{q}(\omega) = 1\right\} \cup \left\{\omega \in J, \frac{p}{q}(\omega) = -\lambda\right\}$$

and

$$E^-(p,q) = \left\{\omega \in I, \frac{p}{q}(\omega) = -1\right\} \cup \left\{\omega \in J, \frac{p}{q}(\omega) = \lambda\right\}.$$

In Fig. 2.1, ten "extreme" points (6 in E^+ and 4 in E^-) are plotted.

Definition 2.1.3 *A sequence of consecutive points ($\omega_1 < \omega_2 < \cdots < \omega_k$) is called "alternant" if its points belong alternatively to the sets $E^+(p,q)$ and $E^-(p,q)$.*

In Fig. 2.1, an alternant sequence of nine consecutive points can be found (points A and B belong to the same set and cannot therefore appear consecutively in an alternating sequence). "Extreme" points allow us to determine whether a function is the solution of Problem (2.6) or not. Indeed, the following holds (the proof is given in the next chapter):

Theorem 2.1.4 *The maximization problem (2.6) admits a unique solution. Furthermore, $\frac{P}{Q}$ is an optimal solution of "full rank" if and only if there exists a sequence of $N+2$ "alternant" frequency points with $N = m + n$.*

The alternant sequence is therefore a proof of optimality for a given filtering function.

In the single band case, the characterization we gave is equivalent to the classical equiripple property in the passbands and stopbands. However, in the multi-band case, this is no longer true in general. Indeed, look again at Fig. 2.2: we can check that there are nine alternant points, seven in the pass-bands for which the value of the function is ± 1 and two in the stop-bands at -0.2 and -1.1 (see Fig. 2.3). Therefore, this function is the solution of the given signed problem, but is not equiripple.

We now give another example where the solution is not equiripple. Fig. 2.4 shows the optimal $6-4$ function (considering all the possible combinations of sign) for the stopbands

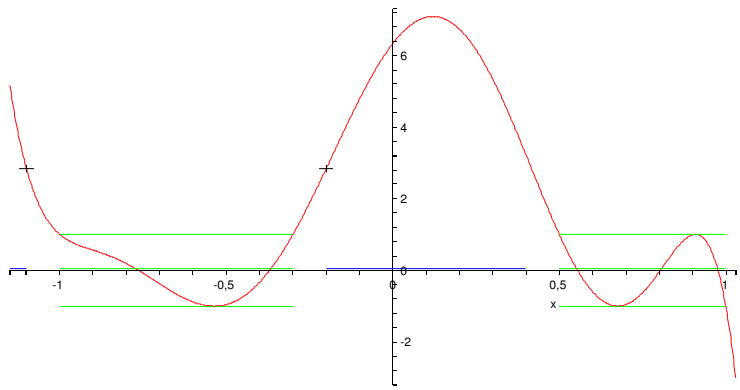

Figure 2.3: Optimal polynomial of degree seven!

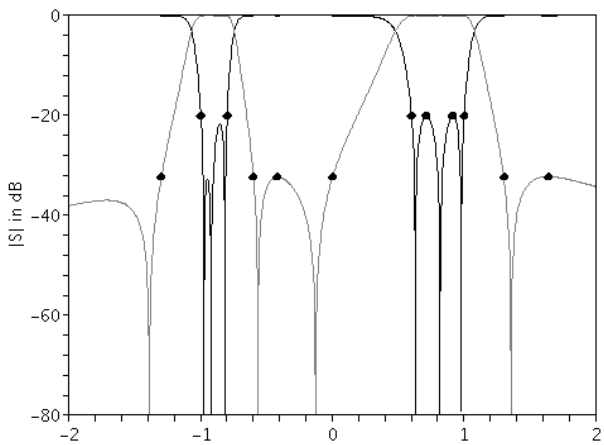

Figure 2.4: Optimal but non-equiripple filtering function with 6 poles and 4 zeros (transmission in grey, reflection in black).

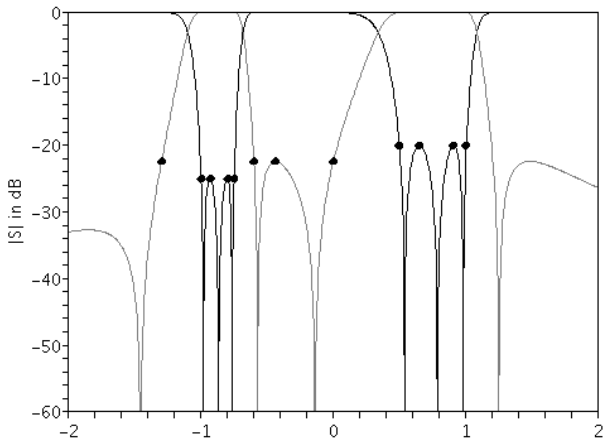

Figure 2.5: Optimal 6-4 response with enlarged passbands and unequal return loss levels in the passbands.

$[-2; -1.3]$, $[-0.6; 0]$, $[1.3; 2]$ and for the passbands $[-1; -0.8]$, $[0.6; 1]$. The attenuation level attained in the stop bands is of 32.2 dB whereas the return loss is set to 20 dB. The twelve "extreme" points confirm that this $6-4$ non-equiriple function is the optimal solution (at least for the considered combination of sign) with respect to the specifications. However, one might enlarge a bit the pass-bands and try to obtain an equiripple response with different return loss levels in the passbands. This was done by solving the problem with the following passbands $[-1; -0.75]$ and $[0.5; 1]$ and return loss levels of respectively 25 dB and 20 dB. As shown on Fig. 2.5, the optimal frequency response for these new specifications is equirriple. These new specifications are harder to meet that the preceding ones (larger passbands and higher return loss in one passband) and result in a poorer optimal attenuation level of 22.4 dB. Here again, twelve "extreme" points confirm the optimality of the response (for the considered choice of signs).

Another non intuitive result is that the degree of the solution is not always maximum. We give an example in the polynomial case ($m = 0$). Take the intervals:

- $I = [-\sqrt{3}; -1] \cup [1; \sqrt{3}]$,
- $J^- = [-3; -\sqrt{3.6}] \cup [\sqrt{3.6}; 3]$,
- $J^+ = [0; \sqrt{0.4}]$.

and look at the polynomial $-x^2 + 2$ (Figure 2.6).

This function has seven alternant points, therefore it is the solution of the problem for $2 \leq n \leq 5$. This shows that the degree of the solution is not always maximum.

2.2 Algorithms

In this section, we focus on computing the solution of the signed problem for a given signed function σ. We recall that in order to obtain the solution of the original problem,

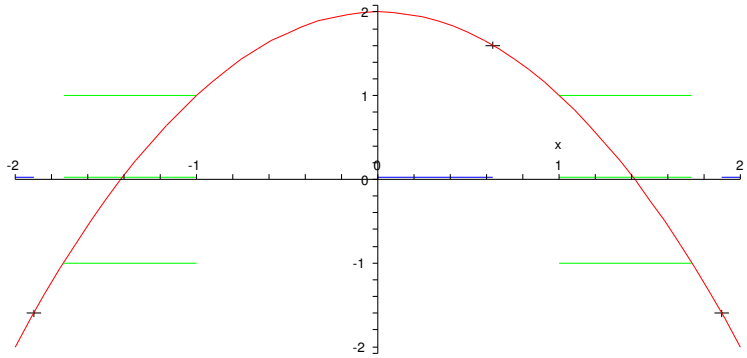

Figure 2.6: solution with a non maximal degree.

$2^{\text{number of bands - 2}}$ such sub-problems have to be solved. Two different algorithms, which are adaptation of classical techniques used in rational approximation ([Cheney, 1998], [Braess, 1986]) are presented here. The first one is a Remes-like algorithm which can only handle all poles functions (i.e. polynomials), but is really effective in this specific case. It is only based on the alternation property verified by the solution. The second algorithm, which is a differential-correction-like algorithm, works in the general case. It uses linear programming.

2.2.1 A Remes-like algorithm for the all pole case

We are now interested in computing the solution when the functions are polynomials. We first enounce the previous results in this particular case. The Zolotarev problem in the all pole case is

$$\text{solve : } \max_{\{p\in\mathcal{P}_n, \|p\|_I\leq 1\}} \min_{\omega\in J} |p(\omega)|$$

Therefore, for a given signed function σ, the sub-problem has the form:

$$\text{find } p \text{ solution of: } \max_{\{p\in\mathcal{P}_n(\mathbb{R}), \|p\|_I\leq 1\}} \min_{\omega\in J} \sigma(\omega)p(\omega). \tag{2.7}$$

We recall that the number of such problems to solve is $2^{\text{number of stop-bands - 1}}$. The following holds :

- the maximization problem (2.7) admits a unique solution,
- P is an optimal solution of (2.7) if and only if there exists a sequence of $n+2$ frequency points $\omega_1 < \omega_2 < \cdots < \omega_{n+2}$ such that its elements belong alternatively to the sets $E^+(P)$ and $E^-(P)$

with

$$E^+(P) = \{\omega \in I, P(\omega) = 1\} \cup \left\{\omega \in J^-, P(\omega) = -\lambda\right\}$$

and

$$E^-(P) = \{\omega \in I, P(\omega) = -1\} \cup \left\{\omega \in J^+, P(\omega) = \lambda\right\}.$$

Note that even if the solution P of the problem is not of maximal degree (i.e. $d^\circ P < n$), it is characterized by a sequence of $n+2$ alternant points.

The exchange algorithm

We now come to an algorithm to solve problem (2.7) for the general multi-band situation. The latter belongs to the family of exchange algorithms first introduced by Remes ([Remes, 1934]) for polynomial approximation. Its main idea is to determine, in an iterative manner, the location of the $n+2$ alternating frequency points. The algorithm is now given, and is next fully detailed on a simple example.

To initialize the algorithm, choose $n+2$ admissible points $\omega_1^0, \ldots, \omega_{n+2}^0$.

The points $\omega_1^0, \ldots, \omega_{n+2}^0$ are admissible if

- at least one point is in I and one point is in J,
- if $\omega_j^0 \in J^+$ (resp. J^-)then $\omega_{j+1}^0 \notin J^+$ (resp. J^-)

Associate to these points values of alternation $\alpha(\omega_1^0), \ldots, \alpha(\omega_{n+2}^0)$.

- if $\omega_j^0 \in J$ then $\alpha(\omega_j^0) = \sigma(\omega_j^0)$,
- if $\omega_j^0 \in I$ then $\alpha(\omega_j^0) = \pm 1$ with the value taken such that the sequence $\alpha(\omega_1^0), \ldots, \alpha(\omega_{n+2}^0)$ is alternated.

Once the initialization is done, repeat the following steps :

1. *Compute p_k on the reference set $\omega_1^k, \ldots, \omega_{n+2}^k$.*

   Let $p_k(\omega) = \sum_{i=0}^n a_i^k \omega^i$.
   We associate to $\omega_1^k, \ldots, \omega_{n+2}^k$ the system of equations:

   $$\begin{cases} p_k(\omega_i^k) = -\alpha(\omega_i^k) \text{ if } \omega_i^k \in I, \\ p_k(\omega_i^k) = \alpha(\omega_i^k)\lambda_k \text{ if } \omega_i^k \in J. \end{cases} \tag{2.8}$$

   Compute the solution of this system with $n+2$ equations and $n+2$ unknowns (the a_i^k and λ_k). We obtain p_k and λ_k.

2. *Look for the "worst" point ω_{worst}.*

   Let $M_k = \max\left(\max_{\omega \in I} |p_k(\omega)| - 1, \max_{\omega \in J} \lambda_k - \sigma(\omega) p_k(\omega)\right)$.
   If $M_k = 0$, the algorithm stops and returns p_k.
   Else take ω_{worst} associated to M_k (i.e. ω_{worst} is a value for which the max in M_k is obtained) and define $\alpha(\omega_{worst})$ as

   $$\begin{cases} \alpha(\omega_{worst}) = sgn(1 - p_k(\omega_{worst})) \text{ if } \omega_{worst} \in I, \\ \alpha(\omega_{worst}) = \sigma(\omega_{worst}) \text{ if } \omega_{worst} \in J. \end{cases}$$

3. *Define a new sequence of $n+2$ points $\omega_1^{k+1}, \ldots, \omega_{n+2}^{k+1}$ by substituting ω_{worst} to one of the ω_i^k in order to keep an alternated sequence.*

We define the index j such that ω_j^k is substituted by ω_{worst} via the following rule:

- if $\omega_i^k < \omega_{worst} < \omega_{i+1}^k$, then if $\alpha(\omega_i^k) = \alpha(\omega_{worst})$, $j = i$, else $j = i + 1$.
- if $\omega_{worst} < \omega_1^k$, then if $\alpha(\omega_1^k) = \alpha(\omega_{worst})$, $j = 1$, else $j = n + 2$.
- if $\omega_{worst} > \omega_{n+2}^k$, then if $\alpha(\omega_{n+2}^k) = \alpha(\omega_{worst})$, $j = n + 2$, else $j = 1$.

This gives a sequence ω'^{k+1}_i :

$$\begin{cases} \omega'^{k+1}_i = \omega_i^k & \forall i \neq j \\ \omega'^{k+1}_j = \omega_{worst} \end{cases}$$

The new sequence ω_i^{k+1} is obtained by sorting in increasing order the ω'^{k+1}_i.

The Remes algorithm gives a sequence of polynomials $(p_i)_{i\in\mathbb{N}}$ which converges to P, solution of the signed problem (2.7). The proof is given in the next chapter.

A detailed example

We now give a detailed example in order to illustrate the Remes algorithm :

$$n = 2, J_1 = [-1.5, -1.3], I_1 = [-1, -0.5], J_2 = [0, 0.5], I_2 = [1, 2], \sigma(J_1) = 1, \sigma(J_2) = -1$$

i.e. $J^+ = J_1, J^- = J_2$.

Step 0: Initial guess for the reference set
We start with an initial guess for the alternating frequencies, for example

$$\begin{array}{llll} \omega_1 = -1.3, & \omega_2 = -1, & \omega_3 = -0.75, & \omega_4 = -0.5 \\ \alpha(\omega_1) = 1, & \alpha(\omega_2) = -1, & \alpha(\omega_3) = 1, & \alpha(\omega_4) = -1. \end{array}$$

Step 1: Solving problem (2.8) on the reference set
On this simple reference set we solve problem (2.8), which means that we look for the polynomial P_0 of degree 2 that has maximal value, say λ_0, in ω_1 under the requirement to remain bounded (in absolute value) by 1 on the other frequencies $\omega_2, \omega_3, \omega_4$.

The alternation property verified by P_0 yields to the following set of linear equations:

$$P_0(-1.3) = \lambda_0, \quad P_0(-1) = 1, \quad P_0(-0.75) = -1, \quad P_0(-0.5) = 1$$

that can be solved for P_0 and λ_0 and lead to $P_0 = 32\omega^2 + 48\omega + 17$ and $\lambda_0 = 8.68$. The resulting polynomial is shown in Fig. 2.7.

Step 2: Determining the point where the polynomial "deviates most"
Obviously the polynomial P_0 does not satisfy the boundedness condition on I_2. We look for the point where our current polynomial "deviates most" from a valid solution either by exceeding the modulus bound on I or by reaching a minimal value on J that is smaller than the current λ_0. More precisely we use the following rule: Let ω_{max} be the point where $|P_0(\omega)|$ is maximal on I, and let ω_{min} a point of J where the minimum of $P_0(\omega)\sigma(\omega)$ is attained. If $|P_0(\omega_{max})| - 1 > \lambda_0 - P_0(\omega_{min})\sigma(\omega_{min})$, take $\omega_{worst} = \omega_{max}$, else $\omega_{worst} = \omega_{min}$.

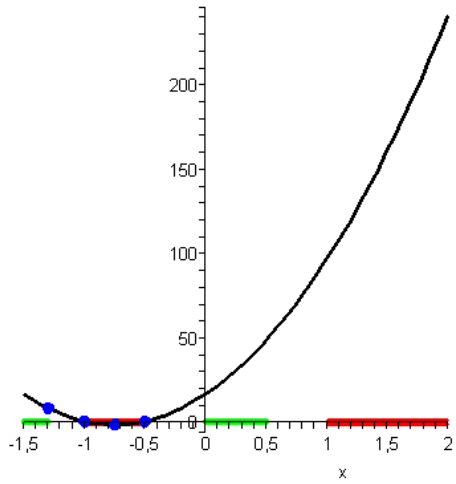

Figure 2.7: Initialization of the exchange algorithm : $\lambda = 8.68, \omega_{max} = 2$.

In the current example $\omega_{worst} = 2$ is selected.

Step 3: Adaptation of the reference set
We now make some change in the reference set $(\omega_1, \omega_2, \omega_3, \omega_4)$ and obtain the following new reference set:

$$\omega_1 = -1.3, \quad \omega_2 = -1, \quad \omega_3 = -0.75, \quad \omega_4 = 2.$$

The inclusion of the new element is performed so as to be able to compute a new alternating polynomial using step 1, see Fig. 2.8.

The latter iterations between step 1 and step 3 are continued until a polynomial P_N is determined that satisfies (in a numerical meaning) the boundedness condition on I and reaches the minimum of (2.7) on J at a frequency point of the reference set (see Fig. 2.9 and 2.10). The polynomial of Figure 2.10 is the optimal solution of the problem (2.7) for the choice of sign σ. The corresponding reference set is

$$\omega_1 = -1.3, \quad \omega_2 = 0, \quad \omega_3 = 1, \quad \omega_4 = 2$$

and satisfies the optimality condition of the preceding section.

In order to determine an optimal solution of the original Zolotarev problem, we also solve the problem (2.7) with the following choice of signs $\sigma(J_1) = 1, \sigma(J_2) = 1$ for which the solution is found to be the constant polynomial 1. Since $1.21 > 1$, the polynomial of Fig. 2.10 is therefore the optimal solution of the original problem.

2.2.2 A differential correction-like algorithm for the rational case

Now we present an algorithm which is an adaptation of the differential-correction algorithm used in rational approximation ([Cheney and Loeb, 1961]). Such an algorithm uses linear programming, which is the topic of the following section.

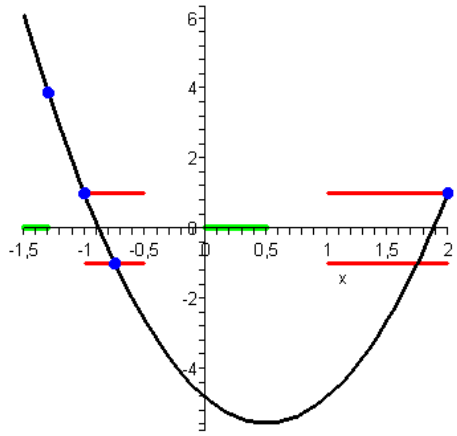

Figure 2.8: Iteration 1 : $\lambda = 3.88, \omega_{max} = 1$.

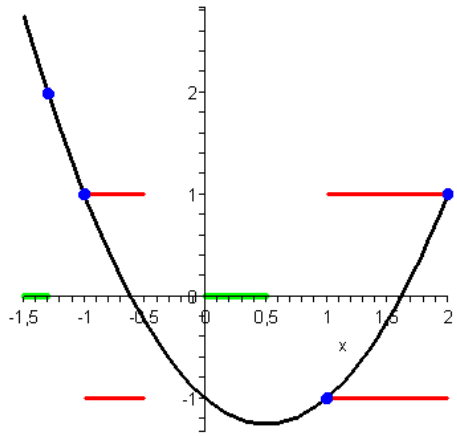

Figure 2.9: Iteration 2 : $\lambda = 1.99, \omega_{min} = 0$.

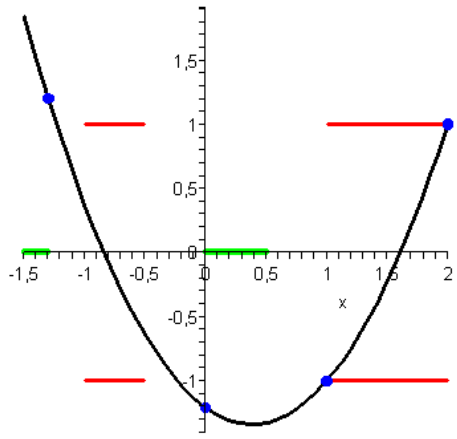

Figure 2.10: Iteration 3 : $\lambda = 1.21$, all constraints satisfied.

Linear programming and polynomial approximation problems

This section is meant as a short tutorial on the use of linear programming in connection with polynomial approximation problems like the one we just stated. Suppose we only have one stop band $J = [1.1, 2]$ and one pass band $I = [-1, 1]$. We are interested in the all-pole filter of order 2 that solves the related Zolotarev problem, i.e. among all polynomials of degree ≤ 2 that are bounded by 1 on I find the one with the fastest growth on J. The solution to this problem is known to be the Chebychev polynomial $P(x) = \cos(2\arccos(x)) = 2x^2 - 1$ (see [Rivlin, 1990]). We will now see that this result can be recovered from a numerical algorithm. The advantage of this procedure is that it will extend to multi-band situations for which closed form formulas are not known. Once a sign has been chosen for the polynomial $P = ax^2 + bx + c$ in J (say positive), the original Zolotarev problem can be formulated as the following optimization problem:

find a, b and c such that μ is maximal, with

$$\begin{cases} \forall x \in J, & \mu \leq ax^2 + bx + c, \quad (i) \\ \forall x \in I, & 1 \geq ax^2 + bx + c, \quad (ii) \\ \forall x \in I, & -1 \leq ax^2 + bx + c. \quad (iii) \end{cases}$$

Here, μ is an auxiliary variable which expresses the minimum of the polynomial over J. Evaluating inequality (i) at sample points in the interval J and inequalities (ii-iii) at sample points in the interval I yields a set of linear inequalities in the variables (a, b, c, μ). In this way, the original Zolotarev problem is cast into a linear optimization problem with linear constraints: a linear program (LP for short). These kinds of problems have been widely studied and efficient software to solve them exists (e.g. Cplex, lp_solve, Matlab, Maple). Using the LP solver of Matlab and taking 100 sample points over the intervals I and J yields the following solution: $a = 2.0002$, $b = 10^{-12}$, $c = -1.0002$. The advantage of this method as compared to closed form formulas is that it can be generalized to any

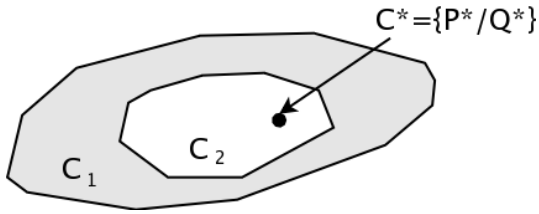

Figure 2.11: Sets $\mathcal{C}(\mu)$ for $\mu_1 < \mu_2 < \mu^*$ ($C_1 := \mathcal{C}(\mu_1)$, $C_2 := \mathcal{C}(\mu_2)$ and $C^* := \mathcal{C}(\mu^*)$).

number and any arrangement of the intervals I and J.
In the following, the general problem of filters with transmission zeros at finite frequencies is tackled. This amounts to dealing with rational fractions instead of polynomials. The general algorithmic framework remains however similar and relies in particular on the use of linear programming.

Geometry of the sub-problem

We will now study problem (2.6) from a geometric point of view. If we denote by μ the value of the criterion min in (2.6) for a given (p, q) (μ can be seen as the rejection level of $\frac{p}{q}$ in the stopbands) then the convex set $\mathcal{C}(\mu)$ defined by

$$\mathcal{C}(\mu) = \{(p, q) \in \mathcal{A}_m^n, \forall \omega \in J : \sigma(\omega)p(\omega) - \mu|q(\omega)| \geq 0\}$$

is in a way the set containing all the functions which have at least a rejection level μ in the stopbands. Let μ^* be the value of the criterion max min in (2.6) (μ^* is the best possible rejection). Then, by definition of the max, $\mathcal{C}(\mu^*)$ is the set of representatives of the optimal function $\frac{P^*}{Q^*}$. The key point for computing the solution of problem (2.6) is that, for $\mu_1 < \mu_2 < \mu^* < \mu_3$, the following holds (see Fig. 2.11) :

- $\mathcal{C}(\mu_3) = \emptyset$,
- $\mathcal{C}(\mu^*) \cong \left\{\frac{P^*}{Q^*}\right\}$ (i.e. $\mathcal{C}(\mu^*)$ is the set of representatives of P^*/Q^*),
- $\mathcal{C}(\mu^*) \subset \mathcal{C}(\mu_2) \subset \mathcal{C}(\mu_1)$.

Indeed, by making an hypothesis on the possible rejection level μ and by checking the emptiness of $\mathcal{C}(\mu)$, the following information on μ^* is known :

- if $\mathcal{C}(\mu)$ is empty, $\mu^* < \mu$,
- if $\mathcal{C}(\mu)$ is non-empty, $\mu^* \geq \mu$.

Therefore, a dichotomy method testing emptiness can be used to compute the optimal rational filtering function. It is crucial to notice that the convexity of the set $\mathcal{C}(\mu)$ allows to check non-emptiness using linear programming.

Detailed Equations for Checking Emptiness

For a criterion μ, let f_μ be the following function:

$$f_\mu(p,q) = \min_{\omega\in J} \left(\sigma(\omega)p(\omega) - \mu|q(\omega)|\right).$$

Note that $\mathcal{C}(\mu) = \{(p,q) \in \mathcal{A}_m^n, \quad f_\mu(p,q) \geq 0\}$. f_μ is continuous and $\mathcal{A}_m^n$ is compact. Therefore, one way of checking emptiness of $\mathcal{C}(\mu)$ is to find (p,q) in $\mathcal{A}_m^n$ which maximizes the function f_μ.

Computation can be done by discretising the I and J intervals. Indeed, in this way, the equations of the constraints in $\mathcal{A}$ become linear in the coefficients of p and q. More precisely, the problem of finding (p,q) is done by solving the LP problem :

$$\text{solve : } \max h \tag{2.9}$$

subject to

$$\begin{cases} \forall y_j, & \sigma(y_j)p(y_j) - \mu q(y_j) \geq h, \\ \forall y_j, & \sigma(y_j)p(y_j) + \mu q(y_j) \geq h, \\ \forall x_j, & \sigma(x_j)q(x_j) \geq 0, \\ \forall x_j, & -\sigma(x_j)q(x_j) \leq p(x_j) \leq \sigma(x_j)q(x_j), \end{cases}$$

where (x_j) (resp. (y_j)) are a discretization of I (resp. J). If the maximum h is positive, then (p,q) in $\mathcal{A}_m^n$ which maximizes f_μ has been computed, therefore the set $\mathcal{C}(\mu)$ is non-empty. Else, if $h < 0$, the set $\mathcal{C}(\mu)$ is empty. Accuracy depends of course of the number and placement of chosen points.

Differential Correction-Like Algorithm

Instead of using dichotomy as suggested previously, we now come to an algorithm which adjusts μ in a more efficient way by using the information gained from solving (2.9).

Initialization : Choose polynomials (p_0, q_0) in $\mathcal{A}_m^n$. Compute

$$\mu_0 = \min_{\omega\in J} \left|\frac{p_0}{q_0}(\omega)\right|.$$

Then repeat :

Compute (p_k, q_k) which solves the LP problem (2.9) for $\mu := \mu_{k-1}$:

$$f_{\mu_{k-1}}(p_k, q_k) = \max_{(p,q)\in\mathcal{A}_m^n} f_{\mu_{k-1}}(p,q).$$

If $f_{\mu_{k-1}}(p_k, q_k) \leq 0$ return (p_{k-1}, q_{k-1}) else compute

$$\mu_k = \min_{\omega\in J} \left|\frac{p_k}{q_k}(\omega)\right|.$$

In our case, as we use a discretization of I and J, the computation is done over finite sets. This ensures that the sequence of criterion $(\mu_k)_k$ converges toward the optimal criterion μ^* (the proof is given in the next chapter).

Chapter 3

A generalized Zolotarev problem

We now give the proofs of the results mentioned in the previous chapter. The Zolotarev problem is extended to a problem with weight and general constraints. This extension allows, in particular, to compute the optimal filtering function with respect to some specifications. We first study the polynomial case, and next, the rational case. The techniques employed are adapted from polynomial and rational approximation (see for example [Rivlin, 1990], [Powell, 1981], [Cheney, 1998] or [Braess, 1986]). Since the problem has been introduced in the previous chapter, we explain very briefly how it is extended, and next, we study the related sub-problems, which are generalizations of (2.6).

3.1 A polynomial Zolotarev problem

In the previous chapter, the polynomial problem was

$$\text{solve:} \max_{\{p\in\mathcal{P}_n(\mathbb{R}),\|p\|_I\leq 1\}} \min_{\omega\in J} |p(\omega)|.$$

This problem can be formulated as

$$\text{solve:} \max_{\{p\in\mathcal{P}_n(\mathbb{R}),-1\leq p(\omega)\leq 1 \text{ for } \omega\in I\}} \min_{\omega\in J} \max(p(\omega)-0, 0-p(\omega)).$$

We now introduce two continuous functions l and u in order to generalize the constraints. We also add a nonnegative "weight" $\frac{1}{|Q|}$, where Q is a given polynomial. The problem becomes

$$\text{solve:} \max_{\{p\in\mathcal{P}_n(\mathbb{R}),\,\forall\omega\in I, l(\omega)\leq\frac{p(\omega)}{|Q|(\omega)}\leq u(\omega)\}} \min_{\omega\in J} \max\left(\frac{p(\omega)}{|Q|(\omega)}-l(\omega), u(\omega)-\frac{p(\omega)}{|Q|(\omega)}\right). \tag{3.1}$$

The notation l (resp. u) is chosen because l (resp. u) is a lower bound (resp. upper bound) for $p/|Q|$ on I. We therefore assume $l\leq u$ on I.
However, we assume that $l\geq u$ on J. As we are going to see, the solution is then bounded above by u on some intervals of J, and bounded below by l on the other intervals of J. Indeed, suppose that the problem (3.1) has a solution p^* such that $\gcd(p^*,Q)=1$ and

$$\min_{\omega\in J} \max\left(\frac{p^*(\omega)}{|Q|(\omega)}-l(\omega), u(\omega)-\frac{p^*(\omega)}{|Q|(\omega)}\right) > 0.$$

Therefore, on each interval of J, either $\frac{p^*}{|Q|} > l$, or $\frac{p^*}{|Q|} < u$. Thus, as mentioned in the previous chapter, the problem can be divided in sub-problems. For each sub-problem, we choose on each interval of J to maximize the minimum of either $\frac{p^*}{|Q|} - l$ or $u - \frac{p^*}{|Q|}$.

We now introduce the notations and the hypotheses made in order to study such a sub-problem.

3.1.1 Notations

We choose to work in the Alexandroff compactification of $\mathbb{R}$, denoted by $\widehat{\mathbb{R}} = \mathbb{R} \cup \{\infty\}$. Therefore, the possibility of intervals of J of infinite length is considered. In order to handle such an extended problem, new considerations have to be done. This is the subject of this section, where all the notations and hypotheses of work are given.

Maximum degree: n.
n is a positive integer.

Pass-bands and stop-bands: I, J.
I, J^+ and J^- are three distinct closed subsets of $\widehat{\mathbb{R}}$ such that

- I is a compact set of $\mathbb{R}$ which contains at least $n+1$ points,
- $J^+ \cap J^- = \emptyset$
- $J = J^+ \cup J^-$ is non-empty,
- $I \cap J \neq I$ and $I \cap J \neq J$,
- $\overbrace{I \cap J}^{\circ} = \emptyset$.

Furthermore, we suppose that the parity of n is in agreement with the "unboundedness" of J^- and J^+. For example, we will never try to compute the optimal polynomial of degree at most 8 with values $-\infty$ at $-\infty$ and $+\infty$ at $+\infty$, because this polynomial cannot be of degree 8. We make the convention that

- if n is even, then J^+ is a compact set of $\mathbb{R}$ or J^- is a compact set of $\mathbb{R}$,
- if n is odd, J^+ and J^- are bounded on the right or on the left (one of them is bounded on the right and the other on the left).

We denote by X the union of I and J.

Note that, the main differences with Chapter 2 is that I and J are not unions of intervals but closed sets, J is not necessary bounded, and the intersection of I and J is not necessary empty. The case $I \cap J \neq \emptyset$ can be considered by a modification of the constraints. One of the interests of such a generalization is, when considering the "simple" problem, to impose a constraint between the intervals, for example a constraint of positivity.

Constraints or specifications: l, u.
l and u are two functions from X into $\overline{\mathbb{R}}$ such that :

- $l(x) < u(x)$ for all x in I,

- $-\infty < l(x) < +\infty$ for all x in J^+,
- $-\infty < u(x) < +\infty$ for all x in J^-,
- $I \cap \{u = +\infty\} \cap \{l = -\infty\} = \emptyset$,
- $I \setminus (\{u = +\infty\} \cup \{l = -\infty\})$ is a compact set of $\mathbb{R}$ which contains at least $n+1$ points,
- l is continuous over J^+ and $I \setminus \{l = -\infty\}$,
- u is continuous over J^- and $I \setminus \{u = +\infty\}$,
- if $x \in I \cap J^+$ then $u(x) = +\infty$, and if $x \in I \cap J^-$ then $l(x) = -\infty$,

We denote by I_l^u the set $I \setminus (\{u = +\infty\} \cup \{l = -\infty\})$, by $I_{-\infty}^u$ the set $I \cap \{l = -\infty\}$ and by $I_l^{+\infty}$ the set $I \cap \{u = +\infty\}$.

Note that in Chapter 2, $u = 1$ and $l = -1$ on I, $u = l = 0$ on J, and $I = I_l^u$. Note also that on I_l^u, the two constraints are active, but on $I \setminus I_l^u$, only one constraint is active.

"Weight" or fixed denominator: $\mathcal{Q}$.
$\mathcal{Q}$ is a function from X into $\mathbb{R}^+$ such that $\mathcal{Q} = |q|g$ with g a positive continuous function over X and q a non-zero polynomial such that:

- $\mathcal{Z}_{\mathcal{Q}} \cap I_l^u = \emptyset$,
- $\overline{J \setminus \mathcal{Z}_{\mathcal{Q}}} = J$,

where $\mathcal{Z}_{\mathcal{Q}}$ is the set of all the roots of q.
When J is unbounded, we also assume that $\mathcal{Q}$, $u\mathcal{Q}$ and $l\mathcal{Q}$ are negligible with respect to $x \mapsto x^n$ (at ∞). The hypothesis on $\mathcal{Q}$ is made to avoid $\lim_{x\to\pm\infty} \left|\frac{p}{\mathcal{Q}}\right| = 0$ for every $p \in \mathcal{P}_n$. The hypotheses on $u\mathcal{Q}$ and $l\mathcal{Q}$ ensure the compactness of the set of "extreme" points.

Note that in Chapter 2, $\mathcal{Q} = 1$.

Admissible polynomials: $\mathcal{A}$.

$$\mathcal{A} = \left\{ p \in \mathcal{P}_n, \forall x \in I, l(x) \leq \frac{p(x)}{\mathcal{Q}(x)} \leq u(x) \right\}.$$

Note that $\mathcal{A}$ is a compact convex set because I_l^u contains at least $n+1$ points.

"distance" of p: $\delta(p)$.
For $p \in \mathcal{P}_n$, we define $\delta(p) : X \longrightarrow \overline{\mathbb{R}}$ by

$$\delta(p)(x) = \begin{cases} \min\left(u(x) - \frac{p(x)}{\mathcal{Q}(x)}, \frac{p(x)}{\mathcal{Q}(x)} - l(x)\right) & \text{if } x \in I_l^u, \\ \frac{p(x)}{\mathcal{Q}(x)} - l(x) & \text{if } x \in J^+ \cup I_l^{+\infty}, \\ u(x) - \frac{p(x)}{\mathcal{Q}(x)} & \text{if } x \in J^- \cup I_{-\infty}^u. \end{cases}$$

Criterion or minimal distance over J from p to the constraints: μ_p.
We denote by $\mu_p \in \overline{\mathbb{R}}$ the value

$$\mu_p = \inf_{x \in J} \delta(p)(x).$$

Note that $p \longmapsto \mu_p$ is concave (but not strictly).

3.1.2 The polynomial problem

We are interested in solving the following problem :

$$\text{Find (whenever it exists) a polynomial } p^* \in \mathcal{A} \text{ such that } \max_{p \in \mathcal{A}} \mu_p = \mu_{p^*} \tag{3.2}$$

We first check the existence of a solution when $\mathcal{A}$ is non-empty.
If for every polynomial p in $\mathcal{A}$, $\mu_p = -\infty$, then every polynomial is solution. We now suppose that a polynomial p_0 with finite criterion exists (i.e. $\mu_{p_0} > -\infty$). We denote by μ^* the upper bound of the set $\{\mu_p, p \in \mathcal{A}\}$ and by $(\mu_{p_i})_{i \in \mathbb{N}}$ a sequence which converges to μ^*. Since $\mathcal{A}$ is a compact set, we can suppose, without loss of generality, that the sequence $(p_i)_{i \in \mathbb{N}}$ converges (this is true for at least a subsequence). We denote by $\check{p}$ the limit of (p_i).
Let $x \in J^+ \setminus \mathcal{Z}_{\mathcal{Q}}$.
As $\delta(p_i)(x) \geq \mu_{p_i}$, we have $p_i(x) \geq (l(x) + \mu_{p_i})\mathcal{Q}(x)$. Therefore, $\check{p}(x) \geq (l(x) + \mu^*)\mathcal{Q}(x)$, and we get $\delta(\check{p})(x) \geq \mu^*$.
If $x \in J^+ \cap \mathcal{Z}_{\mathcal{Q}}$, then either $\check{p}/\mathcal{Q}$ is continuous at x and the result is still true by continuity because $\overline{J \setminus \mathcal{Z}_{\mathcal{Q}}} = J$, or $\check{p}(x)/\mathcal{Q}(x) = +\infty$ and then $\delta(\check{p})(x) = +\infty \geq \mu^*$.
The same argument holds if $x \in J^-$. Thus, we obtain $\mu_{\check{p}} \geq \mu^*$ and we therefore conclude that $\check{p}$ is a solution.

In the following, we assume the non-emptiness of $\mathcal{A}$ and the existence of a polynomial $p \in \mathcal{A}$ such that $\mu_p > -\infty$.

3.1.3 Characterization of the solution

In the case of polynomial uniform approximation, it is well known (e.g. [Cheney, 1998]) that the best approximation of degree n to f, denoted p^*, is characterized by an alternation property, that is by the existence of $n+2$ points $x_1, \ldots, x_{n+2}$ such that

$$f(x_i) - p^*(x_i) = p^*(x_{i-1}) - f(x_{i-1}) = \pm \|f - p^*\|_\infty \quad \text{for } 2 \leq i \leq n+2.$$

As we will see, such kind of alternation property also characterizes the solutions of the Zolotarev problem.

Let $p \in \mathcal{A}$ such that $\mu_p > -\infty$. We associate to p the following sets:

- $E_1^p(u) = \{x \in I, p(x) = u(x)\mathcal{Q}(x)\}$,
- $E_1^p(l) = \{x \in I, p(x) = l(x)\mathcal{Q}(x)\}$,
- $E_1^p = E_1^p(u) \cup E_1^p(l)$,
- $E_2^p = \{x \in J^+, p(x) = (l(x) + \mu_p)\mathcal{Q}(x)\}$,

- $E_3^p = \{x \in J^-, p(x) = (u(x) - \mu_p)\mathcal{Q}(x)\}$,
- $E^p = E_1^p \cup E_2^p \cup E_3^p$.

If J is unbounded and if the degree of p is not maximal (i.e. $d^\circ p < n$), then we add the point ∞, and we denote by $\widehat{E^p}$ the set $E^p \cup \{\infty\}$. If J is bounded or the degree of p is maximal, then $\widehat{E^p} = E^p$.
Note that $\widehat{E^p}$ is a compact of $\hat{\mathbb{R}}$. Indeed, the sets previously defined are the inverse image of $\{0\}$ by continuous functions.

Definition 3.1.1 *An element of $\widehat{E^p}$ is called an extreme point of p.*

We define a map ν_p from $E_1^p \cup J$ into $\{-1, 1\}$ by :

$$\nu_p(x) = \begin{cases} -1 & \text{if } x \in E_1^p(u) \cup J^- \setminus \{\infty\}, \\ 1 & \text{if } x \in E_1^p(l) \cup J^+ \setminus \{\infty\}. \end{cases}$$

and

$$\nu_p(\infty) = \begin{cases} (-1)^n & \text{if } J^+ \quad \text{unbounded on the left}, \\ (-1)^{n+1} & \text{if } J^- \quad \text{unbounded on the left}, \\ +1 & \text{if } J^+ \quad \text{unbounded on the right}, \\ -1 & \text{if } J^- \quad \text{unbounded on the right}. \end{cases}$$

In some way, this function indicates on which direction the polynomial could be improved at its extreme points. For example, if a polynomial in $\mathcal{A}$ reaches the constraint u in I, the only way to locally modify it in order to stay in $\mathcal{A}$ is to decrease its value at this point. This decrease is indicated by the value -1 of ν_p. Similarly, if the reached constraint is l, then the value has to be increased, and this is denoted by $\nu_p = +1$.

We now want to characterize the solutions of the problem (3.2). We introduce to that purpose the following functions.

Definition 3.1.2 *Let $p \in \mathcal{A}$ such that $\mu_p > -\infty$. To each ζ in $E_1^p \cup J$, we associate a map χ_ζ^p from $\mathcal{P}_n(\mathbb{R})$ into $\mathbb{R}$ defined by:*

$$\chi_\zeta^p(h) = \begin{cases} \nu_p(\zeta) \times \frac{h(\zeta)}{1+|\zeta^n|} & \text{if } \zeta \neq \infty, \\ \nu_p(\infty) \times h_n & \text{if } \zeta = \infty \text{ and } h(x) = \sum_{i=0}^n h_i x^i. \end{cases}$$

We call such a map a characterizing function of p at ζ.

Note that the set of all characterizing functions of p is a compact set of the space of linear applications $L(\mathcal{P}_n, \mathbb{R})$.
Indeed, the function $\zeta \mapsto \chi_\zeta^p$ is continuous over $\hat{\mathbb{R}}$ so the sets $\{\chi_\zeta^p, \zeta \in \overline{J} \cap \widehat{E^p}\}$ and $\{\chi_\zeta^p, \zeta \in E_1^p\}$ are compact (as images of compact sets by a continuous function).

Lemma 3.1.3 *Let p^* be a solution of (3.2). There exist distinct points $x_0, \ldots, x_r \in \widehat{E^{p^*}}$ and positive real numbers $\lambda_0, \ldots, \lambda_r$ such that for every polynomial h in $\mathcal{P}_n(\mathbb{R})$,*

$$\sum_{i=0}^r \lambda_i \chi_{x_i}^{p^*}(h) = 0$$

with $r \leq n+1$.

Proof The set of all characterizing functions of p^* is a compact set (see the remark before lemma). Therefore, its convex hull C is compact.

Suppose that $0 \notin C$.
The Hahn-Banach theorem (e.g. [Brezis, 1983]) gives the existence of $\alpha > 0$ and $h \in \mathcal{P}_n \setminus \{0\}$ such that:

$$\forall \zeta \in \widehat{E^{p^*}}, \quad \chi_\zeta^{p^*}(h) > 2\alpha > 0.$$

We next want to show that this hypothesis implies that p^* is not optimal. We construct to that purpose a polynomial $p^* + \epsilon h \in \mathcal{A}$ such that $\mu_{p^*+\epsilon h} > \mu_{p^*}$.

The map $x \longmapsto \frac{h(x)}{1+|x^n|}$ is continuous over $\mathbb{R}$. Therefore, for each $\zeta \in E^{p^*}$, we can take an open interval I_ζ containing ζ and such that

$$\forall x \in I_\zeta, \quad \nu_{p^*}(\zeta)\frac{h(x)}{1+|x^n|} \geq \alpha.$$

For each $\zeta \in \mathcal{Z}_\mathcal{Q} \setminus E^{p^*}$, we can find an open set I_ζ such that $\delta(p^*)(x) \geq 2\alpha$ if $x \in I_\zeta \cap I$ and $\delta(p^*)(x) \geq \mu_{p^*} + 2\alpha$ if $x \in I_\zeta \cap J$.
If the infinity is an extreme point, since $\mathcal{Q}$ is negligible respect to $x \mapsto x^n$ at infinity, we can choose an open set I_∞ on which $\nu_{p^*}(\infty)\frac{h(x)}{\mathcal{Q}(x)} \geq \alpha$.
If the infinity is not an extreme point, p^* is of maximal degree, and we define an interval $I_\infty =]-\infty, a[\cup]b, +\infty[$, $a < 0$, $b > 0$ by

- for $x \in]b, +\infty[\cap J$, $\delta(\sum_{i=0}^n (p_i^* - \frac{p_n^*}{2})x^i) \geq \mu_{p^*} + \alpha$,
- for $x \in]-\infty, a[\cap J$, $\delta(\sum_{i=0}^n (p_i^* - (-1)^i\frac{p_n^*}{2})x^i) \geq \mu_{p^*} + \alpha$.

Let $\beta = \dfrac{1}{3}\inf\limits_{x\in I}(u(x) - l(x)) > 0$.
Using the continuity, we can restrain the intervals I_ζ in order to have

$$\forall \zeta \in E^{p^*}, \forall x \in I_\zeta \cap I, \quad \delta(p^*)(x) \leq \beta.$$

Let $\theta = \bigcup\limits_{\zeta \in E^{p^*} \cup \{\infty\} \cup \mathcal{Z}_\mathcal{Q}} I_\zeta$. θ is an open set, therefore ${}^c\theta \cap I$ and ${}^c\theta \cap J$ are compact. Thus

$$\gamma_1 = \min_{x \in {}^c\theta \cap I} \delta(p^*)(x) > 0 \text{ and } \gamma_2 = \min_{x \in {}^c\theta \cap J} \delta(p^*)(x) > \mu_{p^*}.$$

Let $\omega = \min(\gamma_1, \gamma_2 - \mu_{p^*})$. Then $\omega > 0$, $\omega \leq \gamma_1$, $\mu_{p^*} + \omega \leq \gamma_2$ and

1. $\forall x \in I \cap {}^c\theta$, $\delta(p^*)(x) \geq \omega$,
2. $\forall x \in J \cap {}^c\theta$, $\delta(p^*)(x) \geq \mu_{p^*} + \omega$.

Let $\epsilon > 0$ such that

1. $\epsilon \sup\limits_{x \in X \cap {}^c\theta} \dfrac{|h(x)|}{\mathcal{Q}(x)} < \min(\omega, \beta)$,
2. $\delta(p^* + \epsilon h)(x) \geq \mu_{p^*} + \alpha$ if $x \in I_\zeta \cap J$ and $\zeta \in \mathcal{Z}_\mathcal{Q} \setminus E^{p^*}$,

3. $\delta(p^* + \epsilon h)(x) \geq \alpha$ if $x \in I_\zeta \cap I$ and $\zeta \in \mathcal{Z}_\mathcal{Q} \setminus E^{p^*}$,

4. if the infinity is not an extreme point, ϵ also has to be such that

$$\epsilon \max_{0 \leq i \leq n} |h_i| < \frac{|p_n^*|}{2}.$$

Let us check that $p^* + \epsilon h$ is better than p^*.
The choice of epsilon is such that over each I_ζ, $\zeta \in \mathcal{Z}_\mathcal{Q} \setminus E^{p^*}$, we improve the polynomial p^* by adding ϵh. We will see that this is also true for the other intervals.
On I_∞, we have $\delta(p^* + \epsilon h) \geq \mu_{p^*} + c$ with $c = \min(\alpha, \epsilon\alpha)$. Therefore, on I_∞, we improve p^* by adding an ϵh.
On each I_ζ, $\zeta \in E^{p^*}$, $\nu_{p^*}(\zeta)\dfrac{h(x)}{1+|x^n|} \geq \alpha$, therefore h has the same sign that $\nu_{p^*}(\zeta)$.
If $\zeta \in E_1^{p^*}(u)$, h is negative over I_ζ, therefore $(p^* + \epsilon h)/\mathcal{Q} < p^*/\mathcal{Q} \leq u$ over I_ζ. Furthermore, $u - p^*/\mathcal{Q} \leq \beta$ over $I_\zeta \cap I$, so $p^*/\mathcal{Q} - l \geq 2\beta$. Consequently, $(p^* + \epsilon h)/\mathcal{Q} \geq \beta + l > l$. Thus we get $l < (p^* + \epsilon h)/\mathcal{Q} < u$ over $I_\zeta \cap I$. The result is identical if $\zeta \in E_1^{p^*}(l)$.
If $\zeta \in E_2^{p^*}$, h is positive over I_ζ, therefore $(p^* + \epsilon h)\mathcal{Q} > p^*/\mathcal{Q} \geq l + \mu^*$ over $I_\zeta \cap J$. The same is true if $\zeta \in E_3^{p^*} : (p^* + \epsilon h)/\mathcal{Q} < p^*/\mathcal{Q} \leq u - \mu^*$ over $I_\zeta \cap J$.
If $x \in {}^c\theta \cap I$, we have $\delta(p^* + \epsilon h)(x) > 0$, so $l(x) < (p^*(x) + \epsilon h(x))/\mathcal{Q}(x) < u(x)$. If $x \in {}^c\theta \cap J$, $\delta(p^* + \epsilon h)(x) > \mu_{p^*}$. Thus $(p^* + \epsilon h)/\mathcal{Q}(x) - l(x) > \mu_{p^*}$ over J^+ and $u(x) - (p^* + \epsilon h)/\mathcal{Q}(x) > \mu_{p^*}$ over J^-.

We get $\mu_{p*+\epsilon h} > \mu^*$. This contradicts the maximality of μ^*. Thus $0 \in C$, and using the Carathéodory theorem, we obtain the existence of an integer $r \leq n+1$ such that

$$\exists x_0, \ldots, x_r \text{ distinct } \in \widehat{E^{p^*}}, \lambda_0, \ldots, \quad \lambda_r > 0, \quad \forall h \in \mathcal{P}_n, \sum_{i=0}^{r} \lambda_i \chi_{x_i}^{p^*}(h) = 0.$$

■

We will now see how p^* can be characterized by a "simple alternation property".

Definition 3.1.4 *If Φ is an application with values in $\{-1, 1\}$, the points $x_0, \ldots, x_r$ are called Φ-alternant whenever*

$$\forall i, 0 \leq i \leq r-1, \begin{cases} x_i < x_{i+1} \\ \Phi(x_i) = -\Phi(x_{i+1}) \end{cases}.$$

Proposition 3.1.5 *If p^* is a solution of (3.2), p^* has $n+2$ extreme ν_{p^*}-alternant points.*

Proof Using the previous lemma, we get $r+1$ distinct points $x_0, \ldots, x_r \in \widehat{E^{p^*}}$ and $r+1$ positive real numbers $\lambda_0, \ldots, \lambda_r$ such that for every $h \in \mathcal{P}_n$, $\sum_{i=0}^{r} \lambda_i \chi_{x_i}^{p^*}(h) = 0$ with $r \leq n+1$.

Suppose that $r \leq n$. The existence of $h \in \mathcal{P}_n$ such that $\chi_{x_0}^{p^*}(h) = 1$ and $\chi_{x_i}^{p^*}(h) = 0$ for $1 \leq i \leq r$ implies

$$\sum_{i=0}^{r} \lambda_i \chi_{x_i}^{p^*}(h) = \lambda_0 \neq 0.$$

Consequently, $r = n + 1$.

We now suppose that $x_0, \ldots, x_{n+1}$ are not ν_{p^*}-alternant. Let $\mathcal{I}$ be the set defined by

$$\mathcal{I} = \{i \in \mathbb{N} \ , \ 0 \leq i \leq n \ , \ \nu_{p^*}(x_i) = -\nu_{p^*}(x_{i+1})\} \, .$$

If $\mathcal{I} = \emptyset$, then every constant polynomial contradicts the nullity of the sum. Therefore, we can suppose $1 \leq Card(\mathcal{I}) \leq n$. To each $i \in \mathcal{I}$, we associate a point z_i such that $x_i < z_i < x_{i+1}$. We define h by $h(x) = \nu_{p^*}(x_{n+1}) \prod_{i \in \mathcal{I}} (x - z_i)$.

Then for every i, $0 \leq i \leq n+1$, we get $\chi_{x_i}^{p^*}(h) > 0$ if $x_i \neq \infty$ and $\chi_{\infty}^{p^*}(h) \geq 0$. Therefore

$$\sum_{i=0}^{n+1} \lambda_i \chi_{x_i}^{p^*}(h) > 0.$$

We then deduce that $x_0, \ldots, x_{n+1}$ are ν_{p^*}-alternant.

