Analytical methods for the conception and tuning of microwave devices

Fabien Seyfert

▶ **To cite this version:**

Fabien Seyfert. Analytical methods for the conception and tuning of microwave devices. Automatic. UCA; Université Côte d'Azur, 2019. tel-02444432

HAL Id: tel-02444432
https://inria.hal.science/tel-02444432v1

Submitted on 20 Jan 2020

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Université Côte d'Azur

Mémoire

présenté pour obtenir le diplôme d'

Habilitation à diriger des recherches

en Sciences et Technologies de l'Information et de la Communication

Spécialité : Génie Informatique, Automatique et Traitement du Signal

par

Fabien Seyfert

Sujet: Méthodes analytiques pour la conception et le réglage de dispositifs micro-ondes

Manuscrit examiné par les rapporteurs suivants:

Mr. Jonathan Partington
Mr. Smain Amari
Mr. Jean-Charles Faugère

Soutenu le 6 Février 2019 devant le jury composé de :

Mr.	Yves Rolain	Président,
Mr.	Giuseppe Macchiarella	Examinateur
Mme.	Martine Olivi	Examinatrice
Mr.	Jacques Sombrin	Examinateur
Mr.	Stéphane Bila	Examinateur
Mr.	Laurent Baratchart	Invité

Manuscrit final du 15 Février 2019

Contents

Chapter 0

Introduction

This document is a summary of some of the research activities I have been pursuing since my appointment at Inria in 2001. At this time, inspired by my PhD thesis work, my objective was to improve significantly the de-embedding procedure for microwave filters we had been developing in the Miaou team. The architecture of this procedure is organized in two stages: an analytic completion step and a stable rational approximation one. The completion is rendered necessary by the incomplete nature of harmonic measurements that are, according to the physics of such devices, only performed in a narrow band around the resonating frequencies of the filter. We discuss in chapter 1 why this partial nature of the frequency measurements is problematic: it is essentially due to the ill-posed nature of the stable rational approximation problem, when posed on a strict subset of the boundary of the considered analyticity domain: the circle or the imaginary axis in our case. The completion step aims therefore to furnish an analytic model in the whole right half-plane of the harmonic measurements of the filter that can, in a second step, be safely approached rationally. We explain in chapter 1 why Hardy spaces are a natural class for such extension problems, and why, as shown by previous results obtained by the Miaou team, additional knowledge on the filter's frequency response is needed to perform this step. The main question I had at that time is summarized this way: how much additional information to the harmonic measurements is needed in order to perform a completion procedure in a satisfactory manner ? Answering this question is the main purpose of the first chapter of this document. Without spoiling the suspense around this question, we can roughly answer it this way: as opposed to what the analytic continuation principle may suggest, quite a lot. After introducing elementary bounded extremal problems, we present a mixed norm version of these, where the modulus of the transfer function needs to be specified for frequencies outside the measurement band. This led us to consider a completion problem where additional information is provided, by means of a finite dimensional description of possible extensions of the data. This approach led to a significant improvement of the de-embedding procedure, allowing it to run fully automatically, where before a tedious human assistance was needed to adjust the parameters of the extension procedure for every new data set and this often for poor final results. This in turn triggered the construction of a dedicated toolbox Presto-HF that was transferred to academic, as well as industrial, practitioners of the filter community. The final completion procedure as well as the functioning of Presto-HF are described at the

end of chapter 1.

Working in close connection with filter specialists led us to consider a preliminary stage in the manufacturing process of filters: the synthesis of an ideal frequency response. We contributed here to define a procedure for the computation of multi-band responses, with a guaranteed optimal selectivity at fixed degree and number of transmission zeros. As opposed to the Tchebychev or quasi-elliptic filters, that are the optimal answer to this synthesis problem in the single band case, we showed that a succession of signed sub-problems needs to be solved here to ensure the global optimality of the computed response (see chapter 2). We begin with a description of the Belevitch form, a central mathematical structure common to all questions of filter synthesis and continue with the description of our multi-band synthesis procedure. A specific alternation property is presented in order to characterize optimal solutions of the signed quasi-convex sub-problems, in terms of alternating sequences of extremal points. For the reader familiar with Achieser's result on uniform real valued rational approximation on an interval, our approach consists in an adaptation of the latter to the multi-band case and to the solving of a Zolotariov problem of the third kind.

De-embedding techniques as well as frequency response synthesis procedures have in common that they all end up with a rational 2×2 scattering matrix that needs to be realized as a circuit to proceed further in the filter's tuning or synthesis process. The circuit used here is the low-pass prototype, which consists of ideally coupled resonators. The coupling topology, that is the way resonators are coupled, or not, one to another is crucial here. Whereas the realisation step is relatively simple for canonical topologies, deriving circuits with complex coupling architecture had long been considered as a complicated task. Optimization technique based on the determination of similarity transforms, or direct circuital optimization techniques, were used to tackle these problems: with no guarantee on the outcome, nor on its exhaustivity. At the time we started to work on this problem, it had been observed, that some coupling topologies admit multiple circuit realisations of the same response. We developed an algebraic approach tailored for this problem, together with an abstract framework clarifying the compatibility conditions between coupling topology and class of frequency responses. It is this approach and framework that we detail in chapter 3. We do this in much more details than for preceding chapters, as many of the proofs we present here, are published here for the first time. For this coupling matrix synthesis problem we have favoured a pragmatic and engineering approach, yielding eventually the tool Dedale-HF that is now widely used among practitioners to synthesise their circuits. Formal proofs, at the crossing of circuit theory and algebraic geometry, have often been discussed but never extensively written down: this is now done with this document.

The synthesis of multiplexers is among one of the most difficult problems occurring in the field of microwave device manufacturing. The fact that multiple channels interact together via a common manifold, where resonances might also take place, is one of the multiple complications of the problem. Another one comes from the fact that each channel filter of a multiplexer is plugged on a non purely resistive load: namely the one constituted by the manifold and all other channel filters. Filters can therefore no longer be synthesised with classical procedures where resistive loads at both accesses are assumed. We designed a synthesis procedure for filters

connected at one of their accesses to an unmatched load. The process guarantees perfect matching between filters and load on a set of n discrete points, where n is the degree of the filter. It appears therefore as a possible building block for a recursive approach to the multiplexer synthesis problem. This procedure has a very strong connexion to an aesthetic topic in Schur analysis: Nevanlinna-Pick interpolation. Synthesizing a matching filter with fixed transmission zeros amounts to solving a Nevanlinna-Pick interpolation problem with spectral constraints. Our main result, obtained by combining analyticity properties with Brouwer's invariance of the domain theorem, states that this interpolation problem has a unique solution. In chapter 4 of this document we give a brief description of the long lasting history of matching problems and state our interpolatory result, for its proof we refer to our paper that is reproduced in the bibliographic section.

The document is structured in different chapters that can be seen as introductions and descriptions of our work on bounded extremal problems, multi-band frequency synthesis, coupling matrix synthesis and matching filter synthesis. In each of these chapters we refer to some chosen articles we have written on these topics and that are reproduced in the last chapter of the document, the bibliographic section.

Chapter 1

Bounded extremal problems in Hardy spaces

1.1 Hardy spaces: a natural class for identification

The harmonic measurement of a stable, linear, time invariant, dynamical system is obtained by exciting the later with periodic entries, of the form $u(t) = cos(\omega_0 t)$, and measuring the periodic output signal $v(t) = A_0 cos(\omega_0 t + \phi_0)$, that settles in after a transient state. If H is the frequency response of the system, that is the Laplace transform of its impulse response, the harmonic measurement yields a measure of H at the point $i\omega_0$: we have $H(i\omega_0) = Ae^{i\phi_0}$. This is nearly tautological when looking at the input/output convolution equation ruling such systems,

$$v(t) = \int_0^t h(\tau)u(t-\tau)d\tau$$

which rewrites after setting in $u(t) = cos(\omega_0 t)$,

$$\begin{aligned} v(t) &= \Re\left(\int_0^t h(\tau)e^{i\omega_0(t-\tau)}d\tau\right) \\ &= \Re\left(e^{i\omega_0 t}\int_0^t h(\tau)e^{-i\omega_0\tau}d\tau\right) \\ &\underset{t\to\infty}{=} \Re\left(e^{i\omega_0 t}H(i\omega_0)\right) = Acos(\omega_0 t + \phi_0) \end{aligned}$$

where the convergence necessary for the last equation takes place for example under the system's BIBO stability hypothesis $h \in L^1$. In microwave electronics, this harmonic measurements are performed by a network analyzer, measuring periodic input and output power waves, entering and leaving the system. Thus pointwise measurements of the system's scattering matrix are obtained, which is the transfer function of interest in high-frequency electronics.

Harmonic measurement campaigns provide a discrete set of measurements,

$$(\omega_i, H(\omega_i)_{i=1\ldots n}).$$

System identification amounts to recovering from this partial measurements, a functional representation of the transfer function H. In the case of finite dimensional

systems, that is systems with a finite dimensional state space, the transfer function is a stable rational function, the degree of which relates to the minimal state space size of the system. Indeed BIBO stability is equivalent in this context to exponential stability, which in turn implies that the system's eigenvalue, that are also the poles of the transfer function, belong to $\Pi^- = \{s, \Re(s) < 0\}$. Now suppose that by means of an interpolation process, for example using splines, we are able to pass from the point-wise knowledge of H, to the knowledge of its values on a finite frequency interval I of the imaginary axis. An expansion of H in terms of power series can therefore be determined at some interior point of I, and by the analytic continuation principle, the value of H can be recovered everywhere on $\mathbb{C}$ but at its poles: hence H can be uniquely recovered. This formal mind experiment is however of low use in practice, as measurements are never devoid of errors: electronic noise is for example inevitable, when using a network analyser to measure microwave devices. Note also that the numerical process, passing from measurements at a discrete set of frequencies, to the knowledge of H on an interval, induces itself additional numerical imprecisions. An approximation scheme, taking into account these imperfections, is therefore needed.

Usual, but naive least square rational approximation schemes are problematic as they offer no guaranty on the stability of the obtained rational approximant. The problem here is: when seen as subset of the square integrable functions $L^2(I)$ of the interval I, the set of stable rational functions is not a closed set. This is seen at hand of the sequence $R_k(s) = 1/(s-(2i+1/k))$ of stable rational functions, that converges in $L^2(i[-1,1])$ to $R(s) = 1/(s-2i)$, which is obviously unstable. In particular, when starting from data $(\omega_i, R(\omega_i))$ with $i\omega_i \in I$, the problem of best stable rational approximation on I, has simply no solution; as shown by the minimizing sequence R_k. It is usually argued here, that these kind of weird behaviour never happens in practice, if the data are good enough, meaning originated with sufficient precision from a stable rational function. However, the notion for measurements of being, "the trace of a stable rational function", is a rather versatile one, as seen from following proposition.

Proposition 1.1.1 *Let f be a continuous function defined on a compact interval I of the imaginary axis. Let $\alpha \in \mathbb{C} \setminus I$. For all $\epsilon > 0$ there exists a proper rational function R having all its poles at α such that,*

$$\max_I |f - R| \leq \epsilon$$

.

We give a rapid sketch of the proof because it is instructive.

Proof. We verify that the algebra generated by the family $F = \{\left(\frac{s-1}{s+1}\right)^k, k \in \mathbb{Z}\}$ verifies all properties needed to apply the Stone-Weierstrass theorem. There exists therefore a polynomial p having as monomials the elements of F such $\max_I |f-p| \leq \epsilon/2$. p is holomorphic on $\mathbb{C} \setminus \{1, -1\}$ (an open set containing I) , and there exists therefore by Runge's theorem (on the Riemman Sphere) a proper rational function having its poles in α and such that $\max_I |R - p| \leq \epsilon/2$. □

Practically this means, that harmonic measurements on a finite interval I can either be seen, up to an ϵ, as originated from an unstable or stable rational function.

Eventually note that rational approximation by proper rational functions of fixed maximal degree is not a convex optimization problem: adding two rational functions of same degree, usually results in a rational function of higher degree. Removing the degree bound on approximants in order to obtain the convexity of the approximation set leads us to consider all proper stable rational functions. Now taking the closure of the later for a given norm, in order to avoid the aforementioned problem of sequences of elements converging outside the approximation set, brings us naturally to the notion of Hardy spaces.

Definition 1.1.1 *Let $p \geq 1$ and define H^p to be the space of functions f, holomorphic in the open right half plane and such that,*

$$\forall x > 0,\ \int_{-\infty}^{\infty} |f(x+i\omega)|^p \frac{1}{1+w^2} < \infty.$$

For $p = \infty$ define H^∞ to be the space of functions f, holomorphic in the open right half plane and such that,

$$\forall x > 0,\ \sup_{\omega \in]-\infty,\infty[} f(x+i\omega) < \infty.$$

Eventually define $\mathcal{A}$ the space of holomorphic functions on Π^+ and continuous on $\overline{\Pi^+}$ as well as at ∞ (in the sense that $f(1/z)$ is continuous on a neighborhood of 0 in $\overline{\Pi^+}$).

We define the Hardy spaces of the left half plane mutatis mutandis, and denote them $\overline{H^p}$.

In latter definition we introduced a weight $\frac{d\omega}{1+\omega^2}$ on the imaginary axis, in order to handle proper transfer functions with non-vanishing limits at ∞. From now on, and unless specified otherwise the L^p spaces are the Banach spaces defined according to the corresponding weighted norm, i.e

$$||f||_{L^p(i\mathbb{R})} = \left(\int_{-\infty}^{+\infty} |f(i\omega)|^p \frac{d\omega}{1+\omega^2} \right)^{1/p}.$$

For an interval $I = i[a,b]$ of the imaginary axis, we also note:

$$||f||_{L^p(I)} = \left(\int_{I} |f(i\omega)|^p \frac{d\omega}{1+\omega^2} \right)^{1/p}.$$

If f is a function defined on the imaginary axis, we note $f_{|I}$ its natural restriction to interval I. The elementary properties of these spaces, which can be found in [1], [2], [3] are,

Proposition 1.1.2 *(i) For all p, Hardy functions admit non-tangential limits $\hat{f}$ at almost every point of the imaginary axis. Cauchy (and also Poisson) representation formulas hold and set in a one to one correspondence Hardy functions and their limiting function $\hat{f}$ on the imaginary axis:*

$$For\ s \in \Pi^+,\ f(s) = \frac{1}{2\pi} \int_{-\infty}^{+\infty} \frac{\hat{f}(i\omega)}{i\omega - s} d\omega.$$

These non-tangential limits belong to the corresponding weighted L^p space of the imaginary axis. H^p can therefore be identified as a subspace of L^p, and for $1 < p < \infty$ we have the so called stable-unstable decomposition,

$$L^p = H^p \oplus \overline{H_0^p}$$

where $\overline{H_0^p}$ is the subset of functions of $\overline{H^p}$ that vanish at $z = -1$.

(ii) For all p, H^p is a Banach space. For $p < \infty$, H^p is the closure of the set of proper, rational, stable functions. The closure for the uniform norm L^∞ of the set of proper, rational, stable functions is $\mathcal{A}$.

(iii) If f is in H^p of Π^+ then $f(\frac{1+z}{1-z})$ is in the H^p of the disk (for a definition see [1])

We are now in a position to state our identification problem in the Hardy spaces, that we claim to be "nicer" sets than those of rational functions of given maximal degree. A direct formulation is the following: let $f \in L^p(I)$ be our data obtained by means of harmonic measurements (extended as discussed to a function), seek $g_0 \in H^p$ such that,

$$||(f - g_0)||_{L^p(I)} = \inf_{g \in H^p} ||(f - g)||_{L^p(I)}. \tag{1.1}$$

In view of proposition.1.1.1 this formulation is too naive, as we have that:

Proposition 1.1.3 *For $p < \infty$ we have,*

$$\inf_{g \in H^p} ||(f - g)||_{L^p(I)} = 0$$

and therefore g_0 only exists when f is already the trace of an H^p function.

Proof. For $p < \infty$ and $\epsilon > 0$, there exists a continuous function fc such that $||(f - fc)||_{L^p(I)} \leq \epsilon/2$. Now applying proposition.1.1.1 with $\alpha = -1$ we obtain a stable, proper, rational function R (belonging to every H^p) such that $||(R - fc)||_{L^p(I)} \leq \epsilon/2$. □

At this point the reader is entitled to ask what was really gained by shifting our identification problem, from rational functions, to the class of Hardy spaces. And indeed we moved from one ill posed problem to yet another one. Part of our answer lies in following proposition, which identifies the cause of ill-posedness, and paves the way to a regularized version of our identification problem.

Proposition 1.1.4 *Let J be the complementary interval of I on the imaginary axis, that is $J = i\mathbb{R} \setminus I$. For $1 \leq p \leq \infty$ suppose we have a sequence of functions $g_n \in H^p$ such that,*

$$\lim_{n \to \infty} ||(f - g_n)||_{L^p(I)} = 0$$

then either:

(i) f is exactly the trace of an H^p function on I, or

(ii) $\lim_{n\to\infty} ||g_n||_{L^p(J)} = \infty$

Proof. Suppose that the sequence (g_n) remains bounded on J. For $1 < p < \infty$ the reflexive nature of the considered H^p spaces [4] associated to the Banach-Alaoglu theorem, allows to extract a weakly convergent sequence, converging to an element $g \in H^p$ (H^p is weakly closed in L^p, as a strongly closed subset of the latter). The sub-sequence $g_{n|I}$ converges weakly to $g_{|I}$ in $L^p(I)$, and strongly to f on I, hence $f = g$ *a.e.* on I. For $p = 1, \infty$ a similar reasoning can be done by invoking convergence in the weak-* topology [5, 6, 7]. □

Proposition 1.1.4 clarifies, at least partly the situation, by pointing out, as the main problem, the absence of information relative to the transfer function on the frequency interval J. This appears to be problematic, despite the rigidity of the class of analytic functions we are considering. Adding additional constraints on the transfer function to be recovered, in order to regularize the identification problem, is the topic of the next section.

1.2 Bounded extremal problems

In view of proposition.1.1.4, a canonical way to regularize our identification problem is to bound the norm, on the interval J, of possible functional candidates for our transfer function. Let,

$$C_M = \{g \in H^p,\ ||g_{|J}||_{L^p(J)} \leq M\}$$

we define problem $B(p)$ to be,

$$B(p) : \text{Find } g_0 \in C_M \text{ such that } ||(f - g_{0|I})||_{L^p(I)} = \inf_{g\in C_M} ||(f - g_{|I})||_{L^p(I)}$$

Before going through the extensive literature about these problems, we state the obvious for the case $1 < p < \infty$. The existence of a solution runs exactly as in the proof of proposition 1.1.4: take a minimizing sequence $g_n \in C_M$ which by construction is bounded, extract a weak convergent sequence $g_{\phi(n)}$converging to an element g_0, and note that C_M is weakly closed because it is convex and strongly closed. The function g_0 is therefore in C_M and by definition of weak convergence

$$||(f - g_{0|I})||_{L^p(I)} \leq \lim_{n\to\infty} ||(f - g_{\phi(n)|I})||_{L^p(I)} = \inf_{g\in C_M} ||(f - g_{|I})||_{L^p(I)}.$$

The uniqueness of g_0 is a consequence of the strict convexity of the L^p norms. Proposition.1.1.4 proves eventually, that the constraint is saturated, that is $||g_{0|J}||_{L^p(J)} = M$.

Bounded extremal problems in class of analytical functions where first studied in [8]. The hilbertian case $p = 2$ was first studied in [9], while the cases $1 \leq p < \infty$ are treated in [10]. The more delicate case $p = \infty$ was treated in [5, 6]. First attempts to use this approach for the identification of microwave filters, were reported in [11] and in my thesis [12]. An interesting and slightly different approach where an L^∞ constraint is imposed to the approximant on the whole axis is presented in [13].

For reasons that will become clear later, we add some reference function in the definition of our admissible set C_M. That is for $h \in L^p(J)$ we define,

$$C_M^h = \{g \in H^p,\ ||(g_{|J} - h)||_{L^p(J)} \leq M\}$$

and the associated extremal problem $\hat{B}(p)$, by

$$\hat{B}(p) : \text{Find } g_0 \in C_M^h \text{ such that } ||(f - g_{0|I})||_{L^p(I)} = \inf_{g \in C_M^h} ||(f - g_{|I})||_{L^p(I)}.$$

The introduction of this reference function h is essentially due to ensure necessary conditions for the existence of a unique solution to $\hat{B}(\infty)$. Note that the knowledge of h, in phase and modulus, is however complicate to ensure in practice. We will later introduce a new type of bounded extremal problems, that do not require such demanding additional knowledge.

For problem $\hat{B}(p)$ we have following important results:

Proposition 1.2.1 *We denote by $f \wedge h$ the function equal to f on the interval I and to h on J. Let $C(i\mathbb{R})$ be the set of continuous functions defined on the imaginary axis and at infinity.*

(i) For $1 \leq p < \infty$, there exists a unique solution to $\hat{B}(p)$.

(ii) For $p = \infty$ and provided that $f \wedge h \in H^\infty + C(i\mathbb{R})$, there exists a unique solution to $\hat{B}(\infty)$.

(iii) Assume f is not the trace of an H^p function on I, and that for $p = \infty$ (ii) holds. Define following unconstrained convex optimization problem $\tilde{B}(p)$,

$$\tilde{B}(p) : \textit{Find } g_0 \in H^p \textit{ that minimizes } \psi_{f,h}(g) \stackrel{\text{def}}{=} ||(g-f)||_{L^p(I)} + \lambda ||(g-h)||_{L^p(J)}.$$

For every $M > 0$, let g_0 be the unique solution to $\hat{B}(p)$ then there exists $\lambda > 0$ such that g_0 is the unique solution to $\tilde{B}(p)$. The coefficient λ is the Lagrange multiplier associated to the constrained problem $\hat{B}(p)$.

Proofs can be found in [7] and [12]. Now let ϕ_λ be the outer factor, with modulus 1 on I and $\lambda > 0$ on J, we have,

$$\psi_{f,h}(g) = ||\phi_\lambda (f \wedge h - g)||_{L^p(i\mathbb{R})} = ||(f \wedge h)\phi_\lambda - g\phi_\lambda||_{L^p(i\mathbb{R})}.$$

The outer factor ϕ_λ is invertible in H^p, therefore when g ranges over H^p so does $g\psi_\lambda$. Minimizing $\psi_{f,h}$ over H^p, amounts therefore to find the best approximation in H^p to the function $(f \wedge h)\phi_\lambda$. This is a classical extremal problem in H^p described for example in [4]. For $p < \infty$ there exists a unique best approximation in H^p to any function $l \in L^p$. For $p = \infty$ this also holds provided that $l \in H^\infty + C(i\mathbb{R})$. We therefore note this best approximation operator from $L^p \to H^p$ by P_{H^p} (for $p = \infty$ we restrict the operator to $H^\infty + C(i\mathbb{R})$). If g_0 is the solution to $\hat{B}(p)$, and λ the associated Lagrange multiplier by (iii) of Proposition.1.2.1, it follows that:

$$g_0 = \frac{P_{H^p}\left((f \wedge h)\psi_\lambda\right)}{\phi_\lambda}. \tag{1.2}$$

As we have drastically increased our search space, by extending it from rational functions to functions in Hardy spaces, questions about the regularity of the recovered transfer function are crucial. The next result is of rather negative nature.

Proposition 1.2.2 *Suppose that $f \wedge h$ is 1-Lipschitz on $i\mathbb{R}$, then g_0 given by formula (1.2) in the cases $p = 2, \infty$, is discontinuous for all values of λ but $\lambda = 1$. In this case we have*

$$g_0 = P_{H^p}(f \wedge h),$$

which corresponds to the classical extremal problem in H^p (i.e posed on the whole axis, $i\mathbb{R}$).

Proof. The proof for $p = \infty$ can be found in [7, p.19] and follows from the Carleson-Jacob theorem, asserting that the best H^∞ approximation to a Dini continuous L^∞ function is continuous. For the case $p = 2$ the proof follows exactly the same path, using the similar property for the linear operator $P_{H^2} = P^+$, that can be deduced from properties on the conjugation operator [2, theorem p.108] □

At this stage it seems that we have created more problems than we actually solved. First, the practical choice of the reference function h is an open question. Second, conditions on the choice of λ so as to ensure some regularity on the identified transfer function are not known for general reference functions h. This renders inoperative the procedure that consists in solving $B(2)$ in view to eventually retrieve a rational transfer function. Numerical problems induced by the discontinuity of g are described in [12]. In what follows we will present our contributions to the design of a practical approach to our original identification problem. We first introduce a new type of bounded extremal problems of mixed type and then present, inspired by the philosophy gained from the latter, a class of restricted bounded extremal problems tailored for our purposes.

1.3 Contribution to the identification of transfer functions

1.3.1 Bounded extremal problems of mixed type

We refine the search space associated to the bounded extremal problem by constraining in a point-wise manner the approximant on the interval J. Let $M \in L^2(J)$ be a non-negative function defined on J, we set

$$W_M = \{g \in H^2, \forall s \in J,\ |g(s)| \leq M(s)\}.$$

We suppose moreover that $log(M)$ has a finite weighted L^1 norm on J. This is the least one can ask for, as the log-modulus function of any non-zero function in H^p is integrable [2, theorem 5.3]. Our mixed extremal problem formulates as,

$$E : Find\ g_0 \in W_M\ such\ that\ ||(f - g_0)||_{L^2(I)} = \inf_{g \in W_M} ||(f - g)||_{L^2(I)}.$$

We have chosen to measure the distance to the data in least-squares sense, as this norm is adapted to noisy measurements, while the constraint on the J interval comes usually from more structural considerations. For scattering measurements of passive systems a natural choice is $M = 1$. Indeed the law of energy conservation implies that the modulus of scattering coefficients can not exceed 1.

Problem E can always be normalised to an equivalent problem where $M = 1$. This follows from the observation: $g \in W_M$ if and only if $g\phi_{1/M} \in W_1$. We therefore reformulate a normalised version of problem E, by setting:

$$\tilde{W} = \{g \in H^2, \forall s \in J,\ |g(s)| \leq 1\}.$$

and,

$$\tilde{E} : \textit{Find } g_0 \in \tilde{W} \textit{ such that } ||(f - g_0)||_{L^2(I)} = \inf_{g \in \tilde{W}} ||(f - g)||_{L^2 I}.$$

The detailed study of $\tilde{E}$ is the object of two publications [14, 15], and we reproduce [14] in section 5.1.1. The first paper handles the case where I and J are intervals, while the second tackles the general case where I and J are only supposed to be measurable sets. The second article derives also a Carleson type formula, for the solution g_0, which is the analogue of formula (1.2). We give, in what follows, a summary of these results.