■

We now obtain that the optimal polynomial p^* is totally characterized by the alternant points.

Theorem 3.1.6 *Let $p \in \mathcal{A}$.*
Then p is a solution of (3.2) *if and only if p has $n + 2$ ν_p-alternant extreme points.*

Proof Let p^* be a solution of (3.2) and $p \in \mathcal{A}$ be a polynomial with $n + 2$ ν_p-alternant extreme points, denoted $x_0, \ldots, x_{n+1}$. Let $h = p^* - p$.
If $x_i \in E_1^p(u)$, then $h(x_i) = p^*(x_i) - u(x_i)\mathcal{Q}(x_i) \leq 0$.
If $x_i \in E_1^p(l)$, then $h(x_i) = p^*(x_i) - l(x_i)\mathcal{Q}(x_i) \geq 0$.
If $x_i \in E_2^p$, then $h(x_i) = p^*(x_i) - (l(x_i) + \mu_p)\mathcal{Q}(x_i) \geq 0$.
If $x_i \in E_3^p$, then $h(x_i) = p^*(x_i) - (u(x_i) - \mu_p)\mathcal{Q}(x_i) \leq 0$.
If $x_i = \infty$, and J is unbounded on the right, and p^* is of maximal degree, then

$$\lim_{x \to +\infty} \nu_{p^*}(\infty) p^*(x) = +\infty$$

and therefore $\nu_{p^*} h(x) \geq 0$ for x large enough. This is also true if J is unbounded on the left.

The extreme points being ν_p-alternant, we deduce from what precedes that either h is of degree n and has $n + 1$ roots, or h is of degree less than $n - 1$ and has n roots. So h is the zero polynomial, and $p = p^*$.

■

Looking at the previous proof, the following corollary is immediate.

Corollary 3.1.7 *The problem* (3.2) *has a unique solution.*

3.1.4 A Remes-like algorithm

We will now see how to compute the solution of the problem using an exchange algorithm. This algorithm is an adaptation of the Remes algorithm ([Remes, 1934]), used in polynomial uniform approximation.

We still assume the existence of a polynomial p_0 in $\mathcal{A}$ such that $\mu_{p_0} > -\infty$.

The algorithm

The algorithm consists in solving the problem over a finite number of points. More precisely, if we want to compute the best polynomial of degree at most n, we have to solve the problem over $n+2$ correctly chosen points (see the example in the previous chapter).

In order to assure the validity of these points, we associate to them a value α in the following way:

- if $x \in J^+ \cap I_l^{+\infty}$, $\alpha(x) = 1$,
- if $x \in J^- \cap I_{-\infty}^u$, $\alpha(x) = -1$,
- if $x \in I_l^u$, $\alpha(x) = \pm 1$ (we will see later how to choose the sign).

We say that $n+2$ points $x_1^k, \dots, x_{n+2}^k$ are valid if

- they are in $X \setminus \mathcal{Z}_Q$,
- at least one point is in J,
- they are α-alternant.

We now linearize the problem. We need a new criterion to that purpose:

$$\Lambda_p^h = \max\left(\sup_{x\in J^+} \frac{(l(x)+h)\mathcal{Q}(x) - p(x)}{1+|x^n|}, \sup_{x\in J^-} \frac{p(x) - (u(x)-h)\mathcal{Q}(x)}{1+|x^n|}\right)$$

To initialize the algorithm, we need $n+2$ valid points $x_1^1, \dots, x_{n+2}^1$ such that at least one point is in I (if $x_i^1 \in I_l^u$, we choose the sign of $\alpha(x_i^1)$ in order to obtain an alternated sequence).

The algorithm is iterative. We now detail the kth step.

1. *Compute the solution p_k of the problem over the points $x_1^k, \dots, x_{n+2}^k$.*

   Let $p_k(x) = \sum_{i=0}^n a_i^k x^i$. We associate to x_i^k the equations :

$$\begin{cases} p_k(x_i^k) = \left(\frac{1+\alpha(x_i^k)}{2} l(x_i^k) + \frac{1-\alpha(x_i^k)}{2} u(x_i^k)\right) \mathcal{Q}(x_i^k) \text{ if } x_i^k \in I \setminus J, \\ p_k(x_i^k) = \left(\frac{1+\alpha(x_i^k)}{2} (l(x_i^k)+h_k) + \frac{1-\alpha(x_i^k)}{2} (u(x_i^k)-h_k)\right) \mathcal{Q}(x_i^k) \text{ if } x_i^k \in J, \\ a_n = 0 \text{ if } x_i^k = \pm\infty. \end{cases} \tag{3.3}$$

We obtain a system with $n+2$ equations and $n+2$ unknowns (the a_i^k and h_k) whose solution gives p_k and h_k.

2. *Look for the point y_k in X which most violates the constraints.*

   Let $M_k = \max\left(\max_{x\in I}(p_k(x) - u(x)\mathcal{Q}(x)), \max_{x\in I}(l(x)\mathcal{Q}(x) - p_k(x)), \Lambda_{p_k}^{h_k}\right)$.
   If $M_k = 0$, we stop the algorithm and return p_k.
   Else we choose a point y_k associated to M_k (i.e. a point for which the max in M_k is obtained), and we associate a value $\alpha(y_k)$ to this point.
   If $y_k \in I_l^u$, $\alpha(y_k) = \mathrm{sgn}(u(y_k)\mathcal{Q}(y_k) - p_k(y_k))$.

3. *Substitute y_k to one of the previous x_i^k in order to get a new sequence of $n+2$ points $x_1^{k+1}, \dots, x_{n+2}^{k+1}$.*

   We look for the index j such that x_j^k is replaced by y_k:
   - If $x_i^k < y_k < x_{i+1}^k$, then if $\alpha(x_i^k) = \alpha(y_k)$, $j = i$, else $j = i+1$.
   - If $y_k < x_1^k$, then if $\alpha(x_1^k) = \alpha(y_k)$, $j = 1$, else $j = n+2$.
   - If $y_k > x_{n+2}^k$, then if $\alpha(x_{n+2}^k) = \alpha(y_k)$, $j = n+2$, else $j = 1$.

   We then define a sequence of points ${x'}_i^{k+1}$ by:

$$\begin{cases} {x'}_i^{k+1} = x_i^k & \forall i \neq j \\ {x'}_j^{k+1} = y_k \end{cases}$$

   The x_i^{k+1} are obtained by sorting in increasing order the ${x'}_i^{k+1}$.

This algorithm gives a sequence of polynomials $(p_i)_{i\in\mathbb{N}}$ which converges to the optimal polynomial p^*.

Proof of convergence

We first prove that the system (3.3) at the step 1 of the algorithm always has a solution. Next, we show that the sequence (h_k) of values obtained by solving the system (3.3) is decreasing. Finally, we prove the convergence of the sequence (p_k) to the optimum.

Non-singularity of the system (3.3) We first check that the system (3.3) always has a unique solution.
Suppose that infinity is not an alternant point. For a set W, we denote by $\mathbf{1}_W$ the characteristic function of W (i.e. $\mathbf{1}_W(x) = 1$ if $x \in W$ else $\mathbf{1}_W(x) = 0$). The system (3.3) can be written as :

$$\begin{pmatrix} 1 & x_{n+2} & \cdots & x_{n+2}^n & \mathbf{1}_J(x_{n+2})\alpha(x_{n+2})\mathcal{Q}(x_{n+2}) \\ 1 & x_{n+1} & \cdots & x_{n+1}^n & \mathbf{1}_J(x_{n+1})\alpha(x_{n+1})\mathcal{Q}(x_{n+1}) \\ \vdots & & & & \vdots \\ 1 & x_1 & \cdots & x_1^n & \mathbf{1}_J(x_1)\alpha(x_1)\mathcal{Q}(x_1) \end{pmatrix} \begin{pmatrix} a_0 \\ \vdots \\ a_n \\ -h \end{pmatrix}$$

$$= \begin{pmatrix} \mathcal{Q}(x_{n+2})\left(\frac{1-\alpha(x_{n+2})}{2}u(x_{n+2}) + \frac{1+\alpha(x_{n+2})}{2}l(x_{n+2})\right) \\ \vdots \\ \mathcal{Q}(x_2)\left(\frac{1-\alpha(x_2)}{2}u(x_2) + \frac{1+\alpha(x_2)}{2}l(x_2)\right) \\ \mathcal{Q}(x_1)\left(\frac{1-\alpha(x_1)}{2}u(x_1) + \frac{1+\alpha(x_1)}{2}l(x_1)\right) \end{pmatrix}.$$

Let H be the first matrice of the equality. Since the $\alpha(x_i)$, $1 \leq i \leq n+2$, are alternated, we have

$$H = \begin{pmatrix} 1 & x_{n+2} & \cdots & x_{n+2}^n & \alpha(x_1)(-1)^{n+1}\mathbf{1}_J(x_{n+2})\mathcal{Q}(x_{n+2}) \\ 1 & x_{n+1} & \cdots & x_{n+1}^n & \alpha(x_1)(-1)^{n}\mathbf{1}_J(x_{n+1})\mathcal{Q}(x_{n+1}) \\ \vdots & & & & \vdots \\ 1 & x_1 & \cdots & x_1^n & \alpha(x_1)\mathbf{1}_J(x_1)\mathcal{Q}(x_1) \end{pmatrix}.$$

We denote by $H_{i,j}$ the $n+1 \times n+1$ matrix obtained by removing the $i-th$ row and $j-th$ column of H. We define γ_s, $1 \leq s \leq n+2$, by $\gamma_s = (-1)^s \det(H_{s,n+2})$. Since $H_{s,n+2}$ is a Vandermonde matrix, we get

$$\gamma_s = (-1)^s \det(H_{s,n+2}) = (-1)^s \prod_{\substack{1 \leq i < j \leq n+2 \\ i \neq s, j \neq s}} (x_i - x_j). \tag{3.4}$$

Therefore,

$$|\det(H)| = \left|\sum_{s=1}^{n+2} \mathbf{1}_J(x_s)\mathcal{Q}(x_s)\det(H_{s,n+2})\right| = \left|\sum_{s=1}^{n+2} \mathbf{1}_J(x_s)\mathcal{Q}(x_s) \prod_{\substack{1 \leq i < j \leq n+2 \\ i \neq s, j \neq s}} (x_j - x_i)\right|.$$

$$\text{Suppose that the sequence } (x_i)_i \text{ is such that } \begin{cases} x_1 < x_2 < \cdots < x_{n+2}, \\ \text{at least one point } x_l \text{ is in } \in J \setminus \mathcal{Z}_{\mathcal{Q}}. \end{cases} \tag{3.5}$$

Proposition 3.1.8 *Under the hypothesis (3.5), the system (3.3) has a unique solution which satisfies $h \geq \mu_{p^*}$.*

Proof Since the x_i are sorted by increasing order, for every s, $1 \leq s \leq n+2$, we have $\prod_{\substack{1 \leq i < j \leq n+2 \\ i \neq s, j \neq s}} (x_j - x_i) > 0$. Furthermore, using again the hypothesis, there is at least one point x_l such that $\mathbf{1}_J(x_l)\mathcal{Q}(x_l) \neq 0$. Therefore, the determinant of H is not zero. Thus, the system (3.3) has a unique solution.
We denote by $p := \sum_{i=0}^n a_i x^i$ the solution of the system (3.3). We get

$$p(x_i) = \begin{cases} u(x_i)\mathcal{Q}(x_i) & \text{if } x_i \in I \text{ and } \alpha(x_i) = -1, \\ l(x_i)\mathcal{Q}(x_i) & \text{if } x_i \in I \text{ and } \alpha(x_i) = 1, \\ (l(x_i) + h)\mathcal{Q}(x_i) & \text{if } x_i \in J \text{ and } \alpha(x_i) = 1, \\ (u(x_i) - h)\mathcal{Q}(x_i) & \text{if } x_i \in J \text{ and } \alpha(x_i) = -1. \end{cases}.$$

Suppose $h < \mu_{p^*}$. Then the polynomial $p(x) - p^*(x)$ is such that

$$p(x_i) - p^*(x_i) = \begin{cases} u(x_i)\mathcal{Q}(x_i) - p^*(x_i) \geq 0 & \text{if } x_i \in I \text{ and } \alpha(x_i) = -1, \\ l(x_i)\mathcal{Q}(x_i) - p^*(x_i) \leq 0 & \text{if } x_i \in I \text{ and } \alpha(x_i) = 1, \\ (l(x_i) + h)\mathcal{Q}(x_i) - p^*(x_i) \leq h - \mu_{p^*} \leq 0 & \text{if } x_i \in J \text{ and } \alpha(x_i) = 1, \\ (u(x_i) - h)\mathcal{Q}(x_i) - p^*(x_i) \geq \mu_{p^*} - h \geq 0 & \text{if } x_i \in J \text{ and } \alpha(x_i) = -1. \end{cases}.$$

Therefore, as x_i is a sequence of $n+2$ distinct points sorted in increasing order such that the $\alpha(x_i)$ are alternated, we deduce $p - p^* = 0$. Since $\mu_p < h < \mu_{p^*}$, this leads to a contradiction. Thus, $h \geq \mu_{p^*}$.

■

Now, suppose that the infinity is among the sequence of points, e.g. $x_{n+2} = +\infty$. Then, the system to solve is the same as the previous one, but with n replaced by $n-1$ and with, in addition, the equation $a_n = 0$. Thus the proposition is also true.

We can now study the stop condition of the algorithm.
Note that M_k is always non-negative. Indeed, suppose $M_k < 0$. Then $\max_{x \in I}(p_k(x) - u(x)\mathcal{Q}(x)) < 0$ and $\max_{x \in I}(l(x)\mathcal{Q}(x) - p_k(x)) < 0$, and therefore, p_k is in $\mathcal{A}$. But, since $\Lambda_{p_k}^{h_k} < 0$, we get

$$\begin{cases} \frac{p_k(x)}{\mathcal{Q}(x)} > l(x) + h_k & \forall x \in J^+, \\ \frac{p_k(x)}{\mathcal{Q}(x)} < u(x) - h_k & \forall x \in J^-. \end{cases}$$

Using the previous proposition, we get a contradiction because $h_k \geq \mu_{p^*}$. Therefore, we deduce that $M_k \geq 0$ for every $k \geq 0$.
Suppose now that $M_k = 0$. The same argument implies that p_k is in $\mathcal{A}$ and that its criterion is greater or equal to μ_{p^*}. Therefore, p_k is the solution.

Validity of the points Now, we are going to show that the new points obtained at each iteration are valid (in particular that they satisfy (3.5)).

Suppose that the points x_i^k are valid. By construction, the new points x_i^{k+1} are in X, are sorted by increasing order, and their values α are alternated.
Suppose that all the x_i^{k+1} are in $I \setminus J$. Then $\alpha(x_1)(-1)^i(p_k - p^*)(x_i^{k+1}) \geq 0$. We deduce that $p_k = p^*$, which is a contradiction because $p_k(y_k) \neq p^*(y_k)$ (as $M_k > 0$, $p_k(y_k)$ is not between $l(y_k)$ and $u(y_k)$).
Suppose that y_k is reached at a pole of $\mathcal{Q}$ and that $\Lambda_{p_k}^{h_k}$ is not equal to zero. Then $p_k(y_k) < 0$ if $y_k \in J^+$ and $p_k(y_k) > 0$ if $y_k \in J^-$. We obtain again a contradiction considering $p_k - p^*$.
Finally, suppose that x_i^{k+1} are all in $I \cup \{\infty\}$. We get again that $p_k = p^*$.

Thus, the new sequence of points obtained at each iteration is valid. Therefore, the system (3.3) always has a unique solution.

Decrease of the criterion Suppose that the infinity is not among the points x_i^{k+1}. Consider the matrix H, and replace the last column by $((x_{n+2})^j, (x_{n+1})^j, \ldots, (x_1)^j)^t$ for

$0 \leq j \leq n$. Taking the determinant of the obtained matrix, we get

$$\sum_{i=1}^{n+2} \gamma_i (x_i)^j = 0$$

because the determinant is equal to zero since two columns are equal. Therefore, replacing the points x_i by the points x_i^{k+1} obtained at the $k+1$-st step, we get

$$\sum_{i=1}^{n+2} \gamma_i^{k+1} p(x_i^{k+1}) = 0 \text{ for every } p \in \mathcal{P}_n. \tag{3.6}$$

Using this equality with $p := p_k - p_{k+1}$, we get

$$\sum_{i=1}^{n+2} \gamma_i^{k+1} (p_k(x_i^{k+1}) - p_{k+1}(x_i^{k+1})) = 0.$$

Since the points are all distinct and sorted in increasing order, using (3.4), the γ_i^{k+1} are all different from zero and alternated. The sequence $\alpha(x_i^{k+1})$ being also alternated, we have

$$\sum_{i=1}^{n+2} \alpha(x_i^{k+1}) |\gamma_i^{k+1}| (p_k(x_i^{k+1}) - p_{k+1}(x_i^{k+1})) = 0.$$

Let $L_I = \{i, x_i^{k+1} \in I\}$ and $L_J = \{i, x_i^{k+1} \in J\}$. Using the previous equation, we get

$$\sum_{i \in L_I} \alpha(x_i^{k+1}) |\gamma_i^{k+1}| (p_k(x_i^{k+1}) - p_{k+1}(x_i^{k+1})) = - \sum_{i \in L_J} \alpha(x_i^{k+1}) |\gamma_i^{k+1}| (p_k(x_i^{k+1}) - p_{k+1}(x_i^{k+1})).$$

Suppose that $x_1^k < y_k < x_{n+2}^k$. Let i_0 be the index of the point $x_{i_0}^k$ which is replaced by y_k.

First, suppose $y_k \in I$.
Then, one point in I is changed, and all the others are kept. Therefore,

$$\sum_{i \in L_I} \alpha(x_i^{k+1}) |\gamma_i^{k+1}| (p_k(x_i^{k+1}) - p_{k+1}(x_i^{k+1})) = \alpha(x_{i_0}^{k+1}) |\gamma_{i_0}^{k+1}| (p_k(x_{i_0}^{k+1}) - p_{k+1}(x_{i_0}^{k+1})).$$

Thus, we obtain

$$\alpha(x_{i_0}^{k+1}) |\gamma_{i_0}^{k+1}| (p_k(x_{i_0}^{k+1}) - p_{k+1}(x_{i_0}^{k+1})) = - \sum_{i \in L_J} \alpha(x_i^{k+1}) |\gamma_i^{k+1}| (p_k(x_i^{k+1}) - p_{k+1}(x_i^{k+1})).$$

If $\alpha(x_{i_0}^{k+1}) = -1$, then $p_k(x_{i_0}^{k+1}) > u(x_{i_0}^{k+1})$ and $p_{k+1}(x_{i_0}^{k+1}) = u(x_{i_0}^{k+1})$.
If $\alpha(x_{i_0}^{k+1}) = 1$, then $p_k(x_{i_0}^{k+1}) < l(x_{i_0}^{k+1})$ and $p_{k+1}(x_{i_0}^{k+1}) = l(x_{i_0}^{k+1})$.
Therefore,

$$\alpha(x_{i_0}^{k+1}) |\gamma_{i_0}^{k+1}| (p_k(x_{i_0}^{k+1}) - p_{k+1}(x_{i_0}^{k+1})) = -|\gamma_{i_0}^{k+1}| |p_k(x_{i_0}^{k+1}) - p_{k+1}(x_{i_0}^{k+1})|.$$

If $x_i^{k+1} \in J^+$, then we have $\alpha(x_i^{k+1}) p_k(x_i^{k+1}) = (l(x_i^{k+1}) + h_k) \mathcal{Q}(x_i^{k+1})$ and $\alpha(x_i^{k+1}) p_{k+1}(x_i^{k+1}) = (l(x_i^{k+1}) + h_{k+1}) \mathcal{Q}(x_i^{k+1})$. Therefore, $\alpha(x_i^{k+1})(p_k(x_i^{k+1}) - p_{k+1}(x_i^{k+1})) = (h_k - h_{k+1}) \mathcal{Q}(x_i^{k+1})$.

If $x_i^{k+1} \in J^-$, the equality $\alpha(x_i^{k+1})(p_k(x_i^{k+1}) - p_{k+1}(x_i^{k+1})) = (h_k - h_{k+1})\mathcal{Q}(x_i^{k+1})$ also holds. Combining the previous equations, we obtain

$$|\gamma_{i_0}^{k+1}||p_k(x_{i_0}^{k+1}) - p_{k+1}(x_{i_0}^{k+1})| = \sum_{i \in L_J} |\gamma_i^{k+1}|(h_k - h_{k+1})\mathcal{Q}(x_i^{k+1}).$$

Thus,

$$h_{k+1} = h_k - \frac{|\gamma_{i_0}^{k+1}||p_k(y_k) - p_{k+1}(y_k)|}{\sum_{i \in L_J} |\gamma_i^{k+1}|\mathcal{Q}(x_i^{k+1})} < h_k. \tag{3.7}$$

We now suppose $y_k \in J$.
Then, all the points $(x_i^{k+1}) \in I$ were points of the reference (x_i^k). Therefore,

$$\sum_{i \in L_I} \alpha(x_i^{k+1})|\gamma_i^{k+1}|(p_k(x_i^{k+1}) - p_{k+1}(x_i^{k+1})) = 0.$$

Thus, using the same arguments as above, we get

$$\begin{aligned} -(h_k - h_{k+1}) \sum_{i \in L_J \setminus \{i_0\}} |\gamma_i^{k+1}|\mathcal{Q}(x_i^{k+1}) = &-h_{k+1}|\gamma_{i_0}^{k+1}|\mathcal{Q}(x_{i_0}^{k+1}) \\ &+ |\gamma_{i_0}^{k+1}|\mathbf{1}_{J^+}(y_k)(p_k(y_k) - l(y_k)\mathcal{Q}(y_k)) \\ &+ |\gamma_{i_0}^{k+1}|\mathbf{1}_{J^-}(y_k)(u(y_k)\mathcal{Q}(y_k) - p(y_k)). \end{aligned}$$

This is equivalent to

$$\begin{aligned} h_{k+1} \sum_{i \in L_J} |\gamma_i^{k+1}|\mathcal{Q}(x_i^{k+1}) = h_k &\sum_{i \in L_J} |\gamma_i^{k+1}|\mathcal{Q}(x_i^{k+1}) \\ &+ |\gamma_{i_0}^{k+1}|\mathbf{1}_{J^+}(y_k)(p_k(y_k) - (l(y_k) + h_k)\mathcal{Q}(y_k)) \\ &+ |\gamma_{i_0}^{k+1}|\mathbf{1}_{J^-}((u(y_k) - h_k)\mathcal{Q}(y_k) - p(y_k)). \end{aligned}$$

Thus, we get

$$h_{k+1} = h_k - \frac{|\gamma_{i_0}^{k+1}|\Lambda_{p_k}^{h_k}}{\sum_{i \in L_J} |\gamma_i^{k+1}|\mathcal{Q}(x_i^{k+1})}. \tag{3.8}$$

Note that, since $y_k \in J$, $\Lambda_{p_k}^{h_k}$ is positive. We conclude that $h_{k+1} \leq h_k$. The same holds if $y_k > x_{n+2}^k$ or $y_k < x_1^k$ or if the infinity is among the points (in this case, replace n by $n-1$).

Distance between two points Suppose that the infinity is not among the points. Let $P(k) = \left\{p \in \mathcal{P}_n, l(x_i^k) \leq \frac{p(x_i^k)}{\mathcal{Q}(x_i^k)} \leq u(x_i^k), \forall x_i^k \in I\right\}$. We first show that p_k is maximum for the $n+2$ points x_i^k, that is

$$\sup_{p \in P(k)} \min_{x_i^k \in J} \left(\mathbf{1}_{J^+}(x_i^k)p(x_i^k) - l(x_i^k)\mathcal{Q}(x_i^k) + \mathbf{1}_{J^-}(x_i^k)u(x_i^k)\mathcal{Q}(x_i^k) - p(x_i^k)\right) = h_k.$$

Suppose the existence of $r \in P(k)$ such that :

$$\min_{x_i^k \in J} \left(\mathbf{1}_{J^+}(x_i^k)r(x_i^k) - l(x_i^k)\mathcal{Q}(x_i^k) + \mathbf{1}_{J^-}(x_i^k)u(x_i^k)\mathcal{Q}(x_i^k) - r(x_i^k)\right) \geq h_k.$$

Then $(p_k - r)(x_i^k) = 0$ or $\operatorname{sgn}((p_k - r)(x_i^k)) = -\alpha(x_i^k)$. Therefore, the points x_i^k being alternated, $p_k = r$. So p_k is maximum for the $n+2$ points x_i^k.
If the infinity is among the points, replace n by $n-1$ and consider only the finite points x_i^k.

We now show that one point cannot converge toward another, that is there is a minimum distance between two points x_i^k and x_{i+1}^k. Suppose this is not true. Then X being compact in $\widehat{\mathbb{R}}$, we can extract a sub-sequence of $(\{x_1^k, \dots, x_{n+2}^k\})_k$ which converges to a set $\{x_1, \dots, x_{n+2}\}$ containing at most $n+1$ points.
Let $p \in \mathcal{P}_n$ be a polynomial such that

$$p(x_i) = \begin{cases} \frac{u(x_i)+l(x_i)}{2}\mathcal{Q}(x_i) & \text{if } x_i \in I_l^u, \\ (u(x_i) - 2h_1)\mathcal{Q}(x_i) & \text{if } x_i \in (J^- \cup I_{-\infty}^u) \setminus \mathcal{Z}_{\mathcal{Q}}, \\ (l(x_i) + 2h_1)\mathcal{Q}(x_i) & \text{if } x_i \in (J^+ \cup I_l^{+\infty}) \setminus \mathcal{Z}_{\mathcal{Q}}, \\ 1 & \text{if } x_i \in (J^+ \cup I_l^{+\infty}) \cap \mathcal{Z}_{\mathcal{Q}}, \\ -1 & \text{if } x_i \in (J^- \cup I_{-\infty}^u) \cap \mathcal{Z}_{\mathcal{Q}}, \end{cases}$$

and with a leading coefficient different from zero if the infinity is among the points.
If the x_i^k are close enough to the x_i, then :

$$\begin{cases} l(x_i^k) < \frac{p(x_i^k)}{\mathcal{Q}(x_i^k)} < u(x_i^k) & \text{if } x_i^k \in I_l^u, \\ \frac{p(x_i^k)}{\mathcal{Q}(x_i^k)} - l(x_i^k) > |h_1| > h_k & \text{if } x_i^k \in J^+ \cup I_l^{+\infty}, \\ u(x_i^k) - \frac{p(x_i^k)}{\mathcal{Q}(x_i^k)} > |h_1| > h_k & \text{if } x_i^k \in J^- \cup I_{-\infty}^u. \end{cases}$$

This contradicts the maximality of p_k for the points x_i^k.
Therefore, we get the existence of $d > 0$ such that $x_{i+1}^k - x_i^k > d$ for all i, $1 \le i \le n+1$, and for all $k \ge 0$.

Convergence toward the solution Since there is a minimum distance between two points, we can find a constant $m > 0$ such that $|\gamma_i^k| \ge m$ for all i and all k (see (3.4) for the definition of γ_i^k).
If $y_k \in I$ then we get from equation (3.7)

$$h_k - h_{k+1} = |\gamma_{i_0}^{k+1}|\,|p_k(y_k) - p_{k+1}(y_k)| \tag{3.9}$$

$$\ge m \max\left(\max_{x\in I}(p_k(x) - u(x)\mathcal{Q}(x)), \max_{x\in I}(l(x)\mathcal{Q}(x) - p_k(x))\right) \ge 0. \tag{3.10}$$

If $y_k \in J$, then using equation (3.8), we obtain

$$h_k - h_{k+1} \ge |\gamma_{i_0}^{k+1}|\Lambda_{p_k} \ge m\Lambda_{p_k} \ge 0.$$

Thus we have

$$h_k - h_{k+1} \ge mM_k \ge 0 \text{ for every } k \ge 0.$$

The sequence h_k decreases and μ_{p^*} is a lower bound, therefore the sequence converges and we get $\lim\limits_{k\to\infty} M_k = 0$. Thus,

$$\lim_{k\to\infty} \max\left(\max_{x\in I}(p_k(x) - u(x)\mathcal{Q}(x)), \max_{x\in I}(l(x)\mathcal{Q}(x) - p_k(x))\right) = 0, \tag{3.11}$$

$$\lim_{k\to\infty} \Lambda_{p_k}^{h_k} \le 0. \tag{3.12}$$

We deduce from (3.11) the existence of an integer N such that

$$\forall k \geq N, \forall x \in I, l(x) - 1 \leq \frac{p_k}{\mathcal{Q}}(x) \leq u(x) + 1.$$

Since the set

$$\left\{ p \in \mathcal{P}_n, l(x) - 1 \leq \frac{p(x)}{\mathcal{Q}(x)} \leq u(x) + 1, \forall x \in I \right\}$$

is compact, we can extract a sub-sequence $\left(p_{k_j}\right)$ which converges to a polynomial p. Using again (3.11), we deduce that $p \in \mathcal{A}$. Since p_k is maximum for the points x_i^k (see 3.1.4), $h_k \geq \mu_{p^*}$ for all k. We then deduce from (3.12) that $\mu_p = \lim h_{k_j} \geq \mu_{p^*}$. By definition of p^*, and by uniqueness, we get $p = p^*$. Since all the adherence values of the sequence (p_k) are equal to p^*, (p_k) converges to p^*.

3.2 A rational Zolotarev problem

We are now studying the rational case. We follow the same outline as for the polynomial case.
Let m and n be two positive integers and I, J be two compact subsets of $\mathbb{R}$ such that :

- $I \cap J = \emptyset$,
- I has at least $\max(m, n) + 1$ points,
- J has at least $m + 1$ points.

We denote by X the union of I and J: $X = I \cup J$.
Let f be a continuous function from X into $\mathbb{R}_+^*$ and R_m^n be the following set of rational functions:

$$R_m^n = \left\{ \frac{p}{q}, p \in \mathcal{P}_n, q \in \mathcal{P}_m^* \right\}.$$

Let A_m^n be the set:

$$A_m^n = \{ r \in R_m^n, \forall x \in I, |r(x)| \leq f(x) \}.$$

The problem we are considering is to find (if it exists) a rational function bounded by f over I and which is "as far as possible" from f over J. i.e.

$$\sup_{r \in A_m^n} \min_{x \in J} |r(x)| - f(x) \tag{3.13}$$

For $r \in R_m^n$, we define the criterion μ_r by

$$\mu_r = \min_{x \in J} |r(x)| - f(x).$$

3.2.1 Existence of a solution

We first check the existence of a solution to the above problem.

Proposition 3.2.1 *If I and J have no isolated point, then Problem (3.13) has a solution.*

Proof Let (r_k), $r_k = \dfrac{p_k}{q_k} \in A_m^n$, be a sequence such that $\lim\limits_{k\to\infty} \mu_{r_k} = \sup\limits_{r\in A_m^n} \mu_r$. Since r_k is bounded on I, we choose to normalize q_k by assuming $\|q_k\|_I = 1$. Therefore $\|p_k\|_I \leq \|f\|_I$. As I has at least $\max(m,n)+1$ points, the sets $\{p \in \mathcal{P}_n, \|p\|_I \leq \|f\|_I\}$ and $\{q \in \mathcal{P}_m, \| q \|_I = 1\}$ are compact sets (for every norm because the dimension is finite). Therefore we can extract from $(p_k)_k$ (resp. $(q_k)_k$) a sub-sequence which converges to $p^* \in \mathcal{P}_m$ (resp. $q^* \in \mathcal{P}_n$). As $|p_k(x)| \leq |q_k(x)|f(x)$ for all $x \in I$, we also have $|p^*(x)| \leq |q^*(x)|f(x)$ for all $x \in I$. Since $\|q_k\|_I = 1$, q^* is not the zero polynomial. Therefore, q^* vanishes at a finite number of points. Suppose that $q^*(x) = 0$ at a point $x \in I$. There is an open set $\mathcal{O}$ containing x such that q^* has no root in $\mathcal{O} \setminus \{x\}$ and therefore, $\frac{|p^*(y)|}{|q^*(y)|} \leq f(y)$ for all $y \in I \cap \mathcal{O} \setminus \{x\}$. Since I has no isolated point, $I \cap \mathcal{O} \setminus \{x\} \neq \emptyset$. Thus, p^*/q^* is bounded in a neighborhood of x, and consequently, it is also bounded at x. We then deduce that $\frac{p^*}{q^*}$ is in A_m^n. Furthermore, as $|p_k| \geq (f + \mu_{r_k})|q_k|$ over J, we have $|p^*| \geq |q^*|(f + \sup\limits_{r\in A_m^n} \mu_r)$ over J. Using again the fact there is no isolated point, we get $\sup\limits_{r\in A_m^n} \mu_r = \mu_{\frac{p^*}{q^*}}$.

■

Suppose that Problem (3.13) has a solution r^* such that $\mu_{r^*} > -\inf\limits_{x\in J} f(x)$ (i.e. the solution has a better criterion than the zero function). Write r^* as an irreducible function: $r^* = \frac{p^*}{q^*}$ with $\gcd(p^*, q^*) = 1$. Since $\mu_{r^*} > -\inf\limits_{x\in J} f(x)$, p^* does not vanish on J. Therefore p^* has a constant sign on each connected component of J. Furthermore, as r^* is bounded on I, q^* does not vanish on I, so q^* has a constant sign on each connected component of I.

We therefore divide X in distinct parts over which p and q have a constant sign. Let I^+, I^-, J^+ and J^- be four compact sets of $\mathbb{R}$ such that:

- The intersection of two sets in $\{I^+, I^-, J^+, J^-\}$ is empty,
- $I = I^+ \cup I^-$ has at least $\max(m,n) + 1$ points,
- $J = J^+ \cup J^-$ has at least $m + 1$ points.

We denote by $\mathcal{S}$, $\mathcal{R}$ and $\mathcal{A}$ the sets :

$$\mathcal{S} = \left\{(p,q), p \in \mathcal{P}_n, q \in \mathcal{P}_m^*, p_{|J^+} \geq 0, p_{|J^-} \leq 0, q_{|I^+} \geq 0, q_{|I^-} \leq 0\right\}$$

$$\mathcal{R} = \left\{\frac{p}{q}, (p,q) \in \mathcal{S}\right\}.$$

$$\mathcal{A} = \{r \in \mathcal{R}, \forall x \in I, |r(x)| \leq f(x)\}.$$

We are now interested in finding (when it exists) a solution to the sub-problem defined by

$$\max_{r\in\mathcal{A}} \mu_r. \tag{3.14}$$

We will sometimes use the expression "Let $r = \frac{p}{q} \in \mathcal{R}$ (or $\mathcal{A}$)" for "Let $r \in \mathcal{R}$ (or $\mathcal{A}$) and let $(p,q) \in \mathcal{S}$ such that $r = \frac{p}{q}$", that is, we will always choose p and q which respect the sign constraints.

Proposition 3.2.2 *If I and J have no isolated point, then the signed problem (3.14) has a solution.*

Proof Replacing A_m^n by $\mathcal{A}$, the proof is identical to the previous one (passing to the limit ensures that p^* and q^* have the good signs).

■

3.2.2 Characterization of the solution

In this part, we assume the existence of a solution. As for the polynomial case, we are now going to characterize a solution by a sequence of alternant points.

We assume that a solution R^* is such that R^* *non-constant* and such that $\mu_{R^*} > -\inf_{x\in J} f(x)$. We define a set $\mathcal{S}^*$ by

$$\mathcal{S}^* = \left\{(p,q), p \in \mathcal{P}_n, q \in \mathcal{P}_m^*, p_{|J^+} > 0, p_{|J^-} < 0, q_{|I^+} > 0, q_{|I^-} < 0\right\}.$$

The following lemma shows that a not trivial admissible rational function always has a representative in $\mathcal{S}^*$.

Lemma 3.2.3 *If $R \in \mathcal{A}$ is such that $\mu_R > -\inf_{x\in J} f(x)$, R can be written as $R = \frac{P}{Q}$ with $(P,Q) \in \mathcal{S}^*$.*

Proof Let $(p,q) \in \mathcal{S}$ and $R = \frac{p}{q}$.

Suppose that p has a root z with multiplicity k in J. As p has a constant sign on each connected component of J, k is even or $z \in \partial J$. If k is even, the polynomial $\frac{p(X)}{(X-z)^k}$ has the same sign as p on each connected component of J. If $z \in \partial J$, we can find $z^{\{\epsilon\}}$ such that $[z^{\{\epsilon\}}, z[$ or $]z, z^{\{\epsilon\}}]$ is included in ${}^cX$. Then $\frac{p(X)}{(X-z)^k}(X - z^{\{\epsilon\}})^k$ has the same sign as p on each connected component of J.

As $\mu_R > -\inf_{x\in J} f(x)$, the roots of p over J are also roots of q. We denote by $z_1, \ldots, z_k$ the distinct roots of p in J and $m_1, \ldots, m_k$ their multiplicity. We therefore have $p(X) = p_1(X)\prod_{i=1}^k (X - z_i)^{m_i}$ where p_1 has no root in J, and $q = q_1(X)\prod_{i=1}^k (X - z_i)^{m_i}$. Thus we can write R as

$$R = \frac{p_1(X)\prod_{\text{odd } m_i}(X - z_i)^{m_i}}{q_1(X)\prod_{\text{odd } m_i}(X - z_i)^{m_i}} = \frac{p_1(X)\prod_{\text{odd } m_i}(X - z_i^{\{\epsilon\}})^{m_i}}{q_1(X)\prod_{\text{odd } m_i}(X - z_i^{\{\epsilon\}})^{m_i}}$$

where $p_1(X)\prod_{\text{odd } m_i}(X - z_i^{\{\epsilon\}})^{m_i}$ has the same sign as p on each connected component of J, and has no root in J. Note that $q_1(X)\prod_{\text{odd } m_i}(X - z_i^{\{\epsilon\}})^{m_i}$ also has same sign as q on I, and that its roots in I are the roots of q. Since $|R| \leq f$ over I, the roots of $q_1(X)\prod_{\text{odd } m_i}(X - z_i^{\{\epsilon\}})^{m_i}$ in I are also roots of $p_1(X)\prod_{\text{odd } m_i}(X - z_i^{\{\epsilon\}})^{m_i}$. Therefore,

using the same argument, we obtain two polynomials P and Q such that P has same sign as p on J and no root in J, and Q has same sign as q on I, and no root in I.

■

Note that, conversely, if $(P,Q) \in \mathcal{S}^*$, then $\mu_{P/Q} > -\min_J f$.

We denote by E^r the extreme points of r, i.e. the set

$$E^r = \{x \in I, |r(x)| = f(x)\} \cup \{x \in J, |r(x)| - f(x) = \mu_r\}.$$

Let (P_1,Q_1) and $(P_2,Q_2) \in \mathcal{S}^*$ be such that $R = \frac{P_1}{Q_1} = \frac{P_2}{Q_2} \in \mathcal{A}$. We denote Δ_i a gcd of P_i and Q_i, $i \in \{1,2\}$, and we define two applications ν_R^i by

$$\nu_R^i(x) = \begin{cases} \operatorname{sgn}(R\Delta_i(x)) & \text{if } x \in J \cap E^R, \\ -\operatorname{sgn}(R\Delta_i(x)) & \text{if } x \in I \cap E^R. \end{cases}$$

Note that there is a real constant λ such that

$$\frac{P_1}{P_2} = \frac{Q_1}{Q_2} = \lambda \frac{\Delta_1}{\Delta_2}.$$

Therefore, $\lambda\Delta_1/\Delta_2 \geq 0$ on X, that is Δ_1/Δ_2 has constant sign on X. Thus, we can define a notion of ν_R-alternation in the following way :

Definition 3.2.4 *Let $R \in \mathcal{A}$ such that $\mu_R > -\inf_J f$, (P,Q) a representative of R in $\mathcal{S}^*$ and Δ a gcd of P and Q. We define an application ν_R from E^R into $\{-1,1\}$ by :*

$$\nu_R(x) = \begin{cases} \operatorname{sgn}(R\Delta(x)) & \textit{if } x \in J \cap E^R \\ -\operatorname{sgn}(R\Delta(x)) & \textit{if } x \in I \cap E^R \end{cases}$$

We will say that the extreme points $w_0, \ldots, w_r$ of R are ν_R-alternant if

$$\forall i, 0 \leq i \leq r-1, \begin{cases} w_i < w_{i+1} \\ \nu_R(w_i) = -\nu_R(w_{i+1}) \end{cases}.$$

This notion is independent from the choice of the representative of R and the choice of the gcd.

Lemma 3.2.5 *Assume that Problem (3.14) has a solution. Let $R^* = \frac{P^*}{Q^*}$ be a solution of (3.14), $\frac{\check{P}^*}{\check{Q}^*}$ an irreducible form of R^* and $N = \max(m + d^\circ \check{P}^*, n + d^\circ \check{Q}^*)$.*
There exist distinct extreme points $x_0, \ldots, x_r$ of R^* and real positive numbers $\lambda_0, \ldots, \lambda_r$ such that for all polynomials h in $\mathcal{P}_N$,

$$\sum_{i=0}^{r} \lambda_i \nu_{R^*}(x_i) h(x_i) = 0$$

with $r \leq N+1$.

Proof The proof is similar to the one in the polynomial case. For $\xi = \pm 1$, we define Ψ_ξ by

$$\begin{array}{rccl} \Psi_\xi(x): & \mathcal{P}_n \times \mathcal{P}_m & \longrightarrow & \mathbb{R} \\ & (h_1, h_2) & \longmapsto & \xi(h_1(x)Q^*(x) + h_2(x)P^*(x)) \end{array}$$

and we denote by Ψ the set

$$\Psi = \left\{ \Psi_{\tau(w)}(w), w \in E^{R^*} \right\}$$

where

$$\tau(w) = \begin{cases} \operatorname{sgn}(R^*(w)) & \text{if } w \in J \cap E^{R^*} \\ -\operatorname{sgn}(R^*(w)) & \text{if } w \in I \cap E^{R^*} \end{cases}.$$

Since $\{w \in E^{R^*}, \tau(w) = 1$ and $\{w \in E^{R^*}, \tau(w) = -1$ are compact sets, Ψ is a compact set (as the image of a compact set by a continuous function). Let C be the convex hull of Ψ.

Suppose that $0 \notin C$. Then, using the Hanh-Banach theorem, we can find $\alpha > 0$ and $(P, Q) \in \mathcal{P}_m \times \mathcal{P}_n \setminus \{0\}$ such that :

$$\forall w \in E^{R^*}, \Psi_{\tau(w)}(w)(P, Q) = \tau(w)(Q(w)P^*(w) + P(w)Q^*(w)) > \alpha > 0.$$

Let

$$\Phi = QP^* + PQ^*.$$

We therefore have

$$\forall w \in E^{R^*}, \quad \tau(w)\Phi(w) > \alpha > 0.$$

We now define a rational function R_λ by

$$R_\lambda = \frac{P^* + \lambda P}{Q^* - \lambda Q}$$

where λ is a positive real number. We will choose later the value of λ in order to obtain the following contradiction: $\mu_{R_\lambda > \mu_{R^*}}$.

For $w \in E^{R^*}$, we choose open sets V_w such that:

- if $w \in I$, then $|R^*(x)| \geq \frac{1}{2} \inf_{y \in I} f(y)$ and $\operatorname{sgn}(\Phi(x)) = -\operatorname{sgn}(R^*(x))$ for all $x \in V_w \cap I$,
- if $w \in J$, then $|R^*(x)| \leq f(x) + \mu_{R^*} + c$ and $\operatorname{sgn}(\Phi(x)) = \operatorname{sgn}(R^*(x))$ for all $x \in V_w \cap J$ where c is a positive constant.

We have

$$R_\lambda - R^* = \frac{\lambda \Phi}{Q^*(Q^* - \lambda Q)}.$$

Since Q^* does not vanish in I, we can find C_1' such that

$$\forall \lambda, 0 < \lambda < C_1', \forall x \in I, \operatorname{sgn}(Q^*(x) - \lambda Q(x)) = \operatorname{sgn}(Q^*(x)).$$

If $w \in I \cap E^{R^*}$, since $|R^*| \geq \frac{1}{2}\inf_{x\in I} f(x)$ on $V_w \cap I$, P^* does not vanish on $V_w \cap I$. Therefore, we can find $C_w < C_1'$ such that

$$\forall \lambda, 0 < \lambda < C_w, \operatorname{sgn}(P^* + \lambda P) = \operatorname{sgn}(P^*) \text{ on } V_w \cap I.$$

Thus, for all $x \in V_w \cap I$, and all λ, $0 < \lambda < C_w$, we get $\operatorname{sgn}(R_\lambda(x) - R^*(x)) = -\operatorname{sgn}(R^*(x))$ and $\operatorname{sgn}(R_\lambda(x)) = \operatorname{sgn}(R^*(x))$, and therefore we obtain $|R_\lambda(x)| < |R^*(x)|$.
Let $C_1'' = \min\limits_{w\in I\cap E^{R^*}} C_w$, $0 < \lambda < C_1''$ and $\theta_1 = \bigcup\limits_{w\in I\cap E^{R^*}} V_w$. Then for all $x \in I \cap \theta_1$, $|R_\lambda(x)| < |R^*(x)| \leq f(x)$. Furthermore, as R^* is continuous on I, we can choose $\delta > 0$ such that for all $x \in I \cap \theta_1^c$, $|R^*(x)| < f(x) - \delta$.
Let C_1, $0 < C_1 < C_1''$ such that for all λ, $0 < \lambda < C_1$, $|R_\lambda - R^*| \leq \frac{\delta}{2}$ over I. Then, for all λ, $0 < \lambda < C_1$, and for all $x \in I$:

$$|R_\lambda(x)| < f(x) \text{ and } \operatorname{sgn}(Q^*(x) - \lambda Q(x)) = \operatorname{sgn}(Q^*(x)).$$

Using the same argument, since P^* does not vanish in J, we can find C_2' such that for all $0 < \lambda < C_2'$, $\operatorname{sgn}(P^*(x) + \lambda P(x)) = \operatorname{sgn}(P^*(x))$ for all $x \in J$. Now, note that

$$\frac{1}{R_\lambda} - \frac{1}{R^*} = -\frac{\lambda\Phi}{P^*(P^* + \lambda P)}.$$

If $w \in J \cap E^{R^*}$, as $|R^*| \leq f(x) + \mu_{R^*} + c$ on $V_w \cap I$, Q^* does not vanish. Therefore we can find $C_w < C_2'$ such that for all $0 < \lambda < C_w$, $\operatorname{sgn}(Q^* - \lambda Q) = \operatorname{sgn}(Q^*)$ on $V_w \cap J$. Thus, for all $x \in V_w \cap J$, and all $0 < \lambda < C_w$, $\operatorname{sgn}\left(\frac{1}{R_\lambda(x)} - \frac{1}{R^*(x)}\right) = -\operatorname{sgn}\left(\frac{1}{R^*(x)}\right)$ and $\operatorname{sgn}(R_\lambda(x)) = \operatorname{sgn}(R^*(x))$ and therefore $|R_\lambda(x)| > |R^*(x)|$.
Let $C_2'' = \min\limits_{w\in J\cap E^{R^*}} C_w$, $0 < \lambda < C_2''$ and $\theta_2 = \bigcup\limits_{w\in J\cap E^{R^*}} V_w$. Then for all $x \in J \cap \theta_2$, $|R_\lambda(x)| > |R^*(x)| \geq f(x) + \mu_{R^*}$. Furthermore, since $\frac{1}{R^*}$ is continuous over J, we choose $\delta > 0$ such that for all $x \in J \cap {}^c\theta_2$, $|R^*(x)| \geq f(x) + \mu_{R^*} + \delta$. Let C_2, $0 < C_2 < C_2''$, be such that for all $0 < \lambda < C_2$, $|R_\lambda - R^*| \leq \frac{\delta}{2}$ on J.
Then, for all λ, $0 < \lambda < C_2$, and for all $x \in J$:

$$|R_\lambda(x)| > f(x) + \mu_{R^*} \text{ and } \operatorname{sgn}(P^*(x) + \lambda P(x)) = \operatorname{sgn}(P^*(x)).$$

Taking $\lambda < \min(C_1, C_2)$, we get a contradiction. Thus $0 \in C$, and using the Carathéodory theorem, we obtain the existence of $r'+1$ distinct points $x_0, \ldots, x_{r'} \in E^{p^*}$ and $r'+1$ strictly positive real numbers $\lambda_0', \ldots, \lambda_{r'}'$ such that :

$$\forall (h_1, h_2) \in \mathcal{P}_n \times \mathcal{P}_m, \quad \sum_{i=0}^{r'} \lambda_i' \tau(x_i)(h_1(x_i)Q^*(x_i) + h_2(x_i)P^*(x_i)) = 0 \tag{3.15}$$

with $r' \leq \dim(\mathcal{P}_n Q^* + \mathcal{P}_m P^*)$.

Let Δ be the quotient of P^* by $\check{P}^*$. In order to conclude, we now prove that $\mathcal{P}_n Q^* + \mathcal{P}_m P^* = \mathcal{P}_N \Delta$. First, note that

$$\mathcal{P}_n Q^* + \mathcal{P}_m P^* = \left\{(p\check{Q}^* + q\check{P}^*)\Delta, (p,q) \in \mathcal{P}_n \times \mathcal{P}_m\right\}.$$

Therefore,

$$\mathcal{P}_n Q^* + \mathcal{P}_m P^* \subset \mathcal{P}_N \Delta. \tag{3.16}$$

Furthermore, since $\mathcal{P}_n Q^* + \mathcal{P}_m P^* = (\mathcal{P}_n \check{Q}^* + \mathcal{P}_m \check{P}^*)\Delta$, we get

$$\dim(\mathcal{P}_n Q^* + \mathcal{P}_m P^*) = \dim\left((\mathcal{P}_n \check{Q}^* + \mathcal{P}_m \check{P}^*)\Delta\right) = \dim\left(\mathcal{P}_n \check{Q}^* + \mathcal{P}_m \check{P}^*\right) = N+1.$$

Thus, using (3.16) and the equality of the dimensions, we obtain

$$\mathcal{P}_n Q^* + \mathcal{P}_m P^* = \mathcal{P}_N \Delta.$$

Consequently, if $(h_1, h_2) \in \mathcal{P}_n \times \mathcal{P}_m$, $h_1 Q^* + h_2 P^*$ can be written as $h\Delta$, $h \in \mathcal{P}_N$. Then, equation (3.15) becomes

$$\forall h \in \mathcal{P}_N, \quad \sum_{i=0}^{r'} \lambda_i' \tau(x_i) \Delta(x_i) h(x_i) = 0.$$

Since $\lambda_i' \Delta(x_i)\tau(x_i) = \lambda_i' |\Delta(x_i)| \nu_{R^*}(x_i)$, defining λ_i by $\lambda_i = \lambda_i' |\Delta(x_i)|$, we get

$$\forall h \in \mathcal{P}_N, \quad \sum_{\{i, \lambda_i \neq 0\}} \lambda_i \nu_{R^*}(x_i) h(x_i) = 0.$$

■

If Problem (3.14) has a solution, then it is characterized by a sequence of alternant points:

Theorem 3.2.6 *Let $R \in \mathcal{A}$ and $\frac{\check{P}}{\check{Q}}$ be an irreducible form of R. Assume that R^* is a solution of Problem (3.14). Then :*

$$\mu_R = \mu_{R^*} \Longleftrightarrow R \text{ has } N+2 \text{ extreme } \nu_R\text{-alternant points}$$

with $N = \max(m + d^\circ \check{P}, n + d^\circ \check{Q})$.

Proof Using the same argument as in proposition 3.1.5, one can prove that R^* has $N+2$ extreme ν_R-alternant points.

Suppose that R has $N+2$ extreme ν_R-alternant points $x_1, \ldots, x_{N+2}$. Write R as $R = \Delta\check{P}/\Delta\check{Q}$, $(\Delta\check{P}, \Delta\check{Q}) \in \mathcal{S}^*$. Using (3.6) with $n := N$, we obtain an alternated sequence $(\gamma_i)_{i=1}^{N+2}$ such that $\sum \gamma_i (\check{P}Q^* - \check{Q}P^*)(x_i) = 0$. Therefore, as $\check{Q}$ and Q^* do not vanish on I, and $\check{P}$ and P^* do not vanish on J,

$$\sum_{x_i \in I} \gamma_i \check{Q} Q^*(x_i)(R(x_i) - R^*(x_i)) + \sum_{x_i \in J} \gamma_i \check{P} P^*(x_i) \left(\frac{1}{R^*(x_i)} - \frac{1}{R(x_i)} \right) = 0.$$

For $x_i \in I$, $|R(x_i)| \geq |R^*(x_i)|$, thus $R(x_i) = R^*(x_i)$ or $\operatorname{sgn}(R(x_i) - R^*(x_i)) = \operatorname{sgn}(R(x_i))$. Using the equality $\operatorname{sgn}(\Delta(x_i)\check{Q}(x_i)) = \operatorname{sgn}(Q^*(x_i))$, we get

$$R(x_i) = R^*(x_i) \text{ or } \operatorname{sgn}(\check{Q}(x_i) Q^*(x_i)(R(x_i) - R^*(x_i))) = \operatorname{sgn}(\Delta(x_i) R(x_i)).$$

Similarly, for $x_i \in J$, $\frac{1}{R(x_i)} = \frac{1}{R^*(x_i)}$ or $\operatorname{sgn}(P(x_i)P^*(x_i)(\frac{1}{R^*(x_i)} - \frac{1}{R(x_i)})) = -\operatorname{sgn}(\frac{\Delta(x_i)}{R(x_i)})$, that is :

$$R(x_i) = R^*(x_i) \text{ or } \operatorname{sgn}(P(x_i)P^*(x_i)\left(\frac{1}{R^*(x_i)} - \frac{1}{R(x_i)}\right) = -\operatorname{sgn}(\Delta(x_i)R(x_i)).$$

Since the points are ν_R-alternant, we get that the sign of $\check{P}Q^* - \check{Q}P^*$ alternates at the points $x_1, \ldots, x_{N+2}$, and therefore, $\check{P}Q^* - \check{Q}P^* = 0$. This gives $R = R^*$.

■

The following corollary is immediate:

Corollary 3.2.7 *Problem (3.14) has at most one solution.*

From Proposition 3.2.2, we get:

Corollary 3.2.8 *If I and J have no isolated point, then Problem (3.14) has a unique solution.*

3.2.3 A differential-correction-like algorithm

Two versions of the differential-correction algorithm are known for rational approximation. The first one, the original method, was presented in [Cheney and Loeb, 1961]. A modified version, with guaranteed convergence, was presented by the same authors in 1962 (e.g. [Cheney, 1998]). However, this version seemed to be slower than the original one. Later, it was proven that the original method is globally convergent, and that its rate of convergence is quadratic whenever the solution is of maximal degree (e.g. [Braess, 1986]).

The algorithm presented in section 2.2.2 is akin to the modified version of the differential-correction algorithm. We choose to study here an algorithm which is an adaptation of the original version of the differential-correction algorithm. In practice, this algorithm seems faster than the one presented in section 2.2.2. However, no proof of the rate of convergence is given.

We define the function σ by

$$\sigma(x) = \begin{cases} +1 & \text{if } x \in J^+ \text{ or } x \in I^+, \\ -1 & \text{if } x \in J^- \text{ or } x \in I^-. \end{cases}$$

In order to initialize the algorithm, we need two polynomials P_0 and Q_0 such that $(P_0, Q_0) \in \mathcal{S}^*$ and $\frac{P_0}{Q_0} \in \mathcal{A}$.
The algorithm is iterative. We now detail the kth step:

Let $f_k(P,Q) = \min_{x\in J} \dfrac{\sigma(x)P(x) - (f(x)+\mu_k)|Q(x)|}{|P_k(x)|}$.

Compute $P_{k+1} \in \mathcal{P}_m$ and $Q_{k+1} \in \mathcal{P}_n$ which maximize f_k respect to the constraints:

(i) $|P_{k+1}(x)| \le \sigma(x)Q_{k+1}(x)f(x)$ for $x \in I$,

(ii) $\max_{x\in J} |P_{k+1}(x)| = 1$.

If $f_k(P_{k+1}, Q_{k+1}) \le 0$ return $R_k = \dfrac{P_k}{Q_k}$, else compute $\mu_{k+1} = \min_{x\in J} \left|\dfrac{P_{k+1}(x)}{Q_{k+1}(x)}\right| - f(x)$.

Note that condition (i) implies that $\sigma Q_{k+1} \ge 0$ on I.
Condition (ii) is a choice of normalization of $\frac{P_{k+1}}{Q_{k+1}}$.

We now prove that this algorithm converges to the solution of the Zolotarev problem under different hypotheses.

Theorem 3.2.9 *Let $\mu^*_{m-1,n-1}$ be the optimal criterion for the general Zolotarev problem (3.13) with degrees $(m-1, n-1)$. If $(P_0, Q_0) \in \mathcal{S}^*$, $P_0/Q_0 \in \mathcal{A}$ and*

$$\mu_0 = \min_{x\in J} \left|\frac{P_0(x)}{Q_0(x)}\right| - f(x) \ge \mu^*_{m-1,n-1},$$

and if the signed Zolotarev problem (3.14) with degrees (m,n) has a solution, then the algorithm converges to this solution.

Proof In this proof, we frequently use the following equality: if $a \ge 0$ and $b \ge 0$ then $\min ab \ge \min a \min b$. We denote by μ^* the optimal criterion for the problem of degree n, and by $\frac{P^*}{Q^*} \in \mathcal{S}^*$ the associated optimal function.

1. Let $\mathcal{S}^*_1 = \{p \in \mathcal{P}_n, \sigma p > 0 \text{ on } J\}$.
   We first prove by induction that if $\mu_k < \mu^*$, then $f_k(P_{k+1}, Q_{k+1}) > 0$ and $P_{k+1} \in \mathcal{S}^*$.

   By hypothesis, $P_0 \in \mathcal{S}^*_1$.
   Suppose that $P_k \in \mathcal{S}^*_1$.
   If $\mu_k < \mu^*$, there is a pair (P,Q) in $\mathcal{S}^*$ such that

$$\min_{x\in J} \left|\frac{P(x)}{Q(x)}\right| - f(x) > \mu_k.$$

   We denote by μ the value of the criterion of P/Q, i.e. $\mu = \min_{x\in J} |P(x)/Q(x)| - f(x)$.
   Then :

$$\begin{aligned} f_k(P_{k+1}, Q_{k+1}) &\ge f_k(P,Q) \\ &\ge \min_{x\in J} \left(\left|\frac{P(x)}{Q(x)}\right| - (f(x)+\mu_k) \right) \left|\frac{Q(x)}{P_k(x)}\right| \\ &\ge (\mu - \mu_k) \min_{x\in J} \left|\frac{Q(x)}{P_k(x)}\right| \ge 0. \end{aligned} \tag{3.17}$$

Suppose that $f_k(P_{k+1}, Q_{k+1}) = 0$. Then $f_k(P, Q) = 0$ so it exists $x_0 \in J$ such that

$$|P(x_0)| - (f(x_0) + \mu_k)|Q(x_0)| = 0.$$

If $Q(x_0) \neq 0$, then

$$\frac{P(x_0)}{Q(x_0)} - (f(x_0) + \mu_k) = 0$$

which contradicts $\mu_k < \mu$. If $Q(x_0) = 0$, then we get $P(x_0) = 0$, which contradicts that $(P, Q) \in \mathcal{S}^*$. Therefore, if $\mu_k < \mu^*$, then $f_k(P_{k+1}, Q_{k+1}) > 0$. But the inequality $f_k(P_{k+1}, Q_{k+1}) > 0$ is possible only if $P_{k+1} \in \mathcal{S}_1^*$.

2. We now prove that if $\mu_k < \mu^*$, then $\mu_{k+1} > \mu_k$.

   We remark that

$$\left|\frac{P_{k+1}(x)}{Q_{k+1}(x)}\right| - f(x) = \mu_k + \left|\frac{P_k(x)}{Q_{k+1}(x)}\right| \frac{|P_{k+1}(x)| - (f(x) + \mu_k)|Q_{k+1}(x)|}{|P_k(x)|}. \tag{3.18}$$

   We stated before that if $\mu_k < \mu^*$, then $f_k(P_{k+1}, Q_{k+1}) > 0$ and $P_k \in \mathcal{S}_1^*$. Therefore, using (3.18), and taking the minimum over J, we get

$$\mu_{k+1} \geq \mu_k + \min_J \left|\frac{P_k(x)}{Q_{k+1}(x)}\right| f_k(P_{k+1}, Q_{k+1}) > \mu_k. \tag{3.19}$$

   Note that the previous inequality allows to check that if the algorithm stops, then $\mu_k = \mu^*$. Indeed, suppose that $\mu_k = \mu^*$ and $f_k(P_{k+1}, Q_{k+1}) > 0$. Since $\min_J |P_k(x)| > 0$, we obtain the following contradiction: $\mu_{k+1} > \mu^*$.

3. Finally, we prove that the sequence $\frac{P_k}{Q_k}$ converges to the solution.

   If the algorithm stops at the first iteration, then $\mu_0 = \mu^*$, thus the best rational function $R^* = \frac{P_0}{Q_0}$. Else $\mu_k \geq \mu_1 > \mu_0$ for all $k \geq 1$ so P_k/Q_k is of maximal degree (else P_k/Q_k contradicts the optimality of the criterion for $m-1, n-1$). Let $(P_{\Phi(k)}, Q_{\Phi(k)})$ be a sub-sequence of $(P_k, Q_k)_k$ which converges to (P, Q). We have:

   - $|P(x)| \leq \sigma(x)Q(x)f(x)$ for $x \in I$,
   - $\max_{x \in J} |P(x)| = 1$.

   The last point shows that P is not the zero polynomial. Therefore, using the first point, Q is not the zero polynomial either. Since $(P_k, Q_k) \in \mathcal{S}$ for every k, (P, Q) is also in $\mathcal{S}$.
   We denote by μ the limit of the μ_k:

$$\mu = \lim_{k \to \infty} \mu_k.$$

   As (μ_k) is an increasing sequence such that $\mu_k \geq \mu_1 > \mu_0 \geq \mu^*_{m-1,n-1}$,

$$\mu^*_{m-1,n-1} < \mu.$$

But, $|P| \geq (f+\mu)|Q|$ on J, and therefore, P/Q has better criterion than the best one obtained for all rational functions with degrees $(m-1, n-1)$. Therefore, $\gcd(P,Q) = 1$. Using again the inequality $|P| \geq (f+\mu)|Q|$ on J, we deduce that P has no root in J. Thus, there is no sub-sequence of $(P_k)_k$ which converges to a polynomial with a root in J. This leads to the existence of $\eta > 0$ such that

$$\min_{x \in J} |P_k(x)| \geq \eta \text{ for all } k. \tag{3.20}$$

Furthermore, since for every k, $|P_k| \geq (f+\mu_k)|Q_k|$ on J, using the fact that $\max_{x \in J} |P_k(x)| = 1$, we get

$$\max |Q_k| \leq \frac{1}{\min_J f + \mu_0} \text{ for every } k \geq 1.$$

Thus, using the equation (3.18), we obtain

$$\mu_{k+1} - \mu_k \geq \eta(\min_J f + \mu_0) f_k(P_{k+1}, Q_{k+1}) \geq 0.$$

Since (P_{k+1}, Q_{k+1}) maximizes f_k, we have

$$\mu_{k+1} - \mu_k \geq \eta(\min_J f + \mu_0) f_k(P^*, Q^*) \geq 0.$$

As (μ_k) converges, we get

$$\lim_k f_k(P^*, Q^*) = 0$$

and therefore,

$$\exists y \in J, \quad |P^*(y)| - (f(y)+\mu)|Q^*(y)| = 0.$$

Suppose $|Q^*(y)| = 0$. Then $|P^*(y)| = 0$. Since $(P^*, Q^*) \in \mathcal{S}^*$, we obtain a contradiction. Therefore $|Q^*(y)| \neq 0$, and we get $\mu^* \leq \mu$. Thus, the algorithm converges to the optimal rational function.

■

Since in practice, we use a discretization of J in order to compute (P_{k+1}, Q_{k+1}), the following theorem is important for applications:

Theorem 3.2.10 *If J is a finite set, then the sequence of criterions $(\mu_k)_k$ converges to the optimal criterion μ^* whatever initialization $(P_0, Q_0) \in \mathcal{S}^* \cap \mathcal{A}$ is taken.*

Proof The steps 1 and 2 of the previous proof still hold. Therefore, the sequence (μ_k) is increasing. Since it is bounded by μ^*, it is converging to a limit μ'. Suppose that $\mu' < \mu^*$. Thus, there is a pair $(P,Q) \in \mathcal{S}^*$ such that $\mu > \mu'$. Using (3.19) and the maximality of (P_{k+1}, Q_{k+1}) for f_k, we get

$$\mu_{k+1} \geq \mu_k + \min_J \left| \frac{P_k}{Q_{k+1}} \right| f_k(P,Q).$$

Since $\mu > \mu' \geq mu_k$, we have $f_k(P,Q) \geq \min_J \frac{|P|-(f+\mu')|Q|}{|P_k|} > 0$ for every $k \geq 0$. Using $\max_J |P_k| = 1$, we deduce

$$\mu_{k+1} \geq \mu_k + \min_J \left| \frac{P_k}{Q_{k+1}} \right| \min_J (|P| - (f+\mu')|Q|).$$

Furthermore, since for all k, $\frac{P_{k+1}}{Q_{k+1}} \geq (f+\mu_0)$, we get

$$\mu_{k+1} - \mu_k \geq (\min_J f + \mu_0) \min_J (|P| - (f+\mu')|Q|) \min_J \left| \frac{P_k}{P_{k+1}} \right| \geq 0.$$

Therefore, passing to the limit, since $(\min_J f + \mu_0) \min_J (|P| - (f+\mu')|Q|) > 0$, we obtain

$$\lim_{k\to\infty} \min_J \left| \frac{P_k}{P_{k+1}} \right| = 0.$$

Thus, we find a sequence $(y_k)_k \in J^{\mathbb{N}}$ such that

$$\lim_{k\to\infty} \frac{P_k(y_k)}{P_{k+1}(y_k)} = 0. \tag{3.21}$$

Using again the fact that (P_{k+1}, Q_{k+1}) maximizes f_k, we have

$$\min_J \left| \frac{P_{k+1}}{P_k} \right| \geq f_k(P_{k+1}, Q_{k+1}) \geq f_k(P,Q) \geq \min_J (|P| - (f+\mu')|Q|). \tag{3.22}$$

Now we suppose that J is a finite set of N points :

$$J = \{x_1, x_2, \ldots, x_N\}.$$

We define c by

$$c = \min_J (|P| - (f+\mu')|Q|).$$

Using (3.21), we get the existence of an integer K such that

$$|P_{k+1}(y_k)| \geq 2c^{-N+1} |P_k(y_k)| \text{ for every } k \geq K. \tag{3.23}$$

By (3.22), we get

$$|P_{k+1}(x)| \geq c|P_k(x)| \text{ for every } x \in J.$$

Using the last inequality for $x \neq y_k$, and combining it with (3.23), we obtain for every $k \geq K$

$$\prod_{i=1}^N |P_{k+1}(x_i)| \geq 2\frac{c^{N-1}}{c^{N-1}} \prod_{i=1}^N |P_k(x_i)| \geq 2\prod_{i=1}^N |P_k(x_i)|.$$

Since for every $k \geq 0$, P_k has no root in J, the previous inequality shows that there is a $k \geq K$ such that $\prod_{i=1}^N |P_k(x_i)| > 1$. This contradicts the fact that $|P_k|$ is bounded by 1.

■

Chapter 4

Design examples

In this chapter, we present some multi-band microwave filters manufactured by the XLIM institute (Limoges, France). The theoretical filtering functions were computed using the previous theory. As the number of cavities is proportional to the degree of the filtering function, we keep each time the filtering function with the smallest degree that meet the specifications. The first two examples were presented in [Lunot et al., 2008]. We show that for both of them, the parity of the degree seems to be important. Technicals details are added for specialists of microwave filters.

4.1 A dual-band filter

The first example has the following electrical specifications: a return loss at 20 dB in the passbands ($I_1 = [-1, -0.625]$ and $I_2 = [0.25, 1]$), a rejection at 15 dB in the lower and upper stopbands ($J_1 =]-\infty, -1.188]$ and $J_3 = [1.212, +\infty[$) and 30-dB in the intermediary stopband ($J_2 = [-0.5, 0.125]$). One may first think of computing a 10-3 filtering characteristic to fit in the latter specifications. Since the differential-correction-like algorithm works on finite intervals, the two "outside" stopbands are set to $[-10, -1.188]$ and $[1.212, 10]$. We obtain the filtering function plotted in Fig. 4.1. Only 9 reflection zeros and 14 "extreme" points appear on the graph which seems at first glance to contradict the theory or to indicate that something is wrong with our numerical implementation. A closer inspection of the obtained function indicates however that the lacking "extreme" point is situated in the left limit of the first stopband, i.e. at $\omega = -10$ together with a reflection zero that was rejected to $\omega = -100$. If we increase the size of the left stopband the reflection zero is rejected further towards infinity. This amounts to saying that the optimal characteristic with at most 10 reflection zeros (resp. at most 3 transmission zeros) is in fact of 9-3 type. In some sense, the optimization process indicates that there is no way to improve this 9-3 filtering function by adding an extra reflection zero. Note that here the ability to guarantee the optimality of the computed filtering function is crucial. Someone using a generic optimizer may insist in finding a better starting point for his optimization process or try by all means to restrict the location of reflection zeros: by the optimality argument this can only yield a poorer result.

The low pass specifications given in Fig. 4.1 correspond to the following passbands and stopbands at microwave frequencies: the two passbands are respectively $I_1 = [8.28, 8.31]$

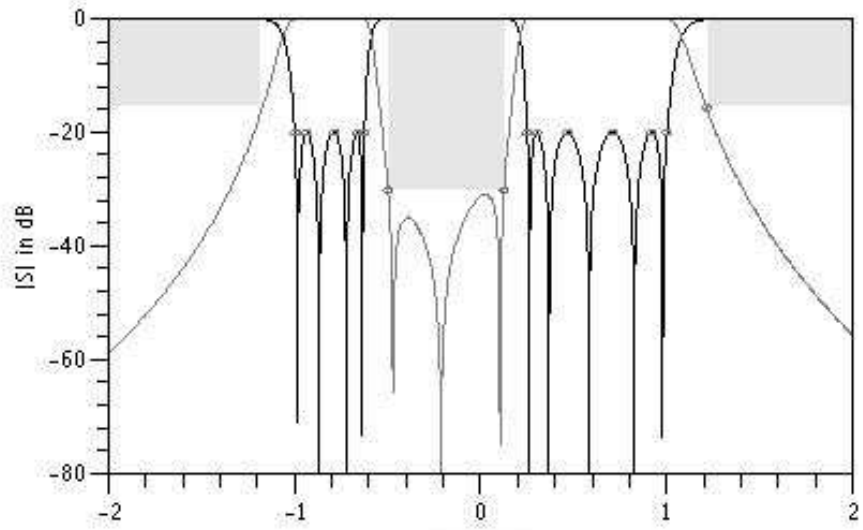

Figure 4.1: Optimal transmission and reflection parameters (example 1).

GHz and $I_2 = [8.38, 8.44]$ GHz and the three stopbands are respectively $J_1 = [0, 8.265]$ GHz, $J_2 = [8.32, 8.37]$ GHz and $J_3 = [8.457, +\infty[$ GHz. From these ideal parameters, a coupled resonator network has to be derived for realizing the desired number of transmission and reflection zeros. The network is chosen to be an extended-box one (see Fig. 4.2) since this topology allows a practical implementation of the filtering function with aligned dual-mode cavities. The technology selected for realizing the microwave filter consists in cylindrical cavities working in their dual-mode TE_{111} and coupled by rectangular irises as shown in Fig. 4.3. Applying an exhaustive coupling matrix synthesis ([Cameron et al., 2005a]), 22 real solutions have been found to realize the optimal function with the extended-box network. A particular solution is then selected and a computer-aided design (CAD) model is tuned, applying a coupling matrix identification at each tuning step ([Bila et al., 2001]). However, in this case, an exhaustive computation of all the solutions to the coupling matrix synthesis problem is necessary for recognizing the solution to be tuned. In case of ambiguity between several identified solutions, the solution that corresponds to the CAD model can be recognized by perturbing some coupling elements (dimensions of irises or screws) and by studying the coherence of the solution modifications (corresponding coupling values). The CAD model is a finite element model. Metallic losses are not considered during CAD tuning to facilite comparison with the synthesized lossless rational function. Moreover, no particular action, i.e. predistortion, is done for compensating losses in the current synthesis. A hardware prototype of the filter has been built with brass. The unloaded quality factor is around 4000 but can be improved using silver plated cavities. However, measured and simulated results are in good agreement as shown in Fig. 4.4. Insertion loss is 2.15 dB in the first passband and 1.45 dB in the second one.

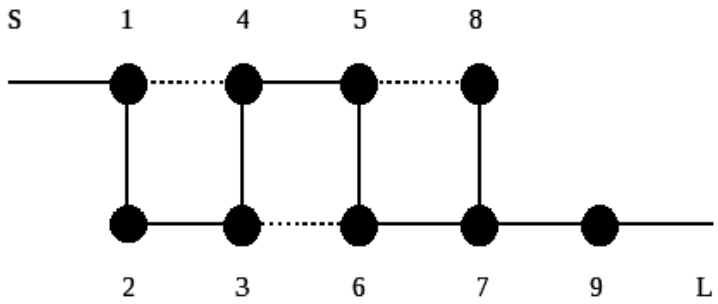

Figure 4.2: Extended-box coupled resonator network for the realization of the ideal 9-3 dual band response in Fig. 4.1.

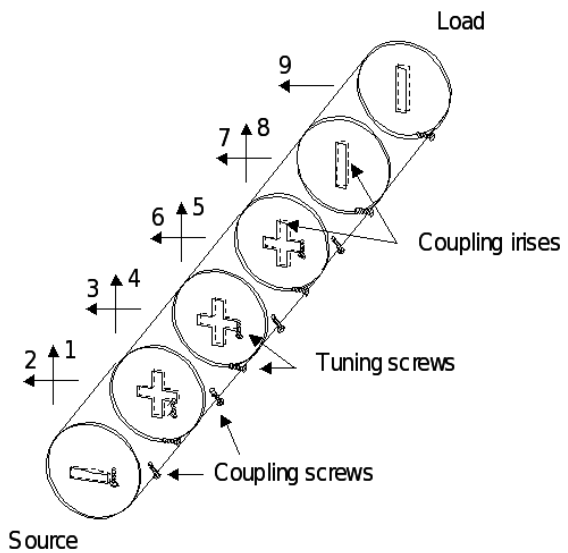

Figure 4.3: Implementation of the 9 pole 3 zero dual-band filter with in-line dual-mode cylindrical cavities, network topology illustrated in Fig. 4.2.

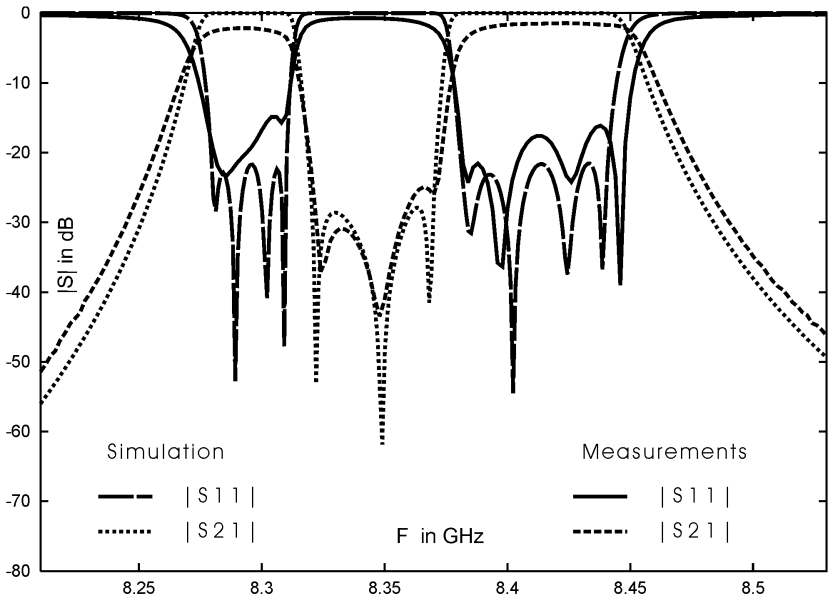

Figure 4.4: Measurements and simulation of the 9 pole 3 zero dual-band filter physically illustrated in Fig. 4.3.

4.2 Another dual-band filter on SPOT5 specifications

The electrical specifications of the second example are defined by: a return loss at 23 dB in the passbands ($I_1 = [-1, -0.383]$ and $I_2 = [0.383, 1]$), in the lower stopband ($J_1 =]-\infty, -1.864]$), the rejection is set at 10 dB in $]-\infty, -1.987]$ and 15 dB in $[-1.987, -1.864]$. The rejection is set at 20 dB in the intermediary stopband ($J_2 = [-0.037, -0.012]$) and 40 dB in the upper stopband ($J_3 = [1.185, +\infty[$). Here again one may think of using an 8-3 characteristic for a realization in extended box topology ([Cameron et al., 2005a]). However, the same phenomenon as in the first example occurs, and the optimal solution appears to be of type 7-3 (Fig. 4.5).

At microwave frequencies, the low pass specifications shown in Fig. 4.5 match into two passbands, respectively at $I_1 = [8.228, 8.278]$ GHz and $I_2 = [8.34, 8.39]$ GHz, and three stopbands, at $J_1 =]0, 8.158]$ GHz, $J_2 = [8.306, 8.308]$ GHz and $J_3 = [8.405, +\infty[$ GHz. The coupled-resonator network, which is selected for realizing the latter filtering function, is the pseudo extended-box topology presented in Fig. 4.6. This configuration of the coupled-resonator network leads to three real solutions for realizing the ideal filtering characteristic. A solution is chosen for implementation in stacked single-mode rectangular cavities as described in [Bila et al., 2006]. The CAD model and the practical hardware are tuned using an exhaustive coupling matrix identification. Measurement results of the brass-made prototype are compared with simulations in Fig. 4.7. Insertion loss is respectively 1.4 dB and 1.25 dB in the passbands.

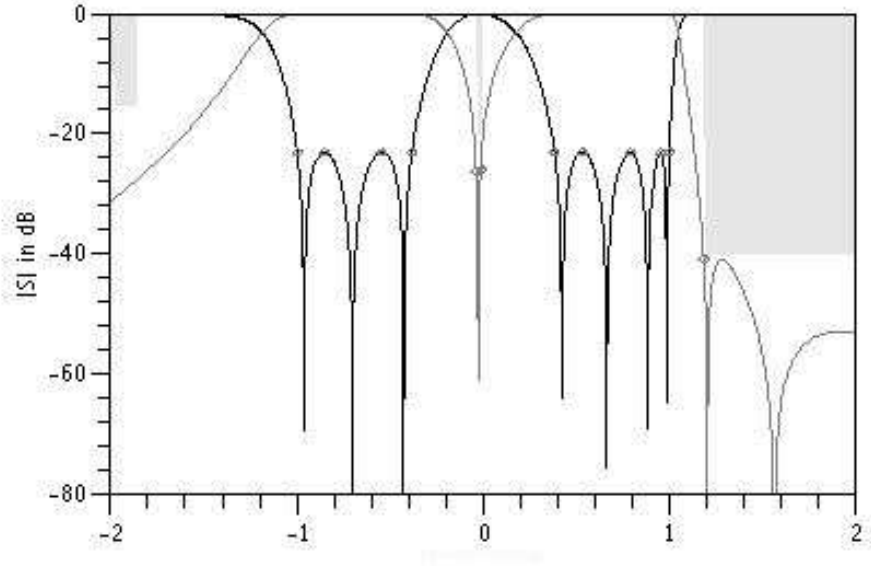

Figure 4.5: Optimal transmission and reflection parameters (example 2).

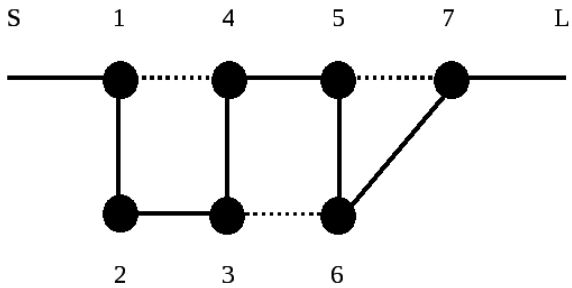

Figure 4.6: Pseudo extended-box coupled resonator network for the realization of the ideal 7-3 dual-band response in Fig. 4.5.

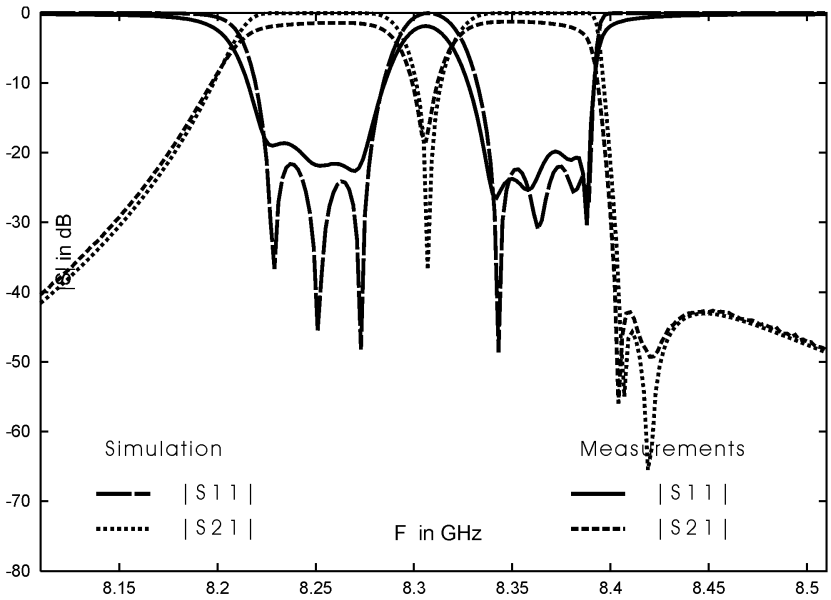

Figure 4.7: Measurements and simulation of the 7-3 dual-band filter, network topology illustrated in Fig. 4.6.

4.3 A tri-band filter

We now consider a tri-band filter whose electrical specifications are given in Fig. 4.8. The optimal filtering function is a 10-8 rational function plotted in Fig. 4.9. This filter has been manufactured, and the measurements are given in Fig. 4.10.

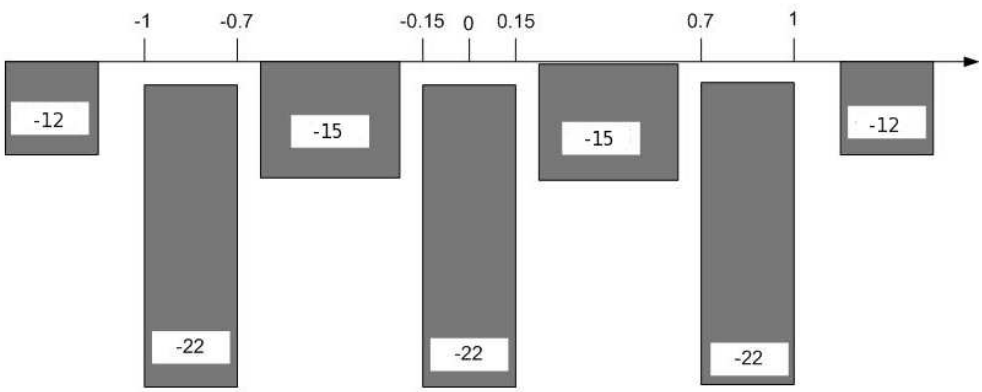

Figure 4.8: Specifications of the tri-band filter.

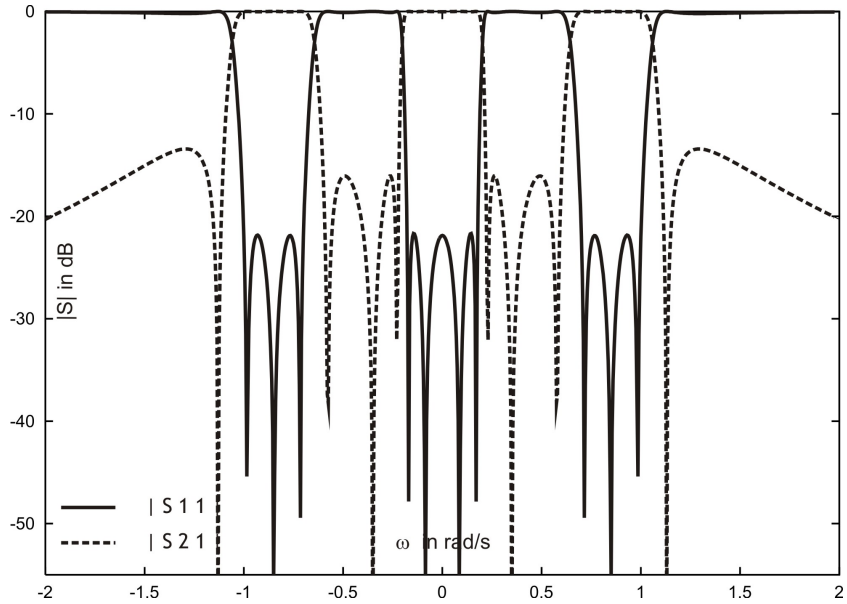

Figure 4.9: Theoretical filtering function of the 10-8 tri-band filter.

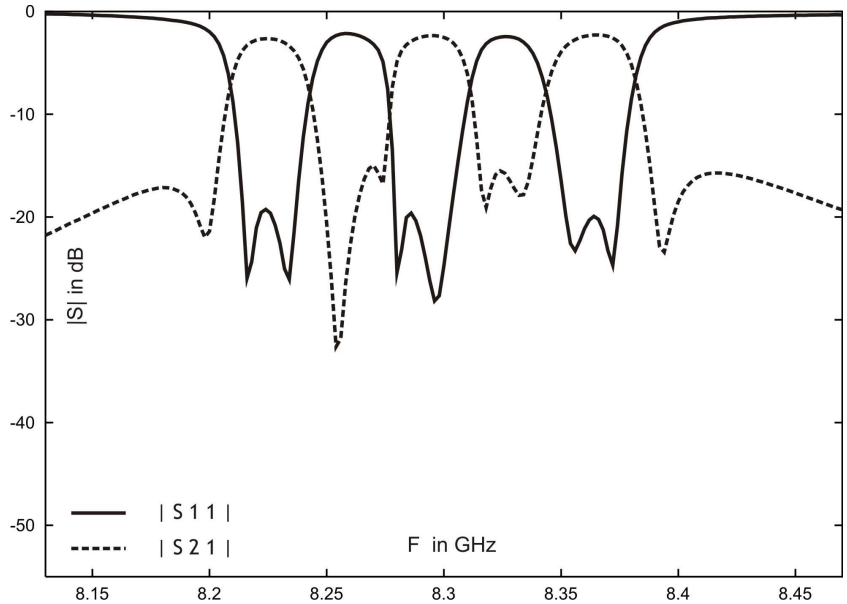

Figure 4.10: Measurements of the 10-8 tri-band filter.

Chapter 5

Conclusion

In Chapters 2 and 3, we presented two algorithms for the computation of the solution to the real Zolotarev sub-problem (2.5). One of them, the Remes-like algorithm, is only for the polynomial case, and the other one, the differential-correction-like algorithm, is for the general case (i.e. rational). These algorithms were used to compute the optimal filtering functions of different multiband microwave filters, presented in Chapter 4. In this chapter, three open problems are presented. The first section gives some clues for the implementation of a rational Remes-like algorithm in order to improve the rate of convergence. The second section is a discussion about the degree of the solution. Finally, in the third section, we explain how the real polynomial Zolotarev problem (2.7) could be extended to a complex Zolotarev problem.

5.1 A rational Remes-like algorithm

In the case of approximation of continuous functions, the Remes algorithm was extended in order to handle rational approximation ([Werner, 1963]). This extended algorithm is proven to be convergent when the best rational approximation is "(m, n)-normal" (i.e. has a numerator degree equal to n or a denominator degree equal to m) and when the starting point of the algorithm is sufficiently close to the best approximation (e.g. [Braess, 1986]). Note that, in practice, the Remes rational algorithm is faster than the differential-correction algorithm, but it only converges if the initialization is "quite good".

Suppose that the rational Remes algorithm could be adapted to our case, and gives a process which is locally convergent when the solution is "(m, n)-normal". Therefore, we could compute a "rough" solution using the differential-correction-like algorithm (by discretizing the intervals with a small number of points), and next, we could refine this solution using the rational Remes process. Combining the differential-correction-like algorithm and the rational Remes algorithm would improve the time of computation of the solution.

We next present what would be an adaptation of such an algorithm to solve our Zolotarev problem. No proof of convergence is given.

The main idea of the rational exchange algorithm is the same as for the polynomial exchange algorithm, that is it consists in computing in an iterative way the alternating points which characterize the solution. The adapted algorithm for solving Problem (2.6)

would be:

Step 1: Initialization
Compute a "rough" solution $\frac{p}{q}$ of problem (2.6) using the differential correction-like algorithm. A criterion λ is found. Determine "extreme" points $\omega_1 < \omega_2 < \cdots < \omega_{m+n+2}$ of $\frac{p}{q}$. Associate to these points signs $s_1 < s_2 < \cdots < s_{n+m+2}$ as follow :

$$\begin{cases} \text{if } \omega_i \in I, s_i = sgn\left(\frac{p}{q}(\omega_i)\right) \\ \text{if } \omega_i \in J, s_i = -sgn\left(\frac{p}{q}(\omega_i)\right) \end{cases} \tag{5.1}$$

Step 2: Adaptation of the reference set
Look for the point where $\frac{p}{q}$ "deviates most" from a valid solution, either by exceeding the modulus bound on I or by reaching a minimal value on J that is smaller than the current λ, i.e. find ω such that

$$\left|\frac{p}{q}(\omega)\right| = \max\left(\max_I \frac{p}{q} - 1, \max_J \lambda - \frac{p}{q}\right).$$

Associate to this point a sign as in (5.1), and include the point in the reference set in order to keep $n+m+2$ alternating points.

Step 3: Solving the problem on the new reference set
Solve the following system of $n+m+2$ equations

$$\begin{aligned} \text{if } \omega_i \in I, p(\omega_i) &= s_i q(\omega_i), \\ \text{if } \omega_i \in J, p(\omega_i) &= s_i \lambda q(\omega_i), \end{aligned} \tag{5.2}$$

with unknowns λ, p and q.

The latter iterations between Step 2 and Step 3 are repeated until a rational function $\frac{p}{q}$ that satisfies the boundedness condition on I is computed. If the initialization is badly chosen, algorithm fails at Step 3 (the system (5.2) does not have any solution). The main difference with the polynomial algorithm is in the computation of the solution (if it exists) of system (5.2). Indeed, system (5.2) is not linear. However, it could be solved thanks to the following observation:
If $\alpha_i = \prod_{j=1, j\neq i}^{n+m+2} \frac{1}{(\omega_j - \omega_i)}$, then $\sum_{i=1}^{n+m+2} \alpha_i g(\omega_i) = 0$ for all polynomials g of degree less than $n+m$. Thus, from equations (5.2), we deduce that

$$\sum_{i\in I} \alpha_i s_i p(\omega_i)\omega_i^k + \sum_{i\in J} \alpha_i s_i \lambda p(\omega_i)\omega_i^k = 0, \quad \forall 0 \leq k \leq m.$$

Therefore, $AP = \lambda BP$, where A and B are $m \times n$ matrices defined by

$$A_{l,j} = \sum_{i\in I} \alpha_i s_i \omega_i^{j+l}$$

and

$$B_{l,j} = \sum_{i \in J} \alpha_i s_i \omega_i^{j+l}$$

for $0 \leq l, j \leq m$. Solving this generalized eigenvalue problem gives $m+1$ possibilities for λ and q. Including these solutions in (5.2) leads to a linear system. If the same argument as for rational approximation could be used ([Werner, 1963]), then at most one solution of this problem would be such that $(p, q) \in A_m^n$.

5.2 Degree of the solution

In this section, we keep the same notations than in Chapter 2.

Throughout this study, we saw that the degree of the solution of the Zolotarev problem is not always maximal. In the "simple" Zolotarev problem (2.5) where I and J are unions of intervals, one could ask whether a slight modification of the boundaries of the intervals of I and J could ensure at least the "(m, n)-normality" of the solution (i.e. that the degree of the numerator is n or the degree of the denominator is m). We have no answer to that question. However, we now show that, in the polynomial case (2.7), a slight modification of the boundaries of the intervals ensure that the degree of the solution is at least $n-1$. We recall that in Problem (2.7), I and J are finite unions of finite intervals and that X is defined by $X = I \cup J$.

Note that, when the degree of the solution is equal to $N < n$, a sequence of $N+2+k$ extreme points exists, $k \geq n - N$.

Lemma 5.2.1 *Let $p \in \mathcal{A}$ be a polynomial of degree $N \geq 1$. If p has $N+2+k$ ν_p-alternant extreme points, $k \geq 0$, then at least $k+3$ of these points are in $\partial X = (\overline{I} \cup \overline{J}) \setminus (\overset{\circ}{I} \cup \overset{\circ}{J})$ and are not root of the derivative of p.*

Proof Note that, if $w \in \overset{\circ}{X}$ is a ν_p-alternant extreme point, then w is a local extremum of p, and therefore the derivative of p vanishes at w. Since $\deg p = N$, the derivative of p vanishes at most $N-1$ times. Thus, the conclusion is immediate.

■

Let $\epsilon > 0$ and $x \in \mathbb{R}$. We denote by $B(x, \epsilon)$ the open interval $]x - \epsilon, x + \epsilon[$.

Definition 5.2.2 *Let $\epsilon > 0$.*
We say that a compact set V is ϵ-close to X if:

- $V \subset X$,
- $X \setminus \bigcup_{x \in \partial K} B(x, \epsilon) \subset V$.

Proposition 5.2.3 *Let $p \in \mathcal{A}$ be a polynomial of degree $n \geq 1$. If p has $n+2+k$ ν_p-alternant extreme points, $k \geq 2$, then there exists $\epsilon > 0$ and V ϵ-close to X such that p has exactly $n+2+\xi$ ν_p-alternant extreme points in V, $\xi \in \{0, 1\}$.*

Proof Let $\mathcal{B}$ be the set of ν_p-alternant extreme points in ∂X such that the derivative of p does not vanish at these points. Since p is not constant, $\mathcal{B}$ is finite, and using the previous proposition, this set contains at least $k+3$ points.

We next see how a slight modification of the boundaries of I and J allows to decrease the number of extreme points. Let $w \in \mathcal{B}$.
If $w \in \mathcal{B} \cap I$, then there exists $\epsilon_w > 0$ such that $B(w, \epsilon_w)$ does not contain other extreme points. Therefore, replacing I by $I \setminus B(w, \epsilon_w)$, the number of extreme points of p on $I \setminus B(w, \epsilon_w) \cup J$ is decreased by 1.
Similarly, if $w \in \mathcal{B} \cap J$, and if there is another extreme point $w_m \neq w \in J$, then there exists $\epsilon_w > 0$ such that $B(w, \epsilon_w)$ does not contain another extreme point. Therefore, replacing J by $J \setminus B(w, \epsilon_w)$, since $\mu_p = p(w_m)$, the minimum of p on $J \setminus B(w, \epsilon_w)$ is equal to the minimum of p on J. Note that the number of extreme points on $I \cup J \setminus B(w, \epsilon_w)$ is decreased by 1.

Using what precedes, applying a slight modification to the boundary of X, we can remove one extreme point. Then, the maximal length of a sequence of ν_p-alternant extreme points decreases by 0, 1 or 2. Since p is not constant, the number of extreme points is finite. Therefore, this process can be repeated until the maximal length of a sequence of extreme points is equal to $n + 2 + \xi$, $\xi \in \{0, 1\}$.

■

The following corollary is then immediate:

Corollary 5.2.4 *For every X, n, and ϵ, if the solution of the polynomial Zolotarev problem (2.7) on X is not constant, then there exists V ϵ-close to X such that degree of the solution of the Zolotarev problem (2.7) on V is at least $n-1$.*

5.3 A complex Zolotarev problem

We are interested in the following problem:

$$\text{find } p^* \text{ solution of: } \max_{\{p \in \mathcal{P}_n(\mathbb{C}), \|p\|_I \leq 1\}} \min_{\omega \in J} \sigma(\omega) p(\omega).$$

This is an extension of Problem (2.7) to the complex case. We recall that here, I and J are a sequence of closed real intervals, non reduced to a point.

Note that $|p^*|^2$ is a real polynomial of degree at most $2n$, positive over $\mathbb{R}$. We next see that this problem can be easily solved using our extended polynomial problem. We define two functions l and u by

$$l(x) = \begin{cases} 1 & \text{over } J \\ 0 & \text{otherwise} \end{cases}$$

and

$$u(x) = \begin{cases} 1 & \text{over } I \\ +\infty & \text{otherwise} \end{cases} .$$

and a sign function σ by $\sigma = +1$ everywhere. Solving the real generalized polynomial Zolotarev problem of degree $2n$ presented in section 3.1, we obtain a polynomial P^*. P^* is a real polynomial positive over $\mathbb{R}$. Therefore there exists a complex polynomial p^* such that $|p^*|^2 = P^*$. It is then straightforward that p^* is a solution to the complex Zolotarev

problem. Note that this process could be extended to a problem with fixed denominator.

Now, recall that the original problem associated to filtering functions is a "mixed" Zolotarev problem (see (2.4)), i.e. a problem with a complex polynomial as numerator, and a real polynomial as denominator. A "not trivial" lower bound to this problem could be obtained the following way. First, compute the solution to the real rational Zolotarev problem (2.6). A real rational function P/Q is obtained. Next, as precedes, compute the solution to the complex polynomial Zolotarev problem with weight $1/|Q|$. This gives a polynomial p^*. The rational function p^*/Q is a function such that $p^* \in \mathcal{P}_n(\mathbb{C})$ and $Q \in \mathcal{P}_m(\mathbb{R})$, and gives a lower bound to the "mixed" problem (2.4).

Note that, if the real rational Zolotarev problem was extended, as the polynomial case, to the entire real axis, we could also obtain a upper bound to the "mixed" problem. Indeed, computing the solution to the problem with degrees $(2n, 2m)$ would give the solution to the complex Zolotarev problem (i.e. the problem with complex polynomials as numerator and denominator).

Computing the solution to the "mixed" Zolotarev problem is still an open problem.

Part II

Schur rational approximation

In this part, we are interested in approximating a Schur function f by a rational function which is also Schur. A Schur function is an analytic function whose modulus is bounded by 1 in the unit disk. This problem of approximation is very important for the synthesis and identification of passive systems. The main idea is to use a generalized multipoint Schur algorithm, that is a Schur algorithm where all the reference points are not taken in 0 but are taken at points $(\alpha_j)_{j\geq 1}$ anywhere in the unit disk. Such an algorithm leads to a sequence of Schur rational functions that we are studying all along this part.

In the first chapter, we introduce the generalized Schur algorithm, and rewrite it as a continued fraction. We then give some basic properties of the convergents of this continued fraction. In particular, the convergents of even order are Schur rational functions which interpolate f at the points (α_j).

In the next chapter, we introduce the orthogonal rational functions on the unit circle and give all the basic results needed on this topic. Our main reference is the book [Bultheel et al., 1999].

The third chapter makes a connection between the Schur algorithm and the orthogonal rational functions. This is a generalization of the Geronimus theorem ([Geronimus, 1944], [Langer and Lasarow, 2004]) which states that the Schur parameters are equal to the Geronimus parameters of the orthogonal polynomials of the measure associated to f by the Herglotz transform.

The first three chapters are in fact all the necessary background to study the asymptotic properties of the convergents of even order. These properties are given in the fourth chapter, and are mainly a generalization of the work of Khrushchev ([Khrushchev, 2001]) who studied the L^2-convergence in the case of the classical algorithm. The difficulty here comes from the fact that we let the points go the circle.
In addition, we obtained a "Szegő condition" and a result of convergence for the Schur functions which seems to be asymptotically very close to a BMO convergence.

Finally, in the fifth chapter, we give some practical ways to approximate a Schur function by a rational function of a given order. We prove that any strictly Schur rational function of degree n can be written as the $2n$-th convergent of the Schur algorithm if the interpolation points are correctly chosen. This leads to a parametrization using the Schur algorithm. We give some details about it, and also explain how to compute effectively the L^2-norm. Some examples are computed using an optimization process, and the results are validated by a comparison with the unconstrained L^2 rational approximation.

Chapter 6

Notations and first definitions

This chapter presents some basic notations and definitions that will be used throughout our study.

We denote by $\mathbb{D}$ the unit disc $\mathbb{D} = \{z \in \mathbb{C}, |z| < 1\}$ and by $\mathbb{T}$ the unit circle $\mathbb{T} = \{z \in \mathbb{C}, |z| = 1\}$.

$H(\mathbb{D})$ and $C(\mathbb{D})$ represent respectively the set of analytic functions and the set of continuous functions over $\mathbb{D}$. We denote by $A(\mathbb{D})$ the disk algebra, i.e. the set of analytic functions in $\mathbb{D}$, continuous on $\overline{\mathbb{D}}$.

For a function f, we define the infinity norm $\|\cdot\|_\infty$ by $\|f\|_\infty = \sup_{z\in\mathbb{D}} |f(z)|$.

Definition 6.0.1 *An analytic function f on $\mathbb{D}$ such that $\|f\|_\infty \leq 1$ is called a Schur function. The set $\mathcal{S}$ of all Schur functions is called the Schur class $\mathcal{S}$.*
If f is an analytic function in $\mathbb{D}$ with $\|f\|_\infty < 1$, we will say that f is strictly Schur.

Let $\{z_n\}$ be a subset of $\mathbb{D} \setminus \{0\}$ and s be a nonnegative integer. A function of the form

$$B(z) = z^s \prod_n \frac{|z_n|}{z_n} \frac{z_n - z}{1 - \bar{z}_n z}$$

is called a Blaschke product. Furthermore, if the set $\{z_n\}$ is finite, it is called a finite Blaschke product.

It is well known (e.g. [Garnett, 2007] or [Rudin, 1987]) that if $\sum_n (1 - |z_n|) < \infty$, then B is in $H^\infty(\mathbb{D})$, the zeros of B are the points z_n (and 0 if $s > 0$) and $|B| = 1$ almost everywhere on $\mathbb{T}$. Therefore, $\sum_n (1 - |z_n|) < \infty$ is a sufficient condition for the existence of a non-zero function in $H^\infty(\mathbb{D})$ with given zeros $\{z_n\}$. In fact, this is also a necessary condition (e.g. [Garnett, 2007] or [Rudin, 1987]): the zeros z_n of a non-zero function in $H^\infty(\mathbb{D})$ satisfy $\sum_n (1 - |z_n|) < \infty$.
We will sometimes use the following corollary: if a function in $H^\infty(\mathbb{D})$ has an infinity of zeros at the points z_n and if $\sum_n (1 - |z_n|) = \infty$, then it is the zero function.

For a sequence $\{\alpha_k\}_{k=0}^\infty \subset \mathbb{D}$ with $\alpha_0 = 0$, we define the elementary Blaschke factors

$$\zeta_k = \frac{z - \alpha_k}{1 - \bar{\alpha}_k z} \quad , k \geq 0 \tag{6.1}$$

and the partial Blaschke products

$$\begin{cases} \mathcal{B}_0(z) = 1 \\ \mathcal{B}_k(z) = \mathcal{B}_{k-1}(z)\zeta_k(z) = \prod_{i=1}^{k} \frac{z-\alpha_i}{1-\bar{\alpha}_i z} \text{ for } k \geq 1. \end{cases} \tag{6.2}$$

The functions $\{\mathcal{B}_0, \mathcal{B}_1, \ldots, \mathcal{B}_n\}$ span the space

$$\mathcal{L}_n = \left\{ \frac{p_n}{\pi_n} \ : \ \pi_n(z) = \prod_{k=1}^{n} (1 - \bar{\alpha}_k z), \quad p_n \in \mathcal{P}_n \right\} \tag{6.3}$$

where $\mathcal{P}_n$ is the space of algebraic polynomials of degree at most n.
In particular, if all the α_k are equal to 0, the space $\mathcal{L}_n$ coincides with the space $\mathcal{P}_n$. Note that a function of $\mathcal{L}_n$ is analytic in $\mathbb{D}$.