Proposition 1.3.1 *Problem $\tilde{E}$ has a unique solution g_0. Assume now that f is not the trace on I of a function in H^2,*

(i) The solution g_0 saturates the constraints everywhere on J, that is

$$\forall s \in J,\ |g(s)| = 1$$

(ii) Critical point equation: g_0 is solution to $\tilde{E}$ if and only if

(1) $\forall s \in J,\ |g(s)| = 1$

(2) there exists a non-negative (real valued) function $\lambda \in L^1(J)$ such that,

$$(g_{0|I} - f) \wedge \lambda g_{0|J} = (g_{0|I} \wedge \lambda g_{0|J} - f \wedge 0) \in \overline{H}^1_0 \tag{1.3}$$

(iii) Optimization problem and Carleson formula: let $\lambda \in L^1(J)$ the functional Lagrange multiplier associated to g_0 as described in ii.2 then,

(1) g_0 is the solution to following unconstrained convex optimization problem:

Minimize function ψ over H^2 given by:

$$\psi(g) = ||(f - g_{|I})||^2_{L^2(I)} + ||\lambda^{1/2} g_{|J}||^2_{L^2(J)} = ||(f \wedge 0)\phi_{\lambda^{1/2}} - g\phi_{\lambda^{1/2}}||^2_{L^2(i\mathbb{R})}$$

(2) Following Carleson formula holds,

$$g_0 = \frac{P_{H_2}\left((f \wedge 0)\phi_{\lambda^{1/2}}\right)}{\phi_{\lambda^{1/2}}}.$$

Proof. The detailed proof can be found in [15, 14]. The strong saturation of the constraint (i) is obtained by means of an original multiplicative variational argument involving tailored outer functions. The main difficulty in the derivation of the critical point equation of $\tilde{E}$ comes from the non-differentiable nature of the infinity norm constraint on the interval J: the boundary of $\tilde{W}$ is not a smooth manifold. In [14], the latter is obtained by a limiting argument where $\tilde{E}$ is first considered on the

set of polynomials of degree less than n. For the latter, a critical point equation in terms of sub-gradient is obtained thanks to the classical Carathedory theorem. Letting $n \to \infty$ yields a limiting critical point equation involving a measure, which thanks to the F. an M. Riesz theorem (which asserts that any analytic or anti-analytic measure is in fact a function), turns out to be our Lagrange multiplier function λ. The assertion (iii.1) is a direct consequence of (ii.2), as the critical point equation associated to the unconstrained optimization problem is exactly equation (1.3). Assertion (iii.2) is a consequence of (iii.1). Eventually a complete version of the proposition is proven in the case of general sets I and J, and a Lagrangian relaxation approach is developed in infinite dimension in [15]. □

1.4 Improved regularity: a finite dimensional approach

Proposition 1.3.1 entails a strong message: every template function $M(s)$ produces a candidate g_0 for the transfer function which has exactly modulus $M(s)$ on J. The a priori knowledge of the modulus of the transfer function on J, is therefore of primary importance. This is intuitively known by practitioners when they mention the necessity to perform harmonic measurements on a frequency band I, "large enough to capture the whole system's dynamic". This is another way of saying that the remaining behaviour of the transfer function, and in particular of its modulus, on the interval J should remain very "simple".

A drawback of the identification problem posed in Hardy classes, is that latter spaces are "too big". When the underlying dynamic is of finite dimension, the associated transfer function is a stable rational function, which on the imaginary axis is C^∞ and in particular continuous. As opposed to this, we have seen that, except for very particular adjustments of the associated Lagrange multiplier (which are for the moment not well understood), the solutions g_0 of bounded extremal problems in H^p are in general discontinuous at the end points of the interval I. This leads to severe approximation problems when rational models are obtained in a second step, via rational approximation of g_0. Rational functions obtained this way will typically exhibit unrealistic poles in Π^-, near the end points of I, that hardly have anything to do with the system's original dynamic.

We have therefore developed a finite dimensional approach for our reconstruction problem. Let A_n be an n dimensional vector space spanned by a basis of n functions of class C^1 defined on the interval J. We also ask that these functions be C^1 on a neighbourhood of ∞, that is write as smooth functions of the variable $1/s$ for s big enough. The space A_n will be the space where we look for possible completions of our partial harmonic measurements represented here by the function f of class C^1, on $I = [i\omega_0, i\omega_1]$. Let $M(i\omega)$ be a non-negative, bounded, continuous function defined on J. We define,

$$K_{M,n} = \{h \in A_n, \text{ such that } h(\omega_0) = f(\omega_0),\, h(\omega_1) = f(\omega_1),\, \forall s \in J,\ h(s) \leq M(s)\},$$

and the associated extremal problem:

$$F : \text{Find } h_0 \in K_{M,n} \text{ such that } ||P_{\overline{H_0^2}}(f \wedge h_0)||_{L^2(i\mathbb{R})} = \inf_{h \in K_{M,n}} ||P_{\overline{H_0^2}}(f \wedge h)||_{L^2(i\mathbb{R})}.$$

We have,

Proposition 1.4.1 *Provided the set $K_{M,n}$ possesses at least one element, problem F has a unique solution. In this case the function $g_0 = P_{H_0^2}(f \wedge h_0)$ is continuous on $i\mathbb{R}$ and at ∞ and*

$$||f - g_0||_{L^2(I)} \leq ||P_{\overline{H_0^2}}(f \wedge h_0)||_{L^2(i\mathbb{R})}.$$

Proof. $K_{M,n}$ is a closed, bounded (M is bounded) subset of the finite dimensional space A_n and therefore compact, for any topology induced by one of the equivalent L^p norms. The function $h \in L^2(J) \to ||P_{\overline{H_0^2}}(f \wedge h)||_{L^2(i\mathbb{R})}$ is continuous and attains therefore its minimum on $K_{M,n}$. Suppose now, that this minimum is attained at two distinct points h_0 and h_1. Suppose first that this minimum is zero: then $||P_{\overline{H_0^2}}(0 \wedge h_0 - h_1)||_{L^2(i\mathbb{R})} = 0$, that is $(0 \wedge (h_0 - h_1)) \in H^2$, which implies $h_0 = h_1$, a contradiction. The minimum is therefore strictly positive, and we have,

$$\begin{aligned} ||P_{\overline{H_0^2}}(f \wedge \frac{1}{2}(h_0 + h_1))||_{L^2(i\mathbb{R})} &= ||P_{\overline{H_0^2}}(\frac{1}{2}(f \wedge h_0)) + P_{\overline{H_0^2}}(\frac{1}{2}(f \wedge h_1))||_{L^2(i\mathbb{R})} \\ &\leq \frac{1}{2}||P_{\overline{H_0^2}}(f \wedge h_0)||_{L^2(i\mathbb{R})} + \frac{1}{2}||P_{\overline{H_0^2}}(f \wedge h_1)||_{L^2(i\mathbb{R})} \\ &= ||P_{\overline{H_0^2}}(f \wedge h_0)||_{L^2(i\mathbb{R})}. \end{aligned} \tag{1.4}$$

As the function $(h_0 + h_1)/2$ is in $K_{M,n}$ the last inequality in (1.4) is an equality. None of the functions $P_{\overline{H_0^2}}(\frac{1}{2}(f \wedge h_0))$ and $P_{\overline{H_0^2}}(\frac{1}{2}(f \wedge h_1))$ is zero as otherwise the minimum of the cost function would be zero. Therefore the strict convexity of the L^2 norm implies,

$$\frac{1}{2}P_{\overline{H_0^2}}(f \wedge h_1) = \frac{1}{2}P_{\overline{H_0^2}}(f \wedge h_0)$$

which again yields, $P_{\overline{H_0^2}}(0 \wedge h_0 - h_1) = 0$, hence a contradiction. The minimizer h_0 is unique.

By construction, the function $f \wedge h_0$ is continuous on $i\mathbb{R}$, C^1 on the intervals I and J, and therefore on $i\mathbb{R}$ and ∞, but on the endpoints $i\omega_0$ and $i\omega_1$. It is therefore Lipschitz on $i\mathbb{R}$ and at ∞, and its analytical projection g_0 is continuous.

Eventually the last inequality comes from,

$$\begin{aligned} ||f - g_{0|I}||_{L^2(I)} &\leq ||f \wedge h_0 - g_0||_{L^2(i\mathbb{R})} \\ &= ||P_{\overline{H_0^2}}(f \wedge h_0)||_{L^2(i\mathbb{R})}. \end{aligned}$$

□

Problem F solves theoretically the discontinuity problem of g_0, at the end-points of the interval I. For microwave applications, where measurements might be perturbed by electronic noise, problem F relies too much on two single measurement points $f(\omega_1)$ and $f(\omega_2)$. We have therefore developed an alternative version of problem F, that is at heart of the software Presto-HF [16]. Suppose that the set of functions A'_n contains now functions with the same regularity as in A_n but defined on a slightly bigger closed interval J' containing neighbourhoods of ω_0 and ω_1. The set $R = I \cap J'$ is therefore made of two compact intervals of the form $R = [i\omega'_0, i\omega_0] \cup [i\omega_1, i\omega'_1]$. We define the set $H_{c,M',n}$ with $c > 0$ to be,

$$H_{c,M',n} = \{h \in A'_n, \forall s \in J',\ h(s) \leq M'(s) \text{ and } ||P_{\overline{H_0^2}}(f \wedge h))||_{L^2(i\mathbb{R})} \leq c\}.$$

and problem F',

$$F' : \text{Find } h_0 \in H_{c,M,n} \text{ such that } ||f - h_0||_{L^2(R)} = \inf_{h \in H_{c,M,n}} ||f - h||_{L^2(R)}.$$

The constant c is adjusted by considering the surrogate problem $\hat{F}'$,

$$\hat{F}' : \text{Find } c_0 \in \mathbb{R} \text{ such that } c_0 = \inf_{h \in K'_{M',n}} ||P_{\overline{H_0^2}}(f \wedge h))||_{L^2(i\mathbb{R})}.$$

where the set $K'_{M',N}$ is defined by,

$$K'_{M,n} = \{h \in A'_n, \text{ such that } \forall s \in J', \ h(s) \leq M'(s)\}.$$

Solving F' is considered for $c \geq c_0$. Problems F' and $\hat{F}'$ are both convex minimzation problems. The strict convexity of the L^2 norm, combined with the convexity of the set $H_{c,M',n}$ lead to the result that F' has a unique solution.

1.5 Practical application

As already mentioned, problem F' is at heart of the functioning of the software Presto-HF [16], dedicated to the identification of microwave filters. The space A'_n is here specialized to be

$$A'_n = span\{1, 1/s \dots 1/s^{n-1}\}.$$

The response's completion is here sought as a finite Taylor expansion at infinity. This is justified, because measurements of the filters are made in the normalized domain on an interval of type $I = [-\omega_0, \omega_0]$ wide enough to capture most of the filter's dynamic: outside the measurement range the reponse behaviour is supposed to be simple. The main advantage of our approach is, that, once problem F' is properly solved, a rational approximation of our data is obtained at nearly no cost using AAK [17, 18, 19] theory and its implementations [20]. The idea is here that once the element

$$g = P_{H^2}(f \wedge h)$$

has been computed as the stable representation of our data, the Hankel operator $\mathcal{H}_g : H^2 \to \overline{H^2}$ defined by $\mathcal{H}_g(v) = P_{\overline{H^2}}(g^* v)$ is considered (here $g^* \in \overline{H^\infty}$ represents the para-conjugate of the H^∞ function g). Kronecker's theorem asserts that when g is rational and stable, $\mathcal{H}_g$ is of finite rank. It is therefore no surprise that a singular value decomposition of the operator $\mathcal{H}_g$ is related to a stable rational approximation of g: this is exactly what AAK theory is about. Computationally, this amounts to a singular value decomposition of the matrix representing $\mathcal{H}_g$ in a proper basis: an operation, that if not simple, is nowadays very efficiently handled by most mathematical systems and numerical libraries on computers. Note also that this technique adapts in a straight-forward manner to matrix valued identification problems, originating from MIMO systems. This step can be followed by an L^2 rational approximation step, for which approximation engines exist as, for example the software RARL2 [21, 11, 22] based on Schur analysis. This overall procedure is detailed in [23], which is reproduced in section 5.1.2. In latter article a de-embedding

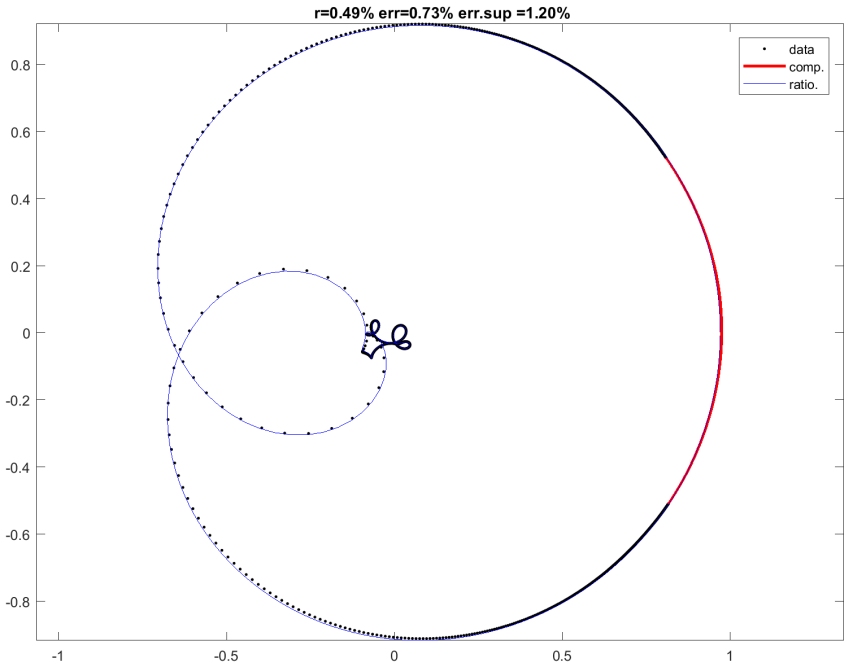


Figure 1.1: Nyquist plot of measurements of an 8^{th} degree S_{11} scattering parameter of a cavity filter. The dots correspond to frequency measurements, red continuous line to the completion computed by solving problem F' and the continuous blue line to a rational approximant computed by AAK theory.

procedure based on problem $\hat{F}'$ is presented to identify delay components, due to access feeds used to measure microwave equipments.

On Figure 1.1 the result of our procedure relying on solving problem F' is illustrated at hand of measurements coming from an 8^{th} order cavity filter. The constant c of problem F' is chosen here to be 0.49 percent of the quantity $||f||_{L^2(I)}$, while the maximal degree of the expansion at infinity is 4 (i.e. $n = 5$). The 2×2 matricial identification of the same measurements is presented on Figure 1.2. The computational time on a computer equipped with a pentium i7 processor, of the whole matricial identification procedure, is in this case of 0.4 *sec.* . This paves the way to applications where an identification procedure running in real time is needed: this is for example the case in fully automised tuning procedures for microwave filters based on the use of tuning robots. Presto-HF is currently in use by the microwave device manufacturers Thales Alenia Space in Toulouse and Madrid, Thales système aéroportés (Paris), Flextronics (US), Inoveos (France) and by our academic partner Xlim. Applications of this techniques to more complex de-embedding problems involving multiplexers have also been considered, in particular during the PhD of Mateo Oldoni that I co-supervised [24], [25], [26], [27].

Eventually, we also reproduce in section 5.1.3 recent published work [28], based on the stable/unstable decomposition of $L^2 = H^2 \oplus \overline{H_0^2}$, and applied to instability detections in power amplifiers.

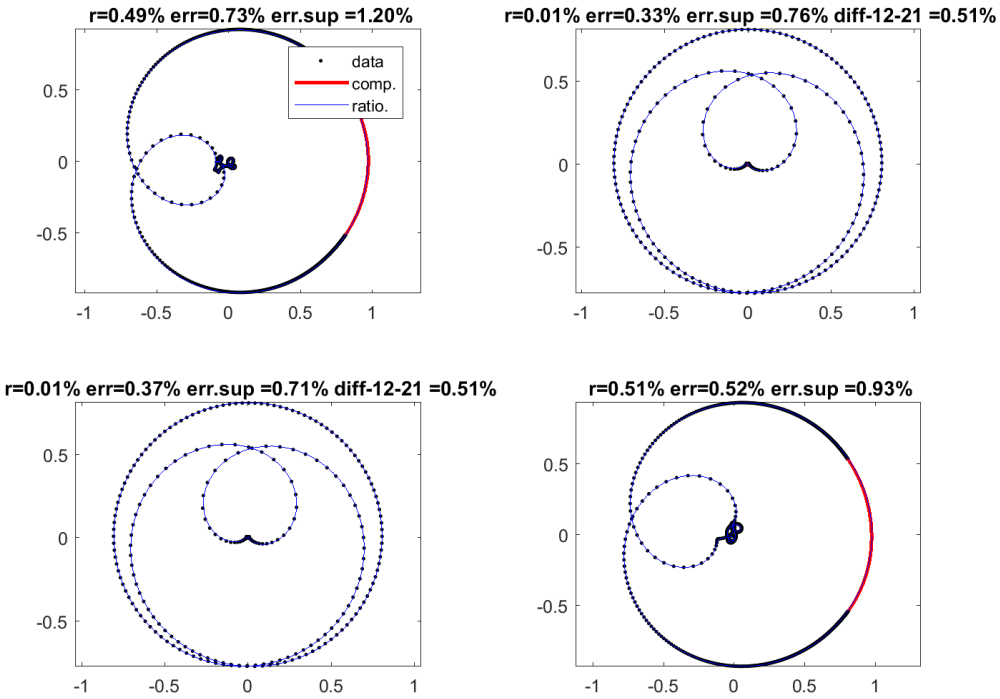


Figure 1.2: Nyquist plot of measurements of an 8^{th} degree, 2×2 scattering matrix, of a cavity filter. The dots correspond to frequency measurements, the red continuous line to the completion computed by solving problem F' for each scattering element and the continuous blue line to a rational, matricial, approximant computed by AAK theory.

Chapter 2

Synthesis of optimal multi-band frequency responses

2.1 The Belevitch form

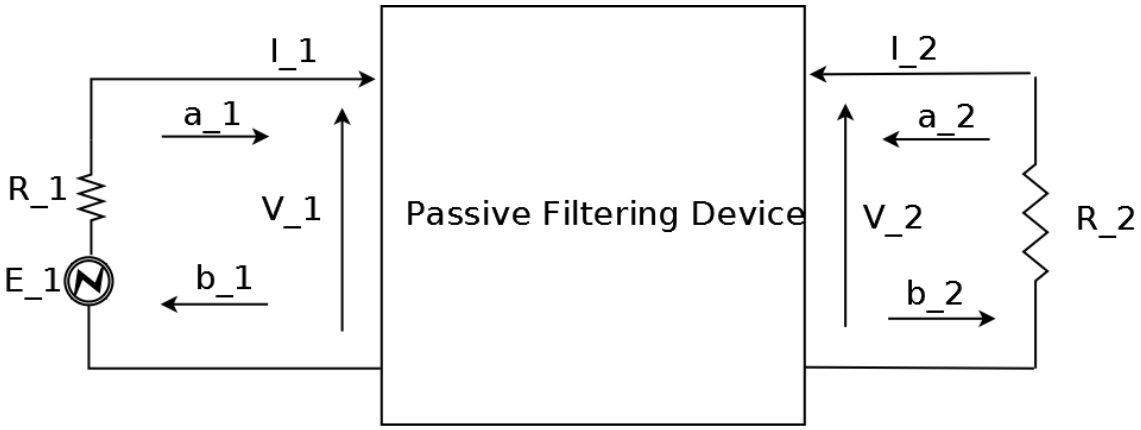


Figure 2.1: Two port filter, as a scattering device. (a_1, a_2) are the incoming power waves, while (b_1, b_2) are the scattered ones.

A microwave filter can be seen as a two port, scattering device. The incident power wave a_1 is partly reflected back two port 1, yielding signal b_1, and partly transmitted to port 2, generating signal b_2, see Figure. 2.1. In the frequency domain, the linear time invariant nature of the device leads to:

$$\begin{aligned} b_1 &= S_{1,1}a_1 \\ b_2 &= S_{2,1}a_1. \end{aligned} \tag{2.1}$$

The coefficient $S_{1,1}$ is called the filter's reflection while $S_{2,1}$ is called the transmission. When port 2 is also excited by a signal a_2 following general scattering equation holds,

$$\begin{pmatrix} b_1 \\ b_2 \end{pmatrix} = \begin{pmatrix} S_{1,1} & S_{1,2} \\ S_{2,1} & S_{2,2} \end{pmatrix} \begin{pmatrix} a_1 \\ a_2 \end{pmatrix}.$$

Filters are usually made of reciprocal materials and this imposes $S_{1,2} = S_{2,1}$: the reciprocity property of the device implies the symmetrical nature of its scattering

matrix. Filters are passive devices, that is, they have no internal source of energy. The scattering elements belong therefore to H^∞ but one can say a bit more. The average scattered electrical power carried by the output signal $b = \begin{pmatrix} b_1 \\ b_2 \end{pmatrix}$ cannot exceed the incoming one, brought in by the input signal $a = \begin{pmatrix} a_1 \\ a_2 \end{pmatrix}$, and as a consequence in terms of matrices we have:

$$\begin{array}{c} \forall \omega \in \mathbb{R},\ Id - \overline{S^t(i\omega)} S(i\omega) \succcurlyeq 0 \\ \forall (i,j) \in \{1,2\} \times \{1,2\},\ S_{i,j} \in H^\infty \end{array} \tag{2.2}$$

In the case where the filter can be modelled as a finite dimensional system, for example when it can be modelled as a coupled resonator network, its S matrix is rational. The MacMillan degree of the scattering matrix, corresponds to the minimal size of the state space necessary to realize it. If in addition the filter is considered to be loss-less, that is balanced in terms of averaged incoming and out-coming power, S becomes an inner matrix that verifies,

$$\forall \omega \in \mathbb{R},\ Id - \overline{S}^t(i\omega) S(i\omega) = 0. \tag{2.3}$$

In this case S admits a particular structure, called the Bellevitch form [29, 30]. We detail the latter and give a formal proof of this result which is central in filter synthesis. In the rest of the document, when G is a rational matrix, G^* represents its para-hermitian conjugate i.e. $G(s) = \overline{G}^t(-s)$, obtained by conjugating the fractions coefficients and changing the variable from s to $-s$.

Proposition 2.1.1 *Let S be a (2×2) rational inner matrix of McMillan degree n, S admits following polynomial structure:*

$$S = \frac{1}{q} \begin{pmatrix} p_1 & -e^{i\theta} p_2^* \\ p_2 & e^{i\theta} p_1^* \end{pmatrix}$$

where $p1, p2, q$ are polynomials and q is the unique stable unitary polynomial of degree n that solves following spectral equation:

$$p_1 p_1^* + p_2 p_2^* = q q^* \tag{2.4}$$

Proof. We give a proof of this result because it is instructive and the result is crucial in the field of filter synthesis.

The condition (2.3) occurring for infinitely many frequencies, and S being a rational function, following structural equality holds (in terms of rational functions):

$$S^* S = Id. \tag{2.5}$$

As S is stable, the function $det(S)$ is also a stable rational function, and equality (2.5) indicates that it takes uni-modular values on the imaginary axis. Hence $det(S)$ is a Blaschke product, that is there exists a stable polynomial q such that

$$det(S) = \frac{q^*}{q}.$$

Using Cramer's rule, we can express $S^* = S^{-1}$ in terms of the matrix $cof(S)$ of the co-factors of S:

$$S^* = cof(S)/det(S) \rightarrow S^* q^* = q.cof(S). \tag{2.6}$$

Poles of the expressions $S^* q^*$ are in Π^-, while poles of $q.cof(s)$ are in Π^+: hence these matrices have no poles, and there exists a polynomial matrix B such that $S^* = B/q^*$. Taking the para-conjugate on both side, we obtain the classical result that for any rational inner matrix S there exists a polynomial matrix N such that:

$$S = \frac{N}{q}$$

where q is a stable polynomial, defined (up to a unimodular constant) by $det(S) = q^*/q$.

The fact that $n = deg(q)$ corresponds to the McMillan degree of S is a bit more technical. As a rational matrix, S admits a Smith McMillan decomposition, that is there exist two square polynomial matrices A, B, that are unimodular (the determinant of which is a non-zero constant), a diagonal matrix D such that: $S = ADB$. The diagonal elements of D are rational functions $d_i = a_i/b_i$, with a_i b_i coprime polynomials, and such that a_i divides a_{i+1} and b_{i+1} divides b_i. The roots of the polynomial b_i are the poles of S, that is points where at least on element of S does not evaluate. The roots of the polynomials a_i are the zeros of S, that is points where S is a singular as a matrix. It is standard system theory [31, 32, 33], that the McMillan degree of S which corresponds to the dimension of any minimal state space realization of S is also equal to the quantity $max(deg(\prod a_i), deg(\prod b_i))$. By definition we have

$$det(S) = \frac{\prod a_i}{\prod bi}.$$

We therefore need to prove that no simplification takes place between $\prod a_i$ and $\prod b_i$. If any, the later takes place for a pair (l, m) of indices with $l \neq m$ (because a_i and b_i are coprime) such that a_l and a_m have a common root z_0. The later is by definition a pole of S and must therefore lie in Π^-. The Smith-McMillan decomposition of S^{-1} is (up to a reordering) given by $S^{-1} = B^{-1} D^{-1} A^{-1}$, which indicates that z_0 is also a pole of S^{-1}. But $S^{-1} = S^*$, and therefore poles of S^{-1} belong to Π^+: a contradiction.

Eventually we will particularize this result to the case where S is of size 2×2. Previous argumentation is to the effect that there exists a polynomial matrix,

$$N = \begin{pmatrix} p_1 & p_3 \\ p_2 & p_4 \end{pmatrix}$$

such that $S = S/q$. Now from Cramer's rule we obtain:

$$\begin{aligned} S_{2,2}/det(S) = S^*_{1,1} \rightarrow \frac{p_4}{q} = \frac{q^*}{q}\frac{p_1^*}{q^*} \rightarrow p_4 = p_1^* \\ -S_{2,1}/det(S) = S^*_{1,2} \rightarrow -\frac{p_3}{q} = \frac{q^*}{q}\frac{p_2^*}{q^*} \rightarrow p_3 = -p_2^* \end{aligned} \tag{2.7}$$

which yields the desired result after a normalization of q (define $det(S) = e^{i\theta} q^*/q$ with q unitary). The spectral equation (2.4), also called Feldkeller equation [34],

derives directly from the unitary condition of the first column of S and the famous spectral theorem that states that any positive polynomial p can be written as $p = uu^*$ with u a stable polynomial. □

In the case of filters the reciprocity condition is usually imposed, that is $S_{1,2} = S_{2,1}^*$ together with a normalization condition $\lim_{s\to\infty} S(s) = Id$. In this case we have the obvious specialization of previous proposition.

Proposition 2.1.2 *If S is a 2×2 reciprocal inner matrix of McMillan degree n, with $\lim_{s\to\infty} S(s) = Id$, then S admits the following form:*

$$S = \frac{1}{q}\begin{pmatrix} p_1 & p_2 \\ p_2 & (-1)^n p_1^* \end{pmatrix}$$

where p_2 is of degree strictly less than n and is auto-reciprocal: it verifies $p_2 = (-1)^{n+1} p_2^*$. The polynomial p_2 is called the transmission polynomial, and its zeros the transmission zeros. These, when not on the imaginary axis, come in pairs that is, if $z_1 = a + ib$ is a transmission zero, so is $z_1' = -(\overline{z_1}) = -a + ib$. The polynomial p_1 is unitary and q is the unique stable unitary polynomial of degree n that solves following spectral equation:

$$p_1 p_1^* + p_2 p_2^* = qq^* \tag{2.8}$$

Proof. Direct adaptation of previous proposition. □

2.2 Frequency design

When designing a microwave filter, one of the important tasks is to design its frequency response. The number of realizable transmission zeros is often related to the complexity of the final hardware implementation. The total degree of the response will have an impact on the footprint of the filter, as well as on electrical losses taking place inside the filter. The design of the filter's frequency response amounts therefore to solving the following problem: given a total number of available transmission zeros, find the frequency response of minimal degree that fulfils some frequency specifications, that is which is admissible for this specifications. In some design applications the locations of the transmission zeros are imposed by the user, in others they are part of the free design parameter. In any of these cases, in order to find the minimal degree for which there exists an admissible response, a simpler problem needs to be recursively solved.

P_1 : for a given degree, and a given total number of transmission zeros (or for transmission zeros with a specified location), does there exist an admissible response ? If yes, compute it.

We will summarize here our contribution to this problem in the case of multi-band frequency design: the corresponding paper [35] is reproduced in section 5.2 and has been part of the work of Vincent's Lunot PhD that I co-supervised [36]. Most classical filter responses are designed by specifications holding on two sets of frequencies: the pass-band where the modulus of the transmission needs to be maximized, and the stop-band where the rejection needs to be maximized or equivalently, when no losses are considered, where the modulus of the transmission needs to be minimized. The model used in first approximation for the filter's scattering responses, is

usually the loss-less rational and reciprocal one, corresponding to proposition 2.1.2. The key remark here is that the modulus of the transmission or reflection terms, can be expressed directly in terms of the transmission and reflexion polynomials p_1 and p_2: no spectral factorisation is involved here, as the polynomial q is not needed to evaluate the modulus of these scattering elements. This comes from,

$$\begin{aligned} \forall \omega \in \mathbb{R},\ |S_{12}(i\omega)|^2 &= \left|\frac{p_2(i\omega)}{q(i\omega)}\right|^2 \\ &= \frac{|p_2(i\omega)|^2}{q(i\omega)q^*(i\omega)} \\ &= \frac{|p_2(i\omega)|^2}{|p_2(i\omega)|^2 + |p_1(i\omega)|^2} \\ &= \frac{1}{1+\left|\frac{p_1(i\omega)}{p_2(i\omega)}\right|^2}. \end{aligned} \tag{2.9}$$

We have the complementary equation for the modulus of the reflexion,

$$\forall \omega \in \mathbb{R},\ |S_{11}(i\omega)|^2 = \frac{1}{1+\left|\frac{p_2(i\omega)}{p_1(i\omega)}\right|^2}. \tag{2.10}$$

The rational function $p1/p2$ is called the filtering function of the filter's loss-less response. In this work we will consider the set of polynomials p that are real valued, or purely imaginary valued on the imaginary axis, that is verify $p^* = \pm p$. Seeing these polynomials as functions of ω, instead of $s = i\omega$, we set $P_1(\omega) = p_1(i\omega)$ and $P_2(\omega) = p_2(i\omega)$. On the real axis, real valued polynomials are just polynomials with real coefficients, that is elements in $\mathbb{R}[X]$. We denote by T_n the subspace of $\mathbb{R}[X]$ of polynomials of maximal degree n. To handle the case of purely imaginary valued polynomials that are needed for the transmission polynomial in the Belevitch form (2.1.2) for n even, we will just multiply by i the real valued polynomial p_2 obtained from our synthesis procedure: expressions of the modulus of the transmission and reflexion term of the filter (2.9) (2.10), are invariant by this operation.

Let I be an union of h_I compact disjoint intervals of the real line, that we will consider to be the pass-bands. Let also J be an union of h_J disjoint compact intervals of the real line, strictly disjoint from the I_k (i.e $I_k \cap J_m = \emptyset$) that will represent our stop-bands. We note

$$I = \bigcup_{i=1}^{h_I} I_i,\ J = \bigcup_{i=1}^{h_J} J_i.$$

Suppose that we want the modulus of transmission $|S_{1,2}|$ to be less than a given value $K < 1$ on the stop-bands. This amounts, by equation (2.10), to impose

$$\forall \omega \in J,\ \left|\frac{P_2(w)}{P_1(w)}\right| \leq \frac{K}{1-K} \stackrel{\text{def}}{=} M. \tag{2.11}$$

Denote by A the set of polynomial pairs (with proper degree bounds on P_1 and P_2) that satisfy previous inequality. The synthesis problem that amounts to find

the response with the best possible worst-case transmission level in the pass-bands under the rejection constraints $|S_{11}| \leq K$ in the stop-bands amounts to solve:

$$\min_{(P1,P2)\in A} \max_{\omega\in I} \left|\frac{P_1(\omega)}{P_2(\omega)}\right|.$$

Remark first, that if $(\hat{P}_1, \hat{P}_2)$ solves latter optimization problem, then $(\hat{P}_1, \hat{P}_2/M)$ solve the problem for $K = 1/2$. We will therefore, without loss of generality, consider solving a normalized version of our synthesis problem for the value $M = 1$. For a precise statement of the latter, let

$$A_{n,m} = \{P_1 \in T_n, P_2 \in T_m,\ \forall\omega \in J,\ \left|\frac{P_2(w)}{P_1(w)}\right| \leq 1\} \tag{2.12}$$

and set,

Problem $\mathcal{P}_1$: Find $(\hat{P}_1, \hat{P}_2) \in A_{n,m}$ such that

$$\max_{\omega\in I} \left|\frac{\hat{P}_1(\omega)}{\hat{P}_2(\omega)}\right| = \min_{(P_1,P_2)\in A_{n,m}} \max_{\omega\in I} \left|\frac{P_1(\omega)}{P_2(\omega)}\right|$$

For any two pairs of polynomials (P_1, P_2) and (P_1', P_2') we will identify them as equivalent if $P_1P_2' - P_1'P_2 = 0$.

Problem $\mathcal{P}_1$ is called a Zolotarev problem of the third kind, by the name of the Russian mathematician, student of Chebychev, who first formulated it. It has an explicit solution, in terms of elliptic functions, in the case $n = m$ and the intervals $I = [-1, 1]$ and $J =]-\infty, 1-\epsilon]\cup]1+\epsilon, \infty[$ and yields the elliptic filters called also Cauer filters or even Zolotarev filters [37, 38]. For the same intervals, and when the polynomial P_2 is fixed (so the search space reduces to a set of possible polynomials P_1), the solution is given by the quasi-elliptic filters, for which a polynomial recurrence scheme is known [34]. For the multiple pass-band problem no explicit solution is known. Our contribution is here the development of a procedure to solve optimally problem $\mathcal{P}_1$.

The existence of an optimal pair $(\hat{P}_1, \hat{P}_2)$ follows from a similar and classical reasoning used for the existence of a best rational approximation in L^∞ norm of a continuous function on a compact interval: see [39, chapter V, p. 107], [40], [36]. Now suppose that $(\hat{P}_1, \hat{P}_2)$ is such a solution where possible numerator denominator simplifications have been performed, that is $\hat{P}_1$ and $\hat{P}_2$ are coprime. The trivial polynomial pair $(1, 1)$, is admissible for $\mathcal{P}_1$, with a cost function of 1. From this we deduce that polynomial $\hat{P}_1$ is of constant sign on each of the intervals J_k and $\hat{P}_2$ is of constant sign on each of the $I_k's$. We therefore define a sign function σ defined on each of the intervals J_k and I_k and valued in $\{-1, 1\}$. We do not know in advance the optimal choice of signs for each of the intervals, but we will cycle through all possible $2^{h_I+h_J}$ choices of such a function. We denote $c_I(\sigma)$ (resp. $c_J(\sigma)$) the number of sign changes of σ on the intervals I_i (resp. J_i). For each such function σ, define

$$\begin{aligned} B_{n,m,\sigma} = A_{n,m} \cap \{P_1 \in T_n,\ \forall\omega \in J,\ \sigma(\omega)P_1(\omega) \geq 0\} \\ \cap\{P_2 \in T_m,\ \forall\omega \in I,\ \sigma(\omega)P_2(\omega) \geq 0\} \end{aligned} \tag{2.13}$$

and

Problem $\mathcal{P}_\sigma$: Find $(\hat{P}_1, \hat{P}_2) \in B_{n,m,\sigma}$ such that

$$\max_{\omega\in I} \frac{|\hat{P}_1(\omega)|}{\sigma(\omega)\hat{P}_2(\omega)} = \min_{(P1,P2)\in B_{n,m,\sigma}} \max_{\omega\in I} \frac{|P_1(\omega)|}{\sigma(\omega)P_2(\omega)}$$

Eventually for any pair $(P_1, P_2) \in A_{n,m}$ with $P_2 \neq 0$ and P_1 and P_2 co-prime, let

$$\lambda(P_1, P_2) = \max_{\omega\in I} \frac{|P_1(\omega)|}{\sigma(\omega)P_2(\omega)}.$$

We define two sets of extremal frequency points:

$$E^+(P_1, P_2) = \{\omega \in I, \frac{P_1(\omega)}{P_2(\omega)} = \lambda(P_1, P_2)\} \cup \{\omega \in J, \frac{P_1(\omega)}{P_2(\omega)} = -1\} \tag{2.14}$$

$$E^-(P_1, P_2) = \{\omega \in I, \frac{P_1(\omega)}{P_2(\omega)} = -\lambda(P_1, P_2)\} \cup \{w \in J, \frac{P_1(\omega)}{P_2(\omega)} = 1\}. \tag{2.15}$$

We call a finite sequence of consecutive points $(\omega_1 < \omega_2 .... \omega_{l-1} < \omega_l)$ alternant if they belong alternatively to $E^+(P_1, P_2)$ and $E^-(P_1, P_2)$. Our main result is,

Proposition 2.2.1 *Suppose $c_I(\sigma) \leq m$ and $c_J(\sigma) \leq n$, then $\mathcal{P}_\sigma$ has a unique solution $(\hat{P}_1, \hat{P}_2)$ of coprime polynomials. For a pair of coprime polynomials $(P_1, P_2) \in B_{n,m,\sigma}$ define $N = \max(deg(\hat{P}_1) + m, deg(\hat{P}_2) + n)$. The pair (P_1, P_2) is the optimal solution to P_σ iff there exists an alternant sequence of extremal points of length $N + 1$, $(\omega_1 < \omega_2 \ldots \omega_{N+1})$. The problem P_σ is quasi-convex and can be efficiently solved using linear programming. In the case where the polynomial P_2 is fixed (with no zeros on I) similar results hold true for the obviously adapted version of problem $\mathcal{P}_\sigma$, this time with $N = n + 1$ (as no numerator-denominator degeneracy occurs) and the resulting convex optimisation problem can be solved using a Remez type algorithm.*

For a detailed proof see [36]. This result can be seen as a version of Achieser's result for uniform rational approximation on a single interval adapted to Zolotariov's third kind multi-intervals problem. Application of this result to the design of multi-band microwave devices are detailed in [35] which is reproduced in section 5.2 and publications [41, 42, 43].

Chapter 3

The coupling matrix synthesis problem

3.1 Low-pass circuit prototype and canonical coupling topologies

One of the purposes of the scattering matrix identification procedure of last section, as well as of the computation of optimal filtering functions, is to provide eventually a circuit representation of the filter to be tuned or to be designed. Indeed, equivalent circuits possess strong connection with the physical filter. If the circuit's topology is well chosen, the circuit's resonators can be identified with physical resonating structures of the hardware under test or to be synthesised. Circuit element values like resonating frequencies and electromagnetic couplings can be linked to tuning elements of the filter, like screws, dimensions of coupling irises, spacing between micro-strip lines etc.... For derivations of these useful analogies between filtering devices and circuit models, and their use in engineering see [44, 45, 46, 47, 48, 34]. We will deal here with the coupled resonators low pass equivalent circuit, see Figure. 3.1, which underpins the synthesis of microwave filters. We give a brief description of this circuit.

- The rectangular blocks between resonators, denoted by $M_{i,k}$ are admittance inverters, also called couplings. They are quadripoles, and if we denote by (I_1, I_2) the currents, and (V_1, V_2) the tensions, at the ports of an inverter of value M we have:
$$\begin{aligned} I_2 &= iMV_1 \\ I_1 &= iMV_2. \end{aligned} \tag{3.1}$$
They are idealised elements, as their value is independent of the frequency and are obtained from a "real" frequency dependent inverter under narrow band hypothesis, see [34, 12].

- Each low-pass resonator is composed of a unit capacitor of admittance $i\omega$, and a purely imaginary susceptance $jM_{i,i}$. The latter represents the shift, with respect to zero, of the resonating frequency of the resonator. Here again this element is obtained via a narrow band hypothesis, around the resonating frequency of a "real" (L, C) resonator.

- By tradition, the left port of the circuit is denoted by $\mathcal{S}$ (Source), while the right port is denoted by $\mathcal{L}$ (Load).

We denote by $X = (U_1, \dots U_N)^t$ the vector of tensions in each resonator, with $U = (U_{\mathcal{S}}, U_{\mathcal{L}})^t$ the vector of tensions and $I = (I_{\mathcal{S}}, I_{\mathcal{L}})^t$ the vector currents at the access ports, by M the $N \times N$ matrix of the couplings $M_{i,j}$, and by

$$B = \begin{pmatrix} M_{\mathcal{S},1} & M_{\mathcal{S},2} & \dots & M_{\mathcal{S},\mathcal{N}} \\ M_{\mathcal{L},1} & M_{\mathcal{L},2} & \dots & M_{\mathcal{L},\mathcal{N}} \end{pmatrix}^t$$

the $N \times 2$ vector of source and load to resonator couplings. Using Kirchhoff's law we have

$$\begin{cases} i(M + \omega Id_N)X + iBU = 0 \\ I = iB^tX \end{cases} \iff \begin{cases} i\omega X = -iMX - iBU \\ I = iB^tX \end{cases} \tag{3.2}$$

and taking iX as the new state, we obtain the symmetric state space system described in the time domain:

$$\begin{cases} \dot{X} = -iMX + BU \\ I = B^tX \end{cases} \tag{3.3}$$

The transfer function of which is the admittance matrix Y of the filter:

$$Y(s = i\omega) = B^t(i\omega Id_N + iM)^{-1}B = -iB^t(\omega Id_N + M)^{-1}B. \tag{3.4}$$

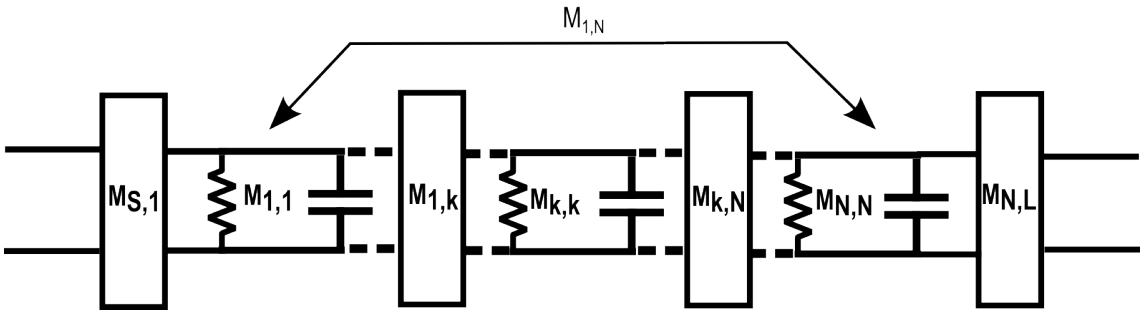


Figure 3.1: Low pass coupled resonators equivalent circuit.

The specification of the non-zero elements of the matrices M and B, that is the description of the coupling structure of the circuit, is called the coupling topology of the filter. In general terms, solving the coupling matrix problem, amounts, when starting from the admittance matrix Y of a filter (or equivalently S, $Y = (I-S)(I+S)^{-1}$), to find matrices B and M with a prescribed coupling topology such that (3.4) holds. Finding matrices B and M, with no specific constraint on their structure, is a classical realisation problem that was formalised by Kalman [49]. In particular for any strictly rational matrix of McMillan degree N the Ho-Kalman algorithm provides complex matrices B, A, C such that,

$$Y(s = i\omega)) = C(i\omega Id_N + iA)^{-1}B$$

holds. In order however to give a circuit interpretation of these matrices, C, A, B need to be real, A symmetric and $B = C^t$. In addition, the filter's implementation needs to have a reasonable complexity, hence the number of couplings to be realized has to be kept at its minimum. We are therefore facing a structured realisation problem.

One of the simplest topologies is the transversal form. It is obtained when M is diagonal and can be computed for any loss-less reciprocal admittance.

Proposition 3.1.1 *Let Y be a loss-less positive-real reciprocal, strictly proper admittance of McMillan degree N. The matrix Y admits the following partial fraction expansion with simple poles,*

$$Y = \sum_{k=1}^{N} \frac{R_k}{s - i\omega_k}.$$

The residues R_k are real, symmetric, non-negative matrices, that verify:

$$\sum_{k=1}^{N} rank(R_k) = N$$

.

The matrix Y admits a circuital realisation B, M, where M is diagonal and contains the values $-\omega_k$ on the diagonal. Each ω_k is repeated $rank(R_k)$ times on the diagonal. If (i, i) and $(i + rank(R_k) - 1, i + rank(R_k) - 1)$ are the corresponding indices between which ω_k is repeated, let B' the submatrix of B defined by (in matlab notations) $B' = B(i : i + rank(R_k) - 1, :)$ then,

$$R_k = (B')^t B'$$

.

Proof. The result is classic: we give a sketch of the proof because it is instructive. The admittance Y is loss-less, so in particular it is positive real in Π^+. Its poles are therefore in $\overline{\Pi}^-$. But the loss-less character of Y imposes that Y is purely imaginary on the imaginary axis, away from its possible poles. This yields,

$$Y + Y^* = 0 \iff Y = -Y^*.$$

But Y^* has its poles in $\overline{\Pi^+}$, hence all the poles of Y lie on the imaginary axis. Suppose without loss of generality that Y has a pole in zero. Near zero Y is essentially equal to its polar expansion of highest order, say m in zero, that is for $s = re^{i\theta}$

$$\begin{aligned} \forall \theta \in [-\pi, \pi],\ Y(re^{i\theta}) + \overline{Y^t}(re^{i\theta}) &= (R_0 + \overline{R_0^t} + o(re^{i\theta}))(\frac{1}{r^m e^{im\theta}} + \frac{1}{r^m e^{-im\theta}}) \\ &= (R_0 + \overline{R_0^t} + o(re^{i\theta}))\frac{2cos(m\theta)}{r^m} \succeq 0. \end{aligned} \tag{3.5}$$

Evaluated at $\theta = 0$ and for r sufficiently small, latter expression implies that R_0 is a non negative hermitian matrix. This in turn implies that $m = 1$, as otherwise the $cos(m\theta)$ term takes all values between $[-1, 1]$ when $\theta \in [-\pi/2, \pi/2]$, and the last inequality in (3.5) cannot hold true. Eventually the reciprocity hypothesis yields

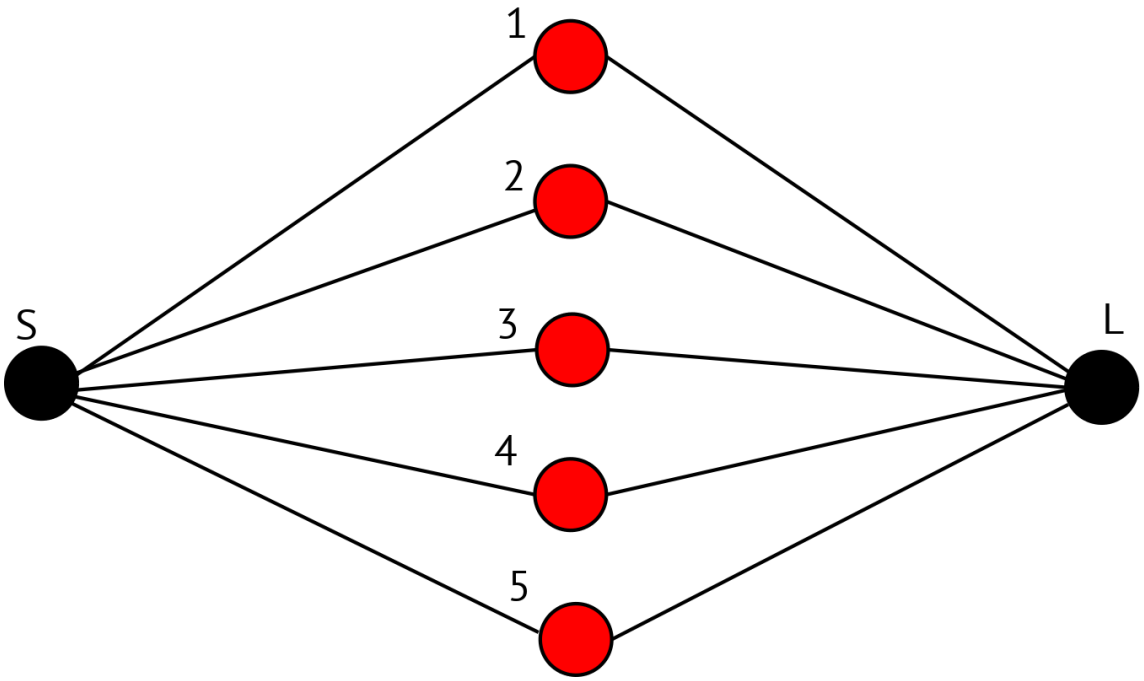


Figure 3.2: Transversal coupling topology.

that R_0 is a symmetric matrix, hence real valued. If $k = rank(R_0)$, the diagonal form of the non-negative quadratic form R_0 yields a real matrix B' of size $k \times 2$ such that $R_0 = (B')^t B'$. The rest of the proof is straightforward.

□

If one notes by a red bullet the resonators, by black ones the source and load nodes, we have proven that every loss-less admittance admits a realisation in transversal form, the schematic of which is described on Figure. 3.2. The transversal topology is however hard to implement in practice when N grows large due to the necessity of simultaneously couple all resonators to the input and output. Engineers have therefore developed coupling topologies that spread couplings among resonators. Some transformations are therefore needed to reconfigure the transversal topology in more practical ones. The transformation on the state space form, leaving the transfer function invariant, are well understood in Kalman's theory: they correspond to a change of basis on the state. In our case they can be further particularised, due to the reciprocity property. We define first what we mean by circuit realisation and describe then the associated transfer invariant transformations.

Definition 3.1.1 *We will say that a state space triple (A, B, C), where A is a complex $N \times N$ matrix, B is $N \times 2$ and C is $2 \times N$ is a circuit realisation if:*

- *A is symmetric, that is $A = A^t$*
- *$B = C^t$.*

In this case we speak of the circuit realisation (A, B). In what precedes, (M, B) is for example a circuit realisation. To every circuit realisation (A, B) we will associate its admittance function $Y(s)$ defined by,

$$Y(i\omega) = -iB^t(\omega Id - A)^{-1}B. \tag{3.6}$$

If (A, B) are real valued, simple arguments (diagonalise A) show that the associated admittance matrix Y is a loss-less positive real matrix function.

Proposition 3.1.2 *Suppose that (A, B) and $(\hat{A}, \hat{B})$ are two minimal circuital realisations of size N, having the same admittance function. There exists a non-singular $N \times N$ matrix P, such that*

$$P^t P = Id,\ \hat{A} = P^t A P,\ \hat{B} = P^t B. \tag{3.7}$$

If (A, B) and $(\hat{A}, \hat{B})$ are real valued, P is real and therefore an orthogonal matrix, also called a similarity transform in the filtering community.

Proof. The proof relies on elementary realisation theory. We give it here because it will be useful in further development. Let $\mathcal{O}_N$ and $\mathcal{C}_N$ (resp. $\hat{\mathcal{O}}_N$ and $\hat{\mathcal{C}}_N$) be the observability and controlability matrices of the realisation triple (A, B, B^t) (resp. $(\hat{A}, \hat{B}, \hat{B}^t)$). That is,

$$\mathcal{C}_N = (B, AB \ldots, A^{N-1}B) \tag{3.8}$$

and $\mathcal{O}_N = \mathcal{C}_N^t$. Equality in the admittance function, implies equality in the Markov parameters $B^t A^k B$ for all $k \geq 0$,which in turn implies

$$\mathcal{O}_N \mathcal{C}_N = \hat{\mathcal{O}}_N \hat{\mathcal{C}}_N. \tag{3.9}$$

Minimality implies that the observability and controlability matrices are full rank, they admit therefore a pseudo inverse $\hat{\mathcal{O}}_N^{\#}$ and $\hat{\mathcal{C}}_N^{\#} = (\hat{\mathcal{O}}_N^{\#})^t$. From (3.9) we obtain,

$$\hat{\mathcal{O}}_N = \mathcal{O}_N \mathcal{C}_N \hat{\mathcal{C}}_N^{\#},$$

and,

$$\hat{\mathcal{C}}_N = \hat{\mathcal{O}}_N^{\#} \mathcal{O}_N \mathcal{C}_N. \tag{3.10}$$

Now we set $P = \mathcal{C}_N \hat{\mathcal{C}}_N^{\#}$ a $N \times N$ matrix, mechanically we have $P^t = \hat{\mathcal{O}}_N^{\#} \mathcal{O}_N$, and we verify that

$$P^t P = \hat{\mathcal{O}}_N^{\#} \mathcal{O}_N \mathcal{C}_N \hat{\mathcal{C}}_N^{\#} = \hat{\mathcal{O}}_N^{\#} \hat{\mathcal{O}}_N \hat{\mathcal{C}}_N \hat{\mathcal{C}}_N^{\#} = Id$$

which proves that P is invertible with inverse P^t. Now (3.10) implies $\hat{B} = P^t B$. Eventually from

$$\mathcal{O}_N A \mathcal{C}_N = \hat{\mathcal{O}}_N \hat{A} \hat{\mathcal{C}}_N$$

we conclude that $\hat{A} = P^t A P$. If the realisations are composed of real valued matrices, the expression for P indicates that it is also the case for P which is therefore an orthogonal matrix. □

In the rest of this chapter we will use the denomination "similarity transform" for what mathematicians understand as "orthogonal matrix", that is a real valued square matrix P such that $P^t P = Id$. The engineering literature is therefore rich in articles that describe in detail the computation of similarity transforms that allow to obtain specific types of coupling topologies, when starting from the transversal one or other canonical topologies. The most famous ones are those of R.J. Cameron, see [50, 51, 52] but also those P. Macchiarella and S. Tamiazzo [53].