For any function f, we introduce the parahermitian conjugate f_* defined by

$$f_*(z) = \overline{f(1/\bar{z})}. \tag{6.4}$$

Two useful and immediate equalities are $\zeta_{n*} = {\zeta_n}^{-1}$ and $\mathcal{B}_{k*} = {\mathcal{B}_k}^{-1}$.

We set for any function $f \in \mathcal{L}_n$:

$$f^* = \mathcal{B}_n f_*. \tag{6.5}$$

It is immediate to check that f^* is also in $\mathcal{L}_n$.
We denote by $\mathcal{B}_{n,i}$ the product $\prod_{k=i}^{k=n} \zeta_k$. If

$$f = a_n \mathcal{B}_n + a_{n-1}\mathcal{B}_{n-1} + \cdots + a_1 \mathcal{B}_1 + a_0$$

then

$$f^* = \bar{a}_0 \mathcal{B}_{n,1} + \bar{a}_1 \mathcal{B}_{n,2} + \cdots + \bar{a}_{n-2}\mathcal{B}_{n,n-1} + \bar{a}_{n-1}\mathcal{B}_{n,n} + \bar{a}_n.$$

Finally, we note that the leading coefficient a_n is given by

$$a_n = \overline{f^*(\alpha_n)}$$

and that

$$a_0 = f(\alpha_1).$$

We denote by m the normalized Lebesgue measure on $\mathbb{T} : m(\mathbb{T}) = 1$.

Now that all the main notations have been presented, we are able to begin with the study of the Schur algorithm.

Chapter 7

The Schur algorithm

Starting from a Schur function f, the classical Schur algorithm ([Schur, 1917]) gives a sequence of Schur functions $(f_k)_{k\in\mathbb{N}}$ and a sequence of complex numbers $(\gamma_k)_{k\in\mathbb{N}}$ as follows:

$$\begin{cases} f_0 = f, \\ \gamma_k = f_k(0), \\ f_{k+1}(z) = \dfrac{1}{z}\dfrac{f_k(z)-\gamma_k}{1-\overline{\gamma_k}f_k(z)}, \end{cases} \quad \text{for } k \geq 0.$$

Note that for every $k \in \mathbb{N}$, $\omega \mapsto \frac{\omega-\gamma_k}{1-\overline{\gamma_k}\omega}$ is a Moebius transform which maps $\mathbb{D}$ onto $\mathbb{D}$, so by the Schwarz lemma ([Garnett, 2007]) f_k is a Schur function for every $k \in \mathbb{N}$. An interesting property ([Bakonyi and Constantinescu, 1992]) of the Schur algorithm is that it realizes a one-to-one correspondence between the Schur class $\mathcal{S}$ and the sequence of complex numbers $(\gamma_k)_{k\in\mathbb{N}}$ having the properties: $|\gamma_k| \leq 1$ for $k \geq 0$, and if for a certain k_0, $|\gamma_{k_0}| = 1$, $f_{k_0}(z) = \gamma_{k_0}$ is a constant function and then $\gamma_k = 0$ for $k > k_0$.

Note that the Schur algorithm extends to operator-valued functions ([Potapov, 1955], [Ceauşescu and Foiaş, 1978]).

7.1 Multipoint Schur algorithm

In the classical algorithm, the Schur parameters γ_n are obtained by evaluating the functions f_n at 0. This process can be extended to more arbitrary evaluation points in $\mathbb{D}$ (e.g. [Jones, 1988], [Langer and Lasarow, 2004]). We next describe such an algorithm.

Let $\{\alpha_k\}_{k=1}^{\infty}$ be a sequence of points in $\mathbb{D}$ and $\{c_k\}_{k=0}^{\infty}$ be a sequence of points in $\mathbb{T}$ with $c_0 = 1$. Then, the generalized Schur algorithm is :

For $k \geq 0$, f_k and γ_k are defined by

$$\begin{cases} f_0 = f \\ \gamma_k = \bar{c}_k f_k(\alpha_{k+1}) \\ f_{k+1} = \dfrac{1}{\zeta_{k+1}}\dfrac{\bar{c}_k f_k - \gamma_k}{1-\bar{\gamma}_k\bar{c}_k f_k} \end{cases} \quad \text{for } k \geq 0,$$

where ζ_k is the Moebius transform defined by (6.1).
If $|\gamma_k| = 1$, the algorithm stops.

The parameters (α_k) are the interpolations points. They are those parameters equal to 0 in the classical Schur algorithm, which are presently taken anywhere in the disk. The parameters (c_k) have modulus equal to 1, and are rotations applied to the f_k at each step of the algorithm. Note that the (c_k) can also be seen as normalization parameters of the Moebius transforms since

$$\begin{aligned} f_{k+1} &= \frac{1}{\zeta_{k+1}} \frac{\bar{c}_k f_k - \gamma_k}{1 - \bar{\gamma}_k \bar{c}_k f_k} \\ &= \frac{\bar{c}_k}{\zeta_{k+1}} \frac{f_k - c_k \gamma_k}{1 - \bar{\gamma}_k \bar{c}_k f_k} \\ &= \frac{1}{c_k \zeta_{k+1}} \frac{f_k - f_k(\alpha_{k+1})}{1 - \overline{f_k(\alpha_{k+1})} f_k}. \end{aligned}$$

As in the classical case, the sequence $(f_n)_{n\in\mathbb{N}}$ is a sequence of Schur functions, therefore the $(\gamma_n)_{n\in\mathbb{N}}$ lie in $\overline{\mathbb{D}}$.

Definition 7.1.1 *The sequence $(\gamma_n)_{n\in\mathbb{N}}$ is called the sequence of Schur parameters of the Schur function f associated to the sequence (α_k).*

The Schur parameters depend only on the values of f and its derivatives $f^{(j)}$ at the points $(\alpha_k)_k$. More precisely,

Proposition 7.1.2 *For $k \in \mathbb{N}$, γ_k depends only on the values $f^{(i)}(\alpha_{j+1})$, $0 \le j \le k$, $0 \le i < m_{j+1}$, where m_{j+1} is the multiplicity of α_{j+1} at the k-th step, i.e. m_{j+1} is the cardinality of the set $\{l,\, 0 \le l \le k,\, \alpha_{l+1} = \alpha_{j+1}\}$.*

Proof Noticing that $f_j(\alpha_j) = f'_{j-1}(\alpha_j)\bar{c}_{j-1}\bar{z}_j \frac{1-|\alpha_j|^2}{1-|f_{j-1}(\alpha_j)|^2}$, the proof is immediate by induction.

■

The Schur algorithm can be reversed in order to express f_{k-1} as a function of f_k. We obtain

$$f_{k-1} = c_{k-1} \frac{\zeta_k f_k + \gamma_{k-1}}{1 + \bar{\gamma}_{k-1}\zeta_k f_k} = c_{k-1}\gamma_{k-1} + \frac{(1 - |\gamma_{k-1}|^2) c_{k-1}\zeta_k}{\bar{\gamma}_{k-1}\zeta_k + \frac{1}{f_k}}. \tag{7.1}$$

We denote by τ_k the map

$$\begin{aligned} \tau_k : \quad \overline{\mathbb{D}} &\longrightarrow \mathcal{S} \\ \omega &\longmapsto \tau_k(\omega) = \begin{cases} c_k\gamma_k + \frac{(1-|\gamma_k|^2)c_k\zeta_{k+1}}{\bar{\gamma}_k\zeta_{k+1} + \frac{1}{\omega}} & \text{if } \omega \neq 0, \\ c_k\gamma_k & \text{if } \omega = 0. \end{cases} \end{aligned}$$

Note that we should write $\tau_k(\omega)(z)$ because $\tau_k(\omega)$ is a Schur function of z through ζ_{k+1}. Much of the recursive complexity of the Schur algorithm lies in the fact that we shall substitute to ω a function of z to make $\tau_k(\omega(z))(z)$ a function of z only. In particular, we have $f_k = \tau_k(f_{k+1})$. Therefore, f is equal to

$$f = \tau_0 \circ \tau_1 \circ \cdots \circ \tau_n(f_{n+1}). \tag{7.2}$$

Proposition 7.1.3 *The Schur algorithm stops if and only if f is a finite Blaschke product.*

Proof For p a polynomial, denote by $\tilde{p}$ the polynomial $z^n\overline{p\left(\frac{1}{\bar{z}}\right)}$ where n is the degree of p.
Suppose that f_n is a Blaschke product of degree n. Then f_n can be expressed as $\frac{p}{\tilde{p}}$ where p has its roots in $\mathbb{D}$, so

$$f_{n+1} = \frac{1-\overline{\alpha_{n+1}}z}{z-\alpha_{n+1}}\frac{\bar{c}_n p - \gamma_n\tilde{p}}{\tilde{p}-\bar{\gamma}_n\bar{c}_n p}.$$

Let $P_0 = \bar{c}_n p - \gamma_n\tilde{p}$. Then $\tilde{P}_0 = c_n(\tilde{p} - \bar{c}_n\bar{\gamma}_n p)$, so $\frac{\bar{c}_n p-\gamma_n\tilde{p}}{\tilde{p}-\bar{\gamma}_n\bar{c}_n p}$ is of the form $\frac{P}{\tilde{P}}$ for some polynomial P. Note that, since $\bar{c}_n f_n(\alpha_{n+1}) = \gamma_n$, P vanishes at α_{n+1}. Therefore f_{n+1} is a Blaschke product of degree $n-1$. Thus, if f is a Blaschke product of degree n, f_n is a Blaschke product of degree 0, i.e. a constant of modulus 1, and the algorithm stops.
Conversely, if $f_k = \frac{p}{\tilde{p}}$ is a Blaschke product of degree $n-k$, then

$$f_{k-1} = c_{k-1}\frac{(z-\alpha_k)p + \gamma_{k-1}\tilde{p}(1-\overline{\alpha_k}z)}{\tilde{p}(1-\overline{\alpha_k}z) + \bar{\gamma}_{k-1}(z-\alpha_k)p},$$

so f_{k-1} is a Blaschke product of degree at most $n-k+1$. In fact, using the first part of the proof, we get that f_{k-1} is exactly of degree $n-k+1$ (otherwise f_k is not of degree $n-k$). Therefore, if f_n is a constant of modulus 1, f is a Blaschke product of degree n.
■

7.2 Continued fractions

In this section, we give a very short introduction to continued fractions. Many good references, such as [Wall, 1948], can be found on this topic.

A continued fraction is an infinite expression of the form

$$b_0 + \frac{a_1}{b_1 + \frac{a_2}{b_2 + \frac{a_3}{b_3 + \frac{a_4}{\cdots}}}}.$$

also denoted for economy of space by

$$b_0 + \frac{a_1}{b_1} \;+\; \frac{a_2}{b_2} \;+\; \frac{a_3}{b_3} \;+\ldots$$

Let $t_0(\omega) = b_0 + \omega$ and

$$t_k(\omega) = \frac{a_k}{b_k + \omega} \quad \text{for } k \geq 1.$$

We call the n-th convergent, and we denote by P_n/Q_n, the fraction

$$\frac{P_n}{Q_n} = t_0 \circ t_1 \circ \cdots \circ t_n(0) = b_0 + \frac{a_1}{b_1 + \frac{a_2}{b_2 + \ddots + \frac{a_n}{b_n}}}.$$

Proposition 7.2.1 *The quantities P_n and Q_n are given by the recurrence relations*

$$\begin{cases} P_{-1} = 1, Q_{-1} = 0, \\ P_0 = b_0, Q_0 = 1, \\ P_{k+1} = b_{k+1}P_k + a_{k+1}P_{k-1} \\ Q_{k+1} = b_{k+1}Q_k + a_{k+1}Q_{k-1} \end{cases}$$

for all non-negative k.
More generally,

$$t_0 \circ t_1 \circ \cdots \circ t_n(\omega) = \frac{P_{n-1}\omega + P_n}{Q_{n-1}\omega + Q_n}.$$

Proof By induction. We have

$$t_0(\omega) = b_0 + \omega = \frac{P_{-1}\omega + P_0}{Q_{-1}\omega + Q_0}.$$

Suppose the statement true for k. Then

$$\begin{aligned} t_0 \circ t_1 \circ \cdots \circ t_{k+1}(\omega) &= \frac{P_{k-1}\frac{a_{k+1}}{b_{k+1}+\omega} + P_k}{Q_{k-1}\frac{a_{k+1}}{b_{k+1}+\omega} + Q_k} \\ &= \frac{P_k\omega + b_{k+1}P_k + a_{k+1}P_{k-1}}{Q_k\omega + b_{k+1}Q_k + a_{k+1}Q_{k-1}} \\ &= \frac{P_k\omega + P_{k+1}}{Q_k\omega + Q_{k+1}}. \end{aligned}$$

This gives the announced result.

■

7.3 Wall rational functions

In this section, we follow the same scheme as in ([Khrushchev, 2001]).

Let $(d_k)_{k\in\mathbb{N}}$ be a sequence of points on the unit circle $\mathbb{T}$, with $d_0 = 1$. We now define the c_k of the Schur algorithm by $c_k = d_k^2$. Let (α_k) be a sequence of points in the unit disk $\mathbb{D}$. Recall from (7.2) that $f = \tau_0 \circ \tau_1 \circ \cdots \circ \tau_n(f_{n+1})$ with

$$\tau_k(\omega) = c_k\gamma_k + \frac{(1 - |\gamma_k|^2)c_k\zeta_{k+1}}{\bar{\gamma}_k\zeta_{k+1} + \frac{1}{\omega}}.$$

A rational Schur function R_n of degree at most n can be obtained by interrupting the Schur algorithm at step n, that is, by replacing f_{n+1} by 0:

$$\begin{aligned} R_n &= \tau_0 \circ \tau_1 \circ \cdots \circ \tau_{n-1} \circ \tau_n(0) \\ &= \tau_0 \circ \tau_1 \circ \cdots \circ \tau_{n-1}(c_n\gamma_n). \end{aligned} \tag{7.3}$$

The rational functions R_n play a key role in what follows. Indeed, we will see later how to approximate f using the sequence (R_n). Therefore, we will now pay a particular attention to the properties of these rational functions. The first one is an interpolation property:

Theorem 7.3.1 *The rational function R_n interpolates f at the points α_k, $1 \leq k \leq n+1$, and has the same $n+1$ first Schur parameters as f.*

Proof Note that $\tau_k(\omega)(\alpha_{k+1})$ is independent of ω. Indeed, $\tau_k(\omega)(\alpha_{k+1}) = c_k\gamma_k$. Let k be an integer such that $0 \leq k \leq n$. Then:

$$\begin{aligned} f(\alpha_{k+1}) &= \tau_0 \circ \cdots \circ \tau_k(\tau_{k+1} \circ \cdots \circ \tau_n \circ f_{n+1})(\alpha_{k+1}) \\ &= \tau_0 \circ \cdots \circ \tau_k(\tau_{k+1} \circ \cdots \circ \tau_n(0))(\alpha_{k+1}) \\ &= R_n(\alpha_{k+1}). \end{aligned}$$

Thus, R_n interpolates f at the point α_{k+1}.
We next prove by induction that f and R_n have the same $n+1$ first Schur parameters. Using what precedes, we get that f and R_n have the same first Schur parameter γ_0. Now, suppose that the k first Schur parameters of f and R_n are equal. Then, if we denote by $R_n^{[1]}, \ldots R_n^{[n]}$ the Schur functions of R_n obtained through the Schur algorithm, $R_n^{[k]}$ is equal to $\tau_{k-1}^{-1} \circ \cdots \circ \tau_0^{-1}(R_n)$. Thus,

$$\begin{aligned} R_n^{[k]}(\alpha_{k+1}) &= \tau_{k-1}^{-1} \circ \cdots \circ \tau_0^{-1}(R_n)(\alpha_{k+1}) \\ &= \tau_{k-1}^{-1} \circ \cdots \circ \tau_0^{-1} \circ \tau_0 \circ \tau_1 \circ \cdots \circ \tau_{n-1}(c_n\gamma_n)(\alpha_{k+1}) \\ &= \tau_k \circ \cdots \circ \tau_{n-1}(c_n\gamma_n)(\alpha_{k+1}) = c_k\gamma_k \end{aligned}$$

since $\tau_k(\omega)(\alpha_{k+1}) = c_k\gamma_k$. Therefore, the $k+1$-st Schur parameter of R_n is equal to the $k+1$-st Schur parameter of f.

■

The previous theorem leads to the existence of a function with given Schur parameters:

Corollary 7.3.2 *Let $\check{\gamma}_i$, $0 \leq i \leq n-1$, be n points in the unit disk $\mathbb{D}$ and c_i, $0 \leq i \leq n-1$, be n points on the unit circle $\mathbb{T}$. Then, there is a Schur function whose n first Schur parameters are the $\check{\gamma}_i$, $0 \leq i \leq n-1$.*

Proof Using the previous theorem, the function

$$\check{R}_n = \check{\tau}_0 \circ \cdots \circ \check{\tau}_{n-1}(c_n\check{\gamma}_n)$$

where

$$\check{\tau}_k(\omega) = c_k\check{\gamma}_k + \frac{(1-|\check{\gamma}_k|^2)c_k\zeta_{k+1}}{\overline{\check{\gamma}_k}\zeta_{k+1} + \frac{1}{\omega}}$$

satisfies the announced condition.

■

We are now going to study the sequence of rational Schur function R_n using continued fractions. We note $\frac{P_n}{Q_n}$ the sequence of convergents associated to the continued fraction

$$c_0\gamma_0 + \frac{(1-|\gamma_0|^2)c_0\zeta_1}{\bar{\gamma}_0\zeta_1} \; + \; \frac{1}{c_1\gamma_1} \; + \; \frac{(1-|\gamma_1|^2)c_1\zeta_2}{\bar{\gamma}_1\zeta_2} \; + \ldots \tag{7.4}$$

so that the R_n are the convergents of even index: $R_n = \frac{P_{2n}}{Q_{2n}}$.

By proposition 7.2.1, for $n \geq 1$:

$$\begin{aligned} P_{2n} &= c_n\gamma_n P_{2n-1} + P_{2n-2} \\ Q_{2n} &= c_n\gamma_n Q_{2n-1} + Q_{2n-2} \\ P_{2n-1} &= \bar{\gamma}_{n-1}\zeta_n P_{2n-2} + (1-|\gamma_{n-1}|^2)c_{n-1}\zeta_n P_{2n-3} \\ Q_{2n-1} &= \bar{\gamma}_{n-1}\zeta_n Q_{2n-2} + (1-|\gamma_{n-1}|^2)c_{n-1}\zeta_n Q_{2n-3} \end{aligned} \tag{7.5}$$

with

$$P_{-1} = 1, \quad P_0 = c_0\gamma_0 = \gamma_0, \quad Q_{-1} = 0, \quad Q_0 = 1.$$

Our purpose is now to give explicit formulas in order to compute R_n, that is formulas for P_{2n} and Q_{2n}. The following lemma expresses the relations between the rational functions of even and odd order. We shall make the convention that $Q^*_{2n} = \mathcal{B}_n Q_{2n*}$ and $Q^*_{2n+1} = \mathcal{B}_{n+1}Q_{2n+1*}$ and similarly for P_{2n} and P_{2n+1}. It will actually follow from the lemma that this convention agrees with definition (6.5), in that we will have $P_{2n+1}, Q_{2n+1} \in \mathcal{L}_{n+1}$ and $P_{2n}, Q_{2n} \in \mathcal{L}_n$ by (7.5).

Lemma 7.3.3 *For $n \geq 0$, we have*

$$P_{2n+1} = C_n\zeta_{n+1}Q^*_{2n}, \quad Q_{2n+1} = C_n\zeta_{n+1}P^*_{2n}$$

where $C_n = \prod_{k=0}^{k=n} c_k \in \mathbb{T}$.

Proof For $n = 0$ we have

$$P_1 = \bar{\gamma}_0\zeta_1 c_0\gamma_0 + (1-|\gamma_0|^2)c_0\zeta_1 = c_0\zeta_1 Q^*_0$$

and

$$Q_1 = \bar{\gamma}_0\zeta_1 = c_0\zeta_1 P^*_0.$$

Assuming the hypothesis is true for all indices smaller than n, we obtain that

$$\begin{aligned} C_n\zeta_{n+1}Q^*_{2n} &= C_n\zeta_{n+1}(c_n\gamma_n Q_{2n-1} + Q_{2n-2})^* \\ &= C_n\zeta_{n+1}(\bar{c}_n\bar{\gamma}_n Q^*_{2n-1} + \zeta_n Q^*_{2n-2}) \\ &= C_{n-1}\zeta_{n+1}(\bar{\gamma}_n Q^*_{2n-1} + c_n\zeta_n Q^*_{2n-2}) \\ &= C_{n-1}\zeta_{n+1}(\bar{\gamma}_n\bar{C}_{n-1}P_{2n-2} + c_n\bar{C}_{n-1}P_{2n-1}) \\ &= \zeta_{n+1}(\bar{\gamma}_n P_{2n-2} + c_n P_{2n-1}) \\ &= \zeta_{n+1}(\bar{\gamma}_n P_{2n} - c_n|\gamma_n|^2 P_{2n-1} + c_n P_{2n-1}) \\ &= P_{2n+1}. \end{aligned}$$

This yields the first relation of the lemma. The proof of the other relation is similar. ■

From (7.5), we have for $n \geq 1$:

$$\begin{aligned} P_{2n+1} &= \bar{\gamma}_n\zeta_{n+1}P_{2n} + (1-|\gamma_n|^2)c_n\zeta_{n+1}P_{2n-1} \\ &= \bar{\gamma}_n\zeta_{n+1}(c_n\gamma_n P_{2n-1} + P_{2n-2}) + (1-|\gamma_n|^2)c_n\zeta_{n+1}P_{2n-1} \\ &= \bar{\gamma}_n\zeta_{n+1}P_{2n-2} + c_n\zeta_{n+1}P_{2n-1} \end{aligned}$$

and similarly $Q_{2n+1} = \bar{\gamma}_n \zeta_{n+1} Q_{2n-2} + c_n \zeta_{n+1} Q_{2n-1}$ so that

$$\begin{aligned}\begin{bmatrix} P_{2n+1} & Q_{2n+1} \\ P_{2n} & Q_{2n} \end{bmatrix} &= \begin{bmatrix} c_n\zeta_{n+1} & \bar{\gamma}_n\zeta_{n+1} \\ \gamma_n c_n & 1 \end{bmatrix} \begin{bmatrix} P_{2n-1} & Q_{2n-1} \\ P_{2n-2} & Q_{2n-2} \end{bmatrix} \\ &= \begin{bmatrix} \zeta_{n+1} & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & \bar{\gamma}_n \\ \gamma_n & 1 \end{bmatrix} \begin{bmatrix} c_n & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} P_{2n-1} & Q_{2n-1} \\ P_{2n-2} & Q_{2n-2} \end{bmatrix}.\end{aligned}$$

Therefore

$$\begin{aligned}&\begin{bmatrix} c_{n+1} & 0 \\ 0 & \zeta_{n+1} \end{bmatrix} \begin{bmatrix} P_{2n+1} & Q_{2n+1} \\ P_{2n} & Q_{2n} \end{bmatrix} \\ &= \tfrac{\zeta_{n+1}}{\zeta_n} \begin{bmatrix} c_{n+1} & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & \bar{\gamma}_n \\ \gamma_n & 1 \end{bmatrix} \begin{bmatrix} \zeta_n & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} c_n & 0 \\ 0 & \zeta_n \end{bmatrix} \begin{bmatrix} P_{2n-1} & Q_{2n-1} \\ P_{2n-2} & Q_{2n-2} \end{bmatrix}.\end{aligned}$$

Thus, using the previous lemma,

$$\begin{aligned}&\begin{bmatrix} C_{n+1} & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} Q^*_{2n} & P^*_{2n} \\ P_{2n} & Q_{2n} \end{bmatrix} \\ &= \begin{bmatrix} c_{n+1} & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & \bar{\gamma}_n \\ \gamma_n & 1 \end{bmatrix} \begin{bmatrix} \zeta_n & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} C_n & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} Q^*_{2n-2} & P^*_{2n-2} \\ P_{2n-2} & Q_{2n-2} \end{bmatrix}.\end{aligned} \tag{7.6}$$

Iterating, we get

$$\begin{aligned}&\begin{bmatrix} C_{n+1}Q^*_{2n} & C_{n+1}P^*_{2n} \\ P_{2n} & Q_{2n} \end{bmatrix} \\ &= \left(\prod_{k=n}^{k=1} \begin{bmatrix} c_{k+1} & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & \bar{\gamma}_k \\ \gamma_k & 1 \end{bmatrix} \begin{bmatrix} \zeta_k & 0 \\ 0 & 1 \end{bmatrix} \right) \begin{bmatrix} c_1 & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & \bar{\gamma}_0 \\ \gamma_0 & 1 \end{bmatrix}.\end{aligned} \tag{7.7}$$

Let $\Sigma_n = \prod_{k=0}^{n} d_k$. Note that, by definition of c_k, we have $\Sigma_n^2 = C_n$. We choose as representative of R_n the rational function $R_n = \frac{A_n}{B_n}$ with $A_n = \bar{\Sigma}_n P_{2n}$ and $B_n = \bar{\Sigma}_n Q_{2n}$.

Definition 7.3.4 *A_n and B_n are called the n-th Wall rational functions associated to the Schur function f and the sequences (α_k) and (d_k).*

As pointed out before, R_n plays a key role in the theory. This role will now be emphasized through the Wall rational functions A_n and B_n.
From what precedes, we have :

Proposition 7.3.5 *The Wall rational functions A_n and B_n are given by the formula*

$$\begin{aligned}&\Sigma_n \begin{bmatrix} c_{n+1} & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} B^*_n & A^*_n \\ A_n & B_n \end{bmatrix} \\ &= \left(\prod_{k=n}^{k=1} \begin{bmatrix} c_{k+1} & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & \bar{\gamma}_k \\ \gamma_k & 1 \end{bmatrix} \begin{bmatrix} \zeta_k & 0 \\ 0 & 1 \end{bmatrix} \right) \begin{bmatrix} c_1 & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & \bar{\gamma}_0 \\ \gamma_0 & 1 \end{bmatrix}\end{aligned}$$

with

$$\Sigma_n = \prod_{k=0}^{k=n} d_k.$$

Corollary 7.3.6 *A_n and B_n have the following properties :*

1. $B_n(z)B_n^*(z) - A_n(z)A_n^*(z) = \mathcal{B}_n(z)\omega_n$,
2. $|B_n(\xi)|^2 - |A_n(\xi)|^2 = \omega_n$ *for* $\xi \in \mathbb{T}$,
3. $f(\alpha_i) = \frac{A_n}{B_n}(\alpha_i) = \frac{B_n^*}{A_n^*}(\alpha_i)$ *for all* $1 \leq i \leq n+1$

with

$$\omega_n = \prod_{k=0}^{k=n} (1 - |\gamma_k|^2).$$

Proof By taking the determinant, we obtain from (7.7) that

$$\begin{aligned} B_n(z)B_n^*(z) - A_n(z)A_n^*(z) &= Q_{2n}(z)Q_{2n}^*(z) - P_{2n}(z)P_{2n}^*(z) \\ &= \mathcal{B}_n(z)\prod_{k=0}^{n}(1 - |\gamma_k|^2). \end{aligned}$$

The conclusion is then immediate.

■

Important properties of the Wall rational functions are:

Proposition 7.3.7 *For all $n \geq 0$:*

1. *B_n is an analytic function which does not vanish on $\overline{\mathbb{D}}$,*
2. *$\frac{A_n^*}{B_n}$ is a Schur function.*

Proof The proof will be given for P_{2n} and Q_{2n}. Since $P_0 = \gamma_0$ and $Q_0 = 1$, P_0 and Q_0 are two analytic functions and Q_0 does not vanish on $\overline{\mathbb{D}}$. Let us assume that these hypothesis are true for n. Then both functions $\frac{P_{2n}}{Q_{2n}}$ and $\frac{P_{2n}^*}{Q_{2n}}$ are analytic on $\overline{\mathbb{D}}$. From corollary 7.3.6, and by the maximum principle, these two functions are Schur. Furthermore, from (7.6), it is immediate that P_{2n+2} and Q_{2n+2} are both analytic in the disk and that

$$\begin{aligned} |Q_{2n+2}(z)| &= |\zeta_{n+1}(z)C_{n+1}\gamma_{n+1}P_{2n}^*(z) + Q_{2n}(z)| \\ &\geq |Q_{2n}(z)|\left(1 - |\gamma_{n+1}|\left|\frac{A_n^*}{B_n}\right|\right) > 0. \end{aligned}$$

■

The Wall rational functions A_n and B_n are related to f by the following formula:

Theorem 7.3.8 *The Wall rational functions A_n and B_n are rational functions $\in \mathcal{L}_n$ such that*

$$f(z) = \frac{A_n(z) + \zeta_{n+1}(z)B_n^*(z)f_{n+1}(z)}{B_n(z) + \zeta_{n+1}(z)A_n^*(z)f_{n+1}(z)}.$$

Proof Proposition 7.2.1 applied to the continuous fraction (7.4) gives us in view of (7.2)

$$f(z) = \frac{P_{2n}\frac{1}{f_{n+1}} + P_{2n+1}}{Q_{2n}\frac{1}{f_{n+1}} + Q_{2n+1}} = \frac{P_{2n} + P_{2n+1}f_{n+1}}{Q_{2n} + Q_{2n+1}f_{n+1}}.$$

But using lemma 7.3.3, we get

$$\begin{aligned} f(z) &= \frac{P_{2n} + C_n\zeta_{n+1}Q^*_{2n}f_{n+1}}{Q_{2n} + C_n\zeta_{n+1}P^*_{2n}f_{n+1}} \\ &= \frac{\overline{C_n}^{1/2}P_{2n} + C_n^{1/2}\zeta_{n+1}Q^*_{2n}f_{n+1}}{\overline{C_n}^{1/2}Q_{2n} + C_n^{1/2}\zeta_{n+1}P^*_{2n}f_{n+1}} \\ &= \frac{A_n + \zeta_{n+1}B^*_n f_{n+1}}{B_n + \zeta_{n+1}A^*_n f_{n+1}}. \end{aligned}$$

■

Chapter 8

Orthogonal rational functions on the unit circle

Orthogonal rational functions have been widely studied ([Djrbashian, 1962], [Pan, 1996], [Bultheel et al., 1999]). We recall here the main aspects of this theory. Its remarkable feature is to make connection with the Schur algorithm as we shall see in the next chapter.

8.1 Reproducing kernel Hilbert spaces

Good references on reproducing kernel Hilbert spaces are [Schwartz, 1964], [Dym, 1989] and [Alpay, 2001]. We recall here, mostly without proof, the properties that will be useful in what follows. We will write RKHS for "Reproducing Kernel Hilbert Space".

A RKHS is a complex-valued function Hilbert space in which pointwise evaluation is a continuous linear function, that is:

Definition 8.1.1 *Let X be an arbitrary set and H be an Hilbert space of complex valued functions on X. H is a RKHS if and only if the linear map $f \mapsto f(x)$ from H to $\mathbb{C}$ is continuous for each $x \in X$.*

From the Riesz-Fréchet theorem ([Rudin, 1987]), for $\omega \in X$ there exists a unique function $k(.,\omega)$ in H such that

$$f(\omega) = \langle f, k(.,\omega)\rangle \quad \forall f \in H.$$

Definition 8.1.2 *The function $(z,\omega) \mapsto k(z,\omega)$ from $X \times X$ to $\mathbb{C}$ such that*

$$f(\omega) = \langle f, k(.,\omega)\rangle \quad \forall f \in H \tag{8.1}$$

is called the reproducing kernel of H. The reproducing kernel is clearly unique.

The reproducing kernel is a Hermitian function, that is

$$\forall z \in X, \forall \omega \in X, k(z,\omega) = \overline{k(\omega,z)}.$$

Since in a Hilbert space of finite dimension pointwise evaluation is always continuous, we have

Proposition 8.1.3 *A Hilbert space of functions of finite dimension is a RKHS.*

The result we mainly use throughout is:

Proposition 8.1.4 *If H is a RKHS, and if (e_n) is an orthonormal basis, then the reproducing kernel k of H is equal to*

$$k(z,w) = \sum_n e_n(z)\overline{e_n(w)}. \tag{8.2}$$

Proof First, note that if $\dim(H) = \infty$, $\sum_n e_n(z)\overline{e_n(w)}$ converges in H. Indeed, we have $\sum_n |e_n(w)|^2 = \sum_n \langle e_n(.), k(.,\omega)\rangle = \|k(.,\omega)\|_2 < +\infty$ because $k(.,\omega) \in H$.

We next prove the equality (8.2). Let f in H. Expressing f in the basis (e_n), we obtain that $f = \sum_n a_n e_n$ for some $a_n \in \mathbb{C}$. Thus,

$$\begin{aligned}
\langle f, \sum_n e_n(.)\overline{e_n(w)}\rangle &= \langle \sum_n a_n e_n(.), \sum_n e_n(.)\overline{e_n(w)}\rangle \\
&= \sum_n \langle a_n e_n(.), \overline{e_n(w)} e_n(.)\rangle \\
&= \sum_n a_n e_n(\omega) \\
&= f(\omega).
\end{aligned}$$

As the reproducing kernel is unique, we get

$$k(z,w) = \sum_n e_n(z)\overline{e_n(w)}.$$

■

8.2 Christoffel-Darboux formulas in $\mathcal{L}_n$

Let μ be a real probability measure on the unit circle $\mathbb{T}$ with infinite support and $L^2(\mu)$ the familiar Hilbert space with inner product

$$\langle f, g\rangle_\mu = \int_{\mathbb{T}} f(\xi)\overline{g(\xi)} d\mu(\xi).$$

The space $\mathcal{L}_n$ endowed with the inner product $< .,. >_\mu$ is a Hilbert space of finite dimension, so it is a RKHS. Therefore, there exists a reproducing kernel $k_n(z,w)$ such that for every point $w \in \mathbb{D}$, $k_n(z,w) \in \mathcal{L}_n$ as a function of z and

$$\forall f \in \mathcal{L}_n, \forall w \in \mathbb{D}, f(w) = \langle f(.), k_n(.,w)\rangle_\mu. \tag{8.3}$$

Let us denote by $\{\phi_0, \phi_1, \ldots, \phi_n\}$ an orthonormal basis for $\mathcal{L}_n$ such that $\phi_0 = 1$ and $\phi_k \in \mathcal{L}_k \setminus \mathcal{L}_{k-1}$. Such a basis is easily obtained by the Gram-Schmidt orthonormalization process applied to $\mathcal{B}_0, \mathcal{B}_1, \ldots, \mathcal{B}_n$. We can write

$$\phi_n = a_{n,n}\mathcal{B}_n + a_{n,n-1}\mathcal{B}_{n-1} + \ldots + a_{n,1}\mathcal{B}_1 + a_{n,0}\mathcal{B}_0, \quad a_{n,n} = \kappa_n. \tag{8.4}$$

Note that $\kappa_n = \overline{\phi_n^*(\alpha_n)}$.

For $0 \leq k \leq n$, $\mathcal{B}_n\phi_{k*}$ is in $\mathcal{L}_n$. Moreover, $\{\mathcal{B}_n\phi_{0*}, \mathcal{B}_n\phi_{1*}, \ldots, \mathcal{B}_n\phi_{n*}\}$ is also an orthonormal basis, since

$$\langle \mathcal{B}_n\phi_{k*}, \mathcal{B}_n\phi_{l*}\rangle_\mu = \int_{\mathbb{T}} |\mathcal{B}_n(\xi)|^2 \overline{\phi_k(\xi)}\phi_l(\xi)d\mu(\xi) = \delta_{k,l}.$$

Using this new basis to compute the reproducing kernel, we get by (8.2) that

$$k_n(z,w) = \mathcal{B}_n(z)\overline{\mathcal{B}_n(w)} \sum_{k=0}^{n} \phi_{k*}(z)\overline{\phi_{k*}(w)}. \tag{8.5}$$

Letting $w \to \alpha_n$, since $\mathcal{B}_n(\alpha_n) = 0$ and no term is singular except if $k = n$, every term in the sum vanishes except for $k = n$, and computing the limit we have

$$\begin{aligned} k_n(z,\alpha_n) &= \mathcal{B}_n(z)\phi_{n*}(z) \lim_{w\to\alpha_n} \overline{\mathcal{B}_n(w)\phi_{n*}(w)} \\ &= \phi_n^*(z)\overline{\phi_n^*(\alpha_n)} \\ &= \kappa_n\phi_n^*(z). \end{aligned} \tag{8.6}$$

In particular, $k_n(\alpha_n,\alpha_n) = |\kappa_n|^2$. From (8.5) we may write

$$\frac{k_n(z,w)}{\mathcal{B}_n(z)\overline{\mathcal{B}_n(w)}} - \frac{k_{n-1}(z,w)}{\mathcal{B}_{n-1}(z)\overline{\mathcal{B}_{n-1}(w)}} = \phi_{n*}(z)\overline{\phi_{n*}(w)}, \quad n \geq 1.$$

Multiplying by $\mathcal{B}_n(z)\overline{\mathcal{B}_n(w)}$ gives the following important relation:

$$k_n(z,w) - \zeta_n(z)\overline{\zeta_n(w)}k_{n-1}(z,w) = \phi_n^*(z)\overline{\phi_n^*(w)}. \tag{8.7}$$

Using (8.2) with the orthonormal basis $(\phi_0, \ldots, \phi_n)$, we also have that

$$k_n(z,w) = k_{n-1}(z,w) + \phi_n(z)\overline{\phi_n(w)}, \quad n \geq 1. \tag{8.8}$$

We may use this relation to replace either $k_n(z,w)$ or $k_{n-1}(z,w)$ in relation (8.7) and then compute the other one. We get this way the following Christoffel-Darboux relations ([Bultheel et al., 1999], Theorem 3.1.3):

Proposition 8.2.1 *For z and w in $\mathbb{C}$ such that z and w do not coincide on $\mathbb{T}$, and for $n \geq 1$, we have*

$$k_{n-1}(z,w) = \frac{\phi_n^*(z)\overline{\phi_n^*(w)} - \phi_n(z)\overline{\phi_n(w)}}{1 - \zeta_n(z)\overline{\zeta_n(w)}} \tag{8.9}$$

$$k_n(z,w) = \frac{\phi_n^*(z)\overline{\phi_n^*(w)} - \zeta_n(z)\overline{\zeta_n(w)}\phi_n(z)\overline{\phi_n(w)}}{1 - \zeta_n(z)\overline{\zeta_n(w)}}. \tag{8.10}$$

A direct application of the Christoffel-Darboux relations is ([Bultheel et al., 1999], Corollary 3.1.4):

Proposition 8.2.2 *For all $n \geq 1$, for all $z \in \mathbb{D}$: $\phi_n^*(z) \neq 0$ and $\left|\frac{\phi_n(z)}{\phi_n^*(z)}\right| < 1$.*

Proof From (8.9), we get for $w = z$ that

$$(1 - |\zeta_n(z)|^2)k_{n-1}(z,z) = |\phi_n^*(z)|^2 - |\phi_n(z)|^2.$$

But

$$k_{n-1}(z,z) = \sum_{k=0}^{n-1} |\phi_k(z)|^2 = 1 + \sum_{k=1}^{n-1} |\phi_k(z)|^2 > 0.$$

Since $k_{n-1}(z,z) > 0$ and $|\zeta_n(z)| < 1$ for $z \in \mathbb{D}$, we deduce that

$$|\phi_n^*(z)| > |\phi_n(z)|$$

and the conclusion is immediate.

■

Using the above proposition, we get $\phi_n^*(\alpha_{n-1}) \neq 0$ for every $n \geq 0$. Therefore, since ϕ_n is uniquely determined up to a multiplicative constant of modulus 1, we can fix ϕ_n uniquely by assuming $\phi_n^*(\alpha_{n-1}) > 0$. In what follows, we denote by ϕ_n the orthogonal rational functions normalized by

$$\phi_n^*(\alpha_{n-1}) > 0. \tag{8.11}$$

Note that this is not the same normalization as in [Bultheel et al., 1999], where it is supposed that $\kappa_n = \phi_n^*(\alpha_n) > 0$.

The Christoffel-Darboux formulas imply a recurrence relation for the ϕ_n, which is the object of the next section.

8.3 Orthogonal rational functions of the first kind

Evaluate (8.9) at $w = \alpha_{n-1}$ and take into account the equality $k_{n-1}(z, \alpha_{n-1}) = \kappa_{n-1}\phi_{n-1}^*(z)$ (see (8.6)). This gives the relation

$$\kappa_{n-1}\phi_{n-1}^*(z) = \frac{\phi_n^*(z)\overline{\phi_n^*(\alpha_{n-1})} - \phi_n(z)\overline{\phi_n(\alpha_{n-1})}}{1 - \zeta_n(z)\overline{\zeta_n(\alpha_{n-1})}}, \quad n \geq 1. \tag{8.12}$$

Then take the superstar conjugate

$$\overline{\kappa_{n-1}}\phi_{n-1}(z) = \frac{\phi_n(z)\phi_n^*(\alpha_{n-1}) - \phi_n^*(z)\phi_n(\alpha_{n-1})}{\zeta_n(z) - \zeta_n(\alpha_{n-1})}$$

and put these equations together into a linear system to obtain

$$\begin{aligned}
&\begin{bmatrix} \phi_n^*(\alpha_{n-1}) & -\phi_n(\alpha_{n-1}) \\ -\overline{\phi_n(\alpha_{n-1})} & \overline{\phi_n^*(\alpha_{n-1})} \end{bmatrix} \begin{bmatrix} \phi_n(z) \\ \phi_n^*(z) \end{bmatrix} \\
&= \begin{bmatrix} \overline{\kappa_{n-1}} & 0 \\ 0 & \kappa_{n-1} \end{bmatrix} \begin{bmatrix} \zeta_n(z) - \zeta_n(\alpha_{n-1}) & 0 \\ 0 & 1 - \overline{\zeta_n(\alpha_{n-1})}\zeta_n(z) \end{bmatrix} \begin{bmatrix} \phi_{n-1}(z) \\ \phi_{n-1}^*(z) \end{bmatrix}
\end{aligned}$$

so that we have the recurrence relations

$$\begin{bmatrix} \phi_n(z) \\ \phi_n^*(z) \end{bmatrix} = T_n(z) \begin{bmatrix} \phi_{n-1}(z) \\ \phi_{n-1}^*(z) \end{bmatrix} \qquad \forall n \geq 1,$$

where T_n is equal to

$$T_n = \frac{|\kappa_{n-1}|}{|\phi_n^*(\alpha_{n-1})|^2 - |\phi_n(\alpha_{n-1})|^2} \begin{bmatrix} \overline{\phi_n^*(\alpha_{n-1})} & \phi_n(\alpha_{n-1}) \\ \overline{\phi_n(\alpha_{n-1})} & \phi_n^*(\alpha_{n-1}) \end{bmatrix}$$
$$\begin{bmatrix} \overline{\kappa_{n-1}}/|\kappa_{n-1}| & 0 \\ 0 & \kappa_{n-1}/|\kappa_{n-1}| \end{bmatrix} \begin{bmatrix} \zeta_n - \zeta_n(\alpha_{n-1}) & 0 \\ 0 & 1 - \overline{\zeta_n(\alpha_{n-1})}\zeta_n \end{bmatrix}.$$

Now, it is easily checked that

$$\begin{aligned} \zeta_n(z) - \zeta_n(\alpha_{n-1}) &= \frac{(1-|\alpha_n|^2)(z-\alpha_{n-1})}{(1-\bar{\alpha}_n\alpha_{n-1})(1-\bar{\alpha}_n z)}, \\ 1 - \overline{\zeta_n(\alpha_{n-1})}\zeta_n(z) &= \frac{(1-|\alpha_n|^2)(1-\bar{\alpha}_{n-1}z)}{(1-\alpha_n\bar{\alpha}_{n-1})(1-\bar{\alpha}_n z)}, \end{aligned}$$

so that

$$\begin{aligned} &\begin{bmatrix} \zeta_n(z) - \zeta_n(\alpha_{n-1}) & 0 \\ 0 & 1 - \overline{\zeta_n(\alpha_{n-1})}\zeta_n(z) \end{bmatrix} \\ &= \frac{(1-|\alpha_n|^2)(1-\bar{\alpha}_{n-1}z)}{(1-\alpha_n\bar{\alpha}_{n-1})(1-\bar{\alpha}_n z)} \begin{bmatrix} \eta_n & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} \zeta_{n-1}(z) & 0 \\ 0 & 1 \end{bmatrix} \end{aligned} \tag{8.13}$$

where

$$\eta_n = \frac{1-\alpha_n\bar{\alpha}_{n-1}}{1-\bar{\alpha}_n\alpha_{n-1}} \in \mathbb{T}. \tag{8.14}$$

Furthermore,

$$\begin{aligned} &\begin{bmatrix} \overline{\phi_n^*(\alpha_{n-1})} & \phi_n(\alpha_{n-1}) \\ \overline{\phi_n(\alpha_{n-1})} & \phi_n^*(\alpha_{n-1}) \end{bmatrix} \begin{bmatrix} \overline{\kappa_{n-1}}/|\kappa_{n-1}| & 0 \\ 0 & \kappa_{n-1}/|\kappa_{n-1}| \end{bmatrix} \begin{bmatrix} \eta_n & 0 \\ 0 & 1 \end{bmatrix} \\ &= \begin{bmatrix} \overline{\phi_n^*(\alpha_{n-1})}\eta_n\overline{\kappa_{n-1}}/|\kappa_{n-1}| & 0 \\ 0 & \phi_n^*(\alpha_{n-1})\kappa_{n-1}/|\kappa_{n-1}| \end{bmatrix} \begin{bmatrix} 1 & -\overline{\tilde{\gamma}_n} \\ -\tilde{\gamma}_n & 1 \end{bmatrix} \end{aligned} \tag{8.15}$$

where

$$\tilde{\gamma}_n = -\eta_n \frac{\overline{\phi_n(\alpha_{n-1})}}{\phi_n^*(\alpha_{n-1})} \frac{\overline{\kappa_{n-1}}}{\kappa_{n-1}}, \quad n \geq 1. \tag{8.16}$$

Note that, by proposition 8.2.2, $\tilde{\gamma}_n$ is well defined in $\mathbb{D}$.

Definition 8.3.1 *We call $\tilde{\gamma}_n \in \mathbb{D}$ the n-th Szegő (or Geronimus) parameter of the measure μ associated to the sequence (α_k).*

Evaluating (8.12) at $z = \alpha_{n-1}$ and taking the square root, we get after a short computation

$$|\kappa_{n-1}| = |1-\bar{\alpha}_n\alpha_{n-1}| \frac{\sqrt{|\phi_n^*(\alpha_{n-1})|^2 - |\phi_n(\alpha_{n-1})|^2}}{\sqrt{1-|\alpha_{n-1}|^2}\sqrt{1-|\alpha_n|^2}},$$

so that, from (8.16),

$$\frac{|\kappa_{n-1}|}{|\phi_n^*(\alpha_{n-1})|^2 - |\phi_n(\alpha_{n-1})|^2} = \frac{|1-\bar{\alpha}_n\alpha_{n-1}|}{\sqrt{1-|\alpha_{n-1}|^2}\sqrt{1-|\alpha_n|^2}|\phi_n^*(\alpha_{n-1})|\sqrt{1-|\tilde{\gamma}_n|^2}}. \tag{8.17}$$

Combining (8.13), (8.15) and (8.17), we finally have that

$$T_n(z) = \sqrt{\frac{1-|\alpha_n|^2}{1-|\alpha_{n-1}|^2}} \frac{1}{\sqrt{1-|\tilde{\gamma}_n|^2}} \frac{1-\bar{\alpha}_{n-1}z}{1-\bar{\alpha}_n z} \begin{bmatrix} \lambda_n & 0 \\ 0 & \bar{\lambda}_n \end{bmatrix} \begin{bmatrix} 1 & -\overline{\tilde{\gamma}_n} \\ -\tilde{\gamma}_n & 1 \end{bmatrix} \begin{bmatrix} \zeta_{n-1}(z) & 0 \\ 0 & 1 \end{bmatrix} \tag{8.18}$$

where

$$\lambda_n = \frac{|1-\bar{\alpha}_n\alpha_{n-1}|}{1-\alpha_n\bar{\alpha}_{n-1}} \frac{\overline{\phi_n^*(\alpha_{n-1})}}{|\phi_n^*(\alpha_{n-1})|} \eta_n \frac{\overline{\kappa_{n-1}}}{|\kappa_{n-1}|} = \frac{1-\alpha_n\bar{\alpha}_{n-1}}{|1-\bar{\alpha}_n\alpha_{n-1}|} \frac{\overline{\kappa_{n-1}}}{|\kappa_{n-1}|} \in \mathbb{T}. \tag{8.19}$$

We have obtained the following result ([Bultheel et al., 1999], Theorem 4.1.1, but with another normalization of the orthogonal rational functions):

Proposition 8.3.2 *The orthogonal rational functions are given by the formula*

$$\begin{bmatrix} \phi_n(z) \\ \phi_n^*(z) \end{bmatrix} = T_n(z) \begin{bmatrix} \phi_{n-1}(z) \\ \phi_{n-1}^*(z) \end{bmatrix} \qquad \forall n \geq 1,$$

with $T_n(z)$ defined as in (8.18).

A first application of this formula is to the location of the roots of the orthogonal rational functions. Note that by proposition 8.2.2, since the set of roots of ϕ_n is the image of the set of roots of ϕ_n^* by the map $z \mapsto 1/\bar{z}$, we already know that the roots are in the closed unit disk $\overline{\mathbb{D}}$.

Corollary 8.3.3 *The orthogonal rational functions ϕ_n have all their roots in $\mathbb{D}$.*

Proof By induction, we show that ϕ_n^* has no roots in $\overline{\mathbb{D}}$. This is clearly true for $n = 0$. If it is true for n, then the function $\frac{\phi_n}{\phi_n^*}$ is analytic in $\overline{\mathbb{D}}$ and by proposition 8.2.2, $\left|\frac{\phi_n}{\phi_n^*}\right| \leq 1$ in $\overline{\mathbb{D}}$. Using the previous recurrence formula on ϕ_{n+1}^*, we obtain that

$$\phi_{n+1}^* = \sqrt{\frac{1-|\alpha_{n+1}|^2}{1-|\alpha_n|^2}} \frac{1}{\sqrt{1-|\tilde{\gamma}_{n+1}|^2}} \bar{\lambda}_{n+1} \frac{1-\bar{\alpha}_n z}{1-\bar{\alpha}_{n+1}z} \phi_n^* \left(1 - \tilde{\gamma}_{n+1}\zeta_n \frac{\phi_n}{\phi_n^*}\right).$$

Using the induction hypothesis, and since $|\tilde{\gamma}_{n+1}\zeta_n \frac{\phi_n}{\phi_n^*}| \leq |\tilde{\gamma}_{n+1}| < 1$ for all $z \in \overline{\mathbb{D}}$, the latter expression does not have any root in $\overline{\mathbb{D}}$.

■

The recurrence relation can be inverted in order to express ϕ_{n-1}, ϕ_{n-1}^* as functions of ϕ_n, ϕ_n^*.

Corollary 8.3.4 *The orthogonal rational functions are given by the reverse recurrence formula*

$$\begin{bmatrix} \phi_{n-1}(z) \\ \phi_{n-1}^*(z) \end{bmatrix} = T_n^{-1}(z) \begin{bmatrix} \phi_n(z) \\ \phi_n^*(z) \end{bmatrix} \forall n \geq 1,$$

with $T_n^{-1}(z)$ equal to

$$T_n^{-1}(z) = \sqrt{\frac{1-|\alpha_{n-1}|^2}{1-|\alpha_n|^2}} \frac{1}{\sqrt{1-|\tilde{\gamma}_n|^2}} \frac{1-\bar{\alpha}_n z}{1-\bar{\alpha}_{n-1}z} \begin{bmatrix} \frac{1}{\zeta_{n-1}(z)} & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & \overline{\tilde{\gamma}_n} \\ \tilde{\gamma}_n & 1 \end{bmatrix} \begin{bmatrix} \overline{\lambda_n} & 0 \\ 0 & \lambda_n \end{bmatrix}.$$

Proof Immediate since λ_n is in $\mathbb{T}$ hence

$$\begin{aligned} T_n(z)^{-1} &= \sqrt{\frac{1-|\alpha_{n-1}|^2}{1-|\alpha_n|^2}}\sqrt{1-|\tilde{\gamma}_n|^2}\frac{1-\bar{\alpha}_n z}{1-\bar{\alpha}_{n-1}z}\begin{bmatrix} \frac{1}{\zeta_{n-1}(z)} & 0 \\ 0 & 1 \end{bmatrix}\begin{bmatrix} 1 & -\overline{\tilde{\gamma}_n} \\ -\tilde{\gamma}_n & 1 \end{bmatrix}^{-1}\begin{bmatrix} \overline{\lambda_n} & 0 \\ 0 & \lambda_n \end{bmatrix} \\ &= \sqrt{\frac{1-|\alpha_{n-1}|^2}{1-|\alpha_n|^2}}\frac{1}{\sqrt{1-|\tilde{\gamma}_n|^2}}\frac{1-\bar{\alpha}_n z}{1-\bar{\alpha}_{n-1}z}\begin{bmatrix} \frac{1}{\zeta_{n-1}(z)} & 0 \\ 0 & 1 \end{bmatrix}\begin{bmatrix} 1 & \overline{\tilde{\gamma}_n} \\ \tilde{\gamma}_n & 1 \end{bmatrix}\begin{bmatrix} \overline{\lambda_n} & 0 \\ 0 & \lambda_n \end{bmatrix}. \end{aligned}$$

■

For $\omega \in \mathbb{D}$, we denote by $P(.,\omega)$ the Poisson kernel

$$P(z,\omega) = \frac{1-|\omega|^2}{|z-\omega|^2}, \quad z \in \mathbb{T}.$$

Note that whenever u is harmonic in $\mathbb{D}$ and continuous on $\overline{\mathbb{D}}$, we have

$$u(\omega) = \int_{\mathbb{T}} u(z)P(z,\omega)dm(z).$$

This we call the Poisson identity for harmonic functions.