We will now describe a second canonical form, which is more suitable for applications and at heart of further developments for the computation of arbitrary topologies.

Proposition 3.1.3 *Let Y be a loss-less 2×2 strictly proper reciprocal admittance matrix of McMillan degree $N \geq 1$. Let*

$$Y = \sum_{k=1}^{\infty} \frac{G_k}{s^k},$$

be the formal development of Y at ∞ where the matrices G_k are called the Markov parameters. Suppose G_1 is diagonal, that is:

$$(G_1)_{1,2} = (G_1)_{2,1} = 0. \tag{3.11}$$

If $N = 1$ we define our canonical form (B, M) to be the transversal one. Condition (3.11) imposes in this case that only one of the elements of Y, a diagonal one $(1,1)$ or $(2,2)$, is non zero of degree one, while all others vanish. Y is therefore essentially a scalar response.

Suppose now that $N \geq 2$. A similar situation to the preceding scalar case occurs, if we suppose that one of the diagonal elements of G_1 vanishes, that is $(G_1)_{1,1} = 0$ or (exclusive) $(G_1)_{2,2} = 0$. The admittance Y admits a real valued circuital minimal realisation (B, M) where:

- *B has a single non vanishing element, $B_{1,1}$ or $B_{N,2}$*
- *The only possibly non-vanishing elements of M are its diagonal, sub and sur-diagonal (elements $M(i, i+1)$ and $M(i+1, i)$)*
- *None of the sur- or sub-diagonal elements vanishes.*
- *(M, B) is unique up to a similarity transform of the form $P = diag(\pm 1, \cdots \pm 1)$ in the sense that it is the only realisation of Y such that the unique non-vanishing element of B is $B_{1,1}$ or (exclusive) $B_{N,2}$ and M non vanishing elements are situated on the diagonal, the sur- and sub-diagonal.*
- *Y has only one non-zero element, a diagonal one, and is therefore essentially scalar.*

Suppose now that none of the diagonal elements of G_1 vanishes. The admittance Y admits a real valued circuital minimal realisation (B, M), where:

- *B has two non vanishing elements, $B_{1,1}$ and $B_{N,2}$*
- *The only possibly non-vanishing elements of M are its diagonal, sub and sur-diagonal (elements $M(i, i+1)$ and $M(i+1, i)$), as well as its last line and column (elements (N, i) and (i, N)).*
- *Only one element of the sub and sur-diagonal can vanish. In this case, say we have $M_{i_0-1,i_0} = M_{i_0,i_0-1} = 0$, then $M(i_0 : N-2, N) = M(N, i_0 : N-2) = 0$.*
- *Suppose the $1 < k \leq N-1$ first Markov parameters are diagonal, then $M(1 : k-1, N) = M(N, 1 : k-1) = 0$. In particular if Y corresponds to a scattering matrix S with no transmission zeros at finite frequencies, then M has a purely inline topology (inline means that the resonators are coupled in a row).*

- *If $Y_{1,2} = Y_{2,1} = 0$ then $M(N, 1 : N-2) = M(1 : N-2 : N) = 0$ and one of the sur- or sub-diagonal elements is 0.*
- *The realisation is "unique" up to a signed diagonal similarity transform of the form $P = diag(\pm 1, \cdots \pm 1)$. Here "unique" means, that the sole circuit realisation (B, M) of Y that has zero elements as previously specified.*

Here is a schematic of the matrix M in the regular case (all $M_{i,i+i} \neq 0$),

$$\begin{pmatrix} + & * & 0 & 0 & 0 & 0 & + \\ * & + & * & \ddots & \ddots & \ddots & + \\ 0 & * & + & \ddots & \ddots & \ddots & + \\ 0 & \ddots & \ddots & \ddots & \ddots & \ddots & + \\ 0 & \ddots & \ddots & \ddots & + & * & + \\ 0 & \ddots & \ddots & \ddots & * & + & * \\ + & + & + & + & + & * & + \end{pmatrix}$$

where the $*$ symbol has been used for non zero elements, and $+$ for possibly non vanishing ones. Because of the arrow shape of the coupling matrix M, this topology is often called the arrow form. Here is a schematic of the degenerate case, where $M_{i_0,i_0-1} = 0$.

$$\begin{pmatrix} + & * & 0 & 0 & 0 & 0 & + \\ * & + & * & \ddots & \ddots & \ddots & + \\ 0 & * & + & \mathbf{0} & \ddots & \ddots & + \\ 0 & \ddots & \mathbf{0} & \ddots & * & \ddots & \mathbf{0} \\ 0 & \ddots & \ddots & * & + & * & \mathbf{0} \\ 0 & \ddots & \ddots & \ddots & * & + & * \\ + & + & + & \mathbf{0} & \mathbf{0} & * & + \end{pmatrix}.$$

Proof. In this proof we will make heavy use of orthogonalisation processes. For two vectors u, v in $\mathbb{C}^k$, we note $u.v = u^t v$. The . operator is an inner product in $\mathbb{C}^k$, and the classical scalar product in $\mathbb{R}^k$. In addition, for a vector v such that $v.v \neq 0$ and $v.v$ is not negative real, we define $\mathcal{N}(v) = v/\sqrt{v.v}$ where the branch of the square root is selected, that yields results with a positive real value.

Suppose $N = 1$. In this case, B is made of two scalars, $B_{1,1}$ and $B_{1,2}$, and condition (3.11) states $B_{1,1}B_{1,2} = 0$. Therefore either $B_{1,1} = 0$ or $B_{1,2} = 0$ Both cannot be zero at the same time because this would imply $Y = 0$, contradicting the hypothesis on the McMillan degree. Inspection of the realisation formula (3.6) yields the announced result in this case.

Suppose now $N \geq 2$, and denote by (B_t, M_t) a transversal realisation of Y. We denote by w_1 and w_N, the first and second real valued column vectors of B_t. Condition (3.11), which indicates that $B_t^t B_t$ is diagonal, is to the effect that

$$w_1.w_N = 0. \tag{3.12}$$

Suppose that one of the diagonal elements of G_1, say $(G_1)_{2,2} = w_N.w_N$ is zero. We have therefore $w_N = 0$. Y is therefore scalar again, with non zero element $Y_{1,1}$. We

will now determine an orthogonal matrix P that allows to transform (B_t, M_t) to a realisation possessing the expected topology. We denote by $[v_1, v_2 \dots v_N] = P$ the column vectors of P, and set $v_1 = \mathcal{N}(w_1)$. To proceed further we define,

$$w_2 = M_t w_1 - \frac{(M_t w_1 . w_1)}{w_1 . w_1} w_1.$$

As the controlability matrix of (M_t, B_t) is full rank, $w_2 \neq 0$ as otherwise $M_t w_1$ would be proportional to w_1, which is in contradiction with $N > 1$. We fix $v_2 = \mathcal{N}(w_2)$, a vector by construction orthogonal to v_1. Invoking again the minimality of (B_t, M_t), we define step by step all w_k till $k = N$, and their normalized version v_k, using the recurrence formula for $i > 2$,

$$w_i = M_t w_{i-1} - \frac{(M_t w_{i-1} . w_{i-1})}{w_{i-1} . w_{i-1}} w_{i-1} - \frac{(M_t w_{i-1} . w_{i-2})}{w_{i-2} . w_{i-2}} w_{i-2}. \tag{3.13}$$

Indeed, if a zero vector occurred at index $i = i_0$, we would find a strict subspace spanned by $\{w_1 \dots w_{i_0-1}\}$, stable by M_t and containing w_1 the sole non-zero column vector of B_t: a contradiction. It remains to prove that the w_i form an orthogonal family. Suppose by induction that $\{w_1, w_2 \dots w_{p-1}\}$ form an orthogonal family ($p \geq 3$). By formula (3.13) w_p is orthogonal to w_{p-1}, w_{p-2}. So if $p = 3$ we are set. If $p > 3$, we compute for $1 < k \leq p - 3$

$$\begin{aligned} w_p . w_{p-2-k} &= M_t w_{p-1} . w_{p-2-k} \\ &= w_{p-1} . M_t w_{p-2-k} \\ &= 0. \end{aligned} \tag{3.14}$$

The first equality in (3.14) is obtained by noting that w_p decomposes as a linear combination of the family $(M_t w_{p-1}, w_{p-1}, w_{p-2})$, the last two vectors of which are orthogonal to w_p by the induction hypothesis. The second equality occurs because M_t is self-adjoint. Eventually the fact that $M_t w_{p-2-k}$ is by construction a linear combination of vectors of the family $\{w_{p-1-k}, w_{p-2-k}, w_{p-3-k}\}$, all vectors orthogonal to w_{p-1}, yield the third equality. Define $M = P^t M_t P$ and $B = P^t B_t$. From the formula $M_{i,j} = v_i . M_t v_j$ and relation (3.13) we conclude that M has the announced topology. The only no vanishing term of B is $B(1,1) = \sqrt{w_1 . w_1}$. For $1 < i \leq N$ using (3.13) we obtain,

$$M_{i,i-1} = v_i . M_t v_{i-1} = \sqrt{\frac{w_i . w_i}{w_{i-1} . w_{i-1}}} > 0$$

which implies the non-vanishing of the sur- and sub-diagonal terms of M. Uniqueness is proven easily when starting from a circuit realisation (M, B) with corresponding topology, and looking for a similarity P yielding another one with same shape. The first vector of P needs, obviously in order to preserve the shape of B, to be of the form $v_1 = (\pm 1, 0, \dots 0)^t$. Then by computing $M v_1$ and removing its projection on v_1 we find that v_2 needs to be of the form $v_2 = (0, \pm 1, 0, \dots 0)^t$. Carrying on, this argument finishes the proof for this case.

Suppose now that $N > 2$ and none of the diagonal elements of G_1 are zero. Proceeding in the same manner as before we set $B = [w_1, w_N]$, where w_1 and w_N

are two non zero vectors, because $w_1.w_1 = (G_1)_{1,1}$ and $w_N.w_N = (G_1)_{2,2}$. Following the same reasoning, we set $v_1 = \mathcal{N}(w_1)$ and $v_N = \mathcal{N}(w_2)$ and,

$$w_2 = M_t w_1 - \frac{(M_t w_1.w_1)}{w_1.w_1} w_1 - \frac{(M_t w_1.w_N)}{w_N.w_N} w_N.$$

Suppose first that w_2 is not zero, we set $v_2 = \mathcal{N}(w_2)$ and continue for $i > 2$ with,

$$w_i = M_t w_{i-1} - \frac{(M_t w_{i-1}.w_{i-1})}{w_{i-1}.w_{i-1}} w_{i-1} - \frac{(M_t w_{i-1}.w_{i-2})}{w_{i-2}.w_{i-2}} w_{i-2} - \frac{(M_t w_{i-1}.w_N)}{w_N.w_N} w_N \quad (3.15)$$

till a zero vector is found, at index $i = i_0$. Following the same scheme of proof as before, it is verified that all w_k (resp. v_k) produced this way form an orthogonal (resp. othonormal) family. The index i_0 is therefore at most N. If this is the case, a complete similarity transform has been found, and one verifies that it produces the expected arrow form, with non-zero sub and sur-diagonals. If $i_0 < N$ we start the orthonormalisation process "from the other side" by setting

$$w_{N-1} = M_t w_N - \sum_{k=1}^{i_0-1} \frac{M_t w_N.w_k}{w_k.w_k} w_k - \frac{M_t w_N.w_N}{w_N.w_N} w_N \quad (3.16)$$

Suppose $w_{N-1} = 0$. This would imply that $(w_1, \dots w_{i0-1}, w_N)$ spans a strict subspace, stable by M_t and containing w_1 and w_N the two line vectors of B_t: hence a contradiction to minimality. We continue like this following the backward recurrence formula for all i such that $i_0 < i < N$,

$$w_{i-1} = M_t w_i - \frac{M_t w_i.w_i}{w_i.w_i} w_i - \frac{M_t w_i.w_{i+1}}{w_{i+1}.w_{i+1}} w_{i+1} \quad (3.17)$$

till all vectors of P are determined. We need now to prove that the newly determined vectors are orthogonal one to another, and orthogonal to previously determined ones. The recurrence formula (3.17) is the backward analog to (3.13) and is initiated with vectors w_{N-1} and w_N. We therefore have,

$$\forall (i,j) \in \{i_0 \dots N\}^2,\ i \neq j,\ w_i.w_j = 0.$$

We state now following recurrence hypothesis $\mathcal{R}(p)$: for a given $p \in \{i_0 \dots N-1\}$ we have

$$\forall l \in \{1 \dots i_0 - 1\},\ w_p.w_l = 0.$$

We will prove that $\mathcal{R}(p)$ implies $\mathcal{R}(p-1)$. Note that $\mathcal{R}(N-1)$ is true by construction by means of formula (3.16). Suppose now that $\mathcal{R}(p)$ is true for a given value $p \in \{i_0 + 1 \dots N-1\}$. For $l \in \{1 \dots i_0 - 1\}$ we have,

$$\begin{aligned} w_{p-1}.w_l &= M_t w_p.w_l \\ &= w_p.M_t w_l \\ &= 0 \end{aligned} \quad (3.18)$$

where the first equality in (3.18) comes from (3.17) and $\mathcal{R}(p)$, while the third one from the fact that $M_t w_l$ is by hypothesis in $span(w_1, w_2 \dots w_{i_0-1}, w_N)$. This concludes the proof by induction, and indicates that matrix P constructed from the

normalized version of the vectors w_p is a similarity transform. The fact that the realisation $(M = P^t M_t P, B = P^t B_t)$ has the announced shape, is done exactly as in the previous case, as well as the unicity property.

Suppose now that the $k < N - 1$ first Markov parameters are zero. We have

$$(G_2)_{1,2} = (B^t M B)_{1,2} = B_{1,1} B_{N,2} M_{N,1},$$

which is to effect that $M_{N,N} = 0$. Under this hypothesis

$$(G_2)_{1,2} = B_{1,1} B_{N,2} M_{1,2} M_{N,2}$$

which proves that $M_{N,2} = 0$. Iterating like this $k - 1$ times yields the result and concludes the proof. □

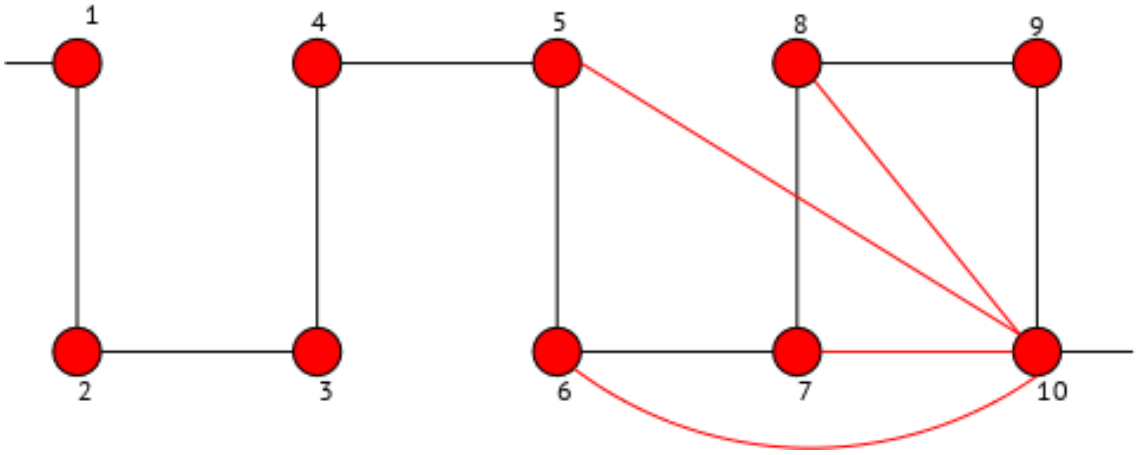


Figure 3.3: Arrow form that can accommodate 4 transmission zeros.

The arrow topology has shown to be very useful because of its uniqueness property that the transversal topology, for which the ordering of resonators is arbitrary, does not have. When all the transmission zeros are at infinity, this form becomes the classical inline topology where one resonator is coupled to the next one, which is probably the topology the most used in practice. However, when high selectivity is necessary, while keeping an overall number of resonators low, finite transmission zeros are inevitable. When one transmission zero is allowed, the arrow form exhibits a triplet, that is a group of resonators coupled in a triangular manner. In this form the triplet necessarily includes the last resonator. Remarkable techniques to move this triplet to other locations in the filter, based on the determination of suitable similarity transforms applied to the arrow form, have been developed [53]. When the number of transmission zeros grows, the arrow form becomes intractable in practice (see Figure 3.3): that is, the corresponding filter is impossible to build. This comes from the high number of couplings the last resonator has to support: in practice these couplings need to be realised by a certain proximity of the pair of resonators to be coupled. Three or four couplings per resonator are therefore an absolute maximum and have led engineers to look for alternative topologies that tend to spread the "coupling effort" across the filter. To understand how the number of transmission zeros relates to the coupling topology, we state a well known property among filter designers. The coupling graph is a non-oriented graph where

the vertices represent the resonators and the edges the couplings. If a coupling is present between resonator i and j, then an edge is drawn between the vertices i and j.

Proposition 3.1.4 *We consider here topologies with N resonators, where the input port excites only one resonator, that we call the input resonator (numbered 1), and the output port excites only one resonator, namely the output resonator (numbered N). If in the coupling graph, the shortest path between the input and output resonator is of length l, then the l first Markov parameters of its admittance are diagonal. This implies that in the Belevitch form associated to the scattering matrix S of such circuits, the transmission polynomial p_2 has at most degree $N-l-1$.*

Proof. A proof of this proposition is available in [54], we give a quick justification here because it is instructive. Let (M, B) be a circuital realisation with a given topology and associated coupling graph. From the definition of the Markov parameter, we want to determine for which integer k the $(1, N)$ term of M^k becomes non-zero. If one performs the symbolic computation of M^k, it is immediately seen that the monomials appearing to express the element (i, j) of M^k are in correspondence with the possible walks of length k from vertex i to vertex j. By the formula for matrix product we have:

$$[M^2]_{i,j} = \sum_{k=1}^{N} M_{i,k} M_{k,j}.$$

So each monomial $M_{i,k}M_{k,j}$ represents a possible walk of length 2 from i to j. In particular if the shortest path between i and j in the coupling graph is of length strictly greater than 2, we have $[M^2]_{i,j}$=0, whatever the values of the couplings involved. Iterating this procedure, we conclude that the polynomial expression of $[M^k]_{1,N}$ in terms of the $M_{i,j}$ is zero, as long as $k < l$, where l is the shortest path between node 1 and N. This yields the announced result. If the l first Markov diagonal parameters of Y vanish (starting from $k = 0$), it is easy to see that this is also the case for those of the associated scattering matrix S. Markov parameters correspond to the formal expansion of the frequency response in powers of $1/s$: the vanishing of the k first elements, indicates a difference of $k+1$ between the numerator and denominator of the transfer function. In the Belevitch form, where the degree of the denominator is taken to be N, this indicates that the transmission polynomial p_2 of formula (2.1.1) is of degree at most $N-l-1$. □

This result shows that there is a strong relation between the coupling topology and the type of responses that are realisable with it: by the last result it is for example hopeless to realize a response with 2 transmission zeros by means of an inline topology. This result is however in no way sufficient: having a coupling topology where the shortest input/output path is shorter than l, does not guarantee that every response with less than $N-l-1$ transmission zeros will be synthesisable this way: intuitively you need also enough parameters in your topology to "control" these zeros. Having this in mind, researchers have designed approaches based on optimization, to realize responses under arbitrary coupling topologies [55, 56, 57, 58, 50]. The optimization involved here is not of convex type. In particular these methods have all in common that they offer no guarantee of success, nor answers

in a certified manner to the questions : can a given response be synthesised by a specified topology and, if yes, in how many ways ?

We therefore developed a framework to solve this problem in a general manner: the latter is based on algebraic geometry techniques and the use of computer algebra. Part of this work has been supported by the Inria ARC Sila that brought together the Apics team and two researchers of the Inria Salsa team: Fabrice Rouillier and Jean-Charles Faugere. We define a coupling topology to be a set of formal electrical parameters X, as well as their associated formal matrices (B, M) with value in the polynomial ring $\mathbb{C}[X]$. A topology σ is therefore a triple (B, M, X). For example the in-line topology is described by the set of formal variables $X = \{B_{1,1}, B_{N,2}, M_{1,1} \dots M_{N,N}, M_{1,2} \dots M_{i,i+1} \dots M_{N-1,N}\}$ as well as its obviously structured matrices B and M with polynomial entries in the variables X. If $x \in \mathbb{C}^r$ (where r is the cardinality of X) we note $B(x), M(x)$ the evaluation at the complex point x of the polynomial matrices B, M. For every topology σ with parameter set X of cardinality r, and parametrised realisation, (B, M), we define following realisation map:

$$\begin{array}{rccl} \pi_\sigma : & \mathbb{C}^r & \mapsto & \left(\mathbb{C}^{2\times 2}\right)^{2N-1} \\ & x & \mapsto & \left(B^t(x)B(x), B^t(x)M(x)B(x) \dots B^t(x)M^{2N-1}(x)B(x)\right). \end{array} \tag{3.19}$$

The name realisation map is justified, by following proposition.

Proposition 3.1.5 *Let $\sigma = (B, M, X)$ and $\sigma' = (B', M', X')$ two topologies of state space size $m \leq N$. Let $(B(x), M(x))$ be a minimal realisation of size $m \leq N$ for a particular $x \in \mathbb{C}^r$, and $(B'(x'), M'(x'))$ a realisation obtained for the parameter values $x' \in \mathbb{C}^{r'}$ of size m and parameter set X' of cardinality r' . Suppose that we have,*

$$\pi_\sigma(x) = \pi_{\sigma'}(x'), \tag{3.20}$$

then:

- *$(B'(x'), M'(x'))$ is minimal*
- *there exists P a $m \times m$ matrix, such that $P^tP = Id$ and $B'(x') = P^tB(x)$ and $M'(x') = P^tM(x)P$*
- *$(B'(x'), M'(x'))$ and $(B(x), M(x))$ have the same transfer function, that is same admittance Y given by (3.6)*

Proof. The proof is very similar to proposition 3.1.2. Let $\mathcal{C}_m$ (resp. $\mathcal{C}'_m$) and $\mathcal{O}_m$ (resp. $\mathcal{O}'_m$) be the controlability and observability matrices associated to $(B(x), M(x))$ and $(B'(x'), M'(x'))$. Relation (3.20) is to the effect that $\mathcal{O}_m\mathcal{C}_m = \mathcal{O}'_m\mathcal{C}'_m$, and the minimality of (B, M) implies $rank(\mathcal{O}_m\mathcal{C}_m) = rank(\mathcal{O}'_m\mathcal{C}'_m) = m$. If $(B'(x'), M'(x'))$ is not minimal, then at least one of the matrices $\mathcal{C}'_m$ or $\mathcal{O}'_m$ is not full rank and $rank(\mathcal{O}'_m\mathcal{C}'_m) < m$: a contradiction. $(B'(x'), M'(x'))$ is therefore minimal. From now on, the proof is exactly the same as in proposition 3.1.2, and yields the existence of the announced similarity transform, which in turn implies equality of the admittances. □

We now define the admissible set of a topology, as well as its non-redundant nature.

Definition 3.1.2 *For a topology $\sigma = (B, M, X)$ with $r = card(X)$ we define $V(\sigma) = \overline{\pi_\sigma(\mathbb{C}^r)}$ to be its admissible set. It is equivalent by prop. 3.1.5, up to the closure operation, to the set of all possible admittances the topology can generate, when its parameters range over $\mathbb{C}^r$.*

Definition 3.1.3 *Let $\sigma = (B, M, X)$ be a topology. As a polynomial map, π_σ has a Jacobian matrix. On an non-empty open Zariski set of $\mathbb{C}[X]$ this Jacobian has constant rank, which is often called its generic rank. We will say that a topology is non-redundant, if the Jacobian of its realisation map is generically full rank, that is of rank $r = card(X)$.*

Solving our coupling matrix synthesis problem is about inverting π_σ on $V(\sigma)$. We have following properties,

Proposition 3.1.6 *Let $\sigma = (B, M, X)$ be a coupling topology, with $r = card(X)$. We have,*

- *$V(\sigma)$ is an irreducible algebraic variety.*
- *If σ is non-redundant, then the dimension of $V(\sigma)$ as an algebraic variety, is r.*
- *If σ is non-redundant there exists an integer $\Theta(\sigma)$ such that all fibers of π_σ are generically of cardinality $\Theta(\sigma)$. More precisely there exists a non-empty Zariski set U open in $V(\sigma)$ such that $\forall y \in U$, $card(\pi_\sigma^{-1}(y)) = \theta(\sigma)$. U is dense in $V(\sigma)$ in both topologies (Zariski and euclidean). We call $\theta(\sigma)$ the order of the topology σ. For topologies with sufficiently off-diagonal couplings (for example the sub-diagonal, like in the arrow form) the order is a multiple of 2^N, due to the invariance of their fibers by a sign matrix. In this case we call $\theta_r(\sigma) = \theta(\sigma)/2^N$ the reduced order of the topology.*
- *If σ_1 and σ_2 are two non-redundant topologies with parameter sets of the same cardinality, and if $V(\sigma_1) \subset V(\sigma_2)$ then $V(\sigma_1) = V(\sigma_2)$*
- *If σ_1 and σ_2 are coupling topologies, and $V(\sigma_1) \cap V(\sigma_2) \neq V(\sigma_1)$, then generically, on a non-empty Zariski open set U of $V(\sigma_1)$, we have*

$$\forall y \in U,\ \pi_{\sigma_2}^{-1}(y) = \emptyset.$$

Proof. The proof of the three first assertions of this proposition is contained in the purely "algebraic geometrical" proposition 3.1.7 to come. The two last assertions are a consequence of a classical theorem: if V_1 is an irreducible variety, and V_2 a strict sub-variety of V_1 then $dim(V_1) > dim(V_2)$ see [59, theorem 2.15]. □

We now come to the proposition proving previous assertions: the latter might be classical to every specialist of algebraic geometry, but is difficult to find as is in the literature. The vocabulary here is the one of algebraic geometry: when we speak of a variety it is a an algebraic one etc...When we say that a property is true generically on an algebraic variety $\mathcal{Y}$, we mean that the set where it is true contains a Zariski open subset of $\mathcal{Y}$ that is Zariski dense in $\mathcal{Y}$.