We now get the orthonormality of $\phi_0, \ldots, \phi_n$ with respect to another measure than μ ([Bultheel et al., 1999], Theorem 6.1.9).

Corollary 8.3.5 *The rational functions $\phi_0, \ldots, \phi_n$ are orthonormal in $L^2\left(\frac{P(.,\alpha_n)}{|\phi_n|^2}dm\right)$.*

Proof Let $N = \int_{\mathbb{T}} \frac{P(.,\alpha_n)}{|\phi_n|^2}dm$. Then $\frac{P(.,\alpha_n)}{N|\phi_n|^2}dm$ is a probability measure. For $n \geq 0$ and $k < n$, we have

$$\begin{aligned} \int_{\mathbb{T}} \sqrt{N}\phi_n\overline{\sqrt{N}\phi_k}\frac{P(.,\alpha_n)}{N|\phi_n|^2}dm &= \int_{\mathbb{T}} \frac{\phi_{k*}}{\phi_{n*}}P(.,\alpha_n)dm \\ &= \int_{\mathbb{T}} \frac{\phi_k^*}{\phi_n^*}\zeta_{k+1}\ldots\zeta_n P(.,\alpha_n)dm \\ &= 0 \end{aligned}$$

because we can apply the Poisson identity since ${\phi_n}^*$ has no zero in $\overline{\mathbb{D}}$. We also have

$$\int_{\mathbb{T}} |\sqrt{N}\phi_n|^2\frac{P(.,\alpha_n)}{N|\phi_n|^2}dm = \int_{\mathbb{T}} P(.,\alpha_n)dm = 1.$$

Therefore, $\sqrt{N}\phi_n$ is orthonormal to $\sqrt{N}\phi_0, \ldots, \sqrt{N}\phi_{n-1}$, that is to $\mathcal{L}_{n-1}$, with respect to the measure $\frac{P(.,\alpha_n)}{N|\phi_n|^2}dm$. But the reverse recurrence formula (corollary 8.3.4) together with (8.16) shows that the first $n-1$ orthogonal rational functions normalized by (8.11) are uniquely determined by the n-th orthogonal rational function and the (α_k). Therefore, the $\sqrt{N}\phi_k$, $0 \leq k \leq n$, are the orthonormal rational functions for the measure $\frac{P(.,\alpha_n)}{N|\phi_n|^2}dm$. In particular,

$$\int_{\mathbb{T}} |\sqrt{N}\phi_0|^2\frac{P(.,\alpha_n)}{N|\phi_n|^2}dm = \int_{\mathbb{T}} \frac{P(.,\alpha_n)}{|\phi_n|^2}dm = 1.$$

Thus, $N = 1$, and the conclusion is immediate.

■

Iterating the recurrence formula, we obtain an expression of ϕ_n.

Corollary 8.3.6 *For $n \geq 1$, ϕ_n and ϕ_n^* are given by the relation:*

$$\begin{bmatrix} \phi_n \\ \phi_n^* \end{bmatrix} = \frac{\sqrt{1-|\alpha_n|^2}}{1-\bar{\alpha}_n z}\frac{1}{\Pi_n}\left(\prod_{k=n}^{k=1}\begin{bmatrix} \lambda_k & 0 \\ 0 & \overline{\lambda_k} \end{bmatrix}\begin{bmatrix} 1 & -\overline{\tilde{\gamma}_k} \\ -\tilde{\gamma}_k & 1 \end{bmatrix}\begin{bmatrix} \zeta_{k-1}(z) & 0 \\ 0 & 1 \end{bmatrix}\right)\begin{bmatrix} 1 \\ 1 \end{bmatrix}$$

with

$$\Pi_n = \prod_{k=n}^{k=1}\sqrt{1-|\tilde{\gamma}_k|^2}.$$

Proof Immediate from proposition 8.3.2 since $\alpha_0 = 0$ and $\phi_0 = \phi_0^* = 1$.

■

8.4 Orthogonal rational functions of the second kind

As in [Bultheel et al., 1999], chapter 4, we now define the sequence (ψ_n) of orthogonal rational functions of the second kind. We shall see later that this sequence satisfies the same recurrence relations as ϕ_n, but with $\tilde{\gamma}_n$ replaced by $-\tilde{\gamma}_n$.

Definition 8.4.1 *Given μ, (α_k) and (ϕ_n) as before, we call orthogonal rational functions of the second kind the sequence ψ_n such that*

$$\begin{cases} \psi_0 = 1 \\ \psi_n(z) = \displaystyle\int_{\mathbb{T}} \frac{t+z}{t-z}\left(\phi_n(t) - \phi_n(z)\right) d\mu(t) \end{cases}.$$

We will see later that the ψ_n are indeed rational functions. The following proposition ([Bultheel et al., 1999], Lemma 4.2.2 and 4.2.3) is very useful for computations.

Proposition 8.4.2 *For $n \geq 1$, the functions (ψ_n) satisfy the formulas:*

$$\psi_n(z)g(z) = \int_{\mathbb{T}} \frac{t+z}{t-z}\left(\phi_n(t)g(t) - \phi_n(z)g(z)\right) d\mu(t)$$

for all g such that $g_* \in \mathcal{L}_{n-1}$, and moreover we have

$$-\psi_n^*(z)h(z) = \int_{\mathbb{T}} \frac{t+z}{t-z}\left(\phi_n^*(t)h(t) - \phi_n^*(z)h(z)\right) d\mu(t)$$

for all h such that $h_* \in \zeta_n\mathcal{L}_{n-1}$.

Proof We first prove the first equality. If g is constant, the result is immediate. We therefore suppose $n \geq 2$. Let $z \in \mathbb{D}$.
If $z = \alpha_k$ for some k, $1 \leq k \leq n-1$, $g(z) = \infty$. By definition, we have

$$\psi_n(\alpha_k) = \int_{\mathbb{T}} \frac{t+\alpha_k}{t-\alpha_k}\left(\phi_n(t) - \phi_n(\alpha_k)\right) d\mu(t).$$

But, since $\frac{t+\alpha_k}{t-\alpha_k} \in \mathcal{L}_{n-1}$ and $n \geq 2$, we get by orthogonality

$$\psi_n(\alpha_k) = -\phi_n(\alpha_k) \int_{\mathbb{T}} \frac{t+\alpha_k}{t-\alpha_k} d\mu(t)$$

which is the announced result when $g(z) = \infty$.
Suppose $z \neq \alpha_k$ for all k, $1 \leq k \leq n-1$. By density, it is enough to prove the result if $g(z)$ is analytic at z with $g(z) \neq 0$. In order to conclude, using the defintion of ψ_n, we just have to check that

$$\int \frac{t+z}{t-z} \phi_n(t) \frac{g(t)}{g(z)} d\mu(t) = \int \frac{t+z}{t-z} \phi_n(t) d\mu(t) \text{ whenever } g_* \in \mathcal{L}_{n-1}.$$

But $\frac{g(t)}{g(z)} - 1$ vanishes for $t = z$, therefore

$$\frac{g(t)}{g(z)} - 1 = (t-z) \frac{p}{\prod_{k=1}^{k=n-1}(t-\alpha_k)}$$

where p is a polynomial in t of degree at most $n-2$. Thus,

$$\begin{aligned}
\int \frac{t+z}{t-z} \phi_n(t) \left(\frac{g(t)}{g(z)} - 1 \right) d\mu(t) &= \int \frac{t+z}{t-z} (t-z) \frac{p(t)}{\prod_{k=1}^{k=n-1}(t-\alpha_k)} \phi_n(t) d\mu(t) \\
&= \int \frac{(t+z)p(t)}{\prod_{k=1}^{k=n-1}(t-\alpha_k)} \phi_n(t) d\mu(t) \\
&= \int \overline{\left(\frac{t^{n-1}\overline{(t+z)p(t)}}{\prod_{k=1}^{k=n-1}(1-\overline{\alpha_k}t)} \right)} \phi_n(t) d\mu(t) \\
&= 0
\end{aligned}$$

because, since $\bar{t} = \frac{1}{t}$ on $\mathbb{T}$ and $\deg p \leq n-2$, we have on $\mathbb{T}$:

$$\frac{t^{n-1}\overline{(t+z)p(t)}}{\prod_{k=1}^{k=n-1}(1-\overline{\alpha_k}t)} = \frac{t^{n-1}\overline{(1/\bar{t}+z)p(1/\bar{t})}}{\prod_{k=1}^{k=n-1}(1-\overline{\alpha_k}t)} \in \mathcal{L}_{n-1}$$

Therefore, the first equality is proved.

Since $\mathcal{B}_n h$ is in $\mathcal{L}_{n-1}$, we get from the latter

$$\psi_n(z)\mathcal{B}_{n*}(z)h_*(z) = \int \frac{t+z}{t-z} \left(\phi_n(t)\mathcal{B}_{n*}(t)h_*(t) - \phi_n(z)\mathcal{B}_{n*}(z)h_*(z) \right) d\mu(t).$$

We conclude by taking the lower-$*$ conjugate in z of this expression.

■

We deduce from the following proposition that ψ_n is indeed a rational function (see [Bultheel et al., 1999], Theorem 4.2.4).

Proposition 8.4.3 *The sequences (ϕ_n) and (ψ_n) satisfy the recurrence relations:*

$$\begin{bmatrix} \phi_n & \psi_n \\ \phi_n^* & -\psi_n^* \end{bmatrix} = \frac{\sqrt{1-|\alpha_n|^2}}{1-\bar{\alpha}_n z} \frac{1}{\Pi_n} \left(\prod_{k=n}^{k=1} \begin{bmatrix} \lambda_k & 0 \\ 0 & \bar{\lambda}_k \end{bmatrix} \begin{bmatrix} 1 & -\overline{\tilde{\gamma}_k} \\ -\tilde{\gamma}_k & 1 \end{bmatrix} \begin{bmatrix} \zeta_{k-1}(z) & 0 \\ 0 & 1 \end{bmatrix} \right) \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix}$$

with

$$\Pi_n = \prod_{k=n}^{k=1} \sqrt{1-|\tilde{\gamma_k}|^2}.$$

In particular, ψ_n is in $\mathcal{L}_n$.

Proof From Corollary 8.3.6, we now that this relation holds for (ϕ_n, ϕ_n^*), so we just have to prove it for (ψ_n, ψ_n^*). We first check that this result for $n = 1$. As $\psi_0 = 1$, we want to prove that

$$\psi_1 = \beta_1 \frac{z + \overline{\tilde{\gamma_1}}}{1 - \overline{\alpha_1} z}$$

with

$$\beta_1 = \sqrt{\frac{1-|\alpha_1|^2}{1-|\tilde{\gamma_1}|^2}} \lambda_1.$$

We have

$$\begin{aligned}
\psi_1(z) &= \int_{\mathbb{T}} \frac{t+z}{t-z} (\phi_1(t) - \phi_1(z)) d\mu(t) \\
&= \beta_1 \int_{\mathbb{T}} \frac{t+z}{t-z} \left(\frac{t - \overline{\tilde{\gamma_1}}}{1 - \overline{\alpha_1} t} - \frac{z - \overline{\tilde{\gamma_1}}}{1 - \overline{\alpha_1} z} \right) d\mu(t) \\
&= \beta_1 \int_{\mathbb{T}} \frac{t+z}{t-z} \left(\frac{(t-z)(1 - \overline{\alpha_1 \tilde{\gamma_1}})}{(1 - \overline{\alpha_1} t)(1 - \overline{\alpha_1} z)} \right) d\mu(t) \\
&= \beta_1 \int_{\mathbb{T}} \frac{(t+z)(1 - \overline{\alpha_1 \tilde{\gamma_1}})}{(1 - \overline{\alpha_1} t)(1 - \overline{\alpha_1} z)} d\mu(t) \\
&= \beta_1 \frac{1 - \overline{\alpha_1 \tilde{\gamma_1}}}{1 - \overline{\alpha_1} z} \int_{\mathbb{T}} \frac{t+z}{1 - \overline{\alpha_1} t} d\mu(t).
\end{aligned}$$

As ϕ_1 is orthogonal to 1, we also have

$$\int_{\mathbb{T}} \frac{t}{1 - \overline{\alpha_1} t} d\mu(t) = \overline{\tilde{\gamma_1}} \int_{\mathbb{T}} \frac{1}{1 - \overline{\alpha_1} t} d\mu(t). \tag{8.20}$$

Therefore

$$\psi_1(z) = \beta_1 \frac{1 - \overline{\alpha_1 \tilde{\gamma_1}}}{1 - \overline{\alpha_1} z} (\overline{\tilde{\gamma_1}} + z) \int_{\mathbb{T}} \frac{1}{1 - \overline{\alpha_1} t} d\mu(t).$$

But, by (8.20),

$$\begin{aligned}
\int_{\mathbb{T}} \frac{1}{1 - \overline{\alpha_1} t} d\mu(t) &= 1 + \overline{\alpha_1} \int_{\mathbb{T}} \frac{t}{1 - \overline{\alpha_1} t} d\mu(t) \\
&= 1 + \overline{\alpha_1} \overline{\tilde{\gamma_1}} \int_{\mathbb{T}} \frac{1}{1 - \overline{\alpha_1} t} d\mu(t)
\end{aligned}$$

thus

$$\int_{\mathbb{T}} \frac{1}{1 - \overline{\alpha_1} t} d\mu(t) = \frac{1}{1 - \overline{\alpha_1} \overline{\tilde{\gamma_1}}}.$$

Therefore,

$$\begin{aligned}\psi_1(z) &= \beta_1 \frac{1-\overline{\alpha_1\tilde{\gamma}_1}}{1-\overline{\alpha_1}z}(\overline{\tilde{\gamma}_1}+z)\frac{1}{1-\overline{\alpha_1}\overline{\tilde{\gamma}_1}} \\ &= \beta_1 \frac{z+\overline{\tilde{\gamma}_1}}{1-\overline{\alpha_1}z}.\end{aligned}$$

which is the result we want.

We now proceed by induction.
Assume $n > 1$. Proposition (8.4.2) gives us with n replaced by $n-1$ and $g = 1$ together with $h = \zeta_{n-1_*}$,

$$\begin{bmatrix} \psi_{n-1}(z) \\ -\psi^*_{n-1}(z) \end{bmatrix} = \int \frac{t+z}{t-z}\left(\begin{bmatrix} \phi_{n-1}(t) \\ \frac{\zeta_{n-1}(z)}{\zeta_{n-1}(t)}\phi^*_{n-1}(t) \end{bmatrix} - \begin{bmatrix} \phi_{n-1}(z) \\ \phi^*_{n-1}(z) \end{bmatrix}\right)d\mu(t).$$

Multiplying by $T_n(z)$ whose definition was given in (8.18), we obtain

$$\begin{aligned}
&T_n(z)\begin{bmatrix} \psi_{n-1}(z) \\ -\psi^*_{n-1}(z) \end{bmatrix} \\
&= \int \frac{t+z}{t-z}\left(T_n(z)\begin{bmatrix} \phi_{n-1}(t) \\ \frac{\zeta_{n-1}(z)}{\zeta_{n-1}(t)}\phi^*_{n-1}(t) \end{bmatrix} - \begin{bmatrix} \phi_n(z) \\ \phi^*_n(z) \end{bmatrix}\right)d\mu(t) \\
&= \int \frac{t+z}{t-z}\left(\frac{(1-\overline{\alpha_n}t)(1-\overline{\alpha_{n-1}}z)}{(1-\overline{\alpha_n}z)(1-\overline{\alpha_{n-1}}t)}T_n(t)\begin{bmatrix} \frac{\zeta_{n-1}(z)}{\zeta_{n-1}(t)} & 0 \\ 0 & 1 \end{bmatrix}\begin{bmatrix} \phi_{n-1}(t) \\ \frac{\zeta_{n-1}(z)}{\zeta_{n-1}(t)}\phi^*_{n-1}(t) \end{bmatrix} - \begin{bmatrix} \phi_n(z) \\ \phi^*_n(z) \end{bmatrix}\right)d\mu(t) \\
&= \int \frac{t+z}{t-z}\left(\frac{(1-\overline{\alpha_n}t)(z-\alpha_{n-1})}{(1-\overline{\alpha_n}z)(t-\alpha_{n-1})}\begin{bmatrix} \phi_n(t) \\ \phi^*_n(t) \end{bmatrix} - \begin{bmatrix} \phi_n(z) \\ \phi^*_n(z) \end{bmatrix}\right)d\mu(t).
\end{aligned}$$

But, by proposition (8.4.2) applied with $g(z) = (1-\bar{\alpha}_n z)/(z-\alpha_{n-1})$, the first row in the right handside of the last term is equal to ψ_n. So it only remains to prove that the second row is equal to $-\psi^*_n$. To this effect, observe that

$$\begin{aligned}
&\int \frac{t+z}{t-z}\left(\frac{z-\alpha_{n-1}}{t-\alpha_{n-1}} - \frac{z-\alpha_n}{t-\alpha_n}\right)\frac{1-\overline{\alpha_n}t}{1-\overline{\alpha_n}z}\phi^*_n(t)d\mu(t) \\
&= \int \frac{t+z}{t-z}\left(\frac{(t-z)(\alpha_n-\alpha_{n-1})}{(t-\alpha_{n-1})(t-\alpha_n)}\right)\frac{1-\overline{\alpha_n}t}{1-\overline{\alpha_n}z}\phi^*_n(t)d\mu(t) \\
&= \int \frac{(1-\overline{\alpha_n}t)(t+z)(\alpha_n-\alpha_{n-1})}{(t-\alpha_n)(t-\alpha_{n-1})(1-\overline{\alpha_n}z)}\phi^*_n(t)d\mu(t) \\
&= \int \mathcal{B}_{n-1}(t)\frac{(t+z)(\alpha_n-\alpha_{n-1})}{(t-\alpha_{n-1})(1-\overline{\alpha_n}z)}\overline{\phi_n(t)}d\mu(t) \\
&= 0
\end{aligned}$$

because

$$\mathcal{B}_{n-1}(t)\frac{(t+z)(\alpha_n-\alpha_{n-1})}{(t-\alpha_{n-1})(1-\overline{\alpha_n}z)} \in \mathcal{L}_{n-1}$$

as a function of t for fixed $z \in \overline{\mathbb{D}}$. Therefore,

$$\begin{aligned}
&\int \frac{t+z}{t-z}\left(\frac{(1-\overline{\alpha_n}t)(z-\alpha_{n-1})}{(1-\overline{\alpha_n}z)(t-\alpha_{n-1})}\phi_n^*(t)-\phi_n^*(z)\right)d\mu(t)\\
&= \int \frac{t+z}{t-z}\left(\frac{(1-\overline{\alpha_n}t)(z-\alpha_n)}{(1-\overline{\alpha_n}z)(t-\alpha_n)}\phi_n^*(t)-\phi_n^*(z)\right)d\mu(t)\\
&= -\psi_n^*(z)
\end{aligned}$$

by proposition (8.4.2) with $h(z) = (1-\bar{\alpha}_n z)/(z-\alpha_n)$. This achieves the induction step. ■

We now show that the sequence (ψ_n) satisfies the same recurrence relations than (ϕ_n), but with $\tilde{\gamma}_n$ replaced by $-\tilde{\gamma}_n$:

Corollary 8.4.4 *The sequence ψ_n satisfies the recurrence relations:*

$$\begin{bmatrix} \psi_n \\ \psi_n^* \end{bmatrix} = \frac{\sqrt{1-|\alpha_n|^2}}{1-\bar{\alpha}_n z}\frac{1}{\Pi_n}\left(\prod_{k=n}^{k=1}\begin{bmatrix} \lambda_k & 0 \\ 0 & \bar{\lambda}_k \end{bmatrix}\begin{bmatrix} 1 & \overline{\tilde{\gamma}_k} \\ \tilde{\gamma}_k & 1 \end{bmatrix}\begin{bmatrix} \zeta_{k-1}(z) & 0 \\ 0 & 1 \end{bmatrix}\right)\begin{bmatrix} 1 \\ 1 \end{bmatrix}$$

Proof Note that, by Proposition 8.4.3,

$$\begin{bmatrix} \psi_n \\ -\psi_n^* \end{bmatrix} = \frac{\sqrt{1-|\alpha_n|^2}}{1-\bar{\alpha}_n z}\frac{1}{\Pi_n}\left(\prod_{k=n}^{k=1}\begin{bmatrix} \lambda_k & 0 \\ 0 & \bar{\lambda}_k \end{bmatrix}\begin{bmatrix} 1 & -\overline{\tilde{\gamma}_k} \\ -\tilde{\gamma}_k & 1 \end{bmatrix}\begin{bmatrix} \zeta_{k-1}(z) & 0 \\ 0 & 1 \end{bmatrix}\right)\begin{bmatrix} 1 \\ -1 \end{bmatrix}.$$

Therefore, since $\begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix}^2 = Id$,

$$\begin{aligned}
\begin{bmatrix} \psi_n \\ \psi_n^* \end{bmatrix} &= \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix}\begin{bmatrix} \psi_n \\ -\psi_n^* \end{bmatrix}\\
&= \frac{\sqrt{1-|\alpha_n|^2}}{1-\bar{\alpha}_n z}\frac{1}{\Pi_n}\\
&\quad\left(\prod_{k=n}^{k=1}\begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix}\begin{bmatrix} \lambda_k & 0 \\ 0 & \bar{\lambda}_k \end{bmatrix}\begin{bmatrix} 1 & -\overline{\tilde{\gamma}_k} \\ -\tilde{\gamma}_k & 1 \end{bmatrix}\begin{bmatrix} \zeta_{k-1}(z) & 0 \\ 0 & 1 \end{bmatrix}\begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix}\right)\begin{bmatrix} 1 \\ 1 \end{bmatrix}\\
&= \frac{\sqrt{1-|\alpha_n|^2}}{1-\bar{\alpha}_n z}\frac{1}{\Pi_n}\left(\prod_{k=n}^{k=1}\begin{bmatrix} \lambda_k & 0 \\ 0 & \bar{\lambda}_k \end{bmatrix}\begin{bmatrix} 1 & \overline{\tilde{\gamma}_k} \\ \tilde{\gamma}_k & 1 \end{bmatrix}\begin{bmatrix} \zeta_{k-1}(z) & 0 \\ 0 & 1 \end{bmatrix}\right)\begin{bmatrix} 1 \\ 1 \end{bmatrix}.
\end{aligned}$$

■

Proposition 8.4.5 *For all z in $\overline{\mathbb{D}}$, it holds that*

$$\phi_n(z)\psi_n^*(z) + \phi_n^*(z)\psi_n(z) = 2\frac{1-|\alpha_n|^2}{(1-\overline{\alpha_n}z)(z-\alpha_n)}z\mathcal{B}_n(z).$$

Proof Taking determinants in the relation of proposition 8.4.3, we get

$$\begin{aligned}
\phi_n(z)\psi_n^*(z) + \phi_n^*(z)\psi_n(z) &= 2\frac{1-|\alpha_n|^2}{(1-\overline{\alpha_n}z)^2}\prod_{k=1}^{k=n}|\lambda_k|^2\zeta_{k-1}(z) \\
&= 2\frac{1-|\alpha_n|^2}{(1-\overline{\alpha_n}z)^2}\, z\mathcal{B}_n(z)\frac{1-\overline{\alpha_n}z}{z-\alpha_n} \\
&= 2\frac{1-|\alpha_n|^2}{(1-\overline{\alpha_n}z)(z-\alpha_n)} z\mathcal{B}_n(z).
\end{aligned}$$

■

In particular, we have:

Corollary 8.4.6 *For $z \in \mathbb{T}$, one has*

$$\phi_n(z)\psi_n^*(z) + \phi_n^*(z)\psi_n(z) = 2\mathcal{B}_n(z)P(z,\alpha_n) \tag{8.21}$$

where $P(z,\alpha_n) = \frac{1-|\alpha_n|^2}{|z-\alpha_n|^2}$ is the Poisson kernel at α_n.

Chapter 9

Link between orthogonal rational functions and Wall rational functions

If we glance at Propositions 7.3.5 and 8.4.3, we see that the recurrence formulas for the Wall rational functions A_n, B_n and for the orthogonal rational functions ϕ_n, ψ_n look quite similar. In this chapter, we will use this similarity to prove a generalized Geronimus theorem (see [Geronimus, 1944] for the original version). We first need to associate to a Schur function f a measure μ: we use for this the Herglotz transform. Next, we prove a Geronimus theorem which states the relation between the Szegő parameters of μ and the Schur parameters of f ([Langer and Lasarow, 2004]).

9.1 The Herglotz transform

We denote by F the Herglotz transform of μ:

$$F(z) = \int_{\mathbb{T}} \frac{\xi + z}{\xi - z} d\mu(\xi). \tag{9.1}$$

We have ([Bultheel et al., 2006], Theorem 3.4):

Proposition 9.1.1 *The Herglotz transform is related to the orthogonal rational functions ϕ_n, ψ_n associated with μ by a relation of the form*

$$F(z) = \frac{\psi_n^*(z)}{\phi_n^*(z)} + \frac{z\mathcal{B}_n(z)u(z)}{\phi_n^*(z)}$$

where u is an analytic function in $\mathbb{D}$.

Proof Proposition 8.4.2 gives us with $h(z) = 1/\mathcal{B}_n(z)$

$$\begin{aligned}
\frac{F(z)\phi_n^*(z) - \psi_n^*(z)}{\mathcal{B}_n(z)} &= \int \frac{t+z}{t-z}\frac{\phi_n^*(z)}{\mathcal{B}_n(z)} d\mu(t) + \int \frac{t+z}{t-z}\left(\frac{\phi_n^*(t)}{\mathcal{B}_n(t)} - \frac{\phi_n^*(z)}{\mathcal{B}_n(z)}\right) d\mu(t) \\
&= \int \frac{t+z}{t-z}\frac{\phi_n^*(t)}{\mathcal{B}_n(t)} d\mu(t).
\end{aligned}$$

This is a Cauchy integral, so it is a holomorphic function of z in $\mathbb{D}$. Evaluating this function at 0, we get

$$\int \frac{\phi_n^*(t)}{\mathcal{B}_n(t)} d\mu(t) = \int \overline{\phi_n(t)} d\mu(t) = 0$$

by orthogonality of ϕ_n and 1. The conclusion is then immediate.

■

The Riesz-Herglotz theorem [Rudin, 1987] states that the Herglotz transform is a one-to-one mapping between the set of probability measures on $\mathbb{T}$ and the set of analytic functions F in $\mathbb{D}$ satisfying

$$F(0) = 1, \quad ReF(z) > 0, \quad z \in \mathbb{D}.$$

$\frac{F-1}{F+1}$ is a Schur function that vanishes at zero, so the Schwarz lemma implies that

$$f(z) = \frac{1}{z} \frac{F(z) - 1}{F(z) + 1}$$

is also a Schur function. Therefore, we obtain a one-to-one correspondence between probability measures μ on $\mathbb{T}$ and Schur functions f via the relation

$$\int_{\mathbb{T}} \frac{\xi + z}{\xi - z} d\mu(\xi) = \frac{1 + zf(z)}{1 - zf(z)}. \tag{9.2}$$

For fixed $z \in \mathbb{D}$, we denote by Ω_z the map

$$\Omega_z : \omega \mapsto \frac{1}{z} \frac{\omega - 1}{\omega + 1}.$$

Note that $f(z) = \Omega_z(F(z))$.

Definition 9.1.2 *The function f associated to μ through (9.2) will be called the Schur function of μ.*

Applying Fatou's theorem on nontangential limits of harmonic functions ([Garnett, 2007]) to the real part of (9.2), we obtain an expression for the Lebesgue derivative μ' of the measure μ in terms of f:

$$\mu'(\xi) = \frac{1 - |f(\xi)|^2}{|1 - \xi f(\xi)|^2} \text{ a.e. on } \mathbb{T}. \tag{9.3}$$

Since $1 - zf(z)$ is a non-zero function of H^∞, it cannot vanish on a set of positive measure. Therefore, $\mu' > 0$ a.e. on $\mathbb{T}$ if and only if $|f| < 1$ a.e. on $\mathbb{T}$.

The Schur parameters of the function f associated with μ can be computed from the orthogonal rationals functions of μ:

Proposition 9.1.3 *$f(z)$ and $\Omega_z\left(\frac{\psi_n^*(z)}{\phi_n^*(z)}\right)$ have the same first n Schur parameters.*

Proof From Proposition 9.1.1, we get

$$F^{(i)}(z) = \left(\frac{\psi_n^*(z)}{\phi_n^*(z)}\right)^{(i)} + \left(\frac{z\mathcal{B}_n(z)u(z)}{\phi_n^*(z)}\right)^{(i)}, \quad i \geq 0. \tag{9.4}$$

Let j be an integer such that $0 \leq j \leq n-1$. We denote by m_{j+1} the multiplicity of α_{j+1} at the n-th step (see Proposition 7.1.2). Then, if $0 \leq i < m_{j+1}$, since $\mathcal{B}_n(z) = h(z)\prod_{k=1}^{m_{j+1}}(z-\alpha_{j+1})$ with $h \in \mathcal{L}_n$, we have $B_n^{(i)}(\alpha_{j+1}) = 0$. Therefore, using (9.4), we obtain

$$F^{(i)}(\alpha_{j+1}) = \left(\frac{\psi_n^*}{\phi_n^*}\right)^{(i)}(\alpha_{j+1}).$$

Since $f(z) = \Omega_z(F(z))$, we conclude using Proposition 7.1.2.

■

9.2 A Geronimus theorem

Geronimus was the first to express the relation between the classical Schur algorithm applied to the Schur function of a measure μ and the orthogonal polynomials of μ. In [Langer and Lasarow, 2004], the connection between the Geronimus parameters of the orthogonal rational functions and the Schur parameters of a multipoint Schur algorithm is detailed. However, the normalisation of the orthogonal rational functions in this reference is different from ours, so the link is made with a multipoint Schur algorithm without the rotations c_k. We chose to keep our generalized multipoint algorithm and we give below another proof of the Geronimus theorem.

Theorem 9.2.1 *Fix $(\alpha_k)_{k\geq 1} \in \mathbb{D}$ and $f \in \mathcal{S}$.*
We associate with f the measure μ given by (9.2). We denote by $(\tilde{\gamma_k})_{k\geq 1}$ the Geronimus parameters of μ (see (8.16)), and by λ_k the elements of $\mathbb{T}$ defined by (8.19).
If the parameters $(c_k)_{k\geq 1}$ of the multipoint Schur algorithm are defined by

$$c_k = \lambda_k^2, \quad c_0 = 1,$$

then the Geronimus parameters $(\tilde{\gamma_k})_{k\geq 1}$ and the Schur parameters $(\gamma_k)_{k\in\mathbb{N}}$ of f are related by

$$\tilde{\gamma}_{k+1} = \gamma_k \text{ for all } k \geq 0.$$

Proof We first study the connection between the recurrence formulas. From proposition 8.4.3, we have

$$\begin{aligned}
&\begin{bmatrix} \phi_{n+1}(z) & \psi_{n+1}(z) \\ \phi_{n+1}^*(z) & -\psi_{n+1}^*(z) \end{bmatrix} \\
&= \Delta_{n+1}\left(\prod_{k=n+1}^{k=1}\begin{bmatrix} \lambda_k^2 & 0 \\ 0 & 1 \end{bmatrix}\begin{bmatrix} 1 & -\overline{\tilde{\gamma_k}} \\ -\tilde{\gamma_k} & 1 \end{bmatrix}\begin{bmatrix} \zeta_{k-1}(z) & 0 \\ 0 & 1 \end{bmatrix}\right)\begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix} \\
&= \Delta_{n+1}\left(\prod_{k=n+1}^{k=1}\begin{bmatrix} -1 & 0 \\ 0 & 1 \end{bmatrix}\begin{bmatrix} \lambda_k^2 & 0 \\ 0 & 1 \end{bmatrix}\begin{bmatrix} 1 & \overline{\tilde{\gamma_k}} \\ \tilde{\gamma_k} & 1 \end{bmatrix}\begin{bmatrix} \zeta_{k-1}(z) & 0 \\ 0 & 1 \end{bmatrix}\begin{bmatrix} -1 & 0 \\ 0 & 1 \end{bmatrix}\right)\begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix}
\end{aligned}$$

with

$$\Delta_{n+1} = \frac{\sqrt{1-|\alpha_{n+1}|^2}}{1-\bar{\alpha}_{n+1}z}\frac{\prod_{k=1}^{n+1}\bar{\lambda}_k}{\prod_{k=1}^{n+1}\sqrt{1-|\tilde{\gamma_k}|^2}}.$$

Therefore, if the parameters c_k are taken such that $c_k = \lambda_k^2$ for all $k \geq 1$ and if $\frac{U_n}{V_n}$ stands for the n-th convergent of a Schur function with parameters $\gamma_k := \tilde{\gamma}_{k+1}$ for all $k \geq 0$ (such a function exists because of Corollary 7.3.2), we get from Proposition 7.3.5 the following expression of ϕ_n, ψ_n with respect to U_n, V_n,

$$\begin{aligned}
&\begin{bmatrix} \phi_{n+1}(z) & \psi_{n+1}(z) \\ \phi^*_{n+1}(z) & -\psi^*_{n+1}(z) \end{bmatrix} \\
&= \Sigma_n \Delta_{n+1} \begin{bmatrix} -1 & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} c_{n+1} & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} V_n^* & U_n^* \\ U_n & V_n \end{bmatrix} \begin{bmatrix} \zeta_0 & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} -1 & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix} \\
&= \Sigma_n \Delta_{n+1} \begin{bmatrix} -1 & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} c_{n+1} & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} -zV_n^* + U_n^* & -zV_n^* - U_n^* \\ -zU_n + V_n & -zU_n - V_n \end{bmatrix}
\end{aligned}$$

with $\Sigma_n = \prod_{k=1}^n \lambda_k$.
Since

$$\Sigma_n \prod_{k=1}^{n+1} \bar{\lambda}_k = \left(\prod_{k=1}^{n} \lambda_k \right) \prod_{k=1}^{n+1} \bar{\lambda}_k = \left(\prod_{k=1}^{n} |\lambda_k| \right) \bar{\lambda}_{n+1} = \bar{\lambda}_{n+1}$$

and $c_{n+1} = \lambda_{n+1}^2$, we obtain

$$\begin{aligned}
&\begin{bmatrix} -1 & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} \phi_{n+1}(z) & \psi_{n+1}(z) \\ \phi^*_{n+1}(z) & -\psi^*_{n+1}(z) \end{bmatrix} \\
&= \frac{\sqrt{1-|\alpha_{n+1}|^2}}{1-\bar{\alpha}_{n+1}z} \frac{1}{\prod_{k=1}^{n+1} \sqrt{1-|\tilde{\gamma}_k|^2}} \begin{bmatrix} \lambda_{n+1} & 0 \\ 0 & \bar{\lambda}_{n+1} \end{bmatrix} \begin{bmatrix} -zV_n^* + U_n^* & -zV_n^* - U_n^* \\ -zU_n + V_n & -zU_n - V_n \end{bmatrix}.
\end{aligned} \tag{9.5}$$

In particular, we have

$$\frac{\psi^*_{n+1}}{\phi^*_{n+1}} = \frac{1 + z\frac{U_n}{V_n}}{1 - z\frac{U_n}{V_n}} \tag{9.6}$$

so

$$\frac{U_n(z)}{V_n(z)} = \Omega_z \left(\frac{\psi^*_{n+1}(z)}{\phi^*_{n+1}(z)} \right).$$

Then, from proposition 9.1.3, $\frac{U_n}{V_n}$ has the same first $n+1$ Schur parameters as the Schur function f of the measure μ. This gives the expected result.

■

Note that a consequence of the theorem is that the elements U_n and V_n of the proof are equal to the Wall rational functions A_n and B_n of f. In particular, equations (9.5) and (9.6) gives us

$$\begin{aligned}
&\begin{bmatrix} -1 & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} \phi_{n+1}(z) & \psi_{n+1}(z) \\ \phi^*_{n+1}(z) & -\psi^*_{n+1}(z) \end{bmatrix} \\
&= \frac{\sqrt{1-|\alpha_{n+1}|^2}}{1-\bar{\alpha}_{n+1}z} \frac{1}{\prod_{k=1}^{n+1} \sqrt{1-|\tilde{\gamma}_k|^2}} \begin{bmatrix} \lambda_{n+1} & 0 \\ 0 & \bar{\lambda}_{n+1} \end{bmatrix} \begin{bmatrix} -zB_n^* + A_n^* & -zB_n^* - A_n^* \\ -zA_n + B_n & -zA_n - B_n \end{bmatrix}
\end{aligned} \tag{9.7}$$

and

$$\frac{\psi^*_{n+1}}{\phi^*_{n+1}} = \frac{1 + z\frac{A_n}{B_n}}{1 - z\frac{A_n}{B_n}}. \tag{9.8}$$

9.3 Consequences of the Geronimus theorem

The following corollary to Theorem 9.2.1 gives the expression of the measure associated to the Wall rational functions by the Herglotz transform. This is a generalization to the multipoint case of [Khrushchev, 2001], Corollary 5.2.

Corollary 9.3.1 *$\frac{A_n}{B_n}$ is the Schur function of the measure $\frac{P(.,\alpha_{n+1})}{|\phi_{n+1}|^2}dm$.*

Proof Indeed, by (8.21), we have on $\mathbb{T}$:

$$\begin{aligned} Re\left(\frac{\psi^*_{n+1}}{\phi^*_{n+1}}\right) &= \frac{\overline{\mathcal{B}_{n+1}}\left(\psi^*_{n+1}\phi_{n+1}+\phi^*_{n+1}\psi_{n+1}\right)}{2\left|\phi_{n+1}\right|^2} \\ &= \frac{P(.,\alpha_{n+1})}{|\phi_{n+1}|^2}. \end{aligned}$$

Thus, $\frac{\psi^*_{n+1}}{\phi^*_{n+1}}$ and $\int \frac{t+z}{t-z}\frac{P(t,\alpha_{n+1})}{|\phi_{n+1}(t)|^2}dm(t)$ are two analytic functions in $\mathbb{D}$ with the same real part, therefore they are related by

$$\frac{\psi^*_{n+1}}{\phi^*_{n+1}} = \int \frac{t+z}{t-z}\frac{P(t,\alpha_{n+1})}{|\phi_{n+1}(t)|^2}dm(t) + ic$$

where c is a real constant. So by (9.8),

$$\frac{1+z\frac{A_n}{B_n}}{1-z\frac{A_n}{B_n}} = \int \frac{t+z}{t-z}\frac{P(t,\alpha_{n+1})}{|\phi_{n+1}(t)|^2}dm(t) + ic.$$

Evaluating the above expression at 0 gives us

$$1 = \int \frac{P(.,\alpha_{n+1})}{|\phi_{n+1}|^2}dm + ic.$$

Since the integral is real, $c = 0$.

■

In view of Corollary 8.4.4, the Geronimus theorem also leads to another definition of the orthogonal rational functions of the second kind:

Corollary 9.3.2 *Up to a normalization, the orthogonal rational functions of the second kind associated to f (or F) are the orthogonal rational functions of the first kind associated to $-f$ (or $\frac{1}{F}$).*

The following theorem gives a useful relation between the Lebesgue derivative μ' of the measure μ, the Schur functions f_n and the orthogonal rational functions ϕ_n. This is a generalization to the multipoint case of [Khrushchev, 2001], Theorem 2.

Theorem 9.3.3 *Let (ϕ_n) be the orthogonal rational functions of a probability measure μ associated to a sequence (α_n), and (f_n) the Schur functions associated to μ with the choice $c_n = \lambda_n^2$. Then*

$$\mu' = \frac{1-|f_n|^2}{|1-\overline{c_n}\zeta_n\frac{\phi_n}{\phi^*_n}f_n|^2}\frac{P(.,\alpha_n)}{|\phi_n|^2} \quad a.e.\ on\ \mathbb{T}.$$

Proof From Theorem 7.3.8, we have:

$$\begin{aligned} 1-|f|^2 &= 1-\left|\frac{A_n+\zeta_{n+1}B_n^*f_{n+1}}{B_n+\zeta_{n+1}A_n^*f_{n+1}}\right|^2 \\ &= \frac{|B_n+\zeta_{n+1}A_n^*f_{n+1}|^2-|A_n+\zeta_{n+1}B_n^*f_{n+1}|^2}{|B_n+\zeta_{n+1}A_n^*f_{n+1}|^2}. \end{aligned} \tag{9.9}$$

Note that on $\mathbb{T}$, $A_n^*\overline{B_n}=\overline{A_n}B_n^*$ so that

$$\zeta_{n+1}A_n^*f_{n+1}\overline{B_n}+B_n\overline{\zeta_{n+1}A_n^*f_{n+1}}-\overline{A_n}\zeta_{n+1}B_n^*f_{n+1}-A_n\overline{\zeta_{n+1}B_n^*f_{n+1}}=0.$$

Therefore, on expanding (9.9), we find that

$$1-|f|^2=\frac{(|B_n|^2-|A_n|^2)(1-|f_{n+1}|^2)}{|B_n+\zeta_{n+1}A_n^*f_{n+1}|^2}.$$

Furthermore, by Corollary 7.3.6, we obtain

$$1-|f|^2=\frac{\omega_n(1-|f_{n+1}|^2)}{|B_n+\zeta_{n+1}A_n^*f_{n+1}|^2} \tag{9.10}$$

where

$$\omega_n=\prod_{k=0}^{k=n}(1-|\gamma_k|^2).$$

Using again Theorem 7.3.8, we get

$$\begin{aligned} |1-zf|^2 &= \left|1-\frac{zA_n+\zeta_{n+1}zB_n^*f_{n+1}}{B_n+\zeta_{n+1}A_n^*f_{n+1}}\right|^2 \\ &= \left|\frac{B_n-zA_n+\zeta_{n+1}f_{n+1}(A_n^*-zB_n^*)}{B_n+\zeta_{n+1}A_n^*f_{n+1}}\right|^2. \end{aligned}$$

In another connection, we deduce from (9.7) and Theorem 9.2.1 that

$$\begin{cases} zB_n^*-A_n^* &= \frac{1-\bar{\alpha}_{n+1}z}{\sqrt{1-|\alpha_{n+1}|^2}}\sqrt{\omega_n}\overline{\lambda_{n+1}}\phi_{n+1} \\ B_n-zA_n &= \frac{1-\bar{\alpha}_{n+1}z}{\sqrt{1-|\alpha_{n+1}|^2}}\sqrt{\omega_n}\lambda_{n+1}\phi_{n+1}^* \end{cases}$$

and therefore

$$|1-zf|^2 = \left|\frac{1-\bar{\alpha}_{n+1}z}{\sqrt{1-|\alpha_{n+1}|^2}}\sqrt{\omega_n}\right|^2\left|\frac{\lambda_{n+1}\phi_{n+1}^*-\zeta_{n+1}f_{n+1}\overline{\lambda_{n+1}}\phi_{n+1}}{B_n+\zeta_{n+1}A_n^*f_{n+1}}\right|^2 \tag{9.11}$$

$$= \omega_n\frac{|1-\bar{\alpha}_{n+1}z|^2}{1-|\alpha_{n+1}|^2}\left|\frac{\lambda_{n+1}\phi_{n+1}^*-\zeta_{n+1}f_{n+1}\overline{\lambda_{n+1}}\phi_{n+1}}{B_n+\zeta_{n+1}A_n^*f_{n+1}}\right|^2. \tag{9.12}$$

From what precedes, we deduce that

$$\frac{1-|f|^2}{|1-zf|^2}=\frac{1-|f_{n+1}|^2}{|\phi_{n+1}^*-\overline{c_{n+1}}\zeta_{n+1}f_{n+1}\phi_{n+1}|^2}\frac{1-|\alpha_{n+1}|^2}{|1-\bar{\alpha}_{n+1}z|^2}.$$

Since $\mu'(\xi) = \frac{1-|f(\xi)|^2}{|1-\xi f(\xi)|^2}$ a.e. on $\mathbb{T}$ by (9.3) and $|\phi^*_{n+1}| = |\phi_{n+1}|$ on $\mathbb{T}$, we obtain

$$\mu' = \frac{1-|f_{n+1}|^2}{|\phi_{n+1}|^2|1-\overline{c_{n+1}}\zeta_{n+1}\frac{\phi_{n+1}}{\phi^*_{n+1}}f_{n+1}|^2}\frac{1-|\alpha_{n+1}|^2}{|\xi-\alpha_{n+1}|^2} \text{ a.e. on } \mathbb{T}.$$

■

Chapter 10

Some asymptotic properties

In [Khrushchev, 2001], various kinds of convergence for the rational functions $\frac{A_n}{B_n}$ are studied in the case of the classical Schur algorithm. There, it is in particular shown that ([Khrushchev, 2001], Theorem 1):

If $\alpha_k = 0$ for every $k \geq 0$, then $|f| < 1$ a.e. on $\mathbb{T}$ if and only if $\lim_n \int_{\mathbb{T}} |f_n|^2 dm = 0$.

In this chapter, we study asymptotic properties of the Schur functions f_n and of the Wall rational functions A_n/B_n. Except for an "asymptotic-BMO-type" convergence of the Schur functions f_n, these are mainly generalizations of the results of Khrushchev where errors are integrated against the Poisson kernel of α_n rather than the Lebesgue measure. The difficulty here comes from the fact that we let the points go to the circle.

In order to prove the convergence respect to the Poincaré metric, we first need to solve a Szegő-type problem.

10.1 A Szegő-type problem

10.1.1 Generalities

We denote by μ' the Lebesgue derivative of the positive measure μ.

Definition 10.1.1 *A measure μ is called a Szegő measure if $\log(\mu') \in L^1(\mathbb{T})$.*

Let μ be a Szegő measure and let S be the Szegő function of μ:

$$S(z) = \exp\left(\frac{1}{2}\int_{\mathbb{T}} \frac{t+z}{t-z}\log(\mu')dm(t)\right).$$

The Szegő function is outer ([Garnett, 2007]) and satisfies $|S|^2 = \mu'$ almost everywhere on $\mathbb{T}$.

Szegő proved ([Szegő, 1975]) the following relation between the orthonormal polynomials ϕ_n of an absolutely continuous Szegő measure and the Szegő function S :

$$\lim_n \phi_n^*(z)S(z) = 1 \text{ locally uniformly in } \mathbb{D}.$$

This was later extended to non-absolutely continuous Szegő measures (see for example [Nikishin and Sorokin, 1991]).

A generalization of this theorem is given in [Bultheel et al., 1999] (Theorem 9.6.9) for orthogonal rational functions :

If μ is Szegő and if the points (α_n) are compactly included in $\mathbb{D}$, then we have locally uniformly in $\mathbb{D}$

$$\lim_n \left| \frac{S(z)\phi_n^*(z)(1-\bar{\alpha}_n z)}{\sqrt{1-|\alpha_n^2|}} \right| = 1.$$

Szegő also proved that the convergence of the orthonormal polynomials is uniform on the unit circle if the Lebesgue derivative of the measure is everywhere strictly positive on $\mathbb{T}$ and Lipschitz-Dini continuous, i.e. satisfies

$$|\mu'(\theta+\delta) - \mu'(\theta)| < L|\log(\delta)|^{-1-\lambda}$$

where L and λ are fixed positive numbers. Our study is akin to this: indeed, we will prove that if μ is absolutely continuous and Szegő, and if $\sum_{k=0}^{\infty}(1-|\alpha_k|) = \infty$, then the orthogonal rational functions ϕ_n satisfy

$$\lim_n |\phi_n^*(\alpha_n)|^2 |S(\alpha_n)|^2 (1-|\alpha_n|^2) = 1$$

as soon as μ' is strictly positive and Dini-continuous. We do not assume here that the α_n are compactly included in $\mathbb{D}$.
A direct consequence of this result is that, under the above hypotheses and if $\lim_n |\alpha_n| = 1$, then $|\phi_n^*(\alpha_n)|$ diverges at the same rate as $(1-|\alpha_n|^2)^{-1}$.

The main tools we will use are reproducing kernels (see section 8.1) and some facts from rational approximation.

10.1.2 An approximation problem

We recall that π_n is defined in (6.3). We denote by $\mathcal{P}_n\left(\frac{d\mu}{|\pi_n|^2}\right)$ the subspace of $L^2\left(\frac{d\mu}{|\pi_n|^2}\right)$ of polynomials of degree at most n and by $H^2\left(\frac{d\mu}{|\pi_n|^2}\right)$ the closure of the polynomials in $L^2\left(\frac{d\mu}{|\pi_n|^2}\right)$.

The idea here is to express $|\phi_n^*(\alpha_n)|^2|S(\alpha_n)|^2(1-|\alpha_n|^2)$ in terms of reproducing kernels of the spaces $\mathcal{P}_n\left(\frac{d\mu}{|\pi_n|^2}\right)$ and $H^2\left(\frac{d\mu}{|\pi_n|^2}\right)$. In what follows, we will sometime use the notation $d\mu_n$ for $\frac{d\mu}{|\pi_n|^2}$.

Proposition 10.1.2 *Let μ be an absolutely continuous Szegő measure. Then, the reproducing kernel E_n of $H^2\left(\frac{d\mu}{|\pi_n|^2}\right)$ is equal to*

$$E_n(\xi,\omega) = \frac{1}{1-\xi\bar{\omega}} \frac{\pi_n(\xi)\overline{\pi_n(\omega)}}{S(\xi)\overline{S(\omega)}}.$$

Proof First of all, it is clear that $E_n(.,\omega)$ is in $H^2(d\mu_n)$ for a fixed ω in $\mathbb{D}$ because on the one hand, $\frac{\pi_n(\xi)}{1-\xi\bar{\omega}}$ can be uniformly approximated by polynomials in $\overline{\mathbb{D}}$, and in the other

hand, the fact that S is outer implies by the Beurling theorem ([Garnett, 2007]) that there is a sequence (p_k) of polynomials such that $\lim_k \|1 - p_k S\|_{L^2(dm)} = 0$. Then,

$$\begin{aligned}\int_{\mathbb{T}} \left(\frac{1}{S} - p_k\right) \frac{d\mu}{|\pi_n|^2} &= \int_{\mathbb{T}} \left(\frac{1}{S} - p_k\right) \frac{|S|^2 dm}{|\pi_n|^2} = \int_{\mathbb{T}} (1 - p_k S)\, \bar{S} \frac{dm}{|\pi_n|^2} \\ &\leq \frac{\|S\|_{L^2(dm)} \|1 - p_k S\|_{L^2(dm)}}{\inf_{\mathbb{T}} |\pi_n|^2}\end{aligned}$$

by the Schwartz inequality. Therefore, we get $\lim_k \|p_k - 1/S\|_{L^2(d\mu_n)} = 0$.
Next, let q be a polynomial. We have

$$\begin{aligned}\int_{\mathbb{T}} q(t) \overline{\left(\frac{1}{1 - t\bar{\omega}} \frac{\pi_n(t)\overline{\pi_n(\omega)}}{S(t)\overline{S(\omega)}}\right)} \frac{d\mu(t)}{|\pi_n(t)|^2} &= \int_{\mathbb{T}} q(t) \frac{1}{1 - \bar{t}\omega} \frac{\overline{\pi_n(t)}\pi_n(\omega)}{\overline{S(t)}S(\omega)} \frac{|S(t)|^2 dm(t)}{|\pi_n(t)|^2} \\ &= \int_{\mathbb{T}} \frac{q(t)}{t - \omega} t \frac{\pi_n(\omega)}{\pi_n(t)} \frac{S(t)}{S(\omega)} dm(t) \\ &= \frac{\pi_n(\omega)}{S(\omega)} \int_{\mathbb{T}} \frac{q(t)S(t)}{(t - \omega)\pi_n(t)} t dm(t).\end{aligned}$$

As $\frac{qS}{\pi_n}$ is in H^2, we obtain by the Cauchy theorem that

$$\int_{\mathbb{T}} \frac{q(t)S(t)}{(t - \omega)\pi_n(t)} t dm(t) = \frac{q(\omega)S(\omega)}{\pi_n(\omega)}.$$

Thus, we get

$$\int_{\mathbb{T}} q(t) \overline{\left(\frac{1}{1 - t\bar{\omega}} \frac{\pi_n(t)\overline{\pi_n(\omega)}}{S(t)\overline{S(\omega)}}\right)} \frac{d\mu(t)}{|\pi_n(t)|^2} = q(\omega) \quad \text{for every } q \text{ polynomial.}$$

By density, this is true for every f in $H^2(d\mu_n)$. As the reproducing kernel is unique, the conclusion is immediate.