Proposition 3.1.7 *Let f be a polynomial map from $\mathbb{C}^n \mapsto \mathbb{C}^m$, with $m \geq n$. We define $\mathcal{Y} = \overline{f(\mathbb{C}^n)}$, where the closure is taken here in the classical euclidean sense. Assume that the Jacobian of f is generically of full rank n on $\mathbb{C}^n$. We have:*

- *$\mathcal{Y}$ is an irreducible algebraic variety of dimension n*
- *There exists a constant integer $d > 0$, and a non-empty Zariski open subset $U \subset \mathcal{Y}$ such that $\forall y \in U$, the fibers $f^{-1}(y)$ are discrete and of cardinality d*
- *There exists a non-empty open Zariski subset V of $\mathbb{C}^n$ such that at each point $x \in V$, the function f is a local parametrisation of Y at $y = f(x)$ in the language of smooth manifolds [60]. This means that there exists an euclidean open neighbourhood V of y in Y, and U an open neighbourhood of x such that the restriction of f to U, is a diffeomporhism from $U \to V$.*

Proof. We define $\mathcal{Y}_c$ as the Zariski closure of $f(\mathbb{C}^n)$. From the fundamental theorem of elimination theory, we have that $f(\mathbb{C}^n)$ is Zariski non empty and open in $\mathcal{Y}_c$ [61] which in turn is to the effect, by a theorem of Mumford [62, Chap. 10, theorem 1], that it is dense in Y_c in the euclidean topology. We have therefore $\mathcal{Y} = \mathcal{Y}_c$. As the Zariski closure of the image of a polynomial map [61, Prop.5 Chap.4], $\mathcal{Y}$ is irreducible and the first claim of the proposition is proven.

For the second assertion recall, that polynomial map f defines an isomorphic inclusion, between the field of fractions of the coordinate ring of the Zariski closure of its image and the fraction field of the coordinate ring of the variety where its is defined. This map is called the pullback map of f [63, 64, 61], and denoted f^*. In our case,

$$f^* : \mathbb{C}(\mathcal{Y}) \hookrightarrow \mathbb{C}(X_1, \ldots, X_n),$$

where $\mathbb{C}(\mathcal{Y})$ denotes the fraction field of $\mathbb{C}[\mathcal{Y}]$ the coordinate ring of $\mathcal{Y}$, which in turn is isomorphic to $\mathbb{C}[X_1, \ldots, X_m]/I(\mathcal{Y})$, where $I(\mathcal{Y})$ denotes the ideal defining $\mathcal{Y}$. The dimension of $\mathcal{Y}$ is equal to the transcendence degree of $\mathbb{C}(\mathcal{Y})$ over $\mathbb{C}$ and therefore by the inclusion f^*, we have that $dim(\mathcal{Y}) \leq n$. Suppose that the inequality is strict, that is $dim(\mathcal{Y}) = k < n$. By Bertini's second theorem [63, Chap. 6, theorem 2.27] there exists a Zarisky non-empty open subset $W \subset \mathcal{Y}$ such that $\forall y \in \mathcal{Y}$ the fiber $f^{-1}(y)$ is non-singular and of dimension $k - n > 0$ according to our hypothesis. But $f^{-1}(U)$ is Zarisky open in $\mathbb{C}^n$, so that we can chose y such that the Jacobian of f is of full rank n at $x = f^{-1}(y)$. A contradiction, as the function f is constant on the non-singular variety $f^{-1}(y)$ containing x. Hence $dim(\mathcal{Y}) = n$.

From the theorem on fibers of polynomial maps [63, Chap. 6, theorem 1.25] , we have that their dimension is $dim(\mathcal{Y}) - n = 0$, that is they are discrete. The inclusion $f^*(\mathbb{C}(\mathcal{Y})) \subset \mathbb{C}(X_1, \ldots X_N)$ is an algebraic field extension: as the transcendence degrees of both fields are equal, the extension is of finite degree, say d. It is now classical, see [59, theorem 11.1], that there exists a Zariski open set U of $\mathcal{Y}$ such that above every $y \in U$ there exist exactly d points of $\mathbb{C}^n$.

Eventually let W be the set of non-singular points of $\mathcal{Y}$: it is a Zariski open subset, and so is $f^{-1}(W)$. The intersection of the latter with the set of points of $\mathbb{C}^n$ where the Jacobian of f is full rank, yields the set V described in the proposition. □

Let's summarize for the filter's practitioner what we have come up with so far:

- Each non-redundant coupling topology σ has an order $\theta(\sigma)$: it is the number of different circuits with topology σ, with possibly complex couplings values, that realize the same admittance matrix Y. This number is independent of the considered admittance, as long as the admittance stays in the admissible set $V(\sigma)$ of the coupling topology: $\theta(\sigma)$ is an invariant of the coupling topology σ. The number of equivalent circuits with real valued elements is not independent of the considered admittance: it may vary between 0 and $\theta(\sigma)$.

- If two coupling topologies σ_1, σ_2 have two different admissible sets, and one is not included in the other, then generically an element of the admissible set of σ_1 can not be realised with topology σ_2 and vice a versa.

We now define a class of admissible sets, based on the arrow form.

Definition 3.1.4 *Suppose $N > 1$, for $k \leq N-2$ let δ_k^N be the arrow coupling topology with parameter set,*

$$\begin{gathered} X_k = \{B_{1,1}, B_{N,2}\} \cup \{i = 1 \dots N, M_{i,i}\} \cup \\ \{i = 1 \dots N-1, M_{i,i+1}\} \cup \{i = 1 \dots k, M_{N,N-1-i}\} \end{gathered} \tag{3.21}$$

with the associated realisation,

$$B = \begin{pmatrix} B_{1,1} & 0 \\ 0 & 0 \\ \vdots & \vdots \\ 0 & B_{N,2} \end{pmatrix}, \; M = \begin{pmatrix} M_{1,1} & M_{1,2} & 0 & 0 & 0 \\ M_{1,2} & \ddots & \ddots & 0 & M_{N,N-1-k} \\ 0 & \ddots & \ddots & \ddots & \vdots \\ 0 & \ddots & \ddots & \ddots & M_{N,N-1} \\ 0 & M_{N,N-1-k} & \dots & M_{N,N-1} & M_{N,N} \end{pmatrix}. \tag{3.22}$$

By proposition 3.1.3 this topology realizes all loss-less responses with at most k transmission zeros. Conversely by the shortest path rule, all loss-less minimal admittance realised this way have at most k transmission zeros. We denote by $E_k^N = V_{\delta_k^N}$.

Proposition 3.1.8 *The coupling topology δ_k^N is non redundant. We have $\theta_r(\delta_k^N) = 1$ and the dimension of $V(\delta_k^N)$ is $2N + k + 1$.*

Proof. We first prove that the generic McMillan degree of a realisation with coupling topology δ_k^N is N. Suppose that all parameters x_k are different from zero but possibly complex. Then starting from this arrow form $(B(x), M(x))$, and running the orthogonalisation process described in proposition 3.1.3 shows that the canonical base, represented by the matrix $P = Id$, is spanned by vectors of the controlability matrix of $(B(x), M(x))$. This also shows, that we have uniqueness of the representation in arrow form of $B(x), M(x)$ up to a sign matrix, and even here for complex (but non-vanishing) values of the parameters.

Suppose now that δ_k^N is redundant, that is that the generic rank of its jacobian is $j < 2N + k + 1$. A similar reasoning to that of proposition 3.1.7 shows that $V_{\delta_k^N}$ is of dimension j. A repeated application of the theorem on the dimension of fibers yields the existence of a non-empty Zariski open subset U of $V(\delta_k^N)$ such that $\forall y \in U$, the

fibers $\pi^{-1}_{\delta^N_k}(y)$ are of positive dimension $N+k+1-j>0$. Now $\pi^{-1}_{\delta^N_k}(U)$ is an non-empty Zariski open subset dense in $\mathbb{C}^{2N+k+1}$, so that we can certainly find a complex vector $x_0 \in \mathbb{C}^{2N+k+1}$ with non-vanishing components, such that the fiber associated to $y_0 = \pi_{\delta^N_k}(x_0)$ is of positive dimension. But this fiber is made of infinitely distinct arrow forms, similar by proposition (3.1.5) to the minimal realisation $B(x_0), M(x_0)$: a contradiction. Eventually the unicity property (up to a sign matrix) proves that the reduced order of this coupling topology is 1. □

We now come to a result on the compatibility conditions between coupling topologies we have been working for.

Proposition 3.1.9 *Let σ be a non-redundant topology of size N with parameter set X of length $2N+k+1$. Suppose that its coupling graph has a shortest path of length $N-k-1$ between resonator 1 and N, then $V(\sigma) = E_k$. In particular, generically, every loss-less admittance Y with at most k transmission zeros can be realised in $\theta(\sigma)$ different realisations with topology σ. Among these, the number of real valued realisations may vary with Y.*

Proof. First suppose that the generic McMillan degree of σ is N. We pick a real valued parameter set x_0 such that in a euclidean open set U of $\mathbb{C}^{2N+k+1}$ containing x_0 the McMillan degree of $(B(x), M(x))$ is constant and equal to N. Examining again the orthogonalisation process of Proposition 3.1.3, it is easily seen that it can be carried out in the complex in exactly the same manner, unless an isotropic vector w_k is found, that is a complex vector such that:

$$w_k.w_k = 0.$$

All quantities occurring in this orthogonalisation process being rational functions of the parameters in X, that is continuous functions away from their singular locus, we conclude that upon shrinking U, every realisation $(M(x), B(x))$ with $x \in U$ can be put in arrow form δ^N_k. But this indicates, that $\pi_\sigma(U) \subset E^N_k$. Now let $p \in I(E^N_k)$, where $I(E^N_k)$ is the polynomial ideal associated to the variety E^N_k. We have that for all $x \in U$,

$$p(\pi_{\sigma(x)}) = 0,$$

and this is to the effect that $p(\sigma(x))$ vanishes on all $\mathbb{C}^{2n+k+1}$. Hence $p \in I(V(\sigma))$, and $I(E^N_k) \subset I(V(\sigma))$ which in turn implies

$$V(\sigma) \subset E^N_k.$$

But E^N_k and $V(\sigma)$ are irreducible varieties of same dimension, and therefore $V(\sigma) = E^N_k$. Eventually suppose that the generic McMillan degree of σ is $M < N$. The same reasoning yields that $V(\sigma) \subset E^M_k$, and hence $dim(V(\sigma)) \leq 2M+k+1 < 2N+k+1$, hence a contradiction. The generic McMillan degree of σ is N. □

3.1.1 Application

This approach to the coupling matrix synthesis problem has had a significant impact on the field of microwave filters. For the first time it was mathematically proven that some coupling topologies admit multiple solutions and a guaranteed method

was designed to compute them all. This method relies on the use of Groebner basis to compute the fibers $\pi_\sigma^{-1}(y)$ for various topologies and specific responses y. The latter has been possible thanks to the use of the wonderful tool FGb [65] by Jean-Charles Faugère: probably one of the most efficient Gröbner engines in the world, to paraphrase a famous beer add ! Thanks to it most classical coupling topologies have been classified, their reduced order computed, and a reference solution processed. In order to render computations practical, we developed a tool based on continuation techniques called [66, Dedale-Hf]. It uses some database containing, for a specific value y_0 in the admissible set of each topology, the corresponding fiber $\pi_\sigma^{-1}(y_0)$ computed by FGb. At hand of the latter, and for a y specified by the user, Dedale-HF computes by continuation the fiber $\pi_\sigma^{-1}(y)$, and this usually in a few seconds. A heuristic based on homotopy techniques was also derived, as part of Dedale-HF, to tackle, but with no guaranty of exhaustivity, problems where the computation of the associated Gröbner has shown to be intractable.

The first papers describing the approach are [67, 68], and [68] is attached at the end of the document in section 5.3.1. A classification effort and survey on the realisation technique is given in [69] and reproduced in the bibliographic section (see section 5.3.2). Eventually application of our techniques to the synthesis of complex dual-band filters in extended box topology are considered in [70], and also present in the bibliographic section (see section 5.3.3).

As an example, the 10^{th} order extended box topology σ represented on Figure 3.4 has been analysed. The shortest path rule indicates that it can implement filtering characteristics with at most 4 transmission zeros. The parameter set has a cardinality of 25, which is also $dim(E_4^{10}) = 2\times 10+4+1$. We check with a computer system the generic rank of the Jacobian of π_σ, by testing it at a random point. We find that the latter is full rank, and therefore $V(\sigma) = E_4^{10}$. Now using Gröbner basis, running on a finite field, we obtain that $\theta(\sigma) = 384$. Running the Dedale-HF heuristic or FGb (on the fied $\mathbb{Q}$ this time) on a computer with enough memory, we compute a particular fiber $\pi_\sigma^{-1}(y_0)$. Solving the coupling matrix synthesis problem for any other y is done by continuation using Dedale-HF. See [66][Dedale-HF] for a classification of all usual coupling topologies. Eventually synthesis techniques for circuits including frequency dependant couplings have also been studied [71]: for the latter a complete algebraic theory is still missing.

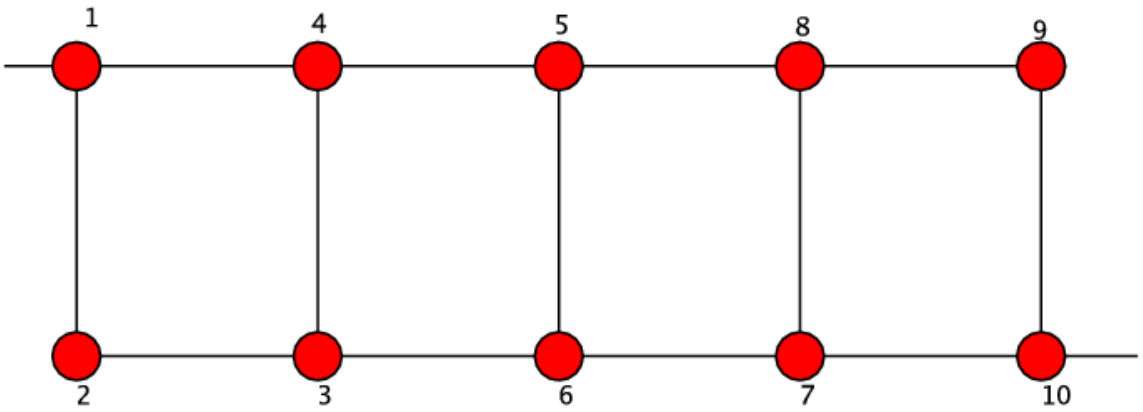


Figure 3.4: Extended box topology accommodating 4 transmission zeros. The reduced order of this topology is 384.

Chapter 4

Matching problems and Nevalinna-Pick interpolation

We now come to the final chapter of this document, relative to matching problems. This section will be relatively short as the publication [72] that we reproduce in the bibliographic section (see section 5.4.1) is rather complete on this topic. Matching problems occur when chaining a 2×2 scattering matrix S to a given load L (see Figure 4.1), which is not perfectly resistive. The intention pursued is usually to transmit the maximum possible power to the load, that is when S is supposed to be loss-less, to minimize the unwanted reflexions obtained after chaining. Matching problems are classically experienced with antennas, the input of which is seldom matched, that is seldom purely resistive. When feeding the antenna with a signal, one wants to get sure that most of it is radiated to free space and not reflected back to the input. But this is not the only situation, where matching is crucial: in multiplexer design for example, each channel filter needs to be matched to the load constituted by the rest of the multiplexer that is the common wave guide and the other channel filters. In a very general setting, matching problems occur each time that two scattering systems are interconnected together and that transmission of power between these systems is of importance. Usually one of these systems is already designed, and the question remains how to design the second one to obtain the best possible power "match", that is the best possible power transfer between both systems.

Consider the case of a loss-less 2×2 scattering matrix S, closed at port 2 on a reflexion coefficientL_{11} (see Figure. 4.1), the reflexion coefficient of which is considered to be strictly dissipative:

$$\forall \omega \in \mathbb{R},\ |L_{1,1}(\omega)| < r < 1.$$

In order to be coherent with paper [72] we consider transfer functions of the real frequency variable ω analytic in the lower half-plane $\mathbb{C}^-$. If we call G_{11} the reflexion

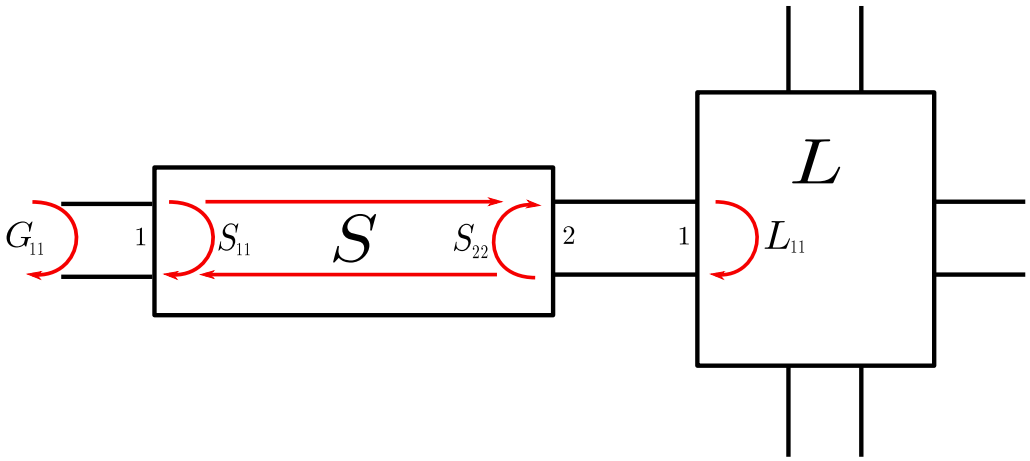


Figure 4.1: 2×2 scattering system plugged to a load L with reflexion coefficient L_{11}

coefficient of the system obtained after closing port 2, we obtain:

$$\begin{aligned} G_{11}(\omega) &= S_{11}(\omega) + \frac{S_{12}(\omega)S_{21}(\omega)L_{11}(\omega)}{1 - S_{22}(\omega)L_{11}(\omega)} \\ &= \frac{S_{11}(\omega) - L_{11}(\omega)\det(S(\omega))}{1 - S_{22}(\omega)L_{11}(\omega)} \\ &= \det(S(\omega))\frac{\overline{S_{22}(\omega)} - L_{11}(\omega)}{1 - S_{22}(\omega)L_{11}(\omega)}. \end{aligned} \tag{4.1}$$

Let I be a compact interval of the frequency axis $\mathbb{R}$. Mathematically, the matching problem can be stated as:

$$P_{\mathcal{H}} : \min_{S\in\mathcal{H}} \max_{\omega\in I} |G_{11}(\omega)| = \min_{S\in\mathcal{H}} \max_{\omega\in I} \left| \frac{\overline{S_{22}(\omega)} - L_{11}(\omega)}{1 - S_{22}(\omega)L_{11}(\omega)} \right| = \min_{S\in\mathcal{H}} \max_{\omega\in I} \delta(\overline{S_{22}(\omega)}, L_{11}(\omega)).$$

where $\mathcal{H}$ is for the moment vaguely determined as a class of loss-less 2×2 scattering matrices where S is being searched for, and δ is the pseudo-hyperbolic distance, defined in the unit disk by:

$$\delta(u, v) = \frac{u - v}{1 - \overline{u}v}.$$

If $\mathcal{H}$ is the class of all loss-less scattering matrices of, say maximal McMillan degree N, $P_{\mathcal{H}}$ is the classical finite dimensional matching problem, for which Bode, Fano and Youla [73, 74], developed one of the most aesthetic theories of electrical engineering in the sixties. Instead of looking for S directly, this approach considers the Darlington 2×2 extension $\hat{G}$, of G_{11}, and $\hat{L}$ the extension of the load L, together with compatibility conditions to ensure that the load $\hat{L}$ can be de-chained from $\hat{G}$. If z_i are the transmission zeros of $\hat{L}$ (the zeros in $\mathbb{C}^-$ of $p_2 p_2^*$ in the Belevitch form [30]), considered here for simplicity to be strictly stable, this compatibility conditions take the form:

$$\hat{G}_{22}(z_i) = \hat{L}_{22}(z_i). \tag{4.2}$$

As $|\hat{G}_{22}(w)| = |\hat{G}_{11}(w)|$, the matching problem reformulates as,

$$P_{\hat{\mathcal{H}}} : \ min_{G_{22}\in\hat{\mathcal{H}}} \max_{\omega\in I} |G_{22}(\omega)|. \tag{4.3}$$

where

$$\hat{\mathcal{H}} = \{G, \ G \text{ is a Schur functions of degree} \leq N \text{ verifying conditions (4.2)}\}.$$

Problem $P_{\hat{\mathcal{H}}}$ is a difficult non-convex optimization problem. When compared with the Zolotarev problem obtained in the previous chapter, the interpolation conditions (4.2) make it impossible to formulate the latter solely in terms of the numerators of the associated Belevitch form. It is the inability to solve this problem in an efficient manner, that led the electrical engineering community to consider it in very specific classes of responses: Tchebychev, Butterworth etc...[30], and with strong restrictions on the degree of the load. But this rigidity and the understanding that Tchebychev responses are by no means optimal in terms of matching, led the community to abandon Fano and Youla's approach and consider approaches based on optimization, like the real frequency technique of Carlin [30]: the latter however, offer no guaranty on optimality. We are currently studying in our team convex relaxations techniques of problem $P_{\hat{\mathcal{H}}}$.

Another important contribution to this problem, is due to the mathematician J. W. Helton. In his reference papers [75] he considers the problem $P_{\mathcal{H}}$ in H^∞. Here the problem becomes,

$$P_{\tilde{\mathcal{H}}} : \ min_{G_{11}\in\tilde{\mathcal{H}}} \max_{\omega\in I} |G_{11}(\omega)|. \tag{4.4}$$

where

$$\tilde{\mathcal{H}} = \{h, \ h \in H^\infty \text{ and } h \text{ is a Schur function}\}.$$

By noting that hyperbolic disks are indeed classical euclidean disks Helton showed that this problem is quasi-convex. He linked it, in a very similar manner to the extremal problems studied in the first chapter of this document, namely to the Nehari problem that solves the best approximation problem from L^∞ to H^∞. It is remarkable however, that even in the H^∞ case, the criterion obtained in $P_{\tilde{\mathcal{H}}}$ is strictly positive in most classical cases. This is easily seen by the fact that perfect matching on a interval I by a function G_0 would mean,

$$\forall\omega \in I, \ G_0^*(\omega) = L_{11}(\omega).$$

But as $L_{11} \in H^\infty$, this would indicate that L_{11} can be continued analytically in $\mathbb{C}\setminus J$, where $J = \mathbb{R} \setminus I$. Such functions can be built by means of Markov functions, but this is a lot to ask for a load L: for example if L_{11} is a Schur non-constant rational function, it can not be analytically continued in $\mathbb{C} \setminus J$, and the criterion in $P_{\tilde{\mathcal{H}}}$ is therefore strictly positive. In terms of microwave circuits, this means that even if a matching circuit is allowed to possess infinitely many resonators, there is, for a given load, a strict bound in terms of the reflexion level $|G_{11}|$ that can be achieved on a given band I. This is a major difference when compared, for example, with the filter synthesis problem considered in chapter 2, for which an arbitrary selectivity level can be obtained, provided no bound on the degree of the filtering function is given. The practical computation of these lower bounds is rendered possible thanks to Helton's theory of broad-band matching [76].

Although infinite dimensional bounds are important, the practical matching problem is usually considered for rather low degrees, say of maximal 10^{th} order when the matching and filtering functions are conceived in the same device. This is essentially due to considerations on ohmic losses occurring inevitably in microwave devices, and increasing with the system's order. In an attempt to find a good starting point for $P_{\mathcal{H}}$ we tackled following interpolation problem,

Problem $\mathcal{P}$: *Given* N distinct real frequencies $(x_1, x_2 \dots x_N)$, and $r \not\equiv 0$ a complex polynomial of degree at most $N-1$, such that $r(x_k) \neq 0$, $k = 1, \dots, N$, *find* (p, q) a pair of monic complex polynomials of degree N such that,

$$\left\{ \begin{array}{rcl} \frac{p}{q}(x_k) & = & \overline{L_{11}(x_k)}, \quad \text{for } k = 1, .., N \\ qq^* - pp^* & = & rr^* \end{array} \right. \tag{4.5}$$

and q has no root in the open lower half-plane $\mathbb{C}^-$ (*i.e.* q is stable in the broad sense).

Problem $\mathcal{P}$ amounts to design a rational matching system S of degree N, which ensures N perfect matching points at specified frequency points $\{x_1, x_2 \dots x_N\}$. The zeros of r, when they are real, can be seen as perfect attenuation points. If synthesized according to these conditions, S can therefore be seen as a matching filter, ensuring a good match of the load on a frequency band, while rejecting the signal for frequencies away from it. The main result of [72] is the following,

Proposition 4.0.1 *Problem* $\mathcal{P}$ *has a unique solution and it can be computed by continuation techniques.*

Proof. See [77]. □

Mathematically, problem $\mathcal{P}$ is a Nevanlinna-Pick interpolation problem [2, 78, 79] under spectral constraints: the transmission polynomial r is fixed. The particularity here is that the interpolation conditions are posed on the boundary of the analyticity domain: in this sense it is a generalisation of [80, 81, 82]. These interpolation problems have a long history, well detailed in [77], and reproduced in section 5.4.1.

Eventually, from discussion with engineers involved in the synthesis of complex multiplexers, it appears that a fair amount of them intuitively uses proposition 4.0.1, by fixing the location of the transmission zeros of each channel filter and imposing, by means of optimisation, some perfect matching points in the form of (4.5). We hope that our result and the associated certified computation techniques will bring some insight and robustness in their procedures. Based on proposition (4.0.1) promising studies are also being pursued in our team to design multiplexers.

Chapter 5

Bibliographic section

The bibliographic section contains reproductions of following articles.

- Laurent Baratchart, Juliette Leblond, and Fabien Seyfert. "Constrained L^2-approximation by polynomials on subsets of the circle". In: *New Trends in Approximation Theory. In Memory of André Boivin.* Ed. by Javad Mashreghi, Myrto Manolaki, and Paul M. Gauthier. Vol. 81. Fields Institute Communications. Springer, 2017, pp. 1–14. URL: https://hal.archives-ouvertes.fr/hal-01671183

- Martine Olivi, Fabien Seyfert, and Jean-Paul Marmorat. "Identification of microwave filters by analytic and rational H^2 approximation". In: *Automatica* (2012). DOI: 10.1016/j.automatica.2012.10.005. URL: http://hal.inria.fr/hal-00753824

- Adam Cooman, Fabien Seyfert, Martine Olivi, Sylvain Chevillard, and Laurent Baratchart. "Model-Free Closed-Loop Stability Analysis: A Linear Functional Approach". In: *IEEE Transactions on Microwave Theory and Techniques* (Sept. 2017). DOI: 10.1109/TMTT.2017.2749222. URL: https://hal.inria.fr/hal-01381731

- V. Lunot, F. Seyfert, S. Bila, and A. Nasser. "Certified computation of optimal multiband filtering functions". In: *IEEE Transactions on Microwave Theory and Techniques* 56.1 (2008), pp. 105–112

- Richard J. Cameron, Jean-Charles Faugère, Fabrice Rouillier, and Fabien Seyfert. "Exhaustive approach to the coupling matrix synthesis problem and application to the design of high degree asymmetric filters". In: *International Journal of RF and Microwave Computer-Aided Engineering* 17.1 (Jan. 2007), pp. 4–12. DOI: 10.1002/mmce.20190. URL: https://hal.inria.fr/hal-00663777

- Fabien Seyfert and Stéphane Bila. "General synthesis techniques for coupled resonator networks". In: *IEEE Microwave Magazine* 8.5 (2007), pp. 98–104. DOI: 10.1109/MMW.2007.4383440. URL: https://hal.inria.fr/hal-00663533

- Philippe Lenoir, Stéphane Bila, Fabien Seyfert, Dominique Baillargeat, and Serge Verdeyme. "Synthesis and design of asymmetrical dual-band bandpass filters based on equivalent network simplification". In: *IEEE Transactions on Microwave Theory and Techniques* 54.7 (2006), pp. 3090–3097. DOI: 10.1109/TMTT.2006.877037. URL: https://hal.inria.fr/hal-00663496

- Laurent Baratchart, Martine Olivi, and Fabien Seyfert. "Boundary Nevanlinna-Pick interpolation with prescribed peak points. Application to impedance matching". In: *SIAM Journal on Mathematical Analysis* (2017). DOI: 10.1137/16M1085577. URL: https://hal.inria.fr/hal-01377782

5.1 Bounded Extremal Problems

5.1.1 Mixed type extremal problems

Following paper is reproduced in this section:

- Laurent Baratchart, Juliette Leblond, and Fabien Seyfert. "Constrained L^2-approximation by polynomials on subsets of the circle". In: *New Trends in Approximation Theory. In Memory of André Boivin.* Ed. by Javad Mashreghi, Myrto Manolaki, and Paul M. Gauthier. Vol. 81. Fields Institute Communications. Springer, 2017, pp. 1–14. URL: https://hal.archives-ouvertes.fr/hal-01671183

Constrained L^2-approximation by polynomials on subsets of the circle

L. Baratchart[*], J. Leblond[*], F. Seyfert[*]

September 11, 2018

To the memory of André Boivin

1 Abstract

We study best approximation to a given function, in the least square sense on a subset of the unit circle, by polynomials of given degree which are pointwise bounded on the complementary subset. We show that the solution to this problem, as the degree goes large, converges to the solution of a bounded extremal problem for analytic functions which is instrumental in system identification. We provide a numerical example on real data from a hyperfrequency filter.

2 Introduction

This paper deals with best approximation to a square summable function, on a finite union I of arcs of the unit circle $\mathbb{T}$, by a polynomial of fixed degree which is bounded by 1 in modulus on the complementary system of arcs $J = \mathbb{T} \setminus I$. This we call, for short, the polynomial problem. We are also concerned with the natural limiting version when the degree goes large, namely best approximation in $L^2(I)$ by a Hardy function of class H^2 which is bounded by 1 on J. To distinguish this issue from the polynomial problem, we term it the analytic problem. The latter is a variant, involving mixed norms, of constrained extremal problems for analytic functions considered in [12, 3, 2, 13, 18]. As we shall see, solutions to the polynomial problem converge to those of the analytic problem as the degree tends to infinity, in a sense to be made precise below. This is why solving for high degree the polynomial problem (which is finite-dimensional) is an interesting way to regularize and approximately solve the analytic problem (which is infinite-dimensional). This is the gist of the present work.

Constrained extremal problems for analytic functions, in particular the analytic problem defined above, can be set up more generally in the context of weighted approximation, *i.e.* seeking best approximation in $L^2(I, w)$ where w is a weight on I. In fact, that kind of generalization is useful for applications as we shall see. As soon as w is invertible in $L^\infty(I)$, though, such a weighted problem turns out to be equivalent to another one with unit weight, hence the present formulation warrants most practical situations. This property allows one to carry the analytic problem over to more general curves than the circle. In particular, in view of the isomorphism between Hardy spaces of the disk and the half-plane arising by composition with a Möbius transform [10, ch. 10], best approximation in $L^2(I)$ from H^2 of the disk can be converted to weighted best approximation in $L^2(\mathfrak{I}, w)$ from the Hardy space $\mathfrak{h}^2$ of a half-plane with $\mathfrak{I}$ a finite union of bounded intervals on the line and w a weight arising from the derivative of the Möbius transform. Since this weight is boundedly invertible on $\mathfrak{I}$, it follows that the analytic problem on the circle and its analog on the line are equivalent. One may also define another Hardy space $\mathcal{H}^2$, say of the right half-plane as the space of analytic functions whose L^2-means over vertical lines are uniformly bounded. Then, best approximation in $L^2(I)$ from H^2 is equivalent to best approximation from $\mathcal{H}^2$ in $L^2(\mathfrak{I})$, *i.e.* weight is no longer needed. Of course, such considerations hold for many other domains and boundary curves than the half-plane and the line, but the latter are of special significance to us as we now explain.

Indeed, on the line, constrained extremal problems for analytic functions naturally arise in Engineering when studying deconvolution issues, in particular those pertaining to system identification and design. This motivation is stressed in [12, 4, 5, 19, 2], whose results are effectively used today to identify microwave

[*]INRIA, BP 93, 06902 Sophia-Antipolis Cedex, FRANCE

devices [1, 14]. More precisely, recall that a linear time-invariant dynamical system is just a convolution operator, hence the Fourier-Laplace transform of its output is that of its input times the Fourier-Laplace transform of its kernel. The latter is called the transfer-function. Now, by feeding periodic inputs to a stable system, one can essentially recover the transfer function pointwise on the line, but typically in a restricted range of frequencies only, corresponding to the passband of the system, say $\mathfrak{I}$ [9]. Here, the type of stability under consideration impinges on the smoothness of the transfer function as well as on the precise kind of recovery that can be achieved, and we refer the reader to [6, Appendix 2] for a more thorough analysis. For the present discussion, it suffices to assume that the system is stable in the L^2 sense, *i.e.* that it maps square summable inputs to square summable outputs. Then, its transfer function lies in H^∞ of the half-plane [15], and to identify it we are led to approximate the measurements on $\mathfrak{I}$ by a Hardy function with a bound on its modulus. Still, on $\mathfrak{I}$, a natural criterion from the stochastic viewpoint is $L^2(\mathfrak{I}, w)$, where the weight w is the reciprocal of the pointwise covariance of the noise assumed to be additive [16]. Since this covariance is boundedly invertible on I, we face an analytic problem on the line upon normalizing the bound on the transfer function to be 1. This stresses how the analytic problem on the line, which can be mapped back to the circle, connects to system identification. Now, this analytic problem is convex but infinite-dimensional. Moreover, as Hardy functions have no discontinuity of the first kind on the boundary [11, ch. II, ex. 7] and since the solution to an analytic problem generically has exact modulus 1 on J, as we prove later on, it will typically oscillate at the endpoints of I, J which is unsuited. One way around these difficulties is to solve the polynomial problem for sufficiently high degree, as a means to regularize and approximately solve the analytic one. This was an initial motivation by the authors to write the present paper, and we provide the reader in Section 6 with a numerical example on real data from a hyperfrequency filter. It must be said that the polynomial problem itself has numerical issues: though it is convex in finitely many variables, bounding the modulus on J involves infinitely many convex constraints which makes it of so-called semi-infinite programming type. A popular technique to handle such problems is through linear matrix inequalities, but we found it easier to approximate from below the polynomial problem by a finite-dimensional one with finitely many constraints, in a demonstrably convergent manner as the number of these constraints gets large.

The organization of the paper is as follows. In section 3 we set some notation and we recall standard properties of Hardy spaces. We state the polynomial and analytic problems in Section 4, where we also show they are well-posed. Section 5 deals with the critical point equations characterizing the solutions, and with convergence of the polynomial problem to the analytic one. Finally, we report on some numerical experiment in Section 6.

3 Notations and preliminaries

Throughout we let $\mathbb{T}$ be the unit circle and $I \subset \mathbb{T}$ a finite union of nonempty open arcs whose complement $J = \mathbb{T} \setminus I$ has nonempty interior. If h_1 (resp. h_2) is a function defined on a set containing I (resp. J), we put $h_1 \vee h_2$ for the concatenated function, defined on the whole of $\mathbb{T}$, which is h_1 on I and h_2 on J.

For $E \subset \mathbb{T}$, we let ∂E and $\overset{\circ}{E}$ denote respectively the boundary and the interior of E when viewed as a subset of $\mathbb{T}$; we also let χ_E for the characteristic function of E and $h_{|_E}$ for the restriction of h to E. Lebesgue measure on $\mathbb{T}$ is just the image of Lebesgue measure on $[0, 2\pi)$ under the parametrization $\theta \mapsto e^{i\theta}$. We denote by $|E|$ the measure of a measurable subset $E \subset \mathbb{T}$, and if $1 \leq p \leq \infty$ we write $L^p(E)$ for the familiar Lebesgue space of (equivalence classes of a.e. coinciding) complex-valued measurable functions on E with norm

$$\|f\|_{L^p(E)} = \left(\frac{1}{2\pi} \int_E |f(e^{i\theta})|^p \, d\theta \right)^{1/p} < \infty \quad \text{if } 1 \leq p < \infty, \qquad \|f\|_{L^\infty(E)} = \operatorname*{ess.sup}_{\theta \in E} |f(e^{i\theta})| < \infty.$$

We sometimes indicate by $L^p_{\mathbb{R}}(E)$ the real subspace of real-valued functions. We also set

$$\langle f, g \rangle_E = \frac{1}{2\pi} \int_E f(e^{i\theta}) \overline{g(e^{i\theta})} \, d\theta \tag{1}$$

whenever $f \in L^p(E)$ and $g \in L^q(E)$ with $1/p + 1/q = 1$. If f and g are defined on a set containing E, we write for simplicity $\langle f, g \rangle_E$ to mean $< f_{|_E}, g_{|_E} >$ and $\|f\|_{L^p(E)}$ to mean $\|f_{|_E}\|_{L^p(E)}$. Hereafter $C(E)$ stands for the space of bounded complex-valued continuous functions on E endowed with the sup norm, while $C_{\mathbb{R}}(E)$ indicates real-valued continuous functions.

Recall that the Hardy space H^p is the closed subspace of $L^p(\mathbb{T})$ consisting of functions whose Fourier coefficients of strictly negative index do vanish. We refer the reader to [11] for standard facts on Hardy

spaces, in particular those recorded hereafter. Hardy functions are the nontangential limits a.e. on $\mathbb{T}$ of functions holomorphic in the unit disk $\mathbb{D}$ having uniformly bounded L^p means over all circles centered at 0 of radius less than 1:

$$\|f\|_{H^p} = \sup_{0\le r<1} \left(\frac{1}{2\pi} \int_0^{2\pi} |f(re^{i\theta})|^p \, d\theta \right)^{1/p} \quad \text{if } 1 \le p < \infty, \qquad \|f\|_{H^\infty} = \sup_{z\in\mathbb{D}} |f(z)|. \tag{2}$$

The correspondence between such a holomorphic function f and its non tangential limit $f^\sharp$ is one-to-one and even isometric, namely the supremum in (2) is equal to $\|f^\sharp\|_p$, thereby allowing us to identify f and $f^\sharp$ and to drop the superscript $\sharp$. Under this identification, we regard members of H^p both as functions in $L^p(\mathbb{T})$ and as holomorphic functions in the variable $z \in \mathbb{D}$, but the argument (which belongs to $\mathbb{T}$ in the former case and to $\mathbb{D}$ in the latter) helps preventing confusion. It holds in fact that $f_r(e^{i\theta}) = f(re^{i\theta})$ converges as $r \to 1^-$ to $f(e^{i\theta})$ in $L^p(\mathbb{T})$ when $f \in H^p$ and $1 \le p < \infty$. It follows immediately from (2) and Hölder's inequality that, whenever $g_1 \in H^{p_1}$ and $g_2 \in H^{p_2}$, we have $g_1 g_2 \in H^{p_3}$ if $1/p_1 + 1/p_2 = 1/p_3$.

Given $f \in H^p$, its values on $\mathbb{D}$ are obtained from its values on $\mathbb{T}$ through a Cauchy as well as a Poisson integral [17, ch. 17, thm 11], namely:

$$f(z) = \frac{1}{2\,i\,\pi} \int_{\mathbb{T}} \frac{f(\xi)}{\xi - z} \, d\xi\,, \quad \text{and also} \quad f(z) = \frac{1}{2\,\pi} \int_{\mathbb{T}} \mathrm{Re}\left\{ \frac{e^{i\theta} + z}{e^{i\theta} - z} \right\} f(e^{i\theta}) \, d\theta\,, \quad z \in \mathbb{D}, \tag{3}$$

where the right hand side of the first equality in (3) is a line integral. The latter immediately implies that the Fourier coefficients of a Hardy function on the circle are the Taylor coefficients of its power series expansion at 0 when viewed as a holomorphic function on $\mathbb{D}$. In this connection, the space H^2 is especially simple to describe: it consists of those holomorphic functions g in $\mathbb{D}$ whose Taylor coefficients at 0 are square summable, namely

$$g(z) = \sum_{k=0}^{\infty} a_k z^k : \quad \|g\|_{H^2}^2 := \sum_{k=0}^{\infty} |a_k|^2 < +\infty, \qquad g(e^{i\theta}) = \sum_{k=0}^{\infty} a_k e^{ik\theta}, \tag{4}$$

where the convergence of the last Fourier series holds in $L^2(\mathbb{T})$ by Parseval's theorem (and also pointwise a.e. by Carleson's theorem but we do not need this deep result). Incidentally, let us mention that for no other value of p is it known how to characterize H^p in terms of the size of its Fourier coefficients.

By the Poisson representation (*i.e.* the second integral in (3)), a Hardy function g is also uniquely represented, up to a purely imaginary constant, by its real part h on $\mathbb{T}$ according to:

$$g(z) = i\mathrm{Im}g(0) + \frac{1}{2\,\pi} \int_{\mathbb{T}} \frac{e^{i\theta} + z}{e^{i\theta} - z} \, h(e^{i\theta}) \, d\theta\,, \quad z \in \mathbb{D}. \tag{5}$$

The integral in (5) is called the *Riesz-Herglotz transform* of h and, whenever $h \in L^1_{\mathbb{R}}(\mathbb{T})$, it defines a holomorphic function in $\mathbb{D}$ which is real at 0 and whose nontangential limit exists a.e. on $\mathbb{T}$ with real part equal to h. Hence the Riesz-Herglotz transform (5) assumes the form $h(e^{i\theta}) + i\widetilde{h}(e^{i\theta})$ a.e. on $\mathbb{T}$, where the real-valued function $\widetilde{h}$ is said to be *conjugate* to h. It is a theorem of M. Riesz [11, chap. III, thm 2.3] that if $1 < p < \infty$, then $\widetilde{h} \in L^p_{\mathbb{R}}(\mathbb{T})$ when $h \in L^p_{\mathbb{R}}(\mathbb{T})$. This neither holds for $p = 1$ nor for $p = \infty$.

A nonzero $f \in H^p$ can be uniquely factored as $f = jw$ where

$$w(z) = \exp\left\{ \frac{1}{2\pi} \int_0^{2\pi} \frac{e^{i\theta} + z}{e^{i\theta} - z} \log |f(e^{i\theta})| \, d\theta \right\} \tag{6}$$

belongs to H^p and is called the *outer factor* of f, while $j \in H^\infty$ has modulus 1 a.e. on $\mathbb{T}$ and is called the *inner factor* of f. That $w(z)$ in (6) is well-defined rests on the fact that $\log|f| \in L^1$ if $f \in H^1 \setminus \{0\}$; it entails that a H^p function cannot vanish on a subset of strictly positive Lebesgue measure on $\mathbb{T}$ unless it is identically zero. For simplicity, we often say that a function is outer (resp. inner) if it is equal, up to a unimodular multiplicative constant, to its outer (resp. inner) factor.

Closely connected to Hardy spaces is the Nevanlinna class N^+, consisting of holomorphic functions in $\mathbb{D}$ that can be factored as jE, where j is an inner function and E an outer function of the form

$$E(z) = \exp\left\{ \frac{1}{2\pi} \int_0^{2\pi} \frac{e^{i\theta} + z}{e^{i\theta} - z} \log \rho(e^{i\theta}) \, d\theta \right\}, \tag{7}$$

with ρ a positive function such that $\log \rho \in L^1(\mathbb{T})$ (though ρ itself may not be summable). Such a function has nontangential limits of modulus ρ a.e. on $\mathbb{T}$. The Nevanlinna class is instrumental in that

$N^+ \cap L^p(\mathbb{T}) = H^p$, see [10, thm 2.11] or [11, 5.8, ch.II]. Thus, formula (7) defines a H^p-function if and only if $\rho \in L^p(\mathbb{T})$.

Let $\hat{\mathbb{C}} = \mathbb{C} \cup \{\infty\}$ be the Riemann sphere. The Hardy space $\bar{H}^p$ of $\hat{\mathbb{C}} \setminus \overline{\mathbb{D}}$ can be given a treatment parallel to H^p upon changing z into $1/z$. Specifically, $\bar{H}^p$ consists of functions in $L^p(\mathbb{T})$ whose Fourier coefficients of strictly positive index do vanish; these are, a.e. on $\mathbb{T}$, the complex conjugates of H^p-functions, and they can also be viewed as nontangential limits of functions analytic in $\hat{\mathbb{C}} \setminus \overline{\mathbb{D}}$ having uniformly bounded L^p means over all circles centered at 0 of radius bigger than 1. We further single out the subspace $\bar{H}^p_0$ of $\bar{H}^p$, consisting of functions vanishing at infinity or, equivalently, having vanishing mean on $\mathbb{T}$. Thus, a function belongs to $\bar{H}^p_0$ if, and only if it is of the form $e^{-i\theta}\overline{g(e^{i\theta})}$ for some $g \in H^p$. For $G \in \bar{H}^p_0$, the Cauchy formula assumes the form :

$$G(z) = \frac{1}{2\,i\,\pi} \int_{\mathbb{T}} \frac{G(\xi)}{z-\xi}\, d\xi\,, \qquad z \in \hat{\mathbb{C}} \setminus \overline{\mathbb{D}}. \tag{8}$$

It follows at once from the Cauchy formula that the duality product $\langle\ ,\ \rangle_{\mathbb{T}}$ makes H^p and $\bar{H}^q_0$ orthogonal to each other, and it reduces to the familiar scalar product when $p = q = 2$. In particular, we have the orthogonal decomposition :

$$L^2(\mathbb{T}) = H^2 \oplus \bar{H}^2_0. \tag{9}$$

For $f \in C(\mathbb{T})$ and $\nu \in \mathcal{M}$, the space of complex Borel measures on $\mathbb{T}$, we set

$$\nu.f = \int_{\mathbb{T}} f(e^{i\theta})\, d\nu(\theta) \tag{10}$$

and this pairing induces an isometric isomorphism between $\mathcal{M}$ (endowed with the norm of the total variation) and the dual of $C(\mathbb{T})$ [17, thm 6.19]. If we let $\mathcal{A} \subset H^\infty$ designate the disk algebra of functions analytic in $\mathbb{D}$ and continuous on $\overline{\mathbb{D}}$, and if $\mathcal{A}_0$ indicates those functions in $\mathcal{A}$ vanishing at zero, it is easy to see that $\mathcal{A}_0$ is the orthogonal space under (10) to those measures whose Fourier coefficients of strictly negative index do vanish. Now, it is a fundamental theorem of F. and M. Riesz that such measures are absolutely continuous, that is have the form $d\nu(\theta) = g(e^{i\theta})\, d\theta$ with $g \in H^1$. The Hahn-Banach theorem implies that H^1 is dual *via* (10) to the quotient space $C(\mathbb{T})/\mathcal{A}_0$ [11, chap. IV, sec. 1]. Equivalently, $\bar{H}^1_0$ is dual to $C(\mathbb{T})/\overline{\mathcal{A}}$ under the pairing arising from the line integral :

$$(\dot{f}, F) = \frac{1}{2i\pi} \int_{\mathbb{T}} f(\xi) F(\xi)\, d\xi\,, \tag{11}$$

where F belongs to $\bar{H}^1_0$ and $\dot{f}$ indicates the equivalence class of $f \in C(\mathbb{T})$ modulo $\overline{\mathcal{A}}$. Therefore, contrary to $L^1(\mathbb{T})$, the spaces H^1 and $\bar{H}^1_0$ enjoy a weak-* compactness property of their unit ball.

We define the analytic and anti-analytic projections $\mathbf{P}_+$ and $\mathbf{P}_-$ on Fourier series by :

$$\mathbf{P}_+\left(\sum_{n=-\infty}^{\infty} a_n e^{in\theta}\right) = \sum_{n=0}^{\infty} a_n e^{in\theta}, \qquad \mathbf{P}_-\left(\sum_{n=-\infty}^{\infty} a_n e^{in\theta}\right) = \sum_{n=-\infty}^{-1} a_n e^{in\theta}\,.$$

It is a theorem of M. Riesz theorem [11, ch. III, sec, 1] that $\mathbf{P}_+ : L^p \to H^p$ and $\mathbf{P}_- : L^p \to \bar{H}^p_0$ are bounded for $1 < p < \infty$, in which case they coincide with the Cauchy projections:

$$\mathbf{P}_+(h)(z) = \frac{1}{2i\pi} \int_{\mathbb{T}} \frac{h(\xi)}{\xi - z}\, d\xi, \quad z \in \mathbb{D}, \qquad \mathbf{P}_-(h)(s) = \frac{1}{2i\pi} \int_{\mathbb{T}} \frac{h(\xi)}{s - \xi}\, d\xi, \quad s \in \hat{\mathbb{C}} \setminus \overline{\mathbb{D}}. \tag{12}$$

When restricted to $L^2(\mathbb{T})$, the projections $\mathbf{P}_+$ and $\mathbf{P}_-$ are just the orthogonal projections onto H^2 and $\bar{H}^2_0$ respectively. Although $\mathbf{P}_\pm(h)$ needs not be the Fourier series of a function when h is merely in $L^1(\mathbb{T})$, it is Abel summable almost everywhere to a function lying in $L^s(\mathbb{T})$ for $0 < s < 1$ and it can be interpreted as a function in the Hardy space of exponent s that we did not introduce [10, cor. to thm 3.2]. To us it will be sufficient, when $h \in L^1$, to regard $\mathbf{P}_\pm(f)$ as the Fourier series of a distribution.

Finally, we let P_n denote throughout the space of complex algebraic polynomials of degree at most n. Clearly, $P_n \subset H^p$ for all p.

4 Two extremal problems

We first state the polynomial problem discussed in Section 2. We call it $PBEP(n)$ for "Polynomial Bounded Extremal Problem":

PBEP(n)

For $f \in L^2(I)$, find $k_n \in P_n$ such that $|k_n(e^{i\theta})| \leq 1$ for a.e. $e^{i\theta} \in J$ and

$$\|f - k_n\|_{L^2(I)} = \inf_{\substack{g \in P_n \\ |g| \leq 1 \text{ a.e. on } J}} \|f - g\|_{L^2(I)} . \tag{13}$$

Next, we state the analytic problem from Section 2 that we call $ABEP$ for "Analytic Bounded Extremal Problem":

ABEP

Given $f \in L^2(I)$, find $g_0 \in H^2$ such that $|g_0(e^{i\theta})| \leq 1$ a.e. on J and

$$\|f - g_0\|_{L^2(I)} = \inf_{\substack{g \in H^2 \\ |g| \leq 1 \text{ a.e. on } J}} \|f - g\|_{L^2(I)} . \tag{14}$$

Note that, in $ABEP$, the constraint $|g| \leq 1$ on J could be replaced by $|g| \leq \rho$ where ρ is a positive function in $L^2(J)$. For if $\log \rho \in L^1(J)$ then, denoting by $w_{1 \vee (1/\rho)}$ the outer factor having modulus 1 on I and $1/\rho$ on J, we find that $g \in H^2$ satisfies $|g| \leq \rho$ on J if and only if $h = g w_{1 \vee (1/\rho)}$ lies in H^2 and satisfies $|h| \leq 1$ on J. It is so because, for g as indicated, h lies in the Nevanlinna class by construction and $|h|_{|I} = |g|_{|I}$ while $|h|_{|J} = |g|_{|J}/\rho$. If, however, $\log \rho \notin L^1(J)$, then we must have $\int_J \log \rho = -\infty$ because $\rho \in L^2(J)$, consequently the set of candidate approximants reduces to $\{0\}$ anyway because a nonzero Hardy function has summable log-modulus. Altogether, it is thus equivalent to consider $ABEP$ for the product f times $(w_{1 \vee \rho^{-1}})_{|I}$. A similar argument shows that we could replace the error criterion $\|.\|_{L^2(I)}$ by a weighted norm $\|.\|_{L^2(I,w)}$ for some weight w which is non-negative and invertible in $L^\infty(I)$. Then, the problem reduces to $ABEP$ for $f(w_{\rho^{1/2} \vee 0})_I$.

Such equivalences do not hold for $PBEP(n)$ because the polynomial character of k_n is not preserved under multiplication by outer factors. Still, the results to come continue to hold if we replace in $PBEP(n)$ the constraint $|k_n| \leq 1$ by $|k_n| \leq \rho$ on J and the criterion $\|.\|_{L^2(I)}$ by $\|.\|_{L^2(I,w)}$, provided that $\rho \in C(J)$ and that w is invertible in $L^\infty(I)$. Indeed, we leave it to the reader to check that proofs go through with obvious modifications.