■

Proposition 10.1.3 *Let R_n be the reproducing kernel of $\mathcal{P}_n\left(\frac{d\mu}{|\pi_n|^2}\right)$. Then*

$$|\pi_n \phi_n^*| = \frac{|R_n(., \alpha_n)|}{\|R_n(., \alpha_n)\|_{L^2(d\mu_n)}}.$$

Proof Let p_{n-1} be a polynomial of degree at most $n - 1$. As ϕ_n is orthogonal to $\mathcal{L}_{n-1}$, we have

$$\int_{\mathbb{T}} \overline{\phi_n} \frac{p_{n-1}}{\pi_{n-1}} d\mu = 0.$$

But

$$\begin{aligned}
\int_{\mathbb{T}} \overline{\phi_n} \frac{p_{n-1}}{\pi_{n-1}} d\mu &= \int_{\mathbb{T}} \overline{\phi_n(t)} \frac{p_{n-1}(t)(1-\bar{\alpha}_n t)}{\pi_n(t)} d\mu(t) \\
&= \int_{\mathbb{T}} \frac{\phi_n^*(t)}{\mathcal{B}_n(t)} \frac{p_{n-1}(t)(1-\bar{\alpha}_n t)}{\pi_n(t)} d\mu(t) \\
&= \int_{\mathbb{T}} \phi_n^*(t) \frac{\pi_n(t)}{t^n \overline{\pi_n(t)}} \frac{p_{n-1}(t)(1-\bar{\alpha}_n t)}{\pi_n(t)} d\mu(t) \\
&= \int_{\mathbb{T}} \pi_n(t) \phi_n^*(t) \bar{t}^{n-1} p_{n-1}(t)(\bar{t}-\bar{\alpha}_n) \frac{d\mu(t)}{|\pi_n(t)|^2} \\
&= \int_{\mathbb{T}} \pi_n(t) \phi_n^*(t) \overline{\left(t^{n-1} \overline{p_{n-1}\left(\frac{1}{\bar{t}}\right)} (t-\alpha_n) \right)} \frac{d\mu(t)}{|\pi_n(t)|^2}.
\end{aligned}$$

Therefore, since $t^{n-1}\overline{p_{n-1}(1/\bar{t})}$ ranges over $\mathcal{P}_{n-1}(z)$ as p_{n-1} ranges over the same set, $\pi_n\phi_n^*$ is μ_n-orthogonal to every polynomial of degree at most n which vanishes at α_n. This is also true for $R_n(.,\alpha_n)$. Thus, $\pi_n\phi_n^*$ and $R_n(.,\alpha_n)$ are proportional. We conclude using the following equality

$$\|\pi_n\phi_n^*\|^2_{L^2(d\mu_n)} = \int_{\mathbb{T}} |\pi_n\phi_n^*|^2 \frac{d\mu}{|\pi_n|^2} = 1 = \left\| \frac{R_n(.,\alpha_n)}{\|R_n(.,\alpha_n)\|_{L^2(d\mu_n)}} \right\|^2_{L^2(d\mu_n)}.$$

■

We now derive an expression of $|\phi_n^*(\alpha_n)|^2|S(\alpha_n)|^2(1-|\alpha_n|^2)$ in terms of the reproducing kernels R_n and E_n.

Corollary 10.1.4 *For every $n \geq 1$,*

$$|\phi_n^*(\alpha_n)|^2|S(\alpha_n)|^2(1-|\alpha_n|^2) = \frac{R_n(\alpha_n,\alpha_n)}{E_n(\alpha_n,\alpha_n)} \leq 1. \tag{10.1}$$

Proof By definition of the reproducing kernel, we have

$$\|R_n(.,\alpha_n)\|^2_{L^2(d\mu_n)} = \int_{\mathbb{T}} R_n(t,\alpha_n)\overline{R_n(t,\alpha_n)} \frac{d\mu(t)}{|\pi_n(t)|^2} = R_n(\alpha_n,\alpha_n).$$

Therefore, by proposition 10.1.3,

$$|\pi_n(\alpha_n)\phi_n^*(\alpha_n)|^2 = \frac{|R_n(\alpha_n,\alpha_n)|^2}{\|R_n(.,\alpha_n)\|^2_{L^2(d\mu_n)}} = R_n(\alpha_n,\alpha_n)$$

and we get the first equality in (10.1) using the fact that, from proposition 10.1.2

$$E_n(\alpha_n,\alpha_n) = \frac{1}{1-|\alpha_n|^2} \frac{|\pi_n(\alpha_n)|^2}{|S(\alpha_n)|^2}. \tag{10.2}$$

Furthermore, as $R_n(.,\omega)$ is the orthogonal projection of $E_n(.,\omega)$ on $\mathcal{P}_n\left(\frac{d\mu}{|\pi_n|^2}\right)$ since $\mathcal{P}_n(d\mu_n) \subset H^2(d\mu_n)$, we have

$$\|R_n(.,\omega)\|_{L^2(d\mu_n)} \leq \|E_n(.,\omega)\|_{L^2(d\mu_n)} \text{ for all } \omega \in \mathbb{D}.$$

Therefore,

$$\|R_n(.,\alpha_n)\|^2_{L^2(d\mu_n)} \leq \|E_n(.,\alpha_n)\|^2_{L^2(d\mu_n)}.$$

As $R_n(\alpha_n,\alpha_n) = \|R_n(.,\alpha_n)\|^2_{L^2(d\mu_n)}$ and $E_n(\alpha_n,\alpha_n) = \|E_n(.,\alpha_n)\|^2_{L^2(d\mu_n)}$, we get

$$\frac{R_n(\alpha_n,\alpha_n)}{E_n(\alpha_n,\alpha_n)} \leq 1.$$

■

We now state our problem in an approximation-theoretic manner.

Because $R_n(.,\alpha_n)$ is the orthogonal projection of $E_n(.,\alpha_n)$ on $\mathcal{P}_n(d\mu_n)$, $R_n(.,\alpha_n)$ is the polynomial of degree at most n which minimizes

$$\min_{r_n\in\mathcal{P}_n} \|E_n(.,\alpha_n) - r_n\|_{L^2(d\mu_n)}.$$

But

$$\begin{aligned}\|E_n(.,\alpha_n) - r_n\|^2_{L^2(d\mu_n)} &= \int_{\mathbb{T}} \left|\frac{1}{1-\overline{\alpha_n}t}\frac{\pi_n(t)\overline{\pi_n(\alpha_n)}}{S(t)\overline{S(\alpha_n)}} - r_n(t)\right|^2 \frac{|S(t)|^2}{|\pi_n(t)|^2}dm(t)\\ &= \int_{\mathbb{T}} \left|\frac{1}{1-\overline{\alpha_n}t}\frac{\overline{\pi_n(\alpha_n)}}{\overline{S(\alpha_n)}} - \frac{r_n(t)S(t)}{\pi_n(t)}\right|^2 dm(t).\end{aligned}$$

Thus, finding the polynomial P_n which minimizes

$$\min_{p_n\in\mathcal{P}_n} \left\|\frac{1}{1-\overline{\alpha_n}t} - \frac{p_n(t)S(t)}{\pi_n(t)}\right\|_{L^2(dm)} \tag{10.3}$$

gives us $R_n(.,\alpha_n)$ by the relation

$$R_n(.,\alpha_n) = \frac{\overline{\pi_n(\alpha_n)}}{\overline{S(\alpha_n)}}P_n.$$

Then, in view of (10.1) and (10.2), the quantity $|\phi_n^*(\alpha_n)|^2|S(\alpha_n)|^2(1-|\alpha_n|^2)$ in which we are interested can be expressed as

$$|\phi_n^*(\alpha_n)|^2|S(\alpha_n)|^2(1-|\alpha_n|^2) = \left|\frac{P_n(\alpha_n)S(\alpha_n)}{\pi_{n-1}(\alpha_n)}\right|. \tag{10.4}$$

Now, for every polynomial p_n

$$\begin{aligned}&\left\|\frac{1}{1-\bar{\alpha}_n t} - \frac{p_n(t)S(t)}{\pi_n(t)}\right\|^2_{L^2(dm)} = \left\|\left(1-\frac{p_n(t)S(t)}{\pi_{n-1}(t)}\right)\frac{1}{1-\bar{\alpha}_n t}\right\|^2_{L^2(dm)}\\ &= \left\|\left(1-\frac{p_n(t)S(t)}{\pi_{n-1}(t)}\right)\frac{1}{t-\alpha_n}\right\|^2_{L^2(dm)}\\ &= \left\|\left(1-\frac{p_n(\alpha_n)S(\alpha_n)}{\pi_{n-1}(\alpha_n)}\right)\frac{1}{t-\alpha_n} + \left(\frac{p_n(\alpha_n)S(\alpha_n)}{\pi_{n-1}(\alpha_n)} - \frac{p_n(t)S(t)}{\pi_{n-1}(t)}\right)\frac{1}{t-\alpha_n}\right\|^2_{L^2(dm)}.\end{aligned}$$

Using the orthogonality between analytic and antianalytic functions and the Cauchy theorem, we get

$$\begin{aligned}&\left\|\frac{1}{1-\bar{\alpha}_n t}-\frac{p_n(t)S(t)}{\pi_n(t)}\right\|^2_{L^2(dm)}\\&=\left|1-\frac{p_n(\alpha_n)S(\alpha_n)}{\pi_{n-1}(\alpha_n)}\right|^2\frac{1}{1-|\alpha_n|^2}+\left\|\left(\frac{p_n(\alpha_n)S(\alpha_n)}{\pi_{n-1}(\alpha_n)}-\frac{p_n(t)S(t)}{\pi_{n-1}(t)}\right)\frac{1}{t-\alpha_n}\right\|^2_{L^2(dm)}.\end{aligned}\tag{10.5}$$

Therefore, if a sequence of polynomials (p_n) exists such that

$$\left\|\frac{1}{1-\bar{\alpha}_n t}-\frac{p_n(t)S(t)}{\pi_n(t)}\right\|^2_{L^2(dm)}=o\left(\frac{1}{1-|\alpha_n|^2}\right),\tag{10.6}$$

then by the definition of P_n (see (10.3)) we also have

$$\left\|\frac{1}{1-\bar{\alpha}_n t}-\frac{P_n(t)S(t)}{\pi_n(t)}\right\|^2_{L^2(dm)}=o\left(\frac{1}{1-|\alpha_n|^2}\right),$$

and using (10.5), we obtain

$$\lim_n\frac{P_n(\alpha_n)S(\alpha_n)}{\pi_{n-1}(\alpha_n)}=1.$$

Then, (10.4) gives

$$\lim_n|\phi_n^*(\alpha_n)|^2|S(\alpha_n)|^2(1-|\alpha_n|^2)=1.$$

Now, suppose that μ' is strictly positive and Dini continuous on $\mathbb{T}$. Then, $\frac{1}{S}$ is an analytic function, continuous on $\mathbb{T}$. If $\sum_{k=0}^n(1-|\alpha_k|)=\infty$, then $\cup_{k=0}^\infty\mathcal{L}_k$ is dense in the disk algebra $A(\mathbb{D})$ ([Achieser, 1992]). Therefore, a sequence of polynomials p_n of degree n exists such that

$$\lim_n\left\|\frac{1}{S}-\frac{p_n}{\pi_n}\right\|_\infty=0.$$

Thus,

$$\lim_n\left\|1-\frac{p_nS}{\pi_n}\right\|_\infty\leq\|S\|_\infty\lim_n\left\|\frac{1}{S}-\frac{p_n}{\pi_n}\right\|_\infty=0.$$

Since by the Cauchy theorem

$$\left\|\frac{1}{1-\bar{\alpha}_n t}-\frac{p_{n-1}(t)S(t)}{\pi_n(t)}\right\|^2_{L^2(dm)}\leq\left\|1-\frac{p_{n-1}S}{\pi_{n-1}}\right\|_\infty\frac{1}{1-|\alpha_n|^2},$$

the sequence (p_{n-1}) satisfies (10.6). We therefore obtained the following theorem :

Theorem 10.1.5 *If μ is an absolutely continuous measure such that μ' is strictly positive and Dini continuous on $\mathbb{T}$, and if $\sum_{k=0}^n(1-|\alpha_k|)=\infty$, then*

$$\lim_n|\phi_n^*(\alpha_n)|^2|S(\alpha_n)|^2(1-|\alpha_n|^2)=1.$$

Note that in our argument, we uniformly approximate the inverse of S. This leads to quite strong hypotheses. In fact, we only need to find a sequence of polynomials which satisfies the problem defined by (10.6). This problem is stated in term of L^2 norm, and without inverse of S. Therefore, the hypotheses could be probably weakened using another argument.

10.2 Convergence of the Schur functions f_n

We first give a L^2-convergence property with respect to a varying weight which is the Poisson kernel taken at the points α_j. This leads to the construction of a sequence of interpolation points for which we obtain an asymptotic-BMO-type convergence.

10.2.1 L^2 convergence with respect to a varying weight

We first show a weak-(*) convergence of the measures $\frac{P(.,\alpha_n)}{|\phi_n|^2}dm$:

Lemma 10.2.1 *If* $\sum_{k=1}^{k=\infty}(1-|\alpha_k|)=\infty$ *then*

$$(*)-\lim_n \frac{P(.,\alpha_n)}{|\phi_n|^2}dm = d\mu.$$

Proof Corollary 8.3.5 states that $\phi_0,\ldots,\phi_n$ are orthonormal in $L^2\left(\frac{P(.,\alpha_n)}{|\phi_n|^2}dm\right)$. Therefore, $\phi_0,\ldots,\phi_n$ are orthonormal in $L^2(d\mu)$ and in $L^2\left(\frac{P(.,\alpha_n)}{|\phi_n|^2}dm\right)$. Thus,

$$\int_{\mathbb{T}} \phi_i\overline{\phi_j}\frac{P(.,\alpha_n)}{|\phi_n|^2}dm = \int_{\mathbb{T}} \phi_i\overline{\phi_j}d\mu$$

for all $0\le i,j\le n$. In particular, for all $0\le i\le n$, we have

$$\int_{\mathbb{T}} \phi_i\frac{P(.,\alpha_n)}{|\phi_n|^2}dm = \int_{\mathbb{T}} \phi_i d\mu.$$

As $(\phi_k)_{0\le k\le n}$ is a basis of $\mathcal{L}_n$, for all $g\in\mathcal{L}_n$, we get

$$\int_{\mathbb{T}} g\frac{P(.,\alpha_n)}{|\phi_n|^2}dm = \int_{\mathbb{T}} g d\mu \tag{10.7}$$

and upon conjugating,

$$\int_{\mathbb{T}} \bar{g}\frac{P(.,\alpha_n)}{|\phi_n|^2}dm = \int_{\mathbb{T}} \bar{g} d\mu. \tag{10.8}$$

But, as $\sum_{k=1}^{k=\infty}(1-|\alpha_k|)=\infty$, $\cup_{k=0}^{k=\infty}\mathcal{L}_k\bigcup\cup_{k=0}^{k=\infty}\overline{\mathcal{L}_k}$ is dense in $C(\mathbb{T})$, the space of continuous functions in $\mathbb{T}$ ([Achieser, 1992]). Therefore,

$$(*)-\lim_n \frac{P(.,\alpha_n)}{|\phi_n|^2}dm = d\mu.$$

■

Note that if the points are compactly included in $\mathbb{D}$ and if I is an open arc on $\mathbb{T}$ such that μ has no mass at the end-points, then we have

$$\lim_n \int_I g\frac{P(.,\alpha_n)}{|\phi_n|^2}dm \le \int_I g d\mu \text{ for every } g\in C(\mathbb{T}). \tag{10.9}$$

Indeed, let $\epsilon > 0$ and let h_I be a continuous positive function such that $h_I(t) = 1$ for every t in I and $\int_{\mathbb{T}} h_I d\mu \leq \mu(I) + \epsilon$. Then, since all the functions are positive, we have

$$\int_I \frac{P(.,\alpha_n)}{|\phi_n|^2} dm \leq \int_{\mathbb{T}} h_I \frac{P(.,\alpha_n)}{|\phi_n|^2} dm.$$

We conclude using the previous lemma since

$$\lim_n \int_{\mathbb{T}} h_I \frac{P(.,\alpha_n)}{|\phi_n|^2} dm = \int_{\mathbb{T}} h_I d\mu \leq \mu(I) + \epsilon.$$

Note also that if $\sum_{k=1}^{k=\infty}(1 - |\alpha_k|) = \infty$, since $P(z,\alpha_n) = z/(z-\alpha_n) + \bar{\alpha_n} z/(1-\bar{\alpha_n} z)$, $P(.,\alpha_n)$ is in $\mathcal{L}_n + \bar{\mathcal{L}_n}$, then we get using (10.7) and (10.8)

$$\int_{\mathbb{T}} P(.,\alpha_n) \frac{P(.,\alpha_n)}{|\phi_n|^2} dm = \int_{\mathbb{T}} P(.,\alpha_n) d\mu. \tag{10.10}$$

If the interpolation points do not tend "too quickly" toward the circle, we have the following L^2-convergence :

Theorem 10.2.2 *Let μ be an absolutely continuous measure. If $\sum_{k=1}^{k=\infty}(1 - |\alpha_k|) = \infty$ and $\lim_k |\alpha_k| = 1$, and if at every point of accumulation of the (α_k) f is continuous and $|f| < 1$, then*

$$\lim_k \int |f_k|^2 P(.,\alpha_k) dm = 0.$$

Proof Suppose that the limit does not converge to 0. Then, there is $\epsilon > 0$, an infinite set $K \subset \mathbb{N}$ and a sub-sequence of (α_k) which converges to $\alpha \in \mathbb{T}$ such that

$$\forall n \in K, \quad \int |f_n|^2 P(.,\alpha_n) dm \geq \epsilon.$$

By theorem 9.3.3, using the elementary equality

$$|1 - \overline{c_n}\zeta_n \frac{\phi_n}{\phi_n^*} f_n|^2 = 1 + |f_n|^2 - 2Re(\overline{c_n}\zeta_n \frac{\phi_n}{\phi_n^*} f_n)$$

we get

$$|\phi_n|^2 \mu'(1 + |f_n|^2 - 2Re(\overline{c_n}\zeta_n \frac{\phi_n}{\phi_n^*} f_n)) = (1 - |f_n|^2) P(.,\alpha_n)$$

and therefore

$$|f_n|^2 = \frac{P(.,\alpha_n) - |\phi_n|^2\mu'}{P(.,\alpha_n) + |\phi_n|^2\mu'} + \frac{2|\phi_n|^2\mu' Re(\overline{c_n}\zeta_n \frac{\phi_n}{\phi_n^*} f_n)}{P(.,\alpha_n) + |\phi_n|^2\mu'}.$$

Thus, we obtain

$$|f_n|^2 = \frac{P(.,\alpha_n) - |\phi_n|^2\mu'}{P(.,\alpha_n) + |\phi_n|^2\mu'} - \frac{P(.,\alpha_n) - |\phi_n|^2\mu'}{P(.,\alpha_n) + |\phi_n|^2\mu'} Re\left(\overline{c_n}\zeta_n \frac{\phi_n}{\phi_n^*} f_n\right) + Re\left(\overline{c_n}\zeta_n \frac{\phi_n}{\phi_n^*} f_n\right).$$

Since $\zeta_n(\alpha_n) = 0$, we get by harmonicity

$$\int Re\left(\overline{c_n}\zeta_n \frac{\phi_n}{\phi_n^*} f_n\right) P(.,\alpha_n)dm = 0.$$

Consequently,

$$\int |f_n|^2 P(.,\alpha_n)dm = \int \frac{P(.,\alpha_n) - |\phi_n|^2\mu'}{P(.,\alpha_n) + |\phi_n|^2\mu'}\left(1 - Re\left(\overline{c_n}\zeta_n \frac{\phi_n}{\phi_n^*} f_n\right)\right) P(.,\alpha_n)dm.$$

But since ζ_n, f_n and $\frac{\phi_n}{\phi_n^*}$ are Schur functions (see proposition 8.2.2),

$$\left|1 - Re\left(\overline{c_n}\zeta_n \frac{\phi_n}{\phi_n^*} f_n\right)\right| \leq 2$$

and we get

$$\int |f_n|^2 P(.,\alpha_n)dm \leq 2\int \left|1 - \frac{2|\phi_n|^2\mu'}{P(.,\alpha_n) + |\phi_n|^2\mu'}\right| P(.,\alpha_n)dm. \tag{10.11}$$

Let

$$g_n = \frac{2|\phi_n|^2\mu'}{P(.,\alpha_n) + |\phi_n|^2\mu'}.$$

Using the inequality

$$\frac{4x^2}{(1+x)^2} \leq x \text{ for all } x \geq 0$$

we deduce

$$\begin{aligned}
\int_{\mathbb{T}} g_n^2 P(.,\alpha_n)dm &= \int_{\mathbb{T}} \frac{4(|\phi_n|^2\mu' P(.,\alpha_n)^{-1})^2}{(1+|\phi_n|^2\mu' P(.,\alpha_n)^{-1})^2} P(.,\alpha_n)dm \\
&\leq \int_{\mathbb{T}} |\phi_n|^2\mu' P(.,\alpha_n)^{-1} P(.,\alpha_n)dm \\
&= \int_{\mathbb{T}} |\phi_n|^2\mu' dm \leq 1
\end{aligned}$$

because of the orthonormality of ϕ_n. By the Schwarz inequality, it follows that

$$\int_{\mathbb{T}} g_n P(.,\alpha_n)dm \leq \left(\int_{\mathbb{T}} g_n^2 P(.,\alpha_n)dm\right)^{1/2} \leq 1. \tag{10.12}$$

Furthermore, we get again by the Schwarz inequality:

$$\begin{aligned}
\int_{\mathbb{T}} \sqrt{\mu'} P(.,\alpha_n)dm &= \int_{\mathbb{T}} \frac{\sqrt{2}|\phi_n|\sqrt{\mu'}\sqrt{P(.,\alpha_n)}}{\sqrt{P(.,\alpha_n) + |\phi_n|^2\mu'}} \frac{\sqrt{P(.,\alpha_n) + |\phi_n|^2\mu'}\sqrt{P(.,\alpha_n)}}{\sqrt{2}|\phi_n|} dm \\
&\leq \left(\int_{\mathbb{T}} g_n P(.,\alpha_n)dm\right)^{1/2} \left(\frac{1}{2}\int_{\mathbb{T}} \left(\frac{P(.,\alpha_n)}{|\phi_n|^2} + \mu'\right) P(.,\alpha_n)dm\right)^{1/2}.
\end{aligned}$$

Using (10.10) and the absolutely continuity of the measure, we get

$$\int_{\mathbb{T}} \sqrt{\mu'} P(.,\alpha_n)dm \leq \left(\int_{\mathbb{T}} g_n P(.,\alpha_n)dm\right)^{1/2} \left(\int_{\mathbb{T}} \mu' P(.,\alpha_n)dm\right)^{1/2}. \tag{10.13}$$

Since by hypothesis, (α_n) converges to $\alpha \in \mathbb{T}$ and μ' is continuous at α, passing to the inferior limit in (10.13), we get

$$\sqrt{\mu'(\alpha)} \leq \sqrt{\mu'(\alpha)} \liminf_n \left(\int_{\mathbb{T}} g_n P(., \alpha_n) dm \right)^{1/2}.$$

Therefore, we obtain

$$\liminf_n \int_{\mathbb{T}} g_n P(., \alpha_n) dm \geq 1.$$

Combining this last inequality with (10.12), we obtain

$$\lim_n \int_{\mathbb{T}} g_n P(., \alpha_n) dm = \lim_n \int_{\mathbb{T}} g_n^2 P(., \alpha_n) dm = 1.$$

It follows that

$$\lim_n \int_{\mathbb{T}} (1-g_n)^2 P(., \alpha_n) dm = \int_{\mathbb{T}} P(., \alpha_n) dm - 2 \lim_n \int_{\mathbb{T}} g_n P(., \alpha_n) dm + \lim_n \int_{\mathbb{T}} g_n^2 P(., \alpha_n) dm = 0.$$

Thus, using the Schwarz inequality and (10.11), we conclude that

$$\lim_n \int_{\mathbb{T}} |f_n|^2 P(., \alpha_n) dm = 0.$$

■

A similar type of convergence is obtained when the (α_n) are compactly in included in $\mathbb{D}$.

Theorem 10.2.3 *If the (α_k) are compactly included in $\mathbb{D}$ and if $|f| < 1$ a.e. on $\mathbb{T}$, then*

$$\lim_n \int |f_n|^2 P(., \alpha_n) dm = 0.$$

Proof We denote by $\alpha \in \mathbb{D}$ an accumulation point of (α_k). Using the same argument as above, equation (10.12) still holds. Now, for any open arc I on $\mathbb{T}$ with no mass at the end-points, we get by the Schwarz inequality:

$$\begin{aligned} \frac{1}{m(I)} \int_I \sqrt{\mu'} dm &= \frac{1}{m(I)} \int_I \frac{\sqrt{2}|\phi_n|\sqrt{\mu'}}{\sqrt{P(., \alpha_n) + |\phi_n|^2 \mu'}} \frac{\sqrt{P(., \alpha_n) + |\phi_n|^2 \mu'}}{\sqrt{2}|\phi_n|} dm \quad (10.14) \\ &\leq \left(\frac{1}{m(I)} \int_I g_n dm \right)^{1/2} \left(\frac{1}{2|I|} \int_I \left(\frac{P(., \alpha_n)}{|\phi_n|^2} + \mu' \right) dm \right)^{1/2} \quad (10.15) \end{aligned}$$

As $g_n = \frac{2|\phi_n|^2 \mu' P(e^{i\theta}, \alpha_n)^{-1}}{1+|\phi_n|^2 \mu' P(e^{i\theta}, \alpha_n)^{-1}}$, we have $0 \leq g_n \leq 2$ a.e on $\mathbb{T}$. Let g be a weak-$(*)$ limit of the bounded sequence $(g_n)_n$ in $L^\infty(\mathbb{T})$. Passing to the limit in (10.15), and using (10.9), we obtain

$$\frac{1}{|I|} \int_I \sqrt{\mu'} dm \leq \left(\frac{1}{|I|} \int_I g dm \right)^{1/2} \left(\frac{1}{2} \frac{\mu(I)}{|I|} + \frac{1}{2|I|} \int_I \mu' dm \right)^{1/2}.$$

Thus, by Lebesgue's theorem on differentiation and by Helly's theorem ([Duren, 1970]),

$$\sqrt{\mu'} \leq \sqrt{g}\left(\frac{1}{2}\mu' + \frac{1}{2}\mu'\right)^{1/2} \leq \sqrt{g}\sqrt{\mu'} \text{ a.e. on } \mathbb{T}.$$

Since $\mu' > 0$ a.e. on $\mathbb{T}$, $g \geq 1$ a.e. on $\mathbb{T}$. Combining this last inequality with (10.12), and using the fact that $\lim_n P(., \alpha_n) = P(., \alpha)$ uniformly on $\mathbb{T}$, we obtain

$$\lim_n \int_{\mathbb{T}} g_n P(., \alpha_n) dm = \lim_n \int_{\mathbb{T}} g_n^2 P(., \alpha_n) dm = 1.$$

It follows that

$$\lim_n \int_{\mathbb{T}} (1-g_n)^2 P(., \alpha_n) dm = \int_{\mathbb{T}} P(., \alpha_n) dm - 2\lim_n \int_{\mathbb{T}} g_n P(., \alpha_n) dm + \lim_n \int_{\mathbb{T}} g_n^2 P(., \alpha_n) dm = 0.$$

Thus, using the Schwarz inequality and (10.11), we conclude that

$$\lim_n \int_{\mathbb{T}} |f_n|^2 P(., \alpha_n) dm = 0.$$

■

Combining the proofs of the two previous theorems, we obtain:

Corollary 10.2.4 *Let μ be an absolutely continuous measure. If $\sum_{k=1}^{k=\infty}(1-|\alpha_k|) = \infty$, if $|f| < 1$ a.e. on $\mathbb{T}$ and if at every point of accumulation of the (α_k) in $\mathbb{T}$, f is continuous and $|f| < 1$, then*

$$\lim_k \int |f_k|^2 P(., \alpha_k) dm = 0.$$

In particular, we obtain a result stated in ([Khrushchev, 2001]) for the classical Schur algorithm:

Corollary 10.2.5 *If $1 \leq p < \infty$, $|f| < 1$ a.e. on $\mathbb{T}$ and $\alpha_k = 0$ for every $k \geq 1$ then*

$$\lim_n \int_{\mathbb{T}} |f_n|^p \, dm = 0.$$

Proof As $\|f_n\|_\infty \leq 1$ for all n, the sequence f_n is in L^p for all $1 \leq p \leq \infty$. But $\|f_n\|_2$ converges to 0, so for every sequence, we can extract a subsequence such that $\lim_k f_k(t) = 0$ a.e. on $\mathbb{T}$. We conclude using Lebesgue's dominated convergence.

■

10.2.2 An asymptotic-BMO-type convergence

In the following, we will construct a sequence of interpolation points for which the sequence f_n tends in L^1 mean to its average on smaller and smaller intervals.

Theorem 10.2.6 *Let $(\epsilon_k)_{k\in\mathbb{N}}$ be a sequence of real numbers such that*

$$\begin{cases} 0 < \epsilon_k \leq \frac{1}{\pi}, \\ \sum_{k=0}^{k=\infty} \epsilon_k = \infty, \\ \lim_{k\to\infty} \epsilon_k = 0, \end{cases}$$

and f be a continuous Schur function such that $|f| < 1$ on $\mathbb{T}$.

Then the points $(\alpha_k)_k$ can be chosen such that

$$\lim_n \sup_{\alpha\in D_n} \int_{\mathbb{T}} |f_n(t) - f_n(\alpha)|\, P(t,\alpha) dm(t) = 0.$$

where D_n denotes the closed disk of radius $1 - \epsilon_n\pi$:

$$D_n = \{z \in \mathbb{C}, |z| \leq 1 - \epsilon_n\pi\}.$$

Proof Recall that

$$|f_{n+1}(e^{i\theta})| = \left| \frac{f_n(e^{i\theta}) - f_n(\alpha_{n+1})}{1 - \overline{f_n(\alpha_{n+1})} f_n(e^{i\theta})} \right|.$$

We denote by $\mathcal{I}_n$ the application from $\mathbb{D}$ to $[0,1]$ such that

$$\mathcal{I}_n(\alpha) = \int_{\mathbb{T}} \left| \frac{f_n(t) - f_n(\alpha)}{1 - \overline{f_n(\alpha)} f_n(t)} \right|^2 P(t,\alpha) dm(t).$$

At each step of the Schur algorithm, we may choose $\alpha_{n+1} \in D_n$ which maximizes $\mathcal{I}_n$. Then we have :

$$\begin{aligned} \int_{\mathbb{T}} |f_{n+1}(t)|^2 P(t,\alpha_{n+1}) dm(t) &= \int_{\mathbb{T}} \left| \frac{f_n(t) - f_n(\alpha_{n+1})}{1 - \overline{f_n(\alpha_{n+1})} f_n(t)} \right|^2 P(t,\alpha_{n+1}) dm(t) \\ &= \sup_{\alpha\in D_n} \int_T \left| \frac{f_n(t) - f_n(\alpha)}{1 - \overline{f_n(\alpha)} f_n(t)} \right|^2 P(t,\alpha) dm(t). \end{aligned}$$

As f_n is Schur, $|1 - \overline{f_n(\alpha)} f_n(t)| \leq 2$. Therefore,

$$2 \int_{\mathbb{T}} |f_{n+1}(t)|^2 P(t,\alpha_{n+1}) dm(t) \geq \sup_{\alpha\in D_n} \int_{\mathbb{T}} |f_n(t) - f_n(\alpha)|^2 P(t,\alpha) dm(t).$$

Using the Schwarz inequality, we get

$$2 \int_{\mathbb{T}} |f_{n+1}(t)|^2 P(t,\alpha_{n+1}) dm(t) \geq \left(\sup_{\alpha\in D_n} \int_{\mathbb{T}} |f_n(t) - f_n(\alpha)|\, P(t,\alpha) dm(t) \right)^2.$$

Thus, corollary 10.2.4 gives

$$\lim_n \sup_{\alpha\in D_n} \int_{\mathbb{T}} |f_n(t) - f_n(\alpha)|\, P(t,\alpha) dm(t) = 0.$$

■

Corollary 10.2.7 *Under the same hypothesis as the previous theorem, the points $(\alpha_k)_k$ can be chosen such that*

$$\lim_{n\to\infty} \sup_{m(I)\geq\epsilon_n} \frac{1}{m(I)} \int_I |f_n - (f_n)_I|\, dm = 0$$

where $(f_n)_I$ is defined by

$$(f_n)_I = \frac{1}{m(I)} \int_I f_n dm.$$

Proof Let I be an arc of $\mathbb{T}$ such that $m(I) \geq \epsilon_n$.
Suppose first that $m(I) \leq \frac{1}{\pi}$ and define by α_I the point of D_n such that $\alpha_I = (1 - m(I)\pi)e^{i\theta_I}$ where $e^{i\theta_I}$ is the center of I. We have

$$\begin{aligned} P(e^{i\theta}, \alpha_I) &= \frac{1 - |\alpha_I|^2}{1 - 2|\alpha_I|\cos(\theta - \theta_I) + |\alpha_I|^2} \\ &= \frac{1 + |\alpha_I|}{1 - |\alpha_I| + 2|\alpha_I|\frac{1-\cos(\theta-\theta_I)}{1-|\alpha_I|}}. \end{aligned}$$

Suppose that $e^{i\theta} \in I$, that is $|\theta - \theta_I| \leq m(I)\pi$. Using the inequality $1 - \cos(x) \leq \frac{x^2}{2}$, we get

$$\begin{aligned} P(e^{i\theta}, \alpha_I) &\geq \frac{1 + |\alpha_I|}{1 - |\alpha_I| + |\alpha_I|\frac{(\theta-\theta_I)^2}{1-|\alpha_I|}} \\ &\geq \frac{1 + |\alpha_I|}{1 - |\alpha_I| + \frac{|\alpha_I|\pi^2 m(I)^2}{1-|\alpha_I|}} \\ &\geq \frac{2 - \pi m(I)}{\pi m(I) + \frac{(1-\pi m(I))\pi^2 m(I)^2}{\pi m(I)}} \\ &\geq \frac{1}{m(I)\pi}. \end{aligned}$$

Therefore, if χ stands for the characteristic function of I and if $\epsilon_n \leq m(I) \leq \frac{1}{\pi}$, then $\frac{\chi(t)}{m(I)} \leq \pi P(t, \alpha_I)$.
Furthermore, if $m(I) > \frac{1}{\pi}$, we have $\pi P(t, 0) = \pi \geq \frac{1}{m(I)}$. Thus, for all arc I of $\mathbb{T}$ such that $m(I) \geq \epsilon_n$, a point α_I in D_n exists such that

$$\frac{\chi(t)}{m(I)} \leq \pi P(t, \alpha_I).$$

Now, remark that $|(f_n)_I - f_n(\alpha_I)| \leq 1/m(I) \int_I |f_n - f_n(\alpha_I)| dm$. Indeed,

$$\begin{aligned} |(f_n)_I - f_n(\alpha_I)| &= \left| \frac{1}{m(I)} \int_I f_n dm - f_n(\alpha_I) \right| = \left| \frac{1}{m(I)} \int_I (f_n - f_n(\alpha_I)) dm \right| \\ &\leq \frac{1}{m(I)} \int_I |f_n - f_n(\alpha_I)| dm. \end{aligned}$$

We conclude using the above theorem and the following inequalities:

$$\begin{aligned}
\sup_{m(I)\geq\epsilon_n} \frac{1}{m(I)} \int_I |f_n - (f_n)_I|\, dm &\leq \sup_{m(I)\geq\epsilon_n} \frac{1}{m(I)} \int_I (|f_n - f_n(\alpha_I)| + |f_n(\alpha_I) - (f_n)_I|)\, dm \\
&\leq 2 \sup_{m(I)\geq\epsilon_n} \frac{1}{m(I)} \int_I |f_n - f_n(\alpha_I)|\, dm \\
&= 2 \sup_{m(I)\geq\epsilon_n} \int \frac{\chi}{m(I)} |f_n - f_n(\alpha_I)|\, dm \\
&\leq 2\pi \sup_{m(I)\geq\epsilon_n} \int |f_n - f_n(\alpha_I)|\, P(., \alpha_I) dm \\
&\leq 2\pi \sup_{\alpha\in D_n} \int |f_n - f_n(\alpha)|\, P(., \alpha) dm
\end{aligned}$$

■

If no constraint is made on the length of the intervals (i.e. $\epsilon_n = 0$ for each n), then the convergence in the previous corollary is called a BMO convergence. Details about BMO can be found in [Garnett, 2007], Chapter 6.

Here, an unsolved question appears: which hypotheses are needed on f in order to obtain a BMO convergence? The difficulty to answer such a question is that the hypotheses made on f have to propagate to every f_n throughout the Schur algorithm.

Note also that we do not obtain a similar result of convergence for the Wall rational functions A_n/B_n. Here, the problem is due to the mean $(f_n)_I$.

10.3 Convergence of the Wall rational functions A_n/B_n

We will now give different kinds of convergence for the Wall rational functions. The first one is convergence on compact subset which is deduced merely from an elementary property satisfied by the zeros of a non-zero function in H^∞. The other three (convergence in the pseudo-hyperbolic distance, the Poincaré metric, and in $L^2(\mathbb{T})$) are implied by the convergence of the Schur functions f_n in $L^2(\mathbb{T})$.

10.3.1 Convergence on compact subsets

Convergence of A_n/B_n on compact subsets of $\mathbb{D}$ is easily obtained, using the fact that the zeros of a non-zero function in H^∞ satisfy the relation $\sum_{k=1}^{\infty}(1-|\alpha_k|) < \infty$ ([Rudin, 1987]).

Theorem 10.3.1 *If $\sum_{k=1}^{k=\infty}(1 - |\alpha_k|) = \infty$, $\frac{A_n}{B_n}$ converges to f uniformly on compact subsets of $\mathbb{D}$.*

Proof As $\left\|\frac{A_n}{B_n}\right\|_\infty \leq 1$ for all $n \in \mathbb{N}$, $\left\{\frac{A_n}{B_n}\right\}$ is a normal family. Therefore, a subsequence that converges uniformly on compact subsets can be extracted. We denote by $\check{f}$ the limit of such a subsequence. As $\frac{A_n}{B_n}(\alpha_k) = f(\alpha_k)$ for all $n \geq k-1$, $f(\alpha_k) = \check{f}(\alpha_k)$ for all k. Thus, the function $f - \check{f}$ belongs to H^∞ and the points α_k are its zeros. As $\sum_{k=1}^{k=\infty}(1-|\alpha_k|) = \infty$,

we conclude that $f = \check{f}$. Thus, f is the only limit point, and A_n/B_n converges to f, locally uniformly in $\mathbb{D}$

■

10.3.2 Convergence with respect to the pseudohyperbolic distance

The pseudohyperbolic distance ρ on $\mathbb{D}$ is defined by ([Garnett, 2007])

$$\rho(z,w) = \left|\frac{z-w}{1-\bar{w}z}\right|.$$

Convergence with respect to the pseudohyperbolic distance is essentially a consequence of the following well-known property.

Property 10.3.2 *The pseudohyperbolic distance is invariant under Moebius transformations.*

Proof Let $\mathcal{M}$ be the Moebius transform defined by

$$\mathcal{M}(z) = \beta\frac{z-\alpha}{1-\bar{\alpha}z} \text{ with } \alpha \in \mathbb{D} \text{ and } \beta \in \mathbb{T}.$$

We have

$$\begin{aligned}\mathcal{M}(z) - \mathcal{M}(\omega) &= \beta\left(\frac{z-\alpha}{1-\bar{\alpha}z} - \frac{\omega-\alpha}{1-\bar{\alpha}\omega}\right)\\ &= \beta\frac{(1-|\alpha|^2)(z-\omega)}{(1-\bar{\alpha}z)(1-\bar{\alpha}\omega)}\end{aligned}$$

and

$$\begin{aligned}1-\overline{\mathcal{M}(z)}\mathcal{M}(\omega) &= 1-\overline{\left(\beta\frac{z-\alpha}{1-\bar{\alpha}z}\right)}\beta\frac{\omega-\alpha}{1-\bar{\alpha}\omega}\\ &= \frac{(1-|\alpha|^2)(1-\bar{z}\omega)}{(1-\alpha\bar{z})(1-\bar{\alpha}\omega)}.\end{aligned}$$

Therefore,

$$\left|\frac{\mathcal{M}(z)-\mathcal{M}(\omega)}{1-\overline{\mathcal{M}(z)}\mathcal{M}(\omega)}\right| = \left|\frac{z-\omega}{1-\bar{z}\omega}\right|.$$

■

The proof of convergence is now immediate ([Khrushchev, 2001], Corollary 2.4 for $\alpha_k = 0$):

Theorem 10.3.3 *If $|f| < 1$ on $\mathbb{T}$, f continuous, and $\sum_{k=1}^{k=\infty}(1-|\alpha_k|) = \infty$ then*

$$\lim_n \int_{\mathbb{T}} \rho\left(f, \frac{A_n}{B_n}\right)^2 P(.,\alpha_{n+1})dm = 0$$

Proof As the pseudohyperbolic distance is invariant under Moebius transformations, we have in view of (7.2) and (7.3),

$$\rho\left(f, \frac{A_n}{B_n}\right) = \rho\left(\tau_0 \circ \cdots \circ \tau_n(f_{n+1}), \tau_0 \circ \cdots \circ \tau_n(0)\right) = \rho(f_{n+1}, 0) = |f_{n+1}|.$$

We conclude using Corollary 10.2.4.

■

10.3.3 Convergence with respect to the Poincaré metric

In the disk, the Poincaré metric is defined by

$$\mathfrak{P}(z,\omega) = \log\left(\frac{1+\rho(z,\omega)}{1-\rho(z,\omega)}\right) \text{ for } z,\omega \in \mathbb{D}.$$

The following theorem is given in the classical case (i.e. $\alpha_k = 0$) in [Khrushchev, 2001], Theorem 2.6.

Theorem 10.3.4 *If μ is an absolutely continuous measure such that μ' is positive and Dini continuous on $\mathbb{T}$ and if $\sum_{k=0}^{n}(1-|\alpha_k|) = \infty$, then*

$$\lim_n \int_{\mathbb{T}} \mathfrak{P}\left(f, \frac{A_n}{B_n}\right) P(., \alpha_{n+1}) dm = 0.$$

In particular, this holds if $|f| < 1$ and f is Dini-continuous on $\mathbb{T}$.

Proof Using again the invariance of the pseudohyperbolic distance under Moebius transformations, we get $\rho\left(f, \frac{A_n}{B_n}\right) = |f_{n+1}|$. This gives

$$\mathfrak{P}\left(f, \frac{A_n}{B_n}\right) = \log\left(\frac{1+|f_{n+1}|}{1-|f_{n+1}|}\right). \tag{10.16}$$

Using Theorem 9.3.3 and the definition of the Szegő function S, since $|\phi_n| = |\phi_n^*|$ on $\mathbb{T}$, we get

$$|\phi_n^*|^2 |S|^2 \frac{|1-\bar{\alpha}_n \xi|^2}{1-|\alpha_n|^2} = \frac{1-|f_n|^2}{|1-\overline{c_n}\zeta_n \frac{\phi_n}{\phi_n^*} f_n|^2} \text{ a.e. on } \mathbb{T}. \tag{10.17}$$

Furthermore, if g is a Schur function, $1-g$ is a function in H^∞ such that $Re(1-g) \geq 0$, and therefore $1-g$ is an outer function (see [Garnett, 2007], Corollary 4.8). Thus,

$$\int_{\mathbb{T}} \log|1-g|^2 P(., \alpha_n) dm = \log(|1-g(\alpha_n)|^2).$$

Consequently, since $\zeta_n(\alpha_n) = 0$,we obtain on putting $g = \overline{c_n}\zeta_n \frac{\phi_n}{\phi_n^*} f_n$ that

$$\int_{\mathbb{T}} \log|1-\overline{c_n}\zeta_n \frac{\phi_n}{\phi_n^*} f_n|^2 P(., \alpha_n) dm = \log(|1-\overline{c_n}\zeta_n(\alpha_n)\frac{\phi_n(\alpha_n)}{\phi_n^*(\alpha_n)} f_n(\alpha_n)|^2) = \log(1) = 0.$$

Using the previous equation and (10.17), we get

$$\int_{\mathbb{T}} \log\left(|\phi_n^*|^2|S|^2\frac{|1-\bar{\alpha}_n\xi|^2}{1-|\alpha_n|^2}\right)P(\xi,\alpha_n)dm(\xi) = \int_{\mathbb{T}} \log(1-|f_n|^2)P(\xi,\alpha_n)dm(\xi).$$

As ϕ_n^*, S^2 and $1-\bar{\alpha}_n\xi$ are outer functions, we obtain

$$\log(|\phi_n^*(\alpha_n)|^2|S(\alpha_n)|^2(1-|\alpha_n|^2)) = \int_{\mathbb{T}} \log(1-|f_n|^2)P(.,\alpha_n)dm,$$

and Theorem 10.1.5 gives us

$$\lim_n \int_{\mathbb{T}} \log(1-|f_n|^2)P(.,\alpha_n)dm = 0. \tag{10.18}$$

Using the inequality $\log(1+x) \leq x$ for $x > -1$, we get

$$0 \leq |f_n|^2 \leq -\log(1-|f_n|^2) \tag{10.19}$$

and

$$0 \leq \log(1+|f_n|) \leq |f_n|. \tag{10.20}$$

Therefore, by (10.19) and (10.18),

$$\lim_n \int_{\mathbb{T}} |f_n|^2 P(.,\alpha_n)dm = 0$$

and, by the previous equation and (10.20),

$$\lim_n \int_{\mathbb{T}} \log(1+|f_n|)P(.,\alpha_n)dm = 0$$

because, by the Schwarz inequality,

$$0 \leq \int_{\mathbb{T}} \log(1+|f_n|)P(.,\alpha_n)dm \leq \int_{\mathbb{T}} |f_n|P(.,\alpha_n)dm \leq \left(\int_{\mathbb{T}} |f_n|^2 P(.,\alpha_n)dm\right)^{1/2}.$$

Since $\log(1-|f_n|^2) = \log(1-|f_n|) + \log(1+|f_n|)$, we also have

$$\lim_n \int_{\mathbb{T}} \log(1-|f_n|)P(.,\alpha_n)dm = 0.$$

We obtain the expected result by (10.16).

■

10.3.4 Convergence in $L^2(\mathbb{T})$

Using the relation between f_{n+1} and $\frac{A_n}{B_n}$ and the L^2 convergence of the Schur functions f_n, we shall directly obtain the L^2 convergence of the Wall rational functions $\frac{A_n}{B_n}$ as follows.

Lemma 10.3.5 *For $t \in \mathbb{T}$, we have*

$$|f_{n+1}(t)| \left| 1 - \frac{A_n}{B_n}(t)\overline{f(t)} \right| = \left| f(t) - \frac{A_n}{B_n}(t) \right|.$$

Proof Proposition 7.3.8 gives

$$f(z) = \frac{A_n(z) + \zeta_{n+1}(z)B_n^*(z)f_{n+1}(z)}{B_n(z) + \zeta_{n+1}(z)A_n^*(z)f_{n+1}(z)}.$$

Therefore,

$$f(z) - \frac{A_n(z)}{B_n(z)} = \zeta_{n+1}(z)f_{n+1}(z)\frac{B_n^*(z) - A_n^*(z)f(z)}{B_n(z)}.$$

Thus, for $t \in \mathbb{T}$,

$$\begin{aligned} \left| f(t) - \frac{A_n(t)}{B_n(t)} \right| &= |\zeta_{n+1}(t)f_{n+1}(t)| \left| \frac{\mathcal{B}_n(t)(\overline{B_n(t)} - \overline{A_n(t)}f(t))}{B_n(t)} \right| \\ &= |f_{n+1}(t)| \left| \frac{\overline{B_n(t)} - \overline{A_n(t)}f(t)}{B_n(t)} \right| \\ &= |f_{n+1}(t)| \left| 1 - \frac{\overline{A_n(t)}}{\overline{B_n(t)}}f(t) \right|. \end{aligned}$$

■

Proposition 10.3.6 *The convergence in L^p, $1 \leq p < \infty$, of f_n to zero with respect to the varying weight $P(., \alpha_n)$ implies the convergence in L^p of $\frac{A_n}{B_n}$ to f with respect to $P(., \alpha_{n+1})$.*

Proof As f and $\frac{A_n}{B_n}$ are two Schur functions, using the previous lemma, we get

$$\left| f(t) - \frac{A_n}{B_n}(t) \right| \leq 2|f_{n+1}(t)| \text{ for } t \in \mathbb{T}.$$

The conclusion is then immediate by dominated convergence.

■

The two following corollaries are direct applications of the previous results.

Corollary 10.3.7 *If $\sum_{k=1}^{k=\infty}(1 - |\alpha_k|) = \infty$, and if $|f| < 1$ and f is continuous on $\mathbb{T}$, then*

$$\lim_n \int_{\mathbb{T}} \left| f - \frac{A_{n-1}}{B_{n-1}} \right|^2 P(., \alpha_n)dm = 0.$$

In particular, we obtain a result given in [Khrushchev, 2001] for the classical Schur algorithm:

Corollary 10.3.8 *If $1 \leq p < +\infty$, $|f| < 1$ a.e. on $\mathbb{T}$, and $\alpha_k = 0$ for every $k \geq 1$, then*

$$\lim_n \int_{\mathbb{T}} \left| f - \frac{A_n}{B_n} \right|^p dm = 0.$$

Chapter 11

Approximation by a Schur rational function of given degree

The goal of this chapter is to give practical means of approximating a function by a Schur rational function. We first show that the Schur algorithm leads to a parametrization of all *strictly* Schur rational functions of given degree. We next explain how to compute efficiently the L^2 norm of a rational function analytic in the unit disk. We then have all the necessary information to implement an optimization process. Examples are given, and compared with L^2 unconstrained approximation.

11.1 Parametrization of strictly Schur rational functions

Below, we parametrize the strictly Schur rational functions of order n by their convergents of order n (see section 7.3). Let $(c_k)_{k\geq 0}$ be a sequence on $\mathbb{T}$ with $c_0 = 1$. We denote by $\mathcal{S}_n$ the set of all strictly Schur rational functions of degree at most n and we define the application Γ by

$$\begin{array}{rccc} \Gamma : & \mathbb{D}^{2n+1} & \longrightarrow & \mathcal{S}_n \\ & (\alpha_1, \ldots, \alpha_n, \gamma_0, \ldots, \gamma_n) & \longmapsto & R_n \end{array}$$

where

$$R_n = \tau_0 \circ \tau_1 \circ \cdots \circ \tau_{n-1} \circ \tau_n(0)$$

with

$$\tau_k(\omega) = c_k\gamma_k + \frac{(1-|\gamma_k|^2)c_k\zeta_{k+1}}{\bar{\gamma}_k\zeta_{k+1} + \frac{1}{\omega}}.$$

The next theorem shows that Γ is surjective.

For h a polynomial of degree n, we denote by $\tilde{h}$ the polynomial of degree n defined by $\tilde{h}(z) = z^n\overline{h(\frac{1}{\bar{z}})}$.