After these preliminaries, we are ready to state a basic existence and uniqueness result.

Theorem 1 . *Problems $PBEP(n)$ and $ABEP$ have a unique solution. Moreover, the solution g_0 to $ABEP$ satisfies $|g_0| = 1$ almost everywhere on J, unless $f = g_{|I}$ for some $g \in H^2$ such that $\|g\|_{L^\infty(J)} \leq 1$.*

Proof. Consider the sets

$$E_n = \{g_{|I} : \ g \in P_n, \ \|g\|_{L^\infty(J)} \leq 1\},$$
$$F = \{g_{|I} : \ g \in H^2, \ \|g\|_{L^\infty(J)} \leq 1\}.$$

Clearly $E_n \subset F$ are convex and nonempty subsets of $L^2(I)$, as they contain 0. To prove existence and uniqueness, it is therefore enough to show they are closed, for we can appeal then to well-known properties of the projection on a closed convex set in a Hilbert space. Since $E_n = P_n \cap F$, it is enough in fact to show that F is closed. For this, let g_m be a sequence in H^2 with $|g_m|_{|J} \leq 1$ and such that $(g_m)_{|I}$ converges in $L^2(I)$. Obviously g_m is a bounded sequence in $L^2(\mathbb{T})$, some subsequence of which converges weakly to $h \in H^2$. We continue to denote this subsequence with g_m. The restrictions $(g_m)_{|I}$ a fortiori converge weakly to $h_{|I}$ in $L^2(I)$, and since the strong and the weak limit must coincide when both exist we find that $(g_m)_{|I}$ converges to $h_{|I}$ in $L^2(I)$. Besides, $(g_m)_{|J}$ is contained in the unit ball of $L^\infty(J)$ which is dual to $L^1(J)$, hence some subsequence (again denoted by $(g_m)_{|J}$) converges weak-* to some $h_1 \in L^\infty(J)$ with $\|h_1\|_{L^\infty(J)} \leq 1$. But since $(g_m)_{|J}$ also converges weakly to $h_{|J}$ in $L^2(J)$, we have that

$$\langle h_1, \varphi \rangle_J = \lim_{m \to \infty} \langle g_m, \varphi \rangle_J = \langle h_{|J}, \varphi \rangle_J$$

for all $\varphi \in L^2(J)$ which is dense in $L^1(J)$. Consequently $h_1 = h_{|J}$, thereby showing that $\|h\|_{L^\infty(J)} \leq 1$, which proves that F is closed.

Assume now that f is not the trace on I of an H^2-function which is less than 1 in modulus on I. To prove that $|g_0| = 1$ a.e. on J, we argue by contradiction. If not, there is a compact set K of positive measure, lying interior to J, such that $\|g_0\|_{L^\infty(K)} \leq 1 - \delta$ for some $0 < \delta < 1$; it is so because, by

hypothesis, J must consist of finitely many closed arcs, of which one at least has nonempty interior. For K' an arbitrary subset of K, consider the Riesz-Herglotz transform of its characteristic function:

$$h_{K'}(z) = \frac{1}{2\,\pi}\int_{K'} \frac{e^{i\theta}+z}{e^{i\theta}-z}\,d\theta\,, \qquad z\in\mathbb{D}, \tag{15}$$

and put $w_t = \exp(t h_{K'})$ for $t\in\mathbb{R}$, which is the outer function with modulus $\exp t$ on K' and 1 elsewhere. By construction, $g_0 w_t$ is a candidate approximant in $ABEP$ for all $t < -\log(1-\delta)$. Thus, the map $t\mapsto \|f-g_0 w_t\|^2_{L^2(I)}$ attains a minimum at $t=0$. Because K is at strictly positive distance from I, we may differentiate this expression with respect to t under the integral sign and equate the derivative at $t=0$ to zero which gives us $2\mathrm{Re}\langle f-g_0, h_{K'}g_0\rangle_I = 0$. Replacing $g_0 w_t$ by $ig_0 w_t$, which is a candidate approximant as well, we get a similar equation for the imaginary part so that

$$0 = \langle f-g_0\,,\, h_{K'}g_0\rangle_I = \langle (f-g_0)\bar{g}_0\,,\, h_{K'}\rangle_I. \tag{16}$$

Let e^{it_0} be a density point of K and I_l the arc centered at e^{it_0} of length l, so that $|I_l\cap K)|/l\to 1$ as $l\to 0$. Since

$$\left|\frac{e^{it}+e^{i\theta}}{e^{it}-e^{i\theta}} - \frac{e^{it_0}+e^{i\theta}}{e^{it_0}-e^{i\theta}}\right| \le \frac{2l}{\mathrm{dist}^2(K,I)} \qquad \text{for } e^{it}\in I_l\cap K,\quad e^{i\theta}\in I, \tag{17}$$

it follows by dominated convergence that

$$\lim_{l\to 0}\frac{1}{|I_l\cap K|}\int_{I_l\cap K}\left|\frac{e^{it}+e^{i\theta}}{e^{it}-e^{i\theta}} - \frac{e^{it_0}+e^{i\theta}}{e^{it_0}-e^{i\theta}}\right| dt = 0, \quad \text{uniformly w.r. to } e^{i\theta}\in I,$$

and therefore that

$$\lim_{l\to 0}\frac{h_{I_l\cap K}(e^{i\theta})}{|I_l\cap K|} = \frac{e^{it_0}+e^{i\theta}}{e^{it_0}-e^{i\theta}} \quad \text{uniformly w.r. to } e^{i\theta}\in I.$$

Applying now (16) with $K' = I_l\cap K$ and taking into account that $(e^{it_0}+e^{i\theta})/(e^{it_0}-e^{i\theta})$ is pure imaginary on I, we find in the limit, as $l\to 0$ that

$$\frac{1}{2\pi}\int_I \frac{e^{it_0}+e^{i\theta}}{e^{it_0}-e^{i\theta}}\Big((f-g_0)\bar{g}_0\Big)(e^{i\theta})\,d\theta = 0. \tag{18}$$

Next, let us consider the function

$$F(z) = \frac{1}{2\pi}\int_I \frac{e^{i\theta}+z}{e^{i\theta}-z}\Big((f-g_0)\bar{g}_0\Big)(e^{i\theta})\,d\theta = -\frac{1}{2\pi}\int_I\Big((f-g_0)\bar{g}_0\Big)(e^{i\theta})d\theta + \frac{1}{i\pi}\int_I \frac{\Big((f-g_0)\bar{g}_0\Big)(\xi)d\xi}{\xi-z}$$

which is the sum of a constant and of twice the Cauchy integral of $(f-(g_0)_{|_I})(\bar{g}_0)_{|_I}\in L^1(I)$, hence is analytic in $\hat{\mathbb{C}}\setminus I$. Equation (18) means that F vanishes at every density point of K, and since a.e. point in K is a density point F must vanish identically because its zeros accumulate in the interior of J. Denoting by F^+ and F^- the nontangential limits of F from sequences of points in $\mathbb{D}$ or $C\setminus\overline{\mathbb{D}}$ respectively, we now get from the Plemelj-Sokhotski formulas [11, ch. III] that

$$0 = F^+(\xi) - F^-(\xi) = (f-g_0)(\xi)\overline{g_0(\xi)}, \qquad a.e.\ \xi\in I.$$

Thus, either g_0 is nonzero a.e. on I, in which case $f=(g_0)_{|_I}$ and we reach the desired contradiction, or else $g_0\equiv 0$. In the latter case, if we put id for the identity map on $\mathbb{T}$, we find that $t\mapsto\|f-t\,\mathrm{id}^k\|^2_{L^2(I)}$ has a minimum at $t=0$ for each integer $k\ge 0$, since $e^{i\theta}\mapsto te^{ik\theta}$ is a candidate approximant for $t\in[-1,1]$. Differentiating with respect to t and expressing that the derivative at $t=0$ is zero, we deduce that all Fourier coefficients of non-negative index of $(f-(g_0)_{|_I})\vee 0$ do vanish. This means this last function lies in $\bar{H}^2$, but as it vanishes on J it is identically zero, therefore $f=(g_0)_{|_I}$ in all cases. ∎

Remark: the theorem shows that the constraint $|g_0|\le 1$ on J is saturated in a very strong sense for problem $ABEP$, namely $|g_0|=1$ a.e. on J unless f is already the trace of the solution on I. In contrast, it is not true that $\|k_n\|_{L^\infty(J)}=1$ unless $f=g_{|_I}$ for some $g\in P_n$ such that $\|g\|_{L^\infty(J)}<1$. To see this, observe that the set E_n is not only closed but compact. Indeed, if we pick distinct points $\xi_1,\cdots,\xi_{n+1}$ in J and form the Lagrange interpolation polynomials $L_j\in P_n$ such that $L_j(\xi_j)=1$ and $L_j(\xi_\ell)=0$ if $\ell\ne j$, we get a basis of P_n in which the coordinates of every $g\in P_n$ meeting $\|g\|_{L^\infty(J)}\le 1$ are bounded by 1 in modulus. Hence E_n is bounded in P_n, and since it is closed by the proof of Theorem 1 it is compact. Thus, each $f\in L^2(I)$ has a best approximant from E_n, and if $(p_n)_{_I}$ is a best approximant to f with $p_n\in P_n$, then for $\lambda>\|p_n\|_{L^\infty(J)}$ we find that p_n/λ is a best approximant to f/λ in $L^2(I)$ which is strictly less than 1 on J. This justifies the remark.

5 Critical point equations and convergence of approximants

At this point, it is worth recalling informally some basic principles from convex optimization, for which the reader may consult [7]. The solution to a strictly convex minimization problem is characterized by a variational inequality expressing that the *criterium* increases under admissible increments of the variable. If the problem is smooth enough, such increments admit a tangent space at the point under consideration (*i.e.* the solution) in the variable space. We term it the tangent space to the constraints, and its orthogonal in the dual space to the variable space is called the orthogonal space to the constraints (at the point under consideration). The variation of the objective function must vanish on the tangent space to the constraints to the first order, thereby giving rise to the so-called *critical point equation*. It says that the gradient of the objective function, viewed as an element of the dual space to the variable space, lies in the orthogonal space to the constraints. If a basis of the latter is chosen, the coordinates of the gradient in this basis are known as the *Lagrange parameters*. More generally, one can form the Lagrangian which is a function of the variable and of the Lagrange parameters, not necessarily optimal ones. It is obtained by adding the gradient of the criterion, at the considered value of the variable, with the member of the orthogonal space to the constraints defined by the chosen Lagrange parameters. By what precedes, the Lagrangian must vanish at the solution for appropriate values of the Lagrange parameters. One can further define a function of the Lagrange parameters only, by minimizing the Lagrangian with respect to the variable. This results in a concave function which gets maximized at the optimal value of the Lagrange parameters for the original problem. This way, one reduces the original constrained convex minimization problem to an unsconstrained concave maximization problem, called the dual problem. In an infinite-dimensional context, the arguments needed to put this program to work may be quite subtle.

Below we derive the critical point equation for $PBEP(n)$ described in (13). For $g \in P_n$ define

$$E(g) = \{x \in J,\ |g(x)| = ||g||_{L^\infty(J)}\},$$

which is the set of extremal points of g on J.

Theorem 2 *A polynomial $g \in P_n$ is the solution to $PBEP(n)$ iff the following two conditions hold:*

- $||g||_{L^\infty(J)} \leq 1$,
- *there exists a set of r distinct points $x_1, \cdots, x_r \in E(g)$ and non-negative real numbers $\lambda_1, \cdots, \lambda_r$, with $0 \leq r \leq 2n+2$, such that*

$$\langle g - f, h\rangle_I + \sum_{j=1}^{r} \lambda_j g(x_j)\overline{h(x_j)} = 0, \qquad \forall h \in P_n. \tag{19}$$

Moreover the λ_j's meet the following bound

$$\sum_{j=1}^{r} \lambda_j \leq 2||f||^2_{L^2(I)}. \tag{20}$$

We emphasize that the set of extremal points $\{x_j,\ j = 1, \ldots, r\}$ is possibly empty (i.e $r = 0$).

Proof. Suppose g verifies the two conditions and differs from the solution k_n. Set $h = k_n - g \in P_n$ and observe that

$$\mathrm{Re}\left(g(x_i)\overline{h(x_i)}\right) = \mathrm{Re}\left(g(x_i)\overline{k_n(x_i)} - 1\right) \leq 0,\ \ i = 1 \ldots r. \tag{21}$$

From the uniqueness and optimality of k_n we deduce that

$$\begin{aligned} ||k_n - f||^2_{L^2(I)} &= ||g - f + h||^2_{L^2(I)} \\ &= ||g - f||^2_{L^2(I)} + ||h||^2_{L^2(I)} + 2\mathrm{Re}\langle g - f, h\rangle_I \\ &< ||g - f||^2_{L^2(I)}. \end{aligned}$$

Consequently $\mathrm{Re}\langle g - f, h\rangle_{L^2(I)} < 0$ which, combined with (21), contradicts (19).

Conversely, suppose that g is the solution to $PBEP(n)$ and let ϕ_0 be the $\mathbb{R}$-linear forms on P_n given by

$$\phi_0(h) = \mathrm{Re}\langle g - f, h\rangle_I, \qquad h \in P_n.$$

For each extremal point $x \in E(g)$, define further a $\mathbb{R}$-linear form ϕ_x by

$$\phi_x(h) = \mathrm{Re}\left(g(x)\overline{h(x)}\right), \quad h \in P_n.$$

Put K for the union of these forms:

$$K = \{\phi_0\} \cup \{\phi_x,\ x \in E(g)\}.$$

If we let $P_n^{\mathbb{R}}$ indicate P_n viewed as a real vector space, K is a subset of the dual $(P_n^{\mathbb{R}})^*$. As J is closed by definition, simple inspection shows that K is closed and bounded in $(P_n^{\mathbb{R}})^*$ (it is in fact finite unless g is a constant), hence it is compact and so is its convex hull $\hat{K}$ as $(P_n^{\mathbb{R}})^*$ is finite-dimensional. Suppose for a contradiction that $0 \notin \hat{K}$. Then, since $(P_n^{\mathbb{R}})^{**} = P_n^{\mathbb{R}}$ because $P_n^{\mathbb{R}}$ is finite-dimensional, there exists by the Hahn-Banach theorem an $h_0 \in P_n$ such that,

$$\phi(h_0) \geq \tau > 0, \quad \forall \phi \in \hat{K}.$$

The latter and the continuity of g and h_0 ensure the existence of a neighborhood V of $E(g)$ on $\mathbb{T}$ such that for x in $U = J \cap V$ we have $\mathrm{Re}\left(g(x)\overline{h_0(x)}\right) \geq \frac{\tau}{2} > 0$, whereas for x in $J \backslash U$ it holds that $|g(x)| \leq 1 - \delta$ for some $\delta > 0$. Clearly, for $\epsilon > 0$ with $\epsilon ||h_0||_{L^\infty(J)} < \delta$, we get that

$$\sup_{J \backslash U} |g(x) - \epsilon h_0(x)| \leq 1. \tag{22}$$

Moreover, assuming without loss of generality that $\epsilon < 1$, it holds for $x \in U$ that

$$\begin{aligned} |g(x) - \epsilon h_0(x)|^2 &= |g(x)|^2 - 2\mathrm{Re}\left(g(x)\overline{h_0(x)}\right) + \epsilon^2 |h_0(x)|^2 \\ &\leq |g(x)|^2 - 2\mathrm{Re}\left(\epsilon g(x)\overline{h_0(x)}\right) + \epsilon^2 |h_0(x)|^2 \\ &\leq 1 - \epsilon\tau + \epsilon^2 ||h_0||^2_{L^\infty(J)}. \end{aligned}$$

The latter combined with (22) shows that, for ϵ sufficiently small, we have

$$||g - \epsilon h_0||_{L^\infty(J)} \leq 1. \tag{23}$$

However, since

$$\begin{aligned} ||f - g - \epsilon h_0||^2_{L^2(J)} &= ||f - g||^2_{L^2(J)} - 2\epsilon\phi_0(h_0) + \epsilon^2 ||h_0||^2_{L^2(J)} \\ &\leq ||f - g||^2_{L^2(J)} - 2\epsilon\tau + \epsilon^2 ||h_0||^2_{L^2(J)}, \end{aligned} \tag{24}$$

we deduce in view of (23) that for ϵ small enough the polynomial $g - \epsilon h_0$ performs better than g in $FBEP$, thereby contradicting optimality. Hence $0 \in \hat{K}$, therefore by Carathedory's theorem [8, ch. 1, sec. 5] there are r' elements γ_j of K, with $1 \leq r' \leq 2(n+1)+1$ (the real dimension of $P_n^{\mathbb{R}}$ plus one), such that

$$\sum_{j=1}^{r'} \alpha_j \gamma_j = 0 \tag{25}$$

for some positive α_j satisfying $\sum \alpha_j = 1$. Of necessity ϕ_0 is a γ_j, otherwise evaluating (25) at g yields the absurd conclusion that

$$0 = \sum_{j=1}^{r'} \alpha_j \gamma_j(g) = \sum_{j=1}^{r'} \alpha_j |g(x_j)|^2 = 1.$$

Equation (25) can therefore be rewritten as

$$\alpha_1 \mathrm{Re}\langle f - g, h\rangle_I + \sum_{j=2}^{r'} \alpha_j \mathrm{Re}(g(x_j)\overline{h(x_j)}) = 0 \quad \forall h \in P_n, \quad \alpha_1 \neq 0.$$

Dividing by α_1 and noting that the last equation is also true with ih instead of h yields (19) with $r = r' - 1$. Finally, replacing h by g in (19) we obtain

$$\sum_{j=1}^{r}|\lambda_j| = \sum_{j=1}^{r}\lambda_j| = \langle f-g, g\rangle_I \leq \ \langle f-g, f-g\rangle_I + |\langle f-g, f\rangle_I|$$
$$\leq ||f-g||^2_{L^2(I)} + ||f-g||_{L^2(I)}||f||_{L^2(I)}$$
$$\leq 2||f||^2_{L^2(I)}$$

where the next to last majorization uses the Schwarz inequality and the last that 0 is a candidate approximant for $PBEP(n)$ whereas g is the optimum. ■

The next result describes the behavior of k_n when n goes to infinity, in connection with the solution g_0 to $ABEP$.

Theorem 3 *Let k_n be the solution to $PBEP(n)$ defined in (13), and g_0 the solution to $ABEP$ described in (14). When $n \to \infty$, the sequence $(k_n)_{|_I}$ converges to $(g_0)_{|_I}$ in $L^2(I)$, and the sequence $(k_n)_{|_J}$ converges to $(g_0)_{|_J}$ in the weak-* topology of $L^\infty(J)$, as well as in $L^p(J)$-norm for $1 \leq p < \infty$ if f is not the race on I of a H^2-function which is at most 1 in modulus on J. Altogether this amounts to:*

$$\lim_{n\to\infty} ||g_0 - k_n||_{L^p(\mathbb{T})} = 0, \quad 1 \leq p \leq 2, \tag{26}$$

$$\lim_{n\to\infty} \langle k_n, h\rangle_J = \langle g_0, h\rangle_J \quad \forall h \in L^1(J), \tag{27}$$

$$\text{if } f \neq g_0 \text{ on } I, \ \lim_{n\to\infty} ||g_0 - k_n||_{L^p(J)} = 0, \quad 1 \leq p < \infty. \tag{28}$$

Proof. Our first objective is to show that g_0 can be approximated arbitrary close in $L^2(I)$ by polynomials that remain bounded by 1 in modulus on J. By hypothesis I is the finite union of $N \geq 1$ open disjoint sub-arcs of $\mathbb{T}$. Without loss of generality, it can thus be written as

$$I = \bigcup_{i=1}^{N}(e^{ia_i}, e^{ib_i}), \qquad 0 = a_1 \leq b_1 \leq a_2 \cdots \leq b_N \leq 2\pi.$$

Let (ϵ_n) be a sequence of positive real numbers decreasing to 0. We define a sequence (v_n) in H^2 by

$$v_n(z) = g_0(z)\exp\left(-\frac{1}{2\pi}\left(\sum_{i=1}^{N}\int_{a_i}^{a_i+\epsilon_n}\frac{e^{it}+z}{e^{it}-z}\log|g_0|dt + \int_{b_i-\epsilon_n}^{b_i}\frac{e^{it}+z}{e^{it}-z}\log|g_0|dt\right)\right)$$

Note that indeed $v_n \in H^2$ for n large enough because then it has the same modulus as g_0 except over the arcs $(a_i, a_i+\epsilon_n)$ and $(b_i-\epsilon_n, b_i)$ where it has modulus 1. *We claim* that $(v_n)_{|_I}$ converges to g_0 in $L^2(I)$ as $n \to \infty$. To see this, observe that v_n converges a.e. on I to g_0, for each $z \in I$ remains at some distance from the sub-arcs $(a_i, a_i+\epsilon_n)$ and $(b_i, b_i+\epsilon_n)$ for all n sufficiently large, hence the argument of the exponential in (29) converges to zero as $n \to \infty$ by absolute continuity of $\log|g_0|dt$. Now, we remark that by construction $|v_n| \leq |g_0| + 1$, hence by dominated convergence, we get that

$$\lim_{n\to\infty} ||g_0 - v_n||_{L^2(I)} = 0.$$

This proves the claim. Now, let $\epsilon > 0$ and $0 < \alpha < 1$ such that $||g_0 - \alpha g_0||_{L^2(I)} \leq \frac{\epsilon}{4}$. Let also n_0 be so large that $||v_{n_0} - g_0||_{L^2(I)} \leq \frac{\epsilon}{4}$. For $0 < r < 1$ define $u_r \in \mathcal{A}$ (the disk algebra) by $u_r(z) = v_{n_0}(rz)$ so that, by Poisson representation,

$$u_r(e^{i\theta}) = \int_{\mathbb{T}} P_r(\theta - t)v_{n_0}(re^{it})dt,$$

where P_r is the Poisson kernel. Whenever $e^{i\phi} \in J$, we note by construction that $|v_n| = 1$ a.e on the sub-arc $(e^{i(\phi-\epsilon_{n_0})}, e^{i(\phi+\epsilon_{n_0})})$. This is to the effect that

$$|u_r(e^{i\phi})| \leq \int_{\mathbb{T}} P_r(\phi - t)|v_{n_0}(re^{it})|dt$$
$$\leq P_r(\epsilon_{n_0})\int_{\mathbb{T}}|v_{n_0}(re^{it})|dt + \int_{-\epsilon_{n_0}}^{+\epsilon_{n_0}} P_r(t)dt$$
$$\leq P_r(\epsilon_{n_0})||v_{n_0}||_{L^1(\mathbb{T})} + 1 \leq P_r(\epsilon_{n_0})||v_{n_0}||_{L^2(\mathbb{T})} + 1$$

by Hölder's inequality. Hence, for r sufficiently close to 1, we certainly have that $|u_r| \leq 1/\alpha^2$ on J and otherwise that $||u_r - v_{n_0}||^2_{L^2(I)} \leq \frac{\epsilon}{4}$ since $u_r \to v_{n_0}$ in H^2. Finally, call q the truncated Taylor expansion of u_r (which converges uniformly to the latter on $\mathbb{T}$), where the order of truncation has been chosen large enough to ensure that $|q| \leq 1/\alpha$ on J and that $||q - u_r||^2_{L^2(I)} \leq \frac{\epsilon}{4}$. Then, we have that

$$\begin{aligned} ||\alpha q - g_0||_{L^2(I)} &\leq \alpha \left(||q - u_r||_{L^2(I)} + ||u_r - v_{n_0}||_{L^2(I)} + ||v_{n_0} - g_0||_{L^2(I)} \right) + ||g_0 - \alpha g_0||_{L^2(I)} \\ &\leq \epsilon. \end{aligned}$$

Thus, we have found a polynomial (namely αq) which is bounded by 1 in modulus on J and close by ϵ to g_0 in $L^2(I)$. By comparison, this immediately implies that

$$\lim_{n \to \infty} ||f - k_n||_{L^2(I)} = ||f - g_0||_{L^2(I)}, \tag{29}$$

from which (26) follows by Hölder's inequality. Moreover, being bounded in H^2, the sequence (k_n) has a weakly convergent sub-sequence. The traces on J of this subsequence are in fact bounded by 1 in $L^\infty(J)$-norm, hence up to another subsequence we obtain (k_{n_m}) converging also in the weak-* sense on J. Let g be the weak limit (H^2 sense) of k_{n_m}, and observe that $g_{|_J}$ is necessarily the weak-* limit of $(k_{n_m})_{|_J}$ in $L^\infty(J)$, as follows by integrating against functions from $L^2(J)$ which is dense in $L^1(J)$. Since balls are weak-* closed in $L^\infty(J)$, we have that $||g||_{L^\infty(J)} \leq 1$, and it follows from (29) that $||f - g||_{L^2(I)} = ||f - g_0||_{L^2(I)}$. Thus, $g = g_0$ by the uniqueness part of Theorem 1. Finally, if $f \neq g_0$ on J, then we know from Theorem 1 that $|g_0| = 1$ a.e. on J. In this case, (29) implies that $\limsup ||k_{n_m}||_{L^2(\mathbb{T})} \leq ||g_0||_{L^2(\mathbb{T})}$, and since the norm of the weak limit is no less than the limit of the norms it follows that $(k_{n_m})_{|_J}$ converges strongly to $(g_0)_{|_J}$ in the strictly convex space $L^2(J)$. The same reasoning applies in $L^p(J)$ for $1 < p < \infty$. Finally we remark that the preceding arguments hold true when k_n is replaced by any subsequence of itself; hence k_n contains no subsequence not converging to g_0 in the sense stated before, which achieves the proof. ■

We come now to an analog of theorem 2 in the infinite dimensional case. We define $H^{2,\infty}_J$ and $H^{2,1}_I$ to be the following vector spaces:

$$H^{2,\infty}_J = \{h \in H^2,\ ||h||_{L^\infty(J)} < \infty\},$$

$$H^{2,1}_I = \{h \in H^1,\ ||h||_{L^2(I)} < \infty\},$$

endowed with the natural norms. We begin with an elementary lemma.