Theorem 11.1.1 *Every* strictly *Schur irreducible rational function* $\frac{p}{q}$ *of degree* n *can be written as a convergent of order* n.
Furthermore, the only possible interpolation points $\alpha_1, \ldots, \alpha_n$ *(counted with multiplicity) are the points in the set*

$$\mathcal{R} = \{z \in \mathbb{D}, (p\tilde{p} - q\tilde{q})(z) = 0\}.$$

Proof We will show that choosing the interpolation points in $\mathcal{R}$ leads to a constant Schur function f_n. We then conclude applying the reverse Schur algorithm.

1. We first prove that $p\tilde{p} - q\tilde{q}$ has n roots in the unit disk $\mathbb{D}$. Suppose that $p\tilde{p} - q\tilde{q}$ is a polynomial of degree $m < 2n$. Then if we put $p = \sum_{k=0}^{n} a_k z^k$ and $q = \sum_{k=0}^{n} b_k z^k$, we have
$$a_{n-k}\bar{a}_k - b_{n-k}\bar{b}_k = 0 \text{ for all } 0 \leq k < 2n - m$$
and therefore, 0 is a root of $\tilde{p}p - \tilde{q}q$ with multiplicity $2n - m$. Suppose now that some root ξ is on the unit circle $\mathbb{T}$. As $\frac{p}{q}$ is Schur and irreducible, $q(\xi) \neq 0$. Then $\frac{p\tilde{p}}{q\tilde{q}}(\xi) = \left|\frac{p}{q}(\xi)\right|^2 = 1$, and therefore, $\frac{p}{q}$ is not strictly Schur, a contradiction. Furthermore, if $\xi \neq 0$ is a root of $p\tilde{p} - q\tilde{q}$, $\frac{1}{\bar{\xi}}$ is also a root of $p\tilde{p} - q\tilde{q}$. Therefore, there are exactly n points (counted with multiplicity) in $\mathcal{R}$.

2. We now show that the degree of f_i decreases at each step of the Schur algorithm if and only if the α_j are taken in $\mathcal{R}$.

   Recall that
$$f_1 = \bar{c}_0 \frac{p - c_0\gamma_0 q}{q - \overline{c_0\gamma_0}p} \frac{1 - \overline{\alpha_1}z}{z - \alpha_1}.$$
   First, note that $p - c_0\gamma_0 q$ and $q - \overline{c_0\gamma_0}p$ are relatively prime. Indeed, if α is a common root, we have $p(\alpha) = c_0\gamma_0 q(\alpha)$ and $q(\alpha) - |\gamma_0|^2 q(\alpha) = 0$. Therefore, $q(\alpha) = 0$ and $p(\alpha) = 0$. This contradicts the irreducibility of $\frac{p}{q}$.
   Note also that, if $\deg(p - c_0\gamma_0 q) \leq n-1$ and $\deg(q - \overline{c_0\gamma_0}p) \leq n-1$, then $\deg p \leq n-1$ and $\deg q \leq n-1$. Indeed, we get $a_n - c_0\gamma_0 b_n = 0$ and $b_n - \overline{c_0\gamma_0}a_n = 0$, and therefore $a_n(1 - |c_0\gamma_0|^2) = 0$ and $b_n(1 - |c_0\gamma_0|^2) = 0$. Since $|c_0\gamma_0| < 1$, we obtain $a_n = b_n = 0$. This contradicts the hypothesis $\deg p/q = n$.
   Thus, the degree of f_1 is equal to $n - 1$ if and only if

   - $z - \alpha_1$ divides $p - c_0\gamma_0 q$, and
   - $1 - \overline{\alpha_1}z$ divides $q - \overline{c_0\gamma_0}p$ if $\alpha_1 \neq 0$, or else the degree of $q - \overline{c_0\gamma_0}p$ is $\leq n - 1$.

   Note that, in this case, $d^o f - d^o f_1 = 1$.

   Suppose $\alpha_1 \in \mathcal{R}$. Then $(p\tilde{p} - q\tilde{q})(\alpha_1) = 0$. As $\frac{p}{q}$ is irreducible and analytic in $\mathbb{D}$, $q(\alpha_1) \neq 0$. Thus
$$\frac{(q\tilde{q} - p\tilde{p})(\alpha_1)}{q(\alpha_1)} = \tilde{q}(\alpha_1) - c_0\gamma_0\tilde{p}(\alpha_1) = 0. \tag{11.1}$$
   If $\alpha_1 \neq 0$, then
$$\overline{\alpha_1}^n q\left(\frac{1}{\overline{\alpha_1}}\right) - \overline{c_0\gamma_0} \cdot \overline{\alpha_1}^n p\left(\frac{1}{\overline{\alpha_1}}\right) = 0.$$
   We deduce that $1 - \overline{\alpha_1}z$ divides $q - \overline{c_0\gamma_0}p$. If $\alpha_1 = 0$, by (11.1), the degree of $q - \overline{c_0\gamma_0}p$ is strictly less than n. Furthermore, by definition of γ_0, $z - \alpha_1$ divides $p - c_0\gamma_0 q$. Thus, $\deg f_1 = n - 1$.

   Conversely, if $\alpha_1 \neq 0$ with $p(\alpha_1) - c_0\gamma_0 q(\alpha_1) = 0$ and $q(\frac{1}{\overline{\alpha_1}}) - \overline{c_0\gamma_0}p(\frac{1}{\overline{\alpha_1}}) = 0$, then $\tilde{q}(\alpha_1) - c_0\gamma_0\tilde{p}(\alpha_1) = 0$, from which it follows that $\alpha_1 \in \mathcal{R}$. If $\alpha_1 = 0$ and $p(0) = c_0\gamma_0 q(0)$ with $\deg(q - \bar{c}_0\bar{\gamma}_0 p) < n$, then $\tilde{q}(0) - c_0\gamma_0\tilde{p}(0) = 0$ and again $\alpha_1 \in \mathcal{R}$.

3. We finally prove that if $f_1 = \frac{p_1}{q_1}$, then the roots of $p_1\tilde{p_1} - q_1\tilde{q_1}$ that lie in the unit disk are the points of $\mathcal{R} \setminus \{\alpha_1\}$ (counting multiplicity). Since

$$\begin{pmatrix} p_1 & \bar{c}_0\tilde{q_1} \\ q_1 & \bar{c}_0\tilde{p_1} \end{pmatrix} = \begin{pmatrix} z-\alpha_1 & 0 \\ 0 & 1-\overline{\alpha_1}z \end{pmatrix}^{-1} \begin{pmatrix} \bar{c}_0 & -\gamma_0 \\ -\overline{c_0\gamma_0} & 1 \end{pmatrix} \begin{pmatrix} p & \tilde{q} \\ q & \tilde{p} \end{pmatrix},$$

taking determinants, we get

$$p_1\tilde{p_1} - q_1\tilde{q_1} = (1-|\gamma_0|^2)\frac{p\tilde{p} - q\tilde{q}}{(z-\alpha_1)(1-\overline{\alpha_1}z)}.$$

Therefore, the set of the roots of $p_1\tilde{p_1} - q_1\tilde{q_1}$ in $\mathbb{D}$ is $\mathcal{R} \setminus \{\alpha_1\}$.

Iterating this process n times, we get $f_n(z) = \gamma_n$. Conclusion is then immediate.

■

We endow the space of rational functions of degree n with the differential structure which is naturally inherited from the coefficients of the numerators and denominators. Then it becomes a smooth submanifold of every Hardy space H^p, $1 < p < \infty$, of the disk of dimension $2n+1$ over $\mathbb{C}$ ([Alpay et al., 1994]).

Theorem 11.1.2 *If $a = (\alpha_1, \ldots, \alpha_n, \gamma_0, \ldots, \gamma_n)$ is such that the points $\alpha_1, \ldots, \alpha_n$ are all distinct and $d^\circ\Gamma(a) = n$, then the derivative $d\Gamma(a)$ at $a \in \mathbb{D}^{2n+1}$ is an isomorphism.*

Proof We give a proof by induction. The result is immediate if $n = 0$. We denote by Γ_i:

$$\Gamma_i(\alpha_{i+1}, \ldots, \alpha_n, \gamma_i, \ldots, \gamma_n) = \tau_i \circ \cdots \circ \tau_n(0).$$

We therefore have

$$\begin{aligned} \Gamma(\alpha_1, \ldots, \alpha_n, \gamma_0, \ldots, \gamma_n) &= \tau_0 \circ \Gamma_1(\alpha_2, \ldots, \alpha_n, \gamma_1, \ldots, \gamma_n) \\ &= \frac{\zeta_1\Gamma_1(\alpha_2, \ldots, \alpha_n, \gamma_1, \ldots, \gamma_n) + \gamma_0}{1 + \overline{\gamma_0}\zeta_1\Gamma_1(\alpha_2, \ldots, \alpha_n, \gamma_1, \ldots, \gamma_n)}. \end{aligned}$$

Note that, in the following, we will just write Γ_1 for $\Gamma_1(\alpha_2, \ldots, \alpha_n, \gamma_1, \ldots, \gamma_n)$. On differentiating if the space of rational functions of degree n is viewed as a submanifold of H^p, $1 < p < \infty$, we have

$$\begin{aligned} \frac{\partial\Gamma}{\partial\gamma_0} &= \frac{1}{1 + \overline{\gamma_0}\zeta_1(z)\Gamma_1(z)} \\ \frac{\partial\Gamma}{\partial\overline{\gamma_0}} &= -\frac{\zeta_1(z)\Gamma_1(z)(\zeta_1(z)\Gamma_1(z) + \gamma_0)}{(1 + \overline{\gamma_0}\zeta_1(z)\Gamma_1(z))^2} \\ \frac{\partial\Gamma}{\partial\alpha_1} &= -\frac{\Gamma_1(z)}{(1 + \overline{\gamma_0}\zeta_1(z)\Gamma_1(z))^2}\frac{1-|\gamma_0|^2}{1-\overline{\alpha_1}z} \\ \frac{\partial\Gamma}{\partial\overline{\alpha_1}} &= \frac{\zeta_1(z)\Gamma_1(z)}{(1 + \overline{\gamma_0}\zeta_1(z)\Gamma_1(z))^2}\frac{(1-|\gamma_0|^2)z}{1-\overline{\alpha_1}z} \end{aligned}$$

and for $k \geq 1$,

$$\begin{aligned}
\frac{\partial \Gamma}{\partial \gamma_k} &= \frac{\zeta_1(z)(1-|\gamma_0|^2)}{(1+\overline{\gamma_0}\zeta_1(z)\Gamma_1(z))^2}\frac{\partial \Gamma_1}{\partial \gamma_k} \\
\frac{\partial \Gamma}{\partial \overline{\gamma_k}} &= \frac{\zeta_1(z)(1-|\gamma_0|^2)}{(1+\overline{\gamma_0}\zeta_1(z)\Gamma_1(z))^2}\frac{\partial \Gamma_1}{\partial \overline{\gamma_k}} \\
\frac{\partial \Gamma}{\partial \alpha_{k+1}} &= \frac{\zeta_1(z)(1-|\gamma_0|^2)}{(1+\overline{\gamma_0}\zeta_1(z)\Gamma_1(z))^2}\frac{\partial \Gamma_1}{\partial \alpha_{k+1}} \\
\frac{\partial \Gamma}{\partial \overline{\alpha_{k+1}}} &= \frac{\zeta_1(z)(1-|\gamma_0|^2)}{(1+\overline{\gamma_0}\zeta_1(z)\Gamma_1(z))^2}\frac{\partial \Gamma_1}{\partial \overline{\alpha_{k+1}}}
\end{aligned}$$

Suppose that the hypothesis is true for $n-1$, that is if $\alpha_2, \ldots, \alpha_n$ are all distinct and $d^\circ\Gamma_1(\hat{a}) = n-1$ then $d\Gamma_1(\hat{a})$ is an isomorphism, with $\hat{a} = (\alpha_2, \ldots, \alpha_n, \gamma_1, \ldots, \gamma_n)$.

Suppose there exists a linear combination such that:

$$\sum_{l=0}^{n-1}\left(\frac{\partial \Gamma}{\partial \gamma_l}d\gamma_l + \frac{\partial \Gamma}{\partial \overline{\gamma_l}}d\overline{\gamma_l} + \frac{\partial \Gamma}{\partial \alpha_{l+1}}d\alpha_{l+1} + \frac{\partial \Gamma}{\partial \overline{\alpha_{l+1}}}d\overline{\alpha_{l+1}}\right) + \frac{\partial \Gamma}{\partial \gamma_n}d\gamma_n + \frac{\partial \Gamma}{\partial \overline{\gamma_n}}d\overline{\gamma_n} = 0$$

Then we have for every z, on multiplying by $(1+\overline{\gamma_0}\zeta_1(z)\Gamma_1(z))^2$,

$$\begin{aligned}
0 = {} & \zeta_1(z)(1-|\gamma_0|^2)\sum_{l=1}^{n-1}\left(\frac{\partial \Gamma_1}{\partial \gamma_l}d\gamma_l + \frac{\partial \Gamma_1}{\partial \overline{\gamma_l}}d\overline{\gamma_l} + \frac{\partial \Gamma_1}{\partial \alpha_{l+1}}d\alpha_{l+1} + \frac{\partial \Gamma_1}{\partial \overline{\alpha_{l+1}}}d\overline{\alpha_{l+1}}\right) \\
& + \zeta_1(z)(1-|\gamma_0|^2)\left(\frac{\partial \Gamma_1}{\partial \gamma_n}d\gamma_n + \frac{\partial \Gamma_1}{\partial \overline{\gamma_n}}d\overline{\gamma_n}\right) \\
& + \zeta_1(z)\Gamma_1(z)\left(\overline{\gamma_0}d\gamma_0 - (\zeta_1(z)\Gamma_1(z)+\gamma_0)d\overline{\gamma_0} + \frac{(1-|\gamma_0|^2)z}{1-\overline{\alpha_1}z}d\overline{\alpha_1}\right) \\
& + d\gamma_0 - \frac{(1-|\gamma_0|^2)\Gamma_1(z)}{1-\overline{\alpha_1}z}d\alpha_1.
\end{aligned} \tag{11.2}$$

Evaluating at α_1, we get

$$d\gamma_0 = \frac{\Gamma_1(\alpha_1)(1-|\gamma_0|^2)}{1-|\alpha_1|^2}d\alpha_1 \tag{11.3}$$

Therefore, the last row in (11.2) can be expressed as :

$$(1-|\gamma_0|^2)\left(\frac{\Gamma_1(\alpha_1)}{1-|\alpha_1|^2} - \frac{\Gamma_1(z)}{1-\overline{\alpha_1}z}\right)d\alpha_1$$

This can be written as

$$(|\gamma_0|^2-1)\zeta_1(z)\left(g_1(z) + \overline{\alpha_1}\frac{\Gamma_1(\alpha_1)}{1-|\alpha_1|^2}\right)d\alpha_1$$

with

$$g_1(z) = \frac{\Gamma_1(z)-\Gamma_1(\alpha_1)}{z-\alpha_1}.$$

A cancellation by ζ_1 in (11.2) gives us:

$$\begin{aligned}0 = (1-|\gamma_0|^2)\sum_{l=1}^{n-1}&\left(\frac{\partial\Gamma_1}{\partial\gamma_l}d\gamma_l + \frac{\partial\Gamma_1}{\partial\overline{\gamma_l}}d\overline{\gamma_l} + \frac{\partial\Gamma_1}{\partial\alpha_{l+1}}d\alpha_{l+1} + \frac{\partial\Gamma_1}{\partial\overline{\alpha_{l+1}}}d\overline{\alpha_{l+1}}\right)\\ &+ (1-|\gamma_0|^2)\left(\frac{\partial\Gamma_1}{\partial\gamma_n}d\gamma_n + \frac{\partial\Gamma_1}{\partial\overline{\gamma_n}}d\overline{\gamma_n}\right)\\ &+ \Gamma_1(z)\left(\overline{\gamma_0}d\gamma_0 - (\zeta_1(z)\Gamma_1(z)+\gamma_0)d\overline{\gamma_0} + \frac{(1-|\gamma_0|^2)z}{1-\overline{\alpha_1}z}d\overline{\alpha_1}\right)\\ &+ (|\gamma_0|^2-1)\left(g_1(z) + \overline{\alpha_1}\frac{\Gamma_1(\alpha_1)}{1-|\alpha_1|^2}\right)d\alpha_1.\end{aligned} \tag{11.4}$$

Γ_1 is a rational irreducible function $\frac{p_1}{q_1}$ of degree $n-1$ by Theorem 11.1.1. Thus $\frac{\partial\Gamma_1}{\partial\square} \in \frac{\mathcal{P}_{2n-2}}{q_1^2}$ where $\square$ denotes any of the variable α_j, γ_j, $\bar{\alpha}_j$ or $\bar{\gamma}_j$. In fact, in the previous expression, all terms are in $\frac{\mathcal{P}_{2n-2}}{q_1^2}$, except perhaps

$$-\zeta_1(z)\Gamma_1^2(z)d\overline{\gamma_0}$$

and

$$(1-|\gamma_0|^2)\frac{z\Gamma_1(z)}{1-\overline{\alpha_1}z}d\overline{\alpha_1}.$$

Using (11.3) and (11.4), we get

$$\left(-\zeta_1(z)\Gamma_1^2(z)\frac{\overline{\Gamma_1(\alpha_1)}}{1-|\alpha_1|^2} + \frac{z\Gamma_1(z)}{1-\overline{\alpha_1}z}\right)d\overline{\alpha_1} \in \frac{\mathcal{P}_{2n-2}}{q_1(z)^2}. \tag{11.5}$$

Note that

$$-\zeta_1(z)\Gamma_1^2(z)\frac{\overline{\Gamma_1(\alpha_1)}}{1-|\alpha_1|^2} + \frac{z\Gamma_1(z)}{1-\overline{\alpha_1}z} = p_1(z)\frac{(1-|\alpha_1|^2)zq_1(z) - \overline{\Gamma_1(\alpha_1)}(z-\alpha_1)p_1(z)}{(1-|\alpha_1|^2)(1-\bar{\alpha}_1 z)q_1(z)^2}. \tag{11.6}$$

Suppose that $d\overline{\alpha_1} \neq 0$.
Then, if $\alpha_1 \neq 0$, combining (11.5) and (11.6), we get

$$p_1(1/\overline{\alpha_1})\left(q_1(1/\overline{\alpha_1}) - \overline{\Gamma_1(\alpha_1)}p_1(1/\overline{\alpha_1})\right) = 0.$$

If $p_1(1/\overline{\alpha_1}) = 0$, then

$$\frac{p(z)}{q(z)} = \frac{(z-\alpha_1)p_1(z) + c_0\gamma_0 q_1(z)(1-\overline{\alpha_1}z)}{q_1(z)(1-\overline{\alpha_1}z) + \overline{c_0\gamma_0}(z-\alpha_1)p_1(z)}$$

has the same degree than $\frac{p_1}{q_1}$ (because $1-\overline{\alpha_1}z$ is a common factor).
If $q_1(1/\overline{\alpha_1}) - \overline{\Gamma_1(\alpha_1)}p_1(1/\overline{\alpha_1}) = 0$, then $(p_1\tilde{p_1} - q_1\tilde{q_1})(\alpha_1) = 0$ and α_1 is a multiple root.
Furthermore, if $\alpha_1 = 0$, we have $zp_1(z)\left(q_1(z) - \overline{\Gamma_1(0)}p_1(z)\right) \in \mathcal{P}_{2n-2}$ if and only if $\deg(zp_1(z)) \leq n-1$ or $\deg(q_1(z) - \overline{\Gamma_1(0)}p_1(z)) \leq n-2$, which is equivalent to $\deg(zp_1(z)) \leq n-1$ or $(p\tilde{p} - q\tilde{q})(0) = 0$.

From what precedes, we deduce that if $\deg p/q = n$ and α_1 is not a multiple root, then the derivative $d\Gamma(a)$ is injective (and therefore surjective counting dimensions).

■

11.2 Computation of the L^2 norm

In order to be able to optimize with respect to the L^2 norm, we will now see how to numerically compute efficiently the Hermitian product $\langle f, g\rangle = \int_{\mathbb{T}} f(t)\overline{g(t)}dm(t)$ for f, g rational functions analytic inside the unit disk. Two kind of methods are presented : the first one uses elementary operations on polynomials, and the other one uses matrix operations.

11.2.1 Two methods using elementary operations on polynomials

The two methods proposed brings the computation of the Hermitian product of two rational functions back to the computation of the Hermitian product of two polynomials. Therefore, they essentially use the elementary property :

Property 11.2.1 *If $p = \sum_{k=0}^{k=m} p_k z^k$ and $q = \sum_{k=0}^{k=n} q_k z^k$ are two polynomials then*

$$\langle p, q\rangle = \sum_{k=0}^{\min(m,n)} p_k \overline{q_k}.$$

The first method is very basic and gives an approximation of the Hermitian product. However, it is quite efficient for Schur rational functions of small degree. It simply consists in approximating f and g by their Taylor polynomials of order N, the Hermitian product is then obtained using the previous property. If N is sufficiently big, the result is very good (for the examples presented in the next section, two hundred Taylor coefficients were taken). The Taylor coefficients are easily obtained using the "long" division with respect to increasing powers.

The second method has the advantage of avoiding any truncation. However, it requires to efficiently compute an extended gcd. For a neater notation, the following computation is done for $\frac{a}{b}$ and $\frac{r}{q}$ rational functions *analytic outside the unit disk*, i.e. the roots of b and q are in the unit disk. This is equivalent to the corresponding problem in the disk upon changing z into $1/z$. Here, for a polynomial q, we denote by $\tilde{q}$ the polynomial $\tilde{q} = z^{d^o q}\overline{q\left(\frac{1}{\bar{z}}\right)}$. As $gcd(b, \tilde{q}) = 1$, there exist u and v such that $ub + v\tilde{q} = 1$. Then, if $r = r_1 q + r_0$ with $d^o r_0 < d^o r$,

$$\begin{aligned}
\left\langle \frac{a}{b}, \frac{q}{r}\right\rangle &= \left\langle \frac{a(ub + v\tilde{q})}{b}, \frac{r}{q}\right\rangle \\
&= \left\langle au, \frac{r}{q}\right\rangle + \left\langle \frac{av\tilde{q}}{b}, \frac{r}{q}\right\rangle \\
&= \langle au, r_1\rangle + \left\langle \frac{av\tilde{q}}{b}, \frac{r}{q}\right\rangle
\end{aligned}$$

where we have taken into account the orthogonality of $H^2(\mathbb{D})$ and $H^2(\mathbb{C} \setminus \overline{\mathbb{D}})$. As $\tilde{q} = z^{d^o q}\overline{q\left(\frac{1}{\bar{z}}\right)}$, we have

$$\left\langle \frac{av\tilde{q}}{b}, \frac{r}{q}\right\rangle = \left\langle \frac{avz^{d^o q}}{b}, r\right\rangle.$$

The euclidean division of $avz^{d^o q}$ by b gives

$$avz^{d^o q} = k_1 b + \rho.$$

Therefore,

$$\left\langle \frac{a}{b}, \frac{q}{r} \right\rangle = \langle au, r_1 \rangle + \langle k_1, r \rangle .$$

Note that the Hermitian product of two rational functions $f = \frac{a_0}{b_0}$ and $g = \frac{r_0}{q_0}$ analytic inside the unit disk is

$$\begin{aligned} \langle f, g \rangle &= \left\langle \frac{a_0}{b_0}, \frac{r_0}{q_0} \right\rangle \\ &= \left\langle \frac{z^{d^o q_0}}{z^{d^o r_0}} \frac{\tilde{r_0}}{\tilde{q_0}}, \frac{z^{d^o b_0}}{z^{d^o a_0}} \frac{\tilde{a_0}}{\tilde{b_0}} \right\rangle \end{aligned}$$

and is therefore obtained as a Hermitian product of two rational function analytic outside the disk.

11.2.2 A method using matrix representations

We now present a method which adopts the matrix point of view. The computation is carried out using a realization of f and g, i.e. by expressing these functions with matrices. More details about realizations and system theory can be found in [Kailath, 1980].

Definition 11.2.2 *A rational function is proper (resp. strictly proper) if the numerator's degree is less or equal (resp. strictly less) than the denominator's degree.*
A matrix is proper rational (resp. strictly proper rational) if its entries are rational proper (resp. strictly proper) functions.

In fact, we will study here how to compute the L^2 norm of proper rational matrices. For this, we first want to express strictly proper rational matrices using 3 complex matrices A, B, C.

Let $H(s)$ be a strictly proper rational matrix $m \times p$ and let $d(s) = s^r + d_1 s^{r-1} + ... + d_r$ be the least common denominator of the entries of $H(s)$. Then $H(s) = \frac{N(s)}{d(s)}$, where $N(s)$ is a matrix $m \times p$ with polynomial entries. As H is strictly proper, there exist complex matrices $m \times p$ $N_1, N_2, ..., N_r$ such that $N(s) = N_1 s^{r-1} + N_2 s^{r-2} + ... + N_r$.
We denote by I_p the $p \times p$ identity matrix.
We define the matrices $A : pr \times pr$, $B : pr \times p$, $C : m \times pr$ by :

$$\begin{cases} A = \begin{bmatrix} -d_1 I_p & -d_2 I_p & \cdots & -d_r I_p \\ I_p & 0 & \cdots & 0 \\ & \ddots & \ddots & \\ (0) & & I_p & 0 \end{bmatrix}, \\ B = \begin{bmatrix} I_p \\ 0 \\ \vdots \\ 0 \end{bmatrix}, \\ C = \begin{bmatrix} N_1 & N_2 & \cdots & N_r \end{bmatrix}. \end{cases}$$

Then:

$$
\begin{aligned}
&(sI-A)\begin{bmatrix} s^{r-1}I_p & s^{r-2}I_p & \cdots & I_p \end{bmatrix}^t \\
&= \begin{bmatrix} (s+d_1)I_p & d_2 I_p & \cdots & d_r I_p \\ -I_p & sI_p & & (0) \\ & \ddots & \ddots & \\ (0) & & -I_p & sI_p \end{bmatrix} \begin{bmatrix} s^{r-1}I_p \\ s^{r-2}I_p \\ \vdots \\ I_p \end{bmatrix} \\
&= \begin{bmatrix} (s^r + d_1 s^{r-1} + d_2 s^{r-2} + ... + d_r)I_p \\ (-s^{r-1} + s^{r-1})I_p \\ \vdots \\ -sI + sI \end{bmatrix} \\
&= d(s) \begin{bmatrix} I_p \\ 0 \\ \vdots \\ 0 \end{bmatrix}.
\end{aligned}
$$

We deduce that $(sI-A)^{-1} = \frac{1}{d(s)} \begin{bmatrix} s^{r-1}I_p & * & \cdots & * \\ s^{r-2}I_p & \vdots & & \vdots \\ \vdots & \vdots & & \vdots \\ I_p & * & \cdots & * \end{bmatrix}$. Therefore,

$$
\begin{aligned}
&C(sI-A)^{-1}B \\
&= \frac{1}{d(s)} \begin{bmatrix} N_1 & N_2 & \cdots & N_r \end{bmatrix} \begin{bmatrix} s^{r-1}I_p & s^{r-2}I_p & \cdots & I_p \end{bmatrix}^t \\
&= \frac{N_1 s^{r-1} + N_2 s^{r-2} + ... + N_r}{d(s)} \\
&= \frac{N(s)}{d(s)} \\
&= H(s).
\end{aligned}
$$

Definition 11.2.3 *Let $H(s)$ be a proper rational matrix. We call realization of H any 4-tuple (A,B,C,D) of complex matrices such that $H(s) = C(sI-A)^{-1}B + D$.*

From what precedes, a realization of a strictly proper rational matrix always exists. Let now H be proper rational and let $D = \lim_{s\to\infty} H$. Then $H - D$ is strictly proper, so there exists (A,B,C) such that $H-D = C(sI-A)^{-1}B$. Therefore, $H = C(sI-A)^{-1}B+D$. Thus, we have obtained a realization for a proper rational matrix. Note that a proper rational matrix does not have a unique realization.

A realization is called a *minimal realization of H* if the size of A is minimal among all the possible realizations of H.

We now briefly explain how to compute the L^2 norm using a minimal realization.

We now suppose that (A,B,C,D) is a minimal realization of a proper rational matrix H whose entries are analytic outside the unit disk and up to the unit circle. It is well-known

that the eigenvalues of A are the poles of H ([Kailath, 1980], [Gohberg et al., 2006]). By analyticity of H, the eigenvalues of A are therefore inside the unit disk. We have

$$(sI - A)^{-1} = s^{-1}\left(I - \frac{A}{s}\right)^{-1} = s^{-1}\sum_{j=0}^{\infty}\left(\frac{A}{s}\right)^{j} = \sum_{j=0}^{\infty} A^j s^{-(j+1)}.$$

Therefore, $H(s) = D + \sum_{j=0}^{\infty} CA^jBs^{-(j+1)}$. Let H_1 and H_2 be two strictly proper rational matrices whose entries are analytical outside the unit disk. From what precedes, we have

$$\begin{cases} H_1(s) = D_1 + \sum_{j=0}^{\infty} C_1A_1^jB_1s^{-(j+1)}, \text{ and} \\ H_2(s) = D_2 + \sum_{j=0}^{\infty} C_2A_2^jB_2s^{-(j+1)}. \end{cases}$$

Thus

$$\begin{aligned} \langle H_1, H_2\rangle &= Tr\left(D_1D_2^* + \sum_{j=0}^{\infty} C_1A_1^jB_1B_2^*(A_2^*)^jC_2^*\right) \\ &= Tr\left(D_1D_2^* + C_1\left(\sum_{j=0}^{\infty} A_1^jB_1B_2^*(A_2^*)^j\right)C_2^*\right). \end{aligned}$$

We denote by P the matrix $P = \sum_{j=0}^{\infty} A_1^jB_1B_2^*(A_2^*)^j$, which is well-defined since A_1 and A_2 have all their eigenvalues in $\mathbb{D}$. It is immediate that P is a solution of the Stein (or Lyapounov) equation: $A_1PA_2^* + B_1B_2^* = P$. Since all the eigenvalues of A_1 and A_2 are in $\mathbb{D}$, no eigenvalue of A_1 is the reciprocal of an eigenvalue of A_2. Therefore, the Stein problem has a unique solution. Since $\langle H_1, H_2\rangle = Tr(D_1D_2^* + C_1PC_2^*)$, solving the Stein problem gives the value of $\langle H_1, H_2\rangle$.
More details about the matrix P and the Stein problem can be found in [Ball et al., 1990].

11.3 Examples

In order to approximate a function f, we have implemented an optimization process using the parametrization presented in section 11.1. The criterion which is minimized is the relative L^2 error

$$e(\alpha_1, \ldots, \alpha_n, \gamma_0, \ldots, \gamma_n) = \frac{\|f - \Gamma(\alpha_1, \ldots, \alpha_n, \gamma_0, \ldots, \gamma_n)\|_2}{\|f\|_2}.$$

In practice, the points of the unit disk $\alpha_1, \ldots, \alpha_n, \gamma_0, \ldots, \gamma_n$ are parametrized by the application

$$\Lambda : \begin{array}{l} \mathbb{R}^2 \longrightarrow \mathbb{D} \\ (x, y) \mapsto \frac{x}{\sqrt{x^2+y^2+1}} + i\frac{y}{\sqrt{x^2+y^2+1}} \end{array}.$$

This allows to do an unconstrained optimization : to compute a Schur rational function of degree n, we would like to optimize

$$\inf_{(x_{\alpha_1}, y_{\alpha_1}, \ldots, x_{\gamma_n}, y_{\gamma_n}) \in \mathbb{R}^{4n+2}} \|f - \Gamma(\Lambda(x_{\alpha_1}, y_{\alpha_1}), \ldots, \Lambda(x_{\gamma_n}, y_{\gamma_n}))\|_2.$$

		Degree 7	Degree 8	Degree 9
L^2 (hyperion)	$\|.\|_\infty$	1.0235	1.0056	1.0014
	error	6.72 e-2	1.16 e-2	1.32 e-3
Schur	error	6.89 e-2	1.19 e-2	1.51 e-3
L^2 normalized	error	7.09 e-2	1.29 e-2	1.99 e-3

Table 11.1: Approximation of the Schur function $p30$: comparison between our Schur process and hyperion

This problem depends of $4n+2$ real parameters. Note that, as the parametrization Γ is not defined for parameters of modulus 1, the infimum is not necessarily attained.

In the following examples, the initialization of the optimization is done using the asymptotic-BMO-type criterion (see section 10.2.2), that is by computing a sequence of points (α_n) such that α_{n+1} minimizes

$$\mathcal{I}_n(\alpha) = \int_{\mathbb{T}} \left| \frac{f_n(t) - f_n(\alpha)}{1 - \overline{f_n(\alpha)} f_n(t)} \right|^2 P(t,\alpha) dm(t).$$

No refined attempts at solving this optimization problem were made: we simply used a grid search.

The results obtained by this "Schur optimization" are compared with the L^2 unconstrained approximation given by the hyperion software[1] ([Grimm, 2000]). In particular, we check that the error of our result s lies between the L^2 error of the result h given by hyperion and the "normalized L^2 error" (i.e. the error of the arl2 function of the hyperion software scaled into the unit disk in order to obtain a Schur function), that is we check that $e(h) \leq e(s) \leq e\left(\frac{h}{\|h\|_\infty}\right)$.

In the following figures, when a function g is plotted, the left graph represents the image by g of the unit circle, and the right graph is the modulus of this image, i.e. we plot:

$$\text{On the left: } t \mapsto g(e^{it}) \quad \text{and on the right: } t \mapsto |g(e^{it})| \qquad \text{for } -\pi \leq t \leq \pi.$$

11.3.1 Approximation of Schur functions

Example 1

We are now interested in approximating a polynomial $p30$ of degree 30 plotted in Fig. 11.1. Note that $p30$ is Schur and $\|p30\|_2 = 0.7852$.

The results given by our optimization process and by hyperion for degrees 7 to 9 are presented in Tab. 11.1. None of the best L^2-unconstrained approximations is Schur.

[1]The hyperion software essential feature is to find a rational approximation of McMillan degree n of a stable transfer function given by incomplete frequency measures. Its development has been abandoned in 2001. The Endymion software, which is still under development, will offer most of the functionalities of hyperion. Note that the author of the hyperion software chose to write "hyperion" in lowercase letters.

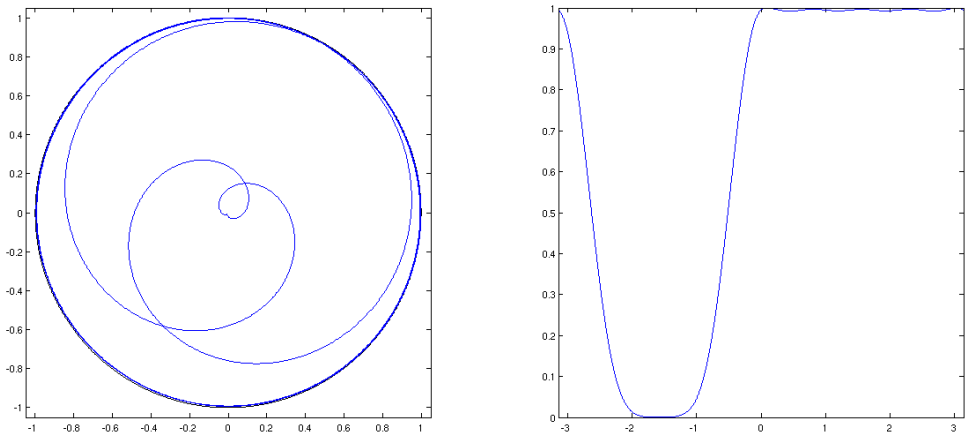

Figure 11.1: Function $p30$, polynomial of degree 30, Schur.

		Degree 7	Degree 8	Degree 9
L^2	$\|.\|_\infty$	1.0053	1.0037	1.0014
(hyperion)	error	2.97 e-2	1.69 e-2	4.5 e-3
Schur	error	3.01 e-2	1.70 e-2	4.7 e-3
L^2 normalized	error	3.02 e-2	1.73 e-2	4.8 e-3

Table 11.2: Approximation of the Schur function $p60$: comparison between our Schur process and hyperion

Fig. 11.2 is a good example of what happens when one approximates a Schur function whose modulus is near 1 on an interval of the unit circle: the L^2 unconstrained approximation oscillates (in modulus) around one. Here, where the approximation computed by hyperion exceeds 1 (in modulus), the Schur approximation "hits" one.

On this example, the initialization points are not very good (see fig. 11.3, 11.5 and 11.7).

Example 2

We are now interested in approximating a polynomial $p60$ of degree 60 plotted in fig. 11.8. Note that $p60$ is Schur and $\|p60\|_2 = 0.9304$.

The approximations of degree 7 to 9 obtained using our Schur process and hyperion are compared in Tab. 11.2. Note that none of the best L^2-unconstrained approximations is Schur.

For the initialization, we first computed points $\alpha_1, \ldots, \alpha_{10}$ using the asymptotic-BMO-type criterion and chose among them. The initial interpolation points at degree 7 are the points $\alpha_2, \ldots, \alpha_8$, at degree 8 they are $\alpha_1, \ldots, \alpha_8$, and at degree 9 they are $\alpha_2, \ldots, \alpha_{10}$. The initializations for the degrees 7 and 8 are quite good (see fig. 11.10 and fig. 11.12).

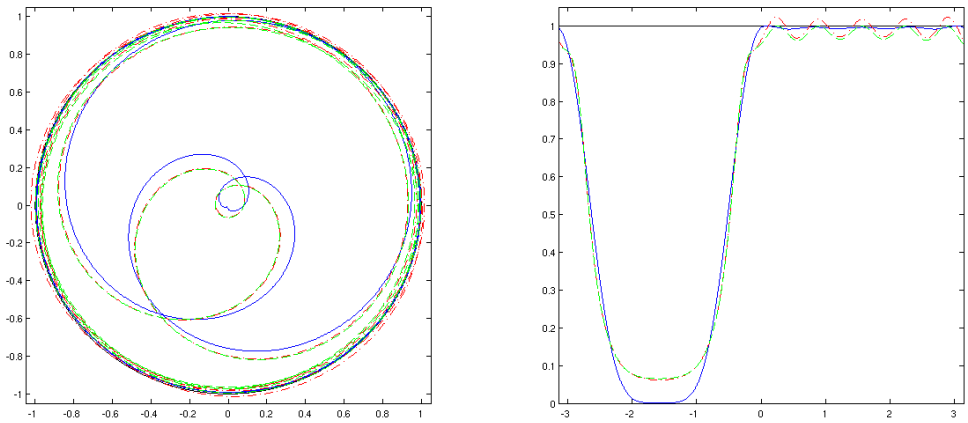

Figure 11.2: Function $p30$ (blue), Schur approximation (green) and L^2 approximation (red) of degree 7.

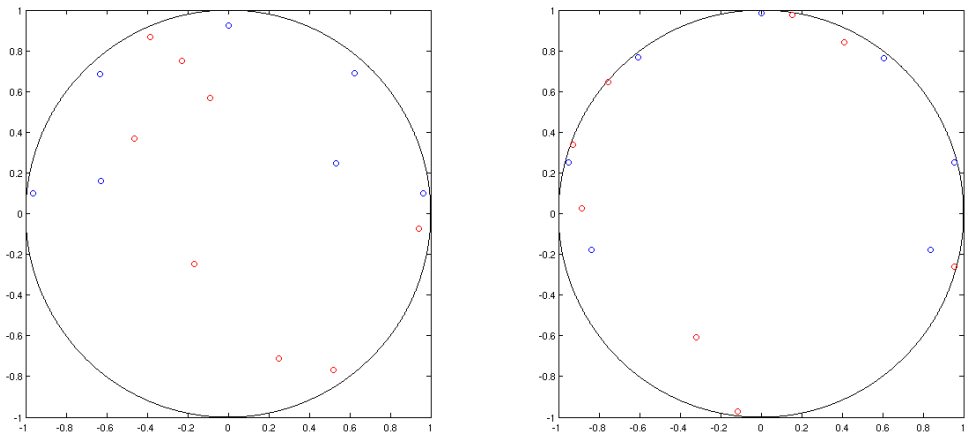

Figure 11.3: Initialization points (left) and optimized points (right) of the Schur function of degree 7 : parameters α (blue) and γ (red).

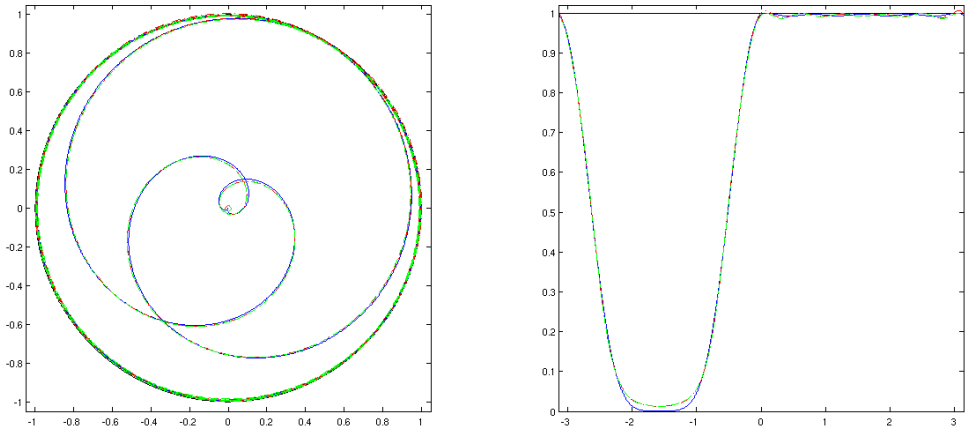

Figure 11.4: Function $p30$ (blue), Schur approximation (green) and L^2 approximation (red) of degree 8.

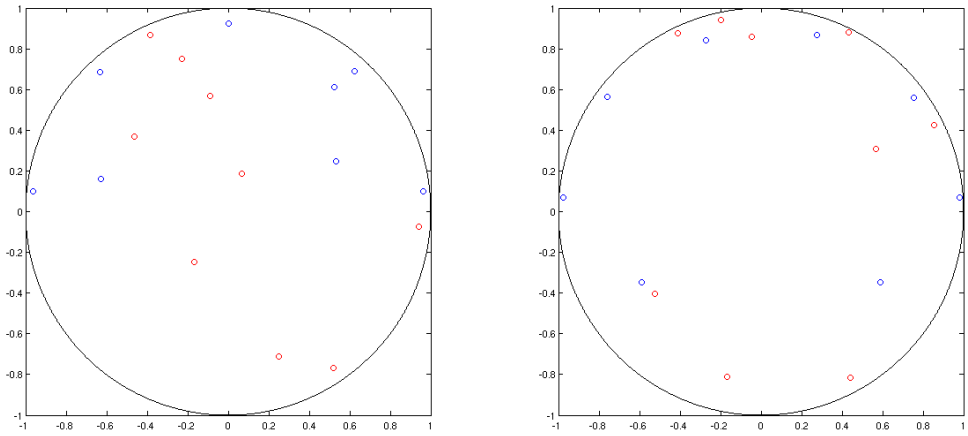

Figure 11.5: Initialization points (left) and optimized points (right) of the Schur function of degree 8 : parameters α (blue) and γ (red).

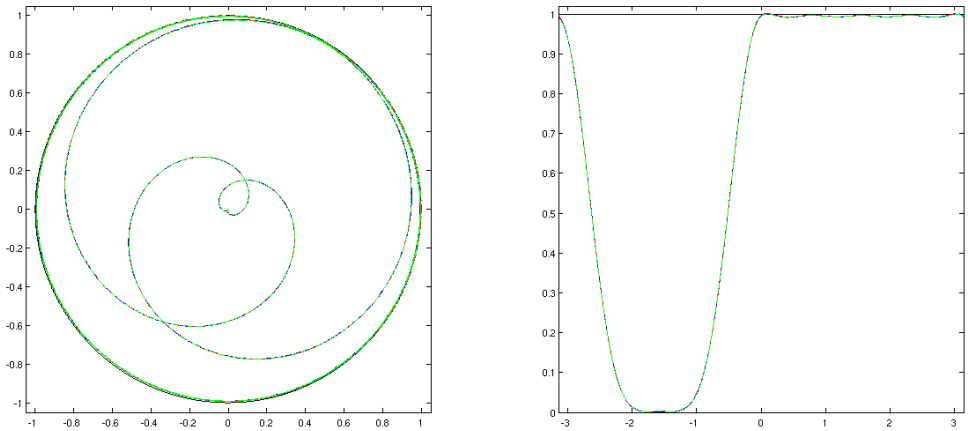

Figure 11.6: Function $p30$ (blue), Schur approximation (green) and L^2 approximation (red) of degree 9.

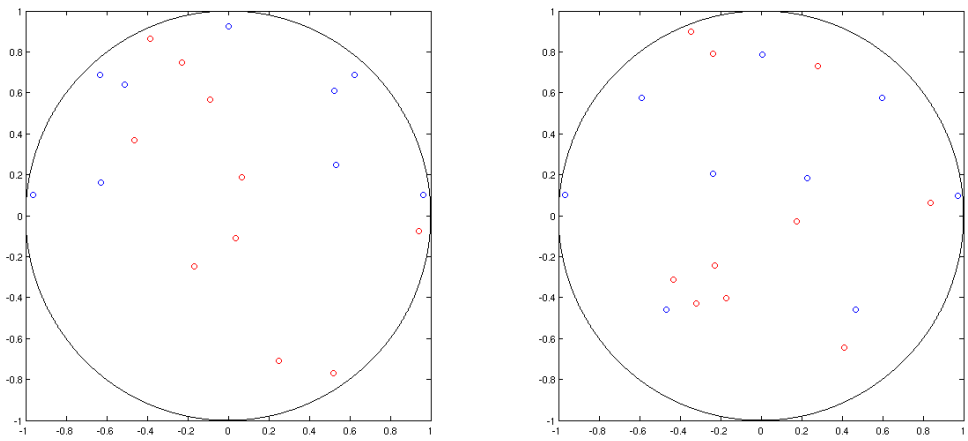

Figure 11.7: Initialization points (left) and optimized points (right) of the Schur function of degree 9 : parameters α (blue) and γ (red).

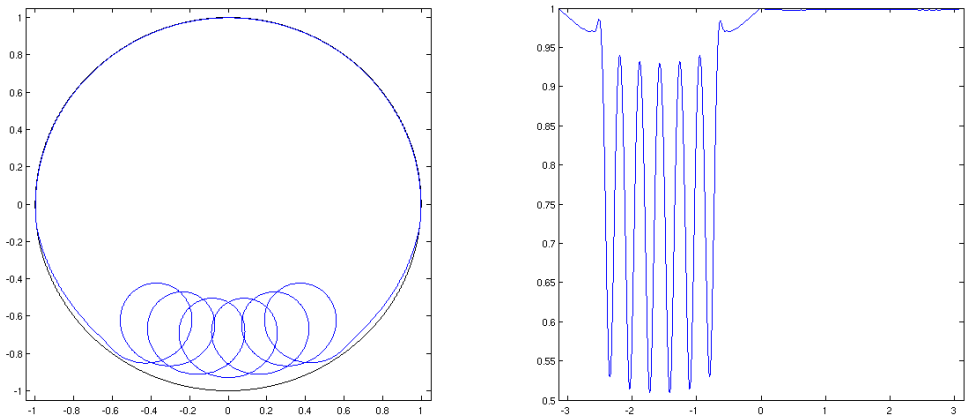

Figure 11.8: Function f, polynomial of degree 60.

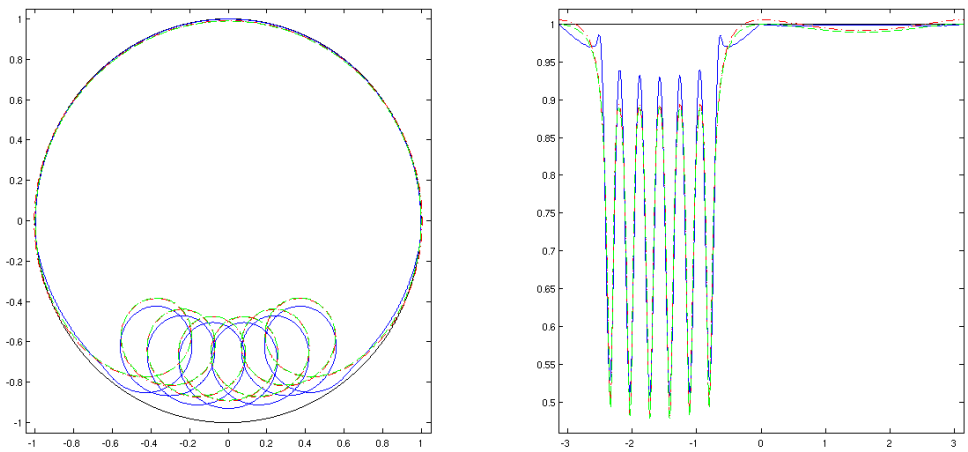

Figure 11.9: Function $p60$ (blue), Schur approximation (green) and L^2 approximation (red) of degree 7.

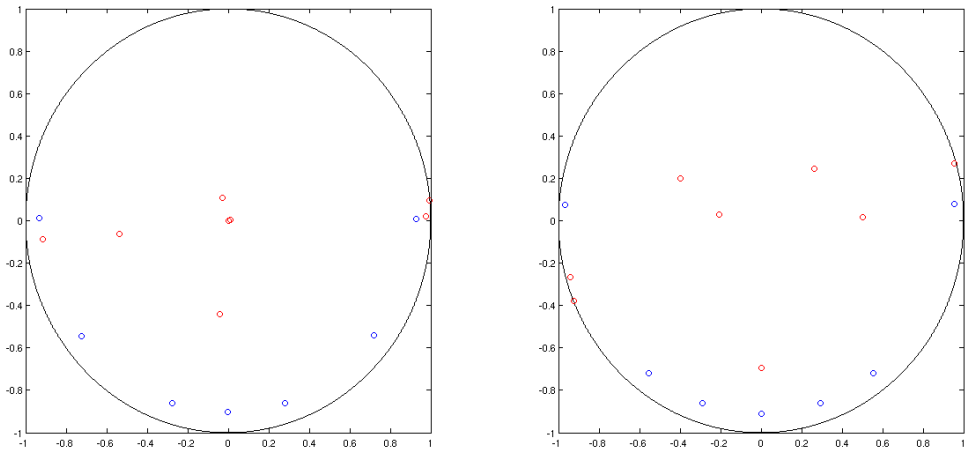

Figure 11.10: Initialization points (left) and optimized points (right) of the Schur function of degree 7 : parameters α (blue) and γ (red).

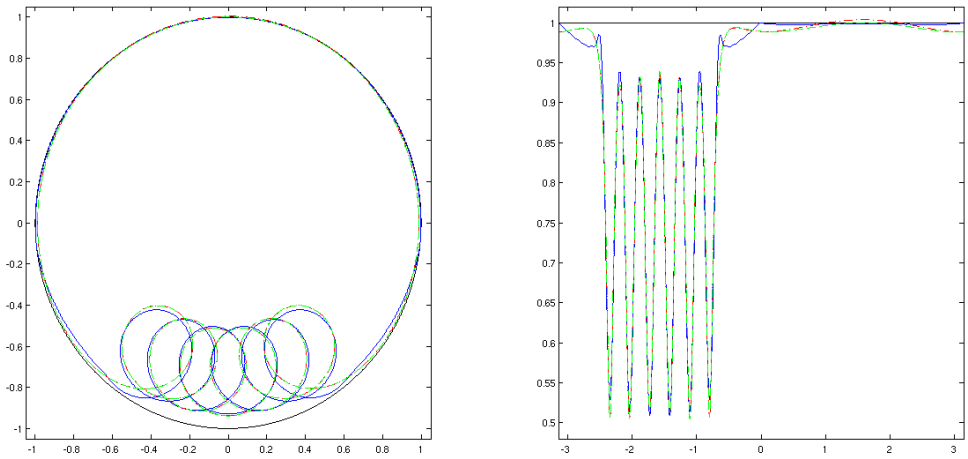

Figure 11.11: Function $p60$ (blue), Schur approximation (green) and L^2 approximation (red) of degree 8.

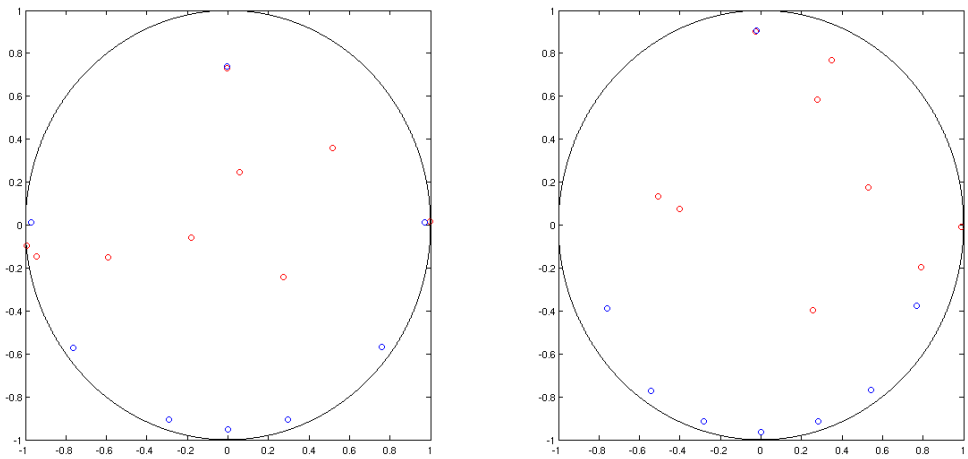

Figure 11.12: Initialization points (left) and optimized points (right) of the Schur function of degree 8 : parameters α (blue) and γ (red).

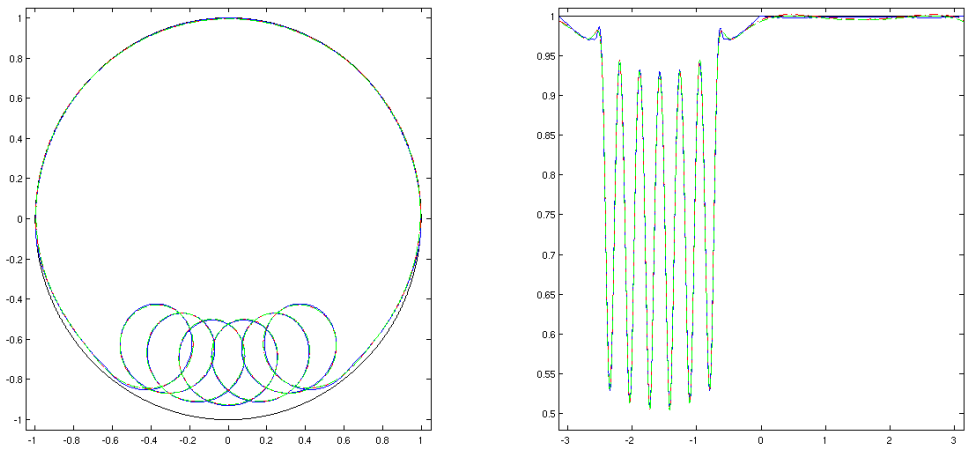

Figure 11.13: Function $p60$ (blue), Schur approximation (green) and L^2 approximation (red) of degree 9.

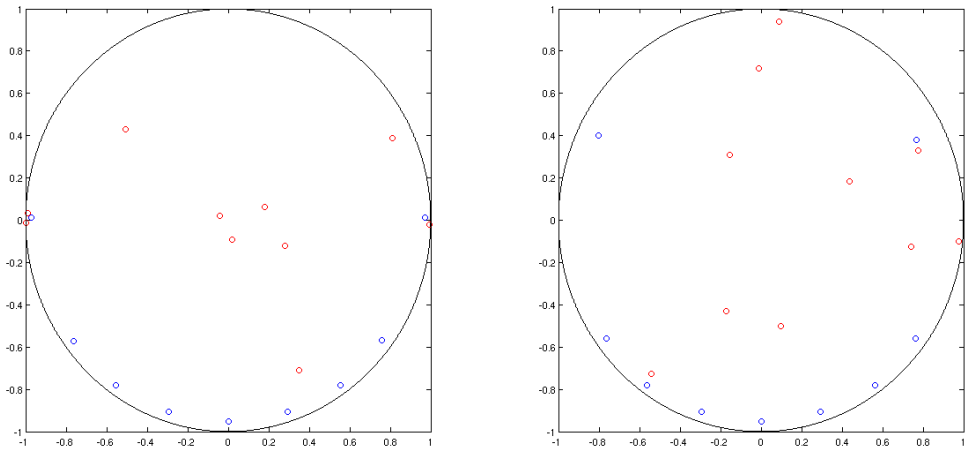

Figure 11.14: Initialization points (left) and optimized points (right) of the Schur function of degree 9 : parameters α (blue) and γ (red).

11.3.2 Approximation of analytic but not Schur functions

In the two following examples, we are interested in approximating analytic, but not Schur, functions. In practice, standard applications arise from the fact that the function is known to be Schur, but some measurement errors occurred and lead to a function with values greater than 1 in modulus at some places.

Example 3

An example is taken of a rational function $r5$ of degree 5 such that $\|r5\|_\infty = 1.01$ and $\|r5\|_2 = 0.6225$. Note that $r5$ is not Schur but is analytic in the unit disk. As the asymptotic-BMO-type criterion can be applied only to Schur functions, the initialization was done upon applying it to the Schur function $r5/\|r5\|_\infty$.

Using our optimization process, we obtain an approximation of degree 5 with an error of $7.89e-3$. Scaling $r5$ into the unit disk (i.e. considering the function $\frac{r5}{\|r5\|_\infty}$) gives an error of $9.90e-3$.

Consider the initial and optimized parameters (see fig. 11.16). In this example, the interpolation points α given by the asymptotic-BMO-type criterion are surprisingly good.

Example 4

We want here to approximate a rational function $r10$ of degree 10, analytic in the unit disk, and such that $\|r10\|_\infty = 1.02$ and $\|r10\|_2 = 0.6772$. The asymptotic-BMO-type criterion applied to $r10/\|r10\|_\infty$ gives a sequence of points with one of multiplicity 3. As such an initialization could numerically leads to some problems, we chose to apply the asymptotic-BMO-type criterion to the strictly Schur function $\frac{r10}{1.05}$. The result is quite good : indeed, only one of the interpolation points α seems to have moved (see fig. 11.17).

The error of approximation is $2.58e-3$ (see fig. 11.18). Scaling $r10$ into the unit disk

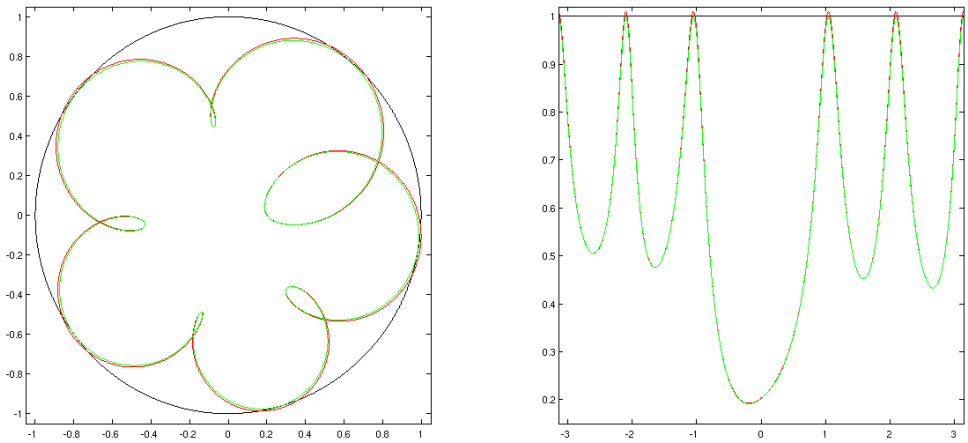

Figure 11.15: Function $r5$ (red) and Schur approximation (green) of degree 5.

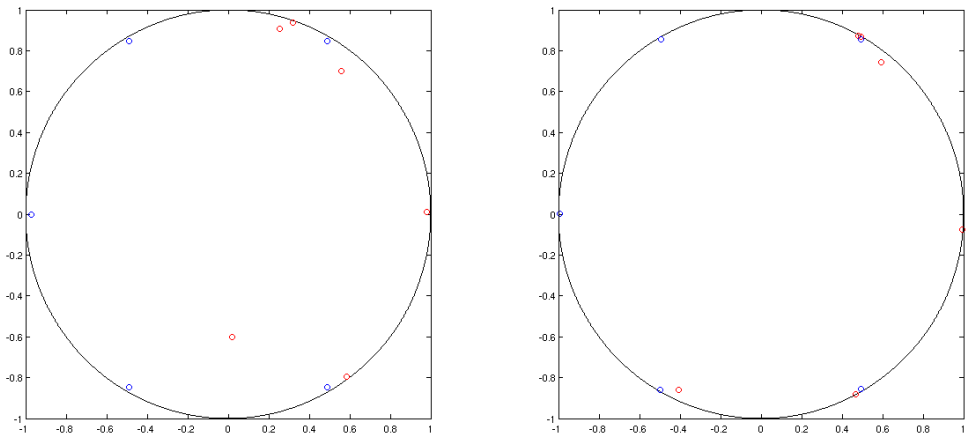

Figure 11.16: Initialization points (left) and optimized points (right) of the Schur function of degree 5 : parameters α (blue) and γ (red).

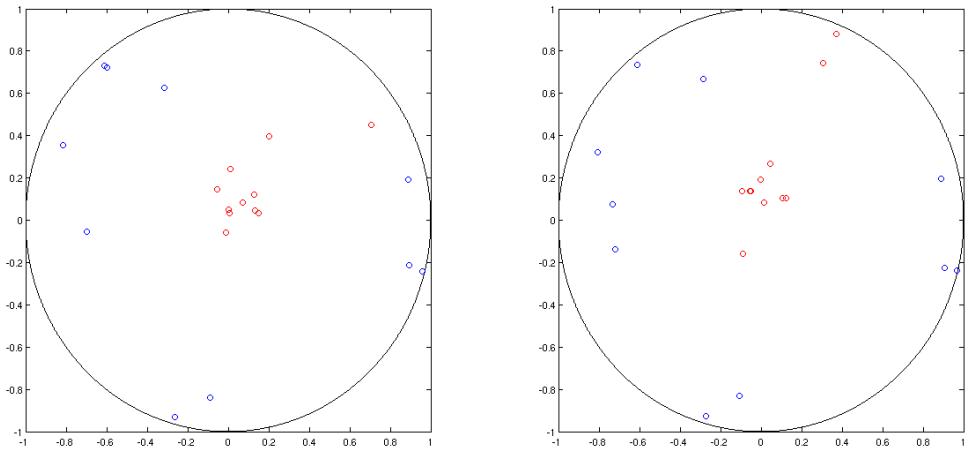

Figure 11.17: Initialization points (left) and optimized points (right) of the Schur function of degree 10 : parameters α (blue) and γ (red).

gives an error of $1.96e-2$.

On the last three examples, at least one initialization for a given degree seems to be quite good. However, all the initial interpolation points of the first example are bad. We chose to compute again an initialization but this time to the scaled strictly Schur function $0.97 \times p30$. This leads to the points plotted in fig. 11.19 for the degree 7. The interpolation points are "in the same directions" than the optimized points of the fig. 11.3.

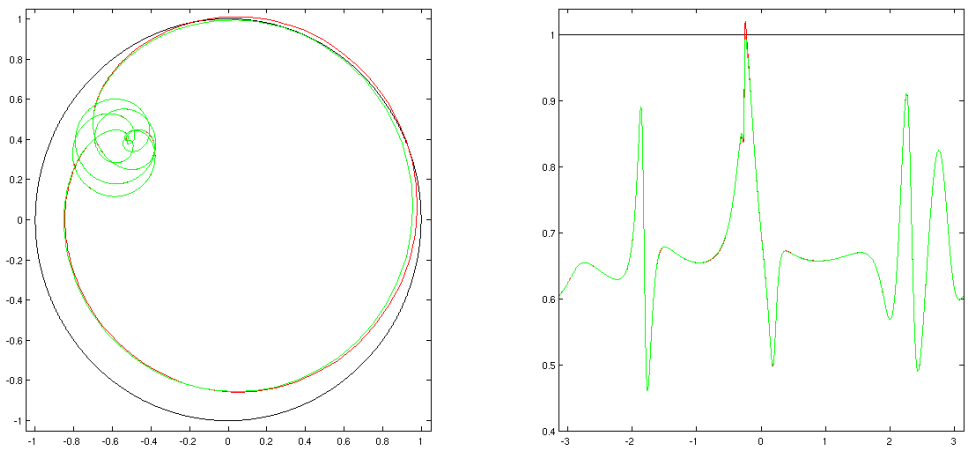

Figure 11.18: Function $r10$ (red) and Schur approximation (green) of degree 10.

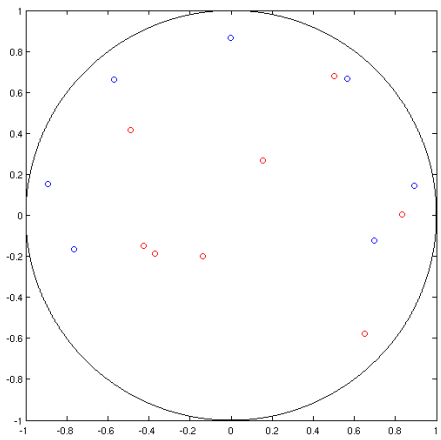

Figure 11.19: Another initialization for the approximation of degree 7 of $p30$: parameters α (blue) and γ (red).

Chapter 12

Conclusion

In the previous chapter, we used a parametrization with Schur parameters of modulus strictly less than 1 only. Using this method, only *strictly* Schur rational functions could be represented. Finding a way to parametrize *all* Schur rational functions of given degree would be a great improvement. This is our attempt in this chapter. We will present an interpolation on the circle, and also another algorithm with Schur parameters strictly less than 1, but which has the advantage to have a limit when the parameters tend toward the circle. How to merge the two types of parametrization into a single one is an open problem as for now.

12.1 J-inner matrices and the Schur algorithm

This section is an introduction to the J-inner matrices and some of their properties.

Definition 12.1.1 *Let* $J = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}$. *A* 2×2 *matrix-valued funtion* θ *is called* J*-inner if it is meromorphic in* $\mathbb{D}$ *and*

- $\theta(z)J\theta(z)^* \leq J$ *at every point* z *of analyticity of* θ *in* $\mathbb{D}$*, and*
- $\theta(z)J\theta(z)^* = J$ *at almost every point* z *of* $\mathbb{T}$.

Many properties of J-inner matrices can be found in ([Dym, 1989]). A basic one is the following:

Proposition 12.1.2 *If* $\theta = \begin{pmatrix} \theta_{11} & \theta_{12} \\ \theta_{21} & \theta_{22} \end{pmatrix}$ *is* 2×2 J*-inner and analytic in* $\mathbb{D}$ *and* g *is a Schur function, then* $(\theta_{21}g + \theta_{22})$ *is invertible in* $\mathbb{D}$*. Furthermore, if* $T_\theta(g)$ *is defined by*

$$T_\theta(g) = (\theta_{11}g + \theta_{12})(\theta_{21}g + \theta_{22})^{-1}$$

then $f = T_\theta(g)$ *is a Schur function.*

The result carries to higher sizes of θ but we will not need it.

Proof The proof can be found in different references, e.g. [Dym, 1989] for the matricial case. However, for a better understanding, we choose to give it again.

We first prove that $\theta_{21}g + \theta_{22}$ is invertible at any point of $\mathbb{D}$. As θ is J-inner, we have $\theta J \theta^* \leq J$ that is

$$\begin{pmatrix} |\theta_{11}|^2 - |\theta_{12}|^2 & \theta_{11}\overline{\theta_{21}} - \theta_{12}\overline{\theta_{22}} \\ \theta_{21}\overline{\theta_{11}} - \theta_{22}\overline{\theta_{12}} & |\theta_{21}|^2 - |\theta_{22}|^2 \end{pmatrix} \leq \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix} \quad \text{in } \mathbb{D}.$$

This leads to $|\theta_{21}|^2 - |\theta_{22}|^2 \leq -1$, which is equivalent to $|\theta_{22}|^2 \geq 1 + |\theta_{21}|^2$. Therefore, θ_{22} is invertible at any point of $\mathbb{D}$. We thus have

$$1 - \left|\frac{\theta_{21}}{\theta_{22}}\right|^2 \geq \frac{1}{|\theta_{22}|^2} > 0$$

that is $\left|\frac{\theta_{21}}{\theta_{22}}\right|^2 < 1$ at any point of $\mathbb{D}$. We then deduce that $\theta_{21}g + \theta_{22} = \theta_{22}(\theta_{22}^{-1}\theta_{21}g + 1)$ is invertible at any point of $\mathbb{D}$.

We now prove that f is Schur. We have:

$$\begin{pmatrix} f \\ 1 \end{pmatrix} = \begin{pmatrix} \theta_{11}g + \theta_{12} \\ \theta_{21}g + \theta_{22} \end{pmatrix} (\theta_{21}g + \theta_{22})^{-1} = \theta \begin{pmatrix} g \\ 1 \end{pmatrix} (\theta_{21}g + \theta_{22})^{-1}$$

and

$$\begin{pmatrix} f \\ 1 \end{pmatrix}^* J \begin{pmatrix} f \\ 1 \end{pmatrix} = |f|^2 - 1.$$

Therefore,

$$\begin{aligned} |f|^2 - 1 &= (\theta_{21}g + \theta_{22})^{-*} \begin{pmatrix} \overline{g} & 1 \end{pmatrix} \theta^* J \theta \begin{pmatrix} g \\ 1 \end{pmatrix} (\theta_{21}g + \theta_{22})^{-1} \\ &\leq (\theta_{21}g + \theta_{22})^{-*}(|g|^2 - 1)(\theta_{21}g + \theta_{22})^{-1} \\ &\leq 0 \end{aligned}$$

and f is Schur.

■

Note that the multipoint Schur algorithm we used is such that

$$f = \frac{\zeta_1 f_1 + \gamma_0}{1 + \bar{\gamma}_0 \zeta_1 f_1}$$

that is $f = T_{\theta_1}(f_1)$ with

$$\theta_1(z) = \frac{1}{\sqrt{1 - |\gamma_0^2|}} \begin{pmatrix} \zeta_1(z) & \gamma_0 \\ \bar{\gamma}_0 \zeta_1(z) & 1 \end{pmatrix}. \tag{12.1}$$

It is easy to check that θ_1 is J-inner. Indeed,

$$\begin{aligned}
J-\theta_1(z)J\theta_1^*(z) &= J-\frac{1}{\sqrt{1-|\gamma_0^2|}}\begin{pmatrix}\zeta_1(z) & \gamma_0\\ \bar\gamma_0\zeta_1(z) & 1\end{pmatrix}J\frac{1}{\sqrt{1-|\gamma_0^2|}}\begin{pmatrix}\overline{\zeta_1(z)} & \gamma_0\overline{\zeta_1(z)}\\ \bar\gamma_0 & 1\end{pmatrix}\\
&= J-\frac{1}{1-|\gamma_0^2|}\begin{pmatrix}\zeta_1(z) & \gamma_0\\ \bar\gamma_0\zeta_1(z) & 1\end{pmatrix}\begin{pmatrix}\overline{\zeta_1(z)} & \gamma_0\overline{\zeta_1(z)}\\ -\bar\gamma_0 & -1\end{pmatrix}\\
&= \frac{1}{1-|\gamma_0^2|}\begin{pmatrix}1-|\zeta_1(z)|^2 & -\gamma_0(|\zeta_1(z)|^2-1)\\ -\bar\gamma_0(|\zeta_1(z)|^2-1) & |\gamma_0|^2(1-|\zeta_1(z)|^2)\end{pmatrix}\\
&= \frac{1-|\zeta_1(z)|^2}{1-|\gamma_0^2|}\begin{pmatrix}1 & \gamma_0\\ \bar\gamma_0 & |\gamma_0|^2\end{pmatrix}\\
&= \frac{1-|\zeta_1(z)|^2}{1-|\gamma_0^2|}\begin{pmatrix}1\\ \bar\gamma_0\end{pmatrix}\begin{pmatrix}1 & \gamma_0\end{pmatrix}\\
&\geq 0 \text{ for } z\in\mathbb{D} \text{ and } = 0 \text{ for } z\in\mathbb{T}.
\end{aligned}$$

The Schur algorithm is based on the following result:

Let f be a Schur function. f satisfies the interpolation property $f(\alpha_1)=\gamma_0$ if and only if $f=T_{\theta_1}(f_1)$ for some Schur function f_1.

This result holds if we replace θ_1 by any J-inner function of the form $\theta_1 H$ where H is a constant matrix satisfying $H^*JH=J$ (such a matrix H is called J-unitary). This is a very particular case of the Nevanlinna-Pick interpolation problem studied for example in [Dym, 1989].

In section 12.3, another choice of J-inner matrix will be proposed.

12.2 Interpolation on the circle

The Schur algorithm studied in the previous chapter falls short of considering points on the unit circle. We now study an algorithm which manages such an interpolation.

The following proposition shows a relation between the value of a Schur function at points of the unit circle, and the value of its angular derivative. The proof can be found in [Ball et al., 1990].

Proposition 12.2.1 *Let α_T and γ_T in $\mathbb{T}$. We denote by $f'(\alpha_T)$ the limit $\lim_{z\to\alpha_T} f'(z)$ where z converges to α_T nontangentially. If f is a Schur function such that $f(\alpha_T)=\gamma_T$, then $f'(\alpha_T)=\rho\bar\alpha_T\gamma_T$ where ρ is a positive real constant.*

We now define a J-inner matrix which leads to an interpolation scheme on the circle.

Proposition 12.2.2 *Let α_T and γ_T be points of the unit circle, ρ be a positive real constant, and x_T be the vector such that $x_T^t=(1\quad\bar\gamma_T)$. Then, the matrix θ_2 defined by*

$$\theta_2(z)=I_2+\frac{1}{2\rho}\frac{z+\alpha_T}{z-\alpha_T}x_Tx_T^*J$$

is J-inner.

Proof We have

$$\begin{aligned} J-\theta_2(z)J\theta_2(z)^* &= J-\left(I_2+\frac{1}{2\rho}\frac{z+\alpha_T}{z-\alpha_T}x_Tx_T^*J\right)J\left(I_2+\frac{1}{2\rho}\overline{\left(\frac{z+\alpha_T}{z-\alpha_T}\right)}Jx_Tx_T^*\right) \\ &= -\frac{1}{2\rho}\frac{z+\alpha_T}{z-\alpha_T}x_Tx_T^*-\frac{1}{2\rho}\overline{\left(\frac{z+\alpha_T}{z-\alpha_T}\right)}x_Tx_T^* \\ &\quad -\frac{1}{(2\rho)^2}\left|\frac{z+\alpha_T}{z-\alpha_T}\right|^2(1-|\gamma_T|^2)x_Tx_T^*. \end{aligned}$$

As $|\gamma_T|=1$, we get

$$J-\theta_2(z)J\theta_2(z)^*=-\frac{1}{2\rho}\left[\frac{z+\alpha_T}{z-\alpha_T}+\overline{\left(\frac{z+\alpha_T}{z-\alpha_T}\right)}\right]x_Tx_T^*=-\frac{1}{\rho}Re\left(\frac{z+\alpha_T}{z-\alpha_T}\right)x_Tx_T^*.$$

But $Re\left(\frac{z+\alpha_T}{z-\alpha_T}\right)=Re\left(\frac{|z|^2+\alpha_T\bar{z}-\bar{\alpha}_Tz-|\alpha_T|^2}{|z-\alpha_T|^2}\right)=\frac{|z|^2-|\alpha_T|^2}{|z-\alpha_T|^2}\leq 0$ for all $z\in\mathbb{D}$, and consequently, $J-\theta_2(z)J\theta_2(z)^*\geq 0$.

■

Proposition 12.2.3 *If g is a Schur function such that $g(\alpha_T)\neq\gamma_T$ then $f=T_{\theta_2}(g)$ is a Schur function such that $f(\alpha_T)=\gamma_T$ and $f'(\alpha_T)=\rho\overline{\alpha_T}\gamma_T$.*

Proof We have

$$\theta_2(z)=\begin{pmatrix} 1+\frac{1}{2\rho}\frac{z+\alpha_T}{z-\alpha_T} & -\frac{\gamma_T}{2\rho}\frac{z+\alpha_T}{z-\alpha_T} \\ \frac{\bar{\gamma}_T}{2\rho}\frac{z+\alpha_T}{z-\alpha_T} & 1-\frac{1}{2\rho}\frac{z+\alpha_T}{z-\alpha_T}\end{pmatrix}$$

so that

$$f(z)=\frac{(2\rho(z-\alpha_T)+(z+\alpha_T))g(z)-\gamma_T(z+\alpha_T)}{\bar{\gamma}_T(z+\alpha_T)g(z)+2\rho(z-\alpha_T)-(z+\alpha_T)}.$$

Therefore

$$f(\alpha_T)=\frac{2\alpha_T(g(\alpha_T)-\gamma_T)}{2\alpha_T(\bar{\gamma}_Tg(\alpha_T)-1)}=\gamma_T$$

because $g(\alpha_T)\neq\gamma_T$.
A direct computation gives

$$\begin{aligned} f'(\alpha_T) &= \frac{((2\rho+1)g(\alpha_T)+2\alpha_Tg'(\alpha_T)-\gamma_T)}{2\alpha_T(\bar{\gamma}_Tg(\alpha_T)-1)} \\ &\quad -\frac{(\bar{\gamma}_Tg(\alpha_T)+2\alpha_T\bar{\gamma}_Tg'(\alpha_T)+2\rho-1)(2\alpha_T(g(\alpha_T)-\gamma_T)}{(2\alpha_T(\bar{\gamma}_Tg(\alpha_T)-1))^2} \\ &= \frac{2\rho(g(\alpha_T)-\gamma_T)}{2\alpha_T(\bar{\gamma}_Tg(\alpha_T)-1)} \\ &= \rho\bar{\alpha}_T\gamma_T. \end{aligned}$$

■

Note that if $f=p/q$, an interpolation point in the circle is always a root of $p\tilde{p}-q\tilde{q}$. We will now show that if we apply the algorithm associated to θ_2 to a Schur rational function

p/q of degree n such that $p/q(\alpha_T) = \gamma_T$ and $(p/q)'(\alpha_T) = \rho\overline{\alpha_T}\gamma_T$, then $g = T_{\theta_2^{-1}}\left(\frac{p}{q}\right)$ is a Schur rational function of degree $n-1$. Indeed,

$$\begin{aligned} g &= \frac{2\rho(z-\alpha_T)\frac{p}{q} - (z+\alpha_T)\left(\frac{p}{q} - \gamma_T\right)}{2\rho(z-\alpha_T) + (z+\alpha_T)\left(1 - \bar{\gamma}_T\frac{p}{q}\right)} \\ &= \frac{2\rho p - (z+\alpha_T)\frac{p-\gamma_T q}{z-\alpha_T}}{2\rho q - (z+\alpha_T)\bar{\gamma}_T\frac{p-\gamma_T q}{z-\alpha_T}}. \end{aligned}$$

But evaluating the numerator and denominator of g at α_T gives

$$2\rho p(\alpha_T) - 2\alpha_T q(\alpha_T) f'(\alpha_T) = 2\rho\gamma_T q(\alpha_T) - 2\alpha_T q(\alpha_T)\rho\bar{\alpha}_T\gamma_T = 0$$

and

$$2\rho q(\alpha_T) - 2\alpha_T\bar{\gamma}_T q(\alpha_T) f'(\alpha_T) = 2\rho q(\alpha_T) - 2\alpha_T\bar{\gamma}_T q(\alpha_T)\rho\bar{\alpha}_T\gamma_T = 0.$$

Therefore, the degree of g is at most $n-1$. Applying the linear transform T_{θ_2} to g increases the degree of at most one. Thus, the degree of g is exactly $n-1$.

12.3 A better algorithm ?

We are now going to study another parametrization whose advantage is to have a limit when points tend towards the circle. The link with the previous Schur algorithm is given.

12.3.1 Another algorithm

Proposition 12.3.1 *Let α and γ be points of the unit disk $\mathbb{D}$, and x be the vector $(1 \quad \bar{\gamma})^t$. Then, the matrix θ_3 defined by*

$$\theta_3(z) = I_2 + \frac{\zeta_\alpha(z) - 1}{1 - |\gamma|^2} xx^* J \tag{12.2}$$

is J-inner.

Proof We have

$$\begin{aligned} J - \theta_3(z) J \theta_3^*(z) &= J - \left(I_2 + \frac{\zeta_\alpha(z)-1}{1-|\gamma|^2} xx^* J\right) J \left(I_2 + Jxx^* \frac{\overline{\zeta_\alpha(z)} - 1}{1-|\gamma|^2}\right) \\ &= -\frac{\zeta_\alpha(z)-1}{1-|\gamma^2|} xx^* - xx^* \frac{\overline{\zeta_\alpha(z)} - 1}{1-|\gamma|^2} - \frac{|\zeta_\alpha(z)-1|^2}{1-|\gamma^2|} xx^* \\ &= -\left(|\zeta_\alpha(z)-1|^2 + \zeta_\alpha(z) - 1 + \overline{\zeta_\alpha(z)} - 1\right) \frac{xx^*}{1-|\gamma|^2} \\ &= -((\zeta_\alpha(z)-1)(\overline{\zeta_\alpha(z)} - 1) + \zeta_\alpha(z) - 1 + \overline{\zeta_\alpha(z)} - 1) \frac{xx^*}{1-|\gamma|^2} \\ &= \frac{1 - |\zeta_\alpha(z)|^2}{1-|\gamma|^2} xx^* \\ &\geq 0 \text{ for all } z \in \mathbb{D}. \end{aligned}$$

■

Proposition 12.3.2 *Let g be a Schur function. Then $f = T_{\theta_3}(g)$ is a Schur function such that $f(\alpha) = \gamma$.*

Proof We have

$$x^* J\theta_3(\alpha) = x^* J - \frac{1}{1-|\gamma|^2} x^* Jxx^* J$$

and $x^* Jx = 1 - |\gamma|^2$, therefore $x^* J\theta_3(\alpha) = 0$. Thus,

$$x^* J \begin{pmatrix} f(\alpha) \\ 1 \end{pmatrix} = x^* J\theta_3(\alpha) \begin{pmatrix} g(\alpha) \\ 1 \end{pmatrix} ((\theta_3)_{21}(\alpha)g(\alpha) + (\theta_3)_{22}(\alpha))^{-1} = 0$$

and we get $f(\alpha) = \gamma$.

■

12.3.2 Relation between the two algorithms

We now show that the J-inner matrix of the "new" algorithm is in fact the J-inner matrix of the previous algorithm multiplied by a constant matrix H.
The proof of the following lemma is immediate.

Lemma 12.3.3 *Let γ in $\mathbb{D}$ and*

$$H(\gamma) = \frac{1}{\sqrt{1-|\gamma|^2}} \begin{pmatrix} 1 & \gamma \\ \bar{\gamma} & 1 \end{pmatrix}.$$

The matrix $H(\gamma)$ has the following properties:

- *$H(\gamma)$ is J-unitary, i.e. $H(\gamma)JH(\gamma)^* = J$,*
- *$H(\gamma)^{-1} = H(-\gamma)$.*

We now give another expression of the J-inner matrix associated to the "new" algorithm ([Hanzon et al., 2006]).

Proposition 12.3.4 *The matrix θ_3 defined by (12.2) is of the form*

$$\theta_3(z) = H(\gamma) \begin{pmatrix} \zeta_\alpha & 0 \\ 0 & 1 \end{pmatrix} H(\gamma)^{-1}.$$

Proof We have

$$\begin{aligned} H(-\gamma)\theta_3(z)H(\gamma) &= H(-\gamma)\left(I_2 + \frac{\zeta_\alpha(z)-1}{1-|\gamma|^2} xx^* J\right) H(\gamma) \\ &= I_2 + \frac{\zeta_\alpha(z)-1}{1-|\gamma|^2} H(-\gamma)xx^* JH(\gamma) \\ &= I_2 + \frac{\zeta_\alpha(z)-1}{1-|\gamma|^2} \sqrt{1-|\gamma|^2} \begin{pmatrix} 1 \\ 0 \end{pmatrix} \sqrt{1-|\gamma|^2} \begin{pmatrix} 1 & 0 \end{pmatrix} \\ &= \begin{pmatrix} \zeta_\alpha & 0 \\ 0 & 1 \end{pmatrix}. \end{aligned}$$

■

Note that the matrix θ_1 defined by (12.1) is of the form

$$\theta_1 = H(\gamma)\begin{pmatrix} \zeta_\alpha & 0 \\ 0 & 1 \end{pmatrix}.$$

Therefore, the link between the matrix θ_3 and θ_1 is given by

$$\theta_3 = \theta_1 H(-\gamma).$$

12.3.3 Toward a parametrization of all Schur rational functions

We now show that when the point α tends to a point α_T of the unit circle, θ_3 tends to θ_2 ([Hanzon et al., 2008]). We have

$$\begin{aligned}
\frac{\zeta_\alpha(z)-1}{1-|f(\alpha)|^2} &= \frac{-\frac{|\alpha|}{\alpha}\frac{z-\alpha}{1-\bar{\alpha}z}-1}{1-|f(\alpha)|^2} \\
&= \frac{\frac{-|\alpha|(z-\alpha)-(\alpha-|\alpha|^2 z)}{\alpha-|\alpha|^2 z}}{1-f(\alpha)\overline{f(\alpha)}} \\
&= \frac{(|\alpha|-1)\frac{\alpha+|\alpha|z}{\alpha(1-\bar{\alpha}z)}}{1-f(\alpha)\overline{f(\alpha)}}.
\end{aligned}$$

Using a Taylor expansion, we get

$$f(\alpha) = f(\alpha_T) + (\alpha-\alpha_T)f'(\alpha_T) + o(|\alpha-\alpha_T|).$$

Therefore,

$$\begin{aligned}
1-f(\alpha)\overline{f(\alpha)} &= -2Re\left[(\alpha-\alpha_T)\overline{f(\alpha_T)}f'(\alpha_T)\right] + o(|\alpha-\alpha_T|) \\
&= -2Re\left[(\alpha-\alpha_T)\bar{\gamma}_T\rho\bar{\alpha}_T\gamma_T\right] + o(|\alpha-\alpha_T|) \\
&= -2Re\left[\rho(\alpha\overline{\alpha_T}-1)\right] + o(|\alpha-\alpha_T|)
\end{aligned}$$

and we get

$$\frac{\zeta_\alpha(z)-1}{1-|f(\alpha)|^2} = \frac{(|\alpha|-1)\frac{z+|\alpha|\alpha}{(|\alpha|^2 z-\alpha)}}{2Re\left[\rho(\alpha\overline{\alpha_T}-1)\right]+o(|\alpha-\alpha_T|)}.$$

It remains to check that $\frac{|\alpha|-1}{2Re(\alpha\overline{\alpha_T}-1)}$ tends toward $\frac{1}{2}$. Let η be a complex number such that $\alpha = \alpha_T + \eta$. Then

$$|\alpha|^2 = |\alpha_T|^2 + 2Re(\eta\bar{\alpha}_T) + |\eta|^2 = 1 + 2Re(\eta\bar{\alpha}_T) + |\eta|^2$$

and we deduce that

$$|\alpha| = 1 + Re(\eta\bar{\alpha}_T) + o(\eta).$$

Thus $|\alpha|-1 = Re(\eta\bar{\alpha}_T)+o(\eta)$. As $2Re\,(\alpha\bar{\alpha_T}-1) = 2Re(\eta\bar{\alpha}_T)$, the conclusion is immediate.

As stated before, only *strictly* Schur rational functions can be represented using the parametrization of the previous chapter. From what precedes, we see that the algorithm associated to θ_3 could be combined with interpolation on the unit circle, and therefore, parameters could be taken in the closed unit disk $\overline{\mathbb{D}}$. This could be a great improvment. However, new questions arise: could this algorithm be related to orthogonal rational functions ? And in practice, when do you choose to take interpolation points on the circle and how could one compute the parameter ρ?

Bibliography

[Achieser, 1992] Achieser, N. I. (1992). *Theory of approximation.* Dover Publications Inc., New York. Translated from the Russian and with a preface by Charles J. Hyman, Reprint of the 1956 English translation.

[Ahlfors, 1973] Ahlfors, L. V. (1973). *Conformal invariants: topics in geometric function theory.* McGraw-Hill Book Co., New York. McGraw-Hill Series in Higher Mathematics.

[Alpay, 2001] Alpay, D. (2001). *The Schur algorithm, reproducing kernel spaces and system theory*, volume 5 of *SMF/AMS Texts and Monographs.* American Mathematical Society, Providence, RI. Translated from the 1998 French original by Stephen S. Wilson.

[Alpay et al., 1994] Alpay, D., Baratchart, L., and Gombani, A. (1994). On the differential structure of matrix-valued rational inner functions. In *Nonselfadjoint operators and related topics (Beer Sheva, 1992)*, volume 73 of *Oper. Theory Adv. Appl.*, pages 30–66. Birkhäuser, Basel.

[Amari, 2000] Amari, S. (Sep 2000). Synthesis of cross-coupled resonator filters using an analytical gradient-based optimization technique. *Microwave Theory and Techniques, IEEE Transactions on*, 48(9):1559–1564.

[Bakonyi and Constantinescu, 1992] Bakonyi, M. and Constantinescu, T. (1992). *Schur's algorithm and several applications*, volume 261 of *Pitman Research Notes in Mathematics Series.* Longman Scientific & Technical, Harlow.

[Ball et al., 1990] Ball, J. A., Gohberg, I., and Rodman, L. (1990). *Interpolation of rational matrix functions*, volume 45 of *Operator Theory: Advances and Applications.* Birkhäuser Verlag, Basel.

[Baratchart et al., 1998] Baratchart, L., Grimm, J., Leblond, J., Olivi, M., Seyfert, S., and Wielonski, F. (1998). Identification d'un filtre hyperfréquences par approximation dans le domaine complexe. *Rapport de Recherche INRIA No. 0219, Mars 1998.*

[Bila et al., 2001] Bila, S., Baillargeat, D., Aubourg, M., Verdeyme, S., Guillon, P., Seyfert, F., Grimm, J., Baratchart, L., Zanchi, C., and Sombrin, J. (Mar 2001). Direct electromagnetic optimization of microwave filters. *Microwave Magazine, IEEE*, 2(1):46–51.

[Bila et al., 2006] Bila, S., Cameron, R. J., Lenoir, P., Lunot, V., and Seyfert, F. (2006). Chebyshev synthesis for multi-band microwave filters. In *2006 IEEE MTT-S International Microwave Symposium Digest*, pages 1221–1224.

[Braess, 1986] Braess, D. (1986). *Nonlinear approximation theory*, volume 7 of *Springer Series in Computational Mathematics*. Springer-Verlag, Berlin.

[Brezis, 1983] Brezis, H. (1983). *Analyse fonctionnelle*. Collection Mathématiques Appliquées pour la Maîtrise. [Collection of Applied Mathematics for the Master's Degree]. Masson, Paris. Théorie et applications. [Theory and applications].

[Bultheel et al., 1999] Bultheel, A., González-Vera, P., Hendriksen, E., and Njåstad, O. (1999). *Orthogonal rational functions*, volume 5 of *Cambridge Monographs on Applied and Computational Mathematics*. Cambridge University Press, Cambridge.

[Bultheel et al., 2006] Bultheel, A., González-Vera, P., Hendriksen, E., and Njåstad, O. (2006). Orthogonal rational functions on the unit circle: from the scalar to the matrix case. In *Orthogonal polynomials and special functions*, volume 1883 of *Lecture Notes in Math.*, pages 187–228. Springer, Berlin.

[Cameron, 1999] Cameron, R. (Apr 1999). General coupling matrix synthesis methods for Chebyshev filtering functions. *Microwave Theory and Techniques, IEEE Transactions on*, 47(4):433–442.

[Cameron et al., 2005a] Cameron, R., Faugere, J., and Seyfert, F. (12-17 June 2005a). Coupling matrix synthesis for a new class of microwave filter configuration. *Microwave Symposium Digest, 2005 IEEE MTT-S International*.

[Cameron et al., 2005b] Cameron, R., Yu, M., and Wang, Y. (Nov. 2005b). Direct-coupled microwave filters with single and dual stopbands. *Microwave Theory and Techniques, IEEE Transactions on*, 53(11):3288–3297.

[Cameron et al., 2007] Cameron, R. J., Mansour, R., and Kudsia, C. M. (2007). *Microwave Filters for Communication Systems: Fundamentals, Design and Applications*. Wiley-Interscience.

[Ceauşescu and Foiaş, 1978] Ceauşescu, Z. and Foiaş, C. (1978). On intertwining dilations. V. *Acta Sci. Math. (Szeged)*, 40(1-2):9–32.

[Cheney, 1998] Cheney, E. W. (1998). *Introduction to approximation theory*. AMS Chelsea Publishing, Providence, RI. Reprint of the second (1982) edition.

[Cheney and Loeb, 1961] Cheney, E. W. and Loeb, H. L. (1961). Two new algorithms for rational approximation. *Numer. Math.*, 3:72–75.

[Collin, 1991] Collin, R. E. (1991). *Field theory of guided waves*. New York IEEE Press.

[Conciauro et al., 2000] Conciauro, G., Guglielmi, M., and Sorrentino, R. (2000). *Advanced modal analysis: CAD techniques for waveguide components and filters*. Wiley.

[Djrbashian, 1962] Djrbashian, M. M. (1962). Orthogonal systems of rational functions on the unit circle with given set of poles. *Soviet Mathematics Doklady*, 3:1794–1798.

[Duren, 1970] Duren, P. L. (1970). *Theory of H^p spaces*. Pure and Applied Mathematics, Vol. 38. Academic Press, New York.

[Dym, 1989] Dym, H. (1989). *J contractive matrix functions, reproducing kernel Hilbert spaces and interpolation*, volume 71 of *CBMS Regional Conference Series in Mathematics*. Published for the Conference Board of the Mathematical Sciences, Washington, DC.

[Fulcheri and Olivi, 1998] Fulcheri, P. and Olivi, M. (1998). Matrix rational H_2 approximation: a gradient algorithm based on Schur analysis. *SIAM J. Control Optim.*, 36(6):2103–2127 (electronic).

[Garnett, 2007] Garnett, J. B. (2007). *Bounded analytic functions*, volume 236 of *Graduate Texts in Mathematics*. Springer, New York, first edition.

[Geronimus, 1944] Geronimus, J. (1944). On polynomials orthogonal on the circle, on trigonometric moment-problem and on allied Carathéodory and Schur functions. *Rec. Math. [Mat. Sbornik] N. S.*, 15(57):99–130.

[Gohberg et al., 2006] Gohberg, I., Lancaster, P., and Rodman, L. (2006). *Invariant subspaces of matrices with applications*, volume 51 of *Classics in Applied Mathematics*. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA. Reprint of the 1986 original.

[Grimm, 2000] Grimm, J. (2000). Rational approximation of transfer functions in the hyperion software. *Rapport de Recherche INRIA No. 4002, Sept. 2000*.

[Hanzon et al., 2006] Hanzon, B., Olivi, M., and Peeters, R. L. M. (2006). Balanced realizations of discrete-time stable all-pass systems and the tangential Schur algorithm. *Linear Algebra Appl.*, 418(2-3):793–820.

[Hanzon et al., 2008] Hanzon, B., Olivi, M., and Peeters, R. L. M. (2008). Boundary interpolation and parametrization of lossless functions. *Rapport de Recherche INRIA*.

[Hong and Lancaster, 2001] Hong, J.-S. and Lancaster, M. (2001). *Microstrip Filters for RF/Microwave Applications*. Wiley-Interscience.

[Jones, 1988] Jones, W. B. (1988). Schur's algorithm extended and Schur continued fractions. In *Nonlinear numerical methods and rational approximation (Wilrijk, 1987)*, volume 43 of *Math. Appl.*, pages 281–298. Reidel, Dordrecht.

[Kailath, 1980] Kailath, T. (1980). *Linear systems*. Prentice-Hall Inc., Englewood Cliffs, N.J. Prentice-Hall Information and System Sciences Series.

[Khrushchev, 2001] Khrushchev, S. (2001). Schur's algorithm, orthogonal polynomials, and convergence of Wall's continued fractions in $L^2(\mathbb{T})$. *J. Approx. Theory*, 108(2):161–248.

[Kimura, 1996] Kimura, H. (1996). *Chain-Scattering Approach to H-Infinity Control*. Birkhauser.

[Koosis, 1998] Koosis, P. (1998). *Introduction to H_p spaces*, volume 115 of *Cambridge Tracts in Mathematics*. Cambridge University Press, Cambridge, second edition. With two appendices by V. P. Havin [Viktor Petrovich Khavin].

[Kurokawa, 1969] Kurokawa, K. (1969). *An introduction to the theory of microwave circuits*. New York Academic Press.

[Langer and Lasarow, 2004] Langer, H. and Lasarow, A. (2004). Solution of a multiple Nevanlinna-Pick problem via orthogonal rational functions. *J. Math. Anal. Appl.*, 293(2):605–632.

[Le Bailly and Thiran, 1998] Le Bailly, B. and Thiran, J. P. (1998). Optimum parameters for the generalized ADI method. *Numer. Math.*, 80(3):377–395.

[Lee and Sarabandi, 2008] Lee, J. and Sarabandi, K. (Jan. 2008). Design of triple-passband microwave filters using frequency transformations. *Microwave Theory and Techniques, IEEE Transactions on*, 56(1):187–193.

[Lunot et al., 2007] Lunot, V., Bila, S., and Seyfert, F. (2007). Optimal synthesis for multi-band microwave filters. In *2007 IEEE MTT-S International Microwave Symposium Digest*, pages 115–118.

[Lunot et al., 2008] Lunot, V., Seyfert, F., Bila, S., and Nasser, A. (2008). Certified computation of optimal multiband filtering functions. *IEEE Transactions on Microwave Theory and Techniques*, 56(1):105–112.

[Macchiarella and Tamiazzo, 2005] Macchiarella, G. and Tamiazzo, S. (Nov. 2005). Design techniques for dual-passband filters. *Microwave Theory and Techniques, IEEE Transactions on*, 53(11):3265–3271.

[Matthaei, 1965] Matthaei, Y. J. (1965). *Microwave filters, impedance matching networks and coupling structures*. New York, Mc Graw Hill.

[Mokhtaari et al., 2006] Mokhtaari, M., Bornemann, J., Rambabu, K., and Amari, S. (Nov. 2006). Coupling-matrix design of dual and triple passband filters. *Microwave Theory and Techniques, IEEE Transactions on*, 54(11):3940–3946.

[Nikishin and Sorokin, 1991] Nikishin, E. M. and Sorokin, V. N. (1991). *Rational approximations and orthogonality*, volume 92 of *Translations of Mathematical Monographs*. American Mathematical Society, Providence, RI. Translated from the Russian by Ralph P. Boas.

[Pan, 1996] Pan, K. (1996). On the convergence of rational functions orthogonal on the unit circle. *J. Comput. Appl. Math.*, 76(1-2):315–324.

[Potapov, 1955] Potapov, V. P. (1955). The multiplicative structure of J-contractive matrix functions. *Trudy Moskov. Mat. Obšč.*, 4:125–236.

[Powell, 1981] Powell, M. J. D. (1981). *Approximation theory and methods*. Cambridge University Press, Cambridge.

[Remes, 1934] Remes, E. (1934). Sur le calcul effectif des polynômes d'approximation de Tchebichef. *Comptes rendus hebdomadaires des séances de l'Académie des Sciences de Paris*, 199:337–340.

[Rivlin, 1990] Rivlin, T. J. (1990). *Chebyshev polynomials*. Pure and Applied Mathematics (New York). John Wiley & Sons Inc., New York, second edition. From approximation theory to algebra and number theory.

[Rudin, 1987] Rudin, W. (1987). *Real and complex analysis*. McGraw-Hill Book Co., New York, third edition.

[Saff and Totik, 1997] Saff, E. B. and Totik, V. (1997). *Logarithmic potentials with external fields*, volume 316 of *Grundlehren der Mathematischen Wissenschaften [Fundamental Principles of Mathematical Sciences]*. Springer-Verlag, Berlin. Appendix B by Thomas Bloom.

[Schur, 1917] Schur, I. (1917). Uber potenzreihen, die im innern des einheitskreises beschrankt sind, i. j. reine angew. math., 147:205–232, 1917. *English translation in: I. Schur methods in operator theory and signal processing (Operator theory: advances and applications OT 18 (1986), Birkh auser Verlag)*.

[Schwartz, 1964] Schwartz, L. (1964). Sous-espaces hilbertiens d'espaces vectoriels topologiques et noyaux associés (noyaux reproduisants). *J. Analyse Math.*, 13:115–256.

[Sombrin, 2002] Sombrin, J. (2002). Analyse, synthèse, identification des paramètres et simulation des filtres hyperfréquences. *Rapport de Recherche CNES No. DTS/AE/TTL 2002-, Août 2002*.

[Szegő, 1975] Szegő, G. (1975). *Orthogonal polynomials*. American Mathematical Society, Providence, R.I., fourth edition. American Mathematical Society, Colloquium Publications, Vol. XXIII.

[Todd, 1988] Todd, J. (1988). A legacy from E. I. Zolotarev (1847–1878). *Math. Intelligencer*, 10(2):50–53.

[Wall, 1948] Wall, H. S. (1948). *Analytic Theory of Continued Fractions*. D. Van Nostrand Company, Inc., New York, N. Y.

[Werner, 1963] Werner, H. (1963). Rationale Tschebyscheff-Approximation, Eigenwerttheorie und Differenzenrechnung. *Arch. Rational Mech. Anal.*, 13:330–347.

Techniques d'approximation rationnelle en synthèse fréquentielle : problème de Zolotarev et algorithme de Schur

Cette thèse présente des techniques d'optimisation et d'approximation rationnelle ayant des applications en synthèse et identification de systèmes passifs.

La première partie décrit un problème de Zolotarev : on cherche à maximiser sur une famille d'intervalles l'infimum du module d'une fonction rationnelle de degré donné, tout en contraignant son module à ne pas dépasser 1 sur une autre famille d'intervalles. On s'intéresse dans un premier temps à l'existence et à la caractérisation des solutions d'un tel problème. Deux algorithmes, de type Remes et correction différentielle, sont ensuite présentés et étudiés. Le lien avec la synthèse de filtres hyperfréquences est détaillé. La théorie présentée permet en fait le calcul de fonctions de filtrage, multibandes ou monobandes, respectant un gabarit fixé. Celle-ci a été appliquée à la conception de plusieurs filtres hyperfréquences multibandes dont les réponses théoriques et les mesures sont données.

La deuxième partie concerne l'approximation rationnelle Schur d'une fonction Schur. Une fonction Schur est une fonction analytique dans le disque unité bornée par 1 en module. On étudie tout d'abord l'algorithme de Schur multipoints, qui fournit un paramétrage des fonctions strictement Schur. Le lien avec les fonctions rationnelles orthogonales, obtenu grâce à un théorème de type Geronimus, est ensuite présenté. Celui-ci permet alors d'établir certaines propriétés d'approximation dans le cas peu étudié où les points d'interpolation tendent vers le bord du disque. En particulier, une convergence en métrique de Poincaré est obtenue grâce à une extension d'un théorème de type Szegő. Une étude numérique sur l'approximation rationnelle Schur à degré fixé est aussi réalisée.

Rational approximation techniques and frequency design: a Zolotarev problem and the Schur algorithm

This thesis presents some rational approximation and optimization techniques with applications to the synthesis and identification of passive systems.

In the first part, we study a Zolotarev-type problem: to maximize on some set of intervals the infimum of the modulus of a rational function of given degree, under the constraint that the modulus of this function is bounded by 1 on another set of intervals. We are first concerned with the existence and the characterization of the solutions to such a problem. Next, a Remes-type algorithm and a differential-correction-type algorithm are studied. The link with the synthesis of microwave filters is carried out in detail. In fact, the theory we present allows one to compute multiband filtering functions with respect to given specifications. From the practical viewpoint, some microwave filters have been designed using this theory, and their theoretical response is compared to the real one.

In the second part, the Schur rational approximation of a Schur function is studied. A Schur function is an analytic function whose modulus is bounded by 1 in the unit disk. First, the multipoint Schur algorithm is presented. It gives a parametrization of all strictly Schur functions. Next, the link with orthogonal rational functions is developed via a Geronimus-type theorem. The latter allows us to prove some approximation properties, where the interpolation points may tend to the unit circle. In particular, a convergence in the Poincaré metric is obtained thanks to an extension of a Szegő-type theorem.
A numerical study for the computation of the Schur approximants of given degree is also presented.