Lemma 1 *Let $v \in L^1(J)$ such that $\mathbf{P}_+(0 \vee v) \in H^{2,1}_I$. Then:*

$$\forall h \in H^{2,\infty}_J,\ \langle \mathbf{P}_+(0 \vee v), h \rangle_{\mathbb{T}} = \langle v, h \rangle_J.$$

Proof. Let u be the function defined on $\mathbb{T}$ by

$$u = (0 \wedge v) - \mathbf{P}_+(0 \vee v).$$

By assumption $u \in L^1(\mathbb{T})$, and by its very definition all Fourier coefficients of u of non-negative index vanish. Hence $u \in \bar{H}^1_0$, and since it is L^2 integrable on I where it coincides with $-\mathbf{P}_+(0 \vee v)$, .we conclude that $\overline{u} \in H^{2,1}_I$ and that $\overline{u}(0) = 0$ Now, for $h \in H^{2,\infty}_J$ we have that

$$\begin{aligned} \langle v\chi_J, h \rangle_{\mathbb{T}} = &\ \langle u, h \rangle_{\mathbb{T}} + \langle \mathbf{P}_+(0 \vee v), h \rangle_{\mathbb{T}} \\ = &\ \overline{u}(0)h(0) + \langle \mathbf{P}_+(0 \vee v), h \rangle_{\mathbb{T}} \\ = &\ \langle \mathbf{P}_+(0 \vee v), h \rangle_{\mathbb{T}} \end{aligned} \tag{30}$$

where the second equality follows from the Cauchy formula because $(\overline{u}h) \in H^1$. ■

Theorem 4 *Suppose that $f \in L^2(I)$ is not the trace on I of a H^2-function of modulus less or equal to 1 a.e on J. Then, $g \in H^2$ is the solution to ABEP iff the following two conditions hold.*

- $|g(e^{i\theta})| = 1$ *for a.e. $e^{i\theta} \in J$,*

- *there exists a nonnegative real function* $\lambda \in L^1_{\mathbb{R}}(J)$ *such that,*

$$\forall h \in H^{2,\infty}_J,\ \langle g - f, h\rangle_I + \langle \lambda g, h\rangle_J = 0. \tag{31}$$

Proof. Suppose g verifies the two conditions and differs from g_0. Set $h = (g_0 - g) \in H^{2,\infty}_J$ and observe that

$$\mathrm{Re}\langle \lambda g, h\rangle_J = \frac{1}{2\pi}\int_J \lambda(\mathrm{Re}(\overline{g} g_0) - 1) \le 0. \tag{32}$$

In another connection, since $-h$ is an admissible increment from g_0, the variational inequality characterizing the projection onto a closed convex set gives us (*cf.* Theorem 1) $\mathrm{Re}\langle g_0 - f, h\rangle_I \le 0$, whence

$$\mathrm{Re}\langle g - f, h\rangle_I = \mathrm{Re}\langle g_0 - f, h\rangle_I - \langle h, h\rangle_I < 0$$

which, combined with (32), contradicts (31).

Suppose now that g is the solution of *ABEP*. The property that $|g| = 1$ on J has been proven in Theorem 1. In order to let n tend to infinity, we rewrite (19) with self-explaining notations as

$$\langle k_n - f, e^{im\theta}\rangle_I + \sum_{j=1}^{r(n)} \lambda_j^n k_n(e^{i\theta_j^n})\overline{e^{im\theta_j^n}} = 0, \quad \forall m \in \{0 \dots n\},\ . \tag{33}$$

We define $(\Lambda_n),\ \ n \in \mathbb{N}$, to be a family of linear forms on $C(J)$ defined as

$$\Lambda_n(u) = \sum_{j=1}^{r(n)} \lambda_j^n k_n(e^{i\theta_j^n}) u(e^{\theta_j^n}), \quad \forall u \in C(J).$$

Equation (20) shows that (Λ_n) is a bounded sequence in the dual $C(J)^*$ which by the Banach-Alaoglu theorem admits a weak-* converging subsequence whose limit we call Λ. Moreover, the Riesz representation theorem ensures the existence of a complex measure μ to represent Λ so that, appealing to Theorem 3 and taking the limit in (33), we obtain

$$\langle g_0 - f, e^{im\theta}\rangle_I + \int_J \overline{e^{im\theta}} d\mu = 0, \quad \forall m \in \mathbb{N}.. \tag{34}$$

Now, the F. and M. Riesz theorem asserts that the measure which is μ on J and $(g_0 - f)d\theta$ on I is absolutely continuous with respect to Lebesgue measure, because its Fourier coefficients of nonnegative index do vanish, by (34). Therefore there is $v \in L^1(J)$ such that,

$$\langle g_0 - f, e^{im\theta}\rangle_I + \langle v, e^{im\theta}\rangle_J = 0, \quad \forall m \in \mathbb{N}\,,$$

which is equivalent to

$$\langle g_0 - f, e^{im\theta}\rangle_I + \langle \lambda g_0, e^{im\theta}\rangle_J = 0, \quad \forall m \in \mathbb{N}, \tag{35}$$

where we have set $\lambda(z) = v(z)\overline{g_0(z)}\ \forall z \in J$. Equation (35) means that

$$\mathbf{P}_+((g_0 - f)\chi_I) = -\mathbf{P}_+(0 \vee \lambda g_0),$$

which indicates that $\mathbf{P}_+(0 \vee \lambda g_0)$ lies in H^2. Thus, thanks to Lemma 1, we get that

$$\langle g_0 - f, u\rangle_I + \langle \lambda g_0, u\rangle_J = 0, \quad \forall u \in H^{2,\infty}_J. \tag{36}$$

In order to prove the realness as well as the nonnegativity of λ, we pick $h \in C^\infty_{c,\mathbb{R}}(I)$, the space of smooth real-valued functions with compact support on I, and we consider its Riesz-Herglotz transform

$$b(z) = \frac{1}{2\,\pi}\int_I \frac{e^{it} + z}{e^{it} - z}\, h(e^{it})\, dt = \frac{1}{2\,\pi}\int_{\mathbb{T}} \frac{e^{it} + z}{e^{it} - z}\, \chi_I(e^{it})\, h(e^{it})\, dt\,. \tag{37}$$

It is standard that b is continuous on $\overline{\mathbb{D}}$ [11, ch. III, thm. 1.3]. For $t \in \mathbb{R}$, define $\omega_t = \exp(tb)$ which is the outer function whose modulus is equal to $\exp th$ on I and 1 on J. The function $g_0\,\omega_\lambda$ is a candidate approximant in problem *ABEP*, hence $t \mapsto \|f - g_0\,\omega_t\|^2_{L^2(I)}$ reaches a minimum at $t = 0$. By

the boundedness of b, we may differentiate this function with respect to t under the integral sign, and equating the derivative to 0 at $t = 0$ yields

$$0 = \mathrm{Re}\langle (f - g_0)\overline{g_0}, b\rangle_I = \mathrm{Re}\langle (f - g_0), bg_0\rangle_I.$$

In view of (36), it implies that

$$0 = \mathrm{Re}\langle \lambda g_0, bg_0\rangle_J = \mathrm{Re}\langle \lambda, b\rangle_J,$$

where we used that $|g_0| \equiv 1$ on J. Remarking that b is pure imaginary on J, this means

$$\langle Im(\lambda), b\rangle_{L^2(J)} = 0, \quad \forall h \in C^\infty_{c,\mathbb{R}}(I).$$

Letting $h = h_m$ range over a sequence of smooth positive functions which are approximate identies, namely of unit $L^1(I)$-norm and supported on the arc $[\theta - 1/m, \theta + 1/m]$ with $e^{i\theta} \in I$, we get in the limit, as $m \to \infty$, that

$$\langle Im(\lambda), (e^{i\theta} + .)/(e^{i\theta} - .)\rangle_J = 0, \quad e^{i\theta} \in I.$$

Then, appealing to he Plemelj-Sokhotski formulas as in the proof of Theorem 1, this time on J, we obtain that $Im(\lambda) = 0$ which proves that λ is real-valued. Note that the argument based on the Plemelj-Sokhotski formulas and the Hahn-Banach theorem together imply that the space generated by $\xi \mapsto (e^{i\theta}+\xi)/(e^{i\theta}-\xi)$, as $e^{i\theta}$ ranges over an infinite compact subset lying interior to J, is dense in $L^p(I)$ for $1 < p < \infty$. In fact using the F. and M. Riesz theorem and the Plemelj-Sokhoski formulas, it is easy to see that such functions are also uniformly dense in $C(\overline{I})$. Then, using that $ABEP$ is a convex problem, we obtain upon differentiating once more that

$$\mathrm{Re}\langle (g_0 - f)\bar{g}_0, b^2\rangle_I \ \geq 0,$$

which leads us by (36) to

$$\mathrm{Re}\langle \lambda, ((e^{i\theta} + .)/(e^{i\theta} - .))^2\rangle_J = \mathrm{Re}\langle \lambda g_0, g_0((e^{i\theta} + .)/(e^{i\theta} - .))^2\rangle_J \ \leq 0, \qquad e^{i\theta} \in I.$$

By the density property just mentioned this implies that $((e^{i\theta} + .)/(e^{i\theta} - .))^2_{|\bar{I}}$ is dense in the set of nonpositive continuous functions on $\overline{I}$, therefore $\lambda \geq 0$. Note also that (35) implies $(f - g_0) \vee \lambda g_0 \in \bar{H}^1$, hence it cannot vanish on a subset of $\mathbb{T}$ of positive measure unless it is the zero function. But this would imply $f = g$ a.e on I which contradicts the hypothesis. This yields $\lambda > 0$ a.e on J. ■

6 A numerical example

For practical applications the continuous constraint of $PBEP$ on the arc J is discretized in $m+1$ points. Suppose that $J = \{e^{it}, t \in [-\theta, \theta]\}$, for some $\theta \in [0, \pi]$. Call J_m the discrete version of the arc J defined by

$$J_m = \{e^{it}, t \in \{-\theta + \frac{2k\theta}{m}, k \in \{0 \dots m\}\}$$

we define following auxiliary extremal problem:

DBEP(n,m)

For $f \in L^2(I)$, find $k_{n,m} \in P_n$ such that $\forall t \in J_m \ |k_{n,m}(t)| \leq 1$ and

$$\|f - k_{n,m}\|_{L^2(I)} = \min_{\substack{g \in P_n \\ |g| \leq 1 \text{ a.e. on } J_m}} \|f - g\|_{L^2(I)}\,. \tag{38}$$

For the discretized problem **DBEP(n,m)**, the following holds.

Theorem 5 *For $\lambda = (\lambda_0, \dots, \lambda_m) \in \mathbb{R}^{m+1}$ and $g \in P_n$ define the Lagrangian*

$$L(\lambda, g) = \|f - g\|_{L^2(I)} + \sum_{k=0}^{m} \lambda_k (|g(e^{i(-\theta + \frac{2k\theta}{m})}|^2 - 1)$$

, then

- *Problem* **DBEP(n,m)** *has a unique solution $k_{n,m}$,*

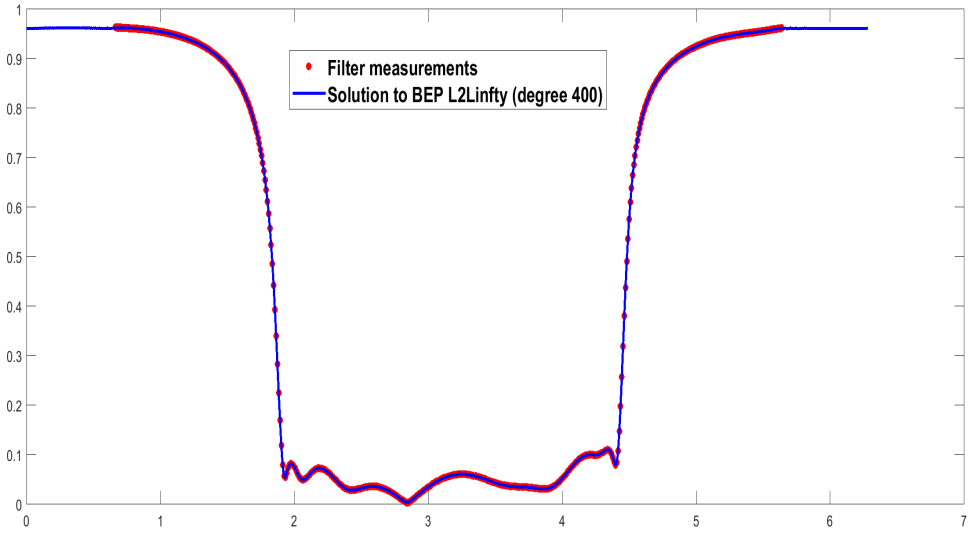


Figure 1: Solution of DBEP at hand of partial scattering measurements of a microwave filter

- *$k_{n,m}$ is also the unique solution of the concave maximisation problem:*

$$\text{to find } g_{opt} \text{ and } \lambda_{opt} \text{ solving for} \quad \max_{\lambda \geq 0} \min_{g \in P_n} L(\lambda, g), \tag{39}$$

  where $\lambda \geq 0$ means that each component of λ is non negative.

- *For a fixed n, $\lim_{m\to\infty} k_{n,m} = k_n$ in P_n.*

The proof of Theorem 5 follows from standard convex optimization theory, using in addition that the *sup*-norm of the derivative of a polynomial of degree n on $\mathbb{T}$ is controlled by the values it assumes at a set of $n+1$ points. This depends on Bernstein's inequality and on the argument using Lagrange interpolation polynomials used in the Remark after Theorem 1.

In the minmax problem (39) , the minimization is a quadratic convex problem. It can be tackled efficiently by solving the critical point equation which is a linear system of equations similar to (19). Eventually, an explicit expression of the gradient and of the hessian of the concave maximization problem (39) allows us for a fast converging computational procedure to estimate $k_{n,m}$.

Figure (1) represents a solution to problem **DBEP(n,m)**, where f is obtained from partial measurement of the scattering reflexion parameter of a wave-guide microwave filter by the CNES (French Space Agency). The problem is solved for $n = 400$ and $m = 800$, while the constraint on J has been renormalized to 0.96 (instead of 1). The modulus of $k_{400,800}$ is plotted as a blue continuous line while the measurements $|f|$ appear as red dots. As the reader can see, the fit is extremely good.

References

[1] L. Baratchart, J. Grimm, J. Leblond, M. Olivi, F. Seyfert, and F. Wielonsky. Identification d'un filtre hyperfréquences par approximation dans le domaine complexe, 1998. INRIA technical report no. 0219.

[2] L. Baratchart, J. Grimm, J. Leblond, and J.R. Partington. Asymptotic estimates for interpolation and constrained approximation in H^2 by diagonalization of toeplitz operators. *Integral equations and operator theory*, 45:269–299, 2003.

[3] L. Baratchart and J. Leblond. Hardy approximation to L^p functions on subsets of the circle with $1 \leq p < \infty$. *Constructive Approximation*, 14:41–56, 1998.

[4] L. Baratchart, J. Leblond, and J.R. Partington. Hardy approximation to L^∞ functions on subsets of the circle. *Constructive Approximation*, 12:423–436, 1996.

[5] L. Baratchart, J. Leblond, and J.R. Partington. Problems of Adamjan–Arov–Krein type on subsets of the circle and minimal norm extensions. *Constructive Approximation*, 16:333–357, 2000.

[6] Laurent Baratchart, Sylvain Chevillard, and Fabien Seyfert. On transfer functions realizable with active electronic components. Technical Report RR-8659, INRIA, Sophia Antipolis, 2014. 36 pages.

[7] J.M. Borwein and A.S. Lewis. *Convex Analysis and Nonlinear Optimization*. CMS Books in Math. Can. Math. Soc., 2006.

[8] E. W. Cheney. *Introduction to approximation theory*. Chelsea, 1982.

[9] J. C. Doyle, B. A. Francis, and A. R. Tannenbaum. *Feedback Control Theory*. Macmillan Publishing Company, 1992.

[10] P.L. Duren. *Theory of H^p spaces*. Academic Press, 1970.

[11] J.B. Garnett. *Bounded analytic functions*. Academic Press, 1981.

[12] M.G. Krein and P.Y. Nudel'man. Approximation of $L^2(\omega_1, \omega_2)$ functions by minimum– energy transfer functions of linear systems. *Problemy Peredachi Informatsii*, 11(2):37–60, 1975. English translation.

[13] J. Leblond and J. R. Partington. Constrained approximation and interpolation in hilbert function spaces. *J. Math. Anal. Appl.*, 234(2):500–513, 1999.

[14] Martine Olivi, Fabien Seyfert, and Jean-Paul Marmorat. Identification of microwave filters by analytic and rational h^2 approximation. *Automatica*, 49(2):317–325, 2013.

[15] Jonathan Partington. *Linear operators and linear systems*. Number 60 in Student texts. London Math. Soc., 2004.

[16] Rik Pintelon, Yves Rollain, and Johan Schoukens. *System Identification: A Frequency Domain Approach*. Wiley, 2012.

[17] W. Rudin. *Real and complex analysis*. McGraw–Hill, 1987.

[18] A. Schneck. Constrained optimization in hardy spaces. Preprint, 2009.

[19] F. Seyfert. Problèmes extrémaux dans les espaces de Hardy. These de Doctorat, Ecole des Mines de Paris, 1998.

5.1.2 De-embedding of microwave filters

Following paper is reproduced in this section:

- Martine Olivi, Fabien Seyfert, and Jean-Paul Marmorat. "Identification of microwave filters by analytic and rational H^2 approximation". In: *Automatica* (2012). DOI: 10.1016/j.automatica.2012.10.005. URL: http://hal.inria.fr/hal-00753824

Identification of microwave filters by analytic and rational H^2 approximation

Martine Olivi [a], Fabien Seyfert [a], Jean-Paul Marmorat [b]

[a] *INRIA - Sophia-Antipolis, BP 93, 06902 Sophia-Antipolis Cedex, FRANCE.*

[b] *CMA - Ecole des Mines de Paris, Rue Claude Daunesse, BP 207,06904 Sophia Antipolis Cedex.*

Abstract

In this paper, an original approach to frequency identification is explained and demonstrated through an application in the domain of microwave filters. This approach splits into two stages: a stable and causal model of high degree is first computed from the data (completion stage); then, model reduction is performed to get a rational low order model. In the first stage the most is made of the data taking into account the expected behavior of the filter. A reduced order model is then computed by rational H^2 approximation. A new and efficient method has been developed, improved over the years and implemented to solve this problem. It heavily relies on the underlying Hilbert space structure and on a nice parametrization of the optimization set. This approach guarantees the stability of the MIMO approximant of prescribed McMillan degree.

Key words: Low-pass filters; system identification; incomplete data; model reduction; analytic approximations; rational approximation; lossless rational matrices; parametrization.

1 Introduction

The microwave filters that we consider are used in telecommunication satellites for channel multiplexing.

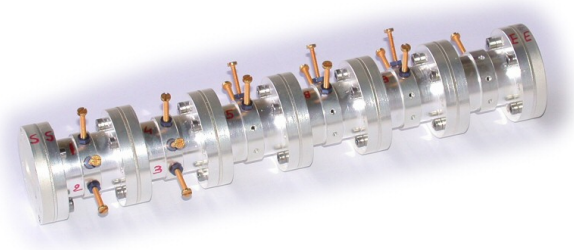


Fig. 1. A microwave filter.

These electromagnetic waveguide filters are made of resonant cavities (see Figure 1) interconnected by coupling irises (orthogonal double slits). Each cavity has 3 screws which allow one to tune the filter. Using a low-pass transformation these high-pass filters are usually modeled by a low-pass electrical circuit (see Figure 2). In this model,

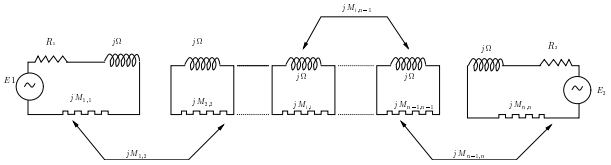


Fig. 2. Low-pass prototype.

Ω is the normalized frequency, each resonant cavity mode is represented by a fictive resonant circuit (frequency M_{kk}) and the coupling between modes (produced by the irises) by impedance inverters (jM_{kl}). In the remainder of the paper we will adopt the mathematical notation, i rather than j, for the square root of -1. Electrical power transfer is then described by a scattering matrix. From a mathematical viewpoint, the scattering matrix R is a rational matrix function with *complex coefficients* which is stable (poles with negative real parts) lossless ($R(i\omega)$ is unitary) and symmetric. The geometry of the filter is characterized by the electrical parameters which appear on a realization in particular form of the scattering matrix. Namely, $R(s) = I + C(s\,I - A)^{-1}B$ with

$$C = \begin{bmatrix} i\sqrt{2r_1} & 0 & \cdots & 0 \\ 0 & \cdots & 0 & i\sqrt{2r_2} \end{bmatrix}, \; B = C^t, \qquad (1)$$
$$A = -R - iM, \quad A = A^t \qquad r = -\tfrac{1}{2}C^tC,$$

Email addresses: martine.olivi@inria.fr (Martine Olivi), fabien.seyfert@inria.fr (Fabien Seyfert), Jean-Paul.Marmorat@mines-paristech.fr (Jean-Paul Marmorat).

where r_1 and r_2 are the input and output loads and the matrix M is the *coupling matrix*. The structure of M (non-zeros entries) specifies the way resonators are coupled to one another. The McMillan degree of R corresponds to the number of circuits, that is the number of resonant modes or else two times the number of cavities.

The problem of extracting coupling parameters from frequency scattering measurements is essential with a view to reducing the cost of hardware and CAD tuning. The direct approach consists in feeding to a generic optimizer the function evaluating the scattering matrix from the coupling parameters, in order to fit the data. However, it often depends on a favorable initial guess and substantial efforts are currently being spent to design more robust methods. Another approach consists in first identifying a rational (linear) model from the data. Then, the coupling parameters are extracted from this rational model using classical design methods. In the filter community, the so-called Cauchy method is widely used to compute the rational model [24], [1]. Let us point out three major problems encountered in this direction, and in many other methods proposed in the literature:

- there is no guarantee on the stability of the rational model, i.e. the derived model can have unstable poles;
- there is no control on the McMillan degree of the model;
- no constraint is imposed to the model outside the frequency band of measurement (broadband), which may result in unrealistic behavior there.

Many toolboxes propose input output identification, while a few deal with frequency data. The software Vector Fitting and its Matrix Fitting extension (see [17], [18] and the bibliography therein) has become popular in the electromagnetic simulation community. However, the convergence towards a stable rational approximant, optimal in some least-square sense, is not guaranteed by this algorithm. Moreover no control is given in the MIMO case on the overall McMillan degree of the result, but only on its number of distinct poles. The same problem arises with the Frequency Domain Identification Toolbox [22] which only deals with SISO systems. This is unacceptable for the application we have in mind, in which the target McMillan degree is prescribed in advance and given by the number of coupled resonators present in the equivalent circuit of the filter.

To overcome these difficulties, we have developed a two stage approach to identify a rational model from the scattering data. A stable and causal model of high degree is first computed from the data (completion stage); then, model reduction is performed to get a model of the prescribed order. The first stage will be addressed in Section 3. Then we will consider the model reduction step. We tackle this problem using rational H^2 approximation and the original approach developed over the years in [6], [14], [28]. We present here the state of the art of this approach which includes an efficient parametrization of balanced output pairs. The exposition is definitely application oriented, so that the emphasis will be put on the effective implementation of the method.

2 The Hilbert space framework

To deal with these completion and model reduction problems, we thus favor an approach based on approximation. A relevant context to deal with approximation is that of a Hilbert space. On the other hand, stability and causality of a rational model are equivalent to the analyticity of the transfer function in the closed right half-plane (poles at finite distance in the open left half-plane). We denote by $\mathbb{C}^+$ and $\mathbb{C}^-$ the open right and left half-planes. To properly handle stability and causality, we embed rational functions in a larger space of analytic functions in $\mathbb{C}^+$, namely a Hardy space naturally endowed with an L^2 norm. Note that, due to the low-pass transformation, the frequency data and the model that we consider do not satisfy the conjugacy requirement. This is why we consider Hardy spaces of *complex* functions.

The usual Hardy space of the half-plane, $H^2(\mathbb{C}^+)$, consists of functions f analytic in $\mathbb{C}^+$, whose L^2-norm remains uniformly bounded on vertical lines,

$$\sup_{x>0} \int_{-\infty}^{\infty} |f(x+i\omega)|^2 d\omega < \infty.$$

The Hardy space of the left half-plane, $H^2(\mathbb{C}^-)$, is defined in a similar way. An important fact is that the Laplace transform gives an isometry from $L^2(\mathbb{R}^\pm)$ onto $H^2(\mathbb{C}^\pm)$. It allows one to consider these Hardy spaces as subspaces of $L^2(i\mathbb{R})$, the image of $L^2(\mathbb{R})$ by the Laplace transform [32]. Moreover,

$$L^2(i\mathbb{R}) = H^2(\mathbb{C}^+) \oplus H^2(\mathbb{C}^-).$$

Each function in $L^2(i\mathbb{R})$ can thus be decomposed as the sum of a function in $H^2(\mathbb{C}^+)$ (stable part) and a function in $H^2(\mathbb{C}^-)$ (anti-stable part).

However, a stable causal function which fails to be strictly proper (to be 0 at ∞) does not belong to $L^2(i\mathbb{R})$. In order to include these functions in our setting, we shall replace the usual Lebesgue measure by the weighted measure $d\mu(w) = \frac{d\omega}{1+\omega^2}$, which also has the advantage to penalize high frequencies. The associated Hardy spaces $H^2_\mu(\mathbb{C}^+)$ and $H^2_\mu(\mathbb{C}^-)$ are defined in a similar way and can be viewed as subspaces of the space $L^2(d\mu)$ of functions defined on the imaginary axis and such that

$$\|f\|^2_\mu = \int_{-\infty}^{\infty} |f(i\omega)|^2 \frac{d\omega}{1+\omega^2} < \infty.$$

However, $H^2_\mu(\mathbb{C}^+)$ and $H^2_\mu(\mathbb{C}^-)$ fail to be orthogonal complements, since their intersection is not empty (it contains for example constant functions). For $f \in L^2(d\mu)$, we denote by $P_+(f)$ its orthogonal projection onto $H^2_\mu(\mathbb{C}^+)$ (stable part) and by $P_-(f)$ its orthogonal projection on the orthogonal complement of $H^2_\mu(\mathbb{C}^+)$ (unstable part). Hardy spaces thus provide an interesting tool to estimate causality and stability of a given transfer function.

3 Compensation of delay components and completion of the data

After the low-pass frequency transformation, we suppose that the harmonic scattering measurements of the filter yield the knowledge of a 2×2 matrix function $\tilde{S}(iw)$ defined on a strict sub-interval J of the imaginary axis. In practice this function is obtained thanks to the interpolation (splines) of a discrete set of measurement points. The mathematical model we want to identify from these measurements is given by

$$\begin{bmatrix} e^{i\frac{\alpha}{2}h(w)} & 0 \\ 0 & e^{i\frac{\beta}{2}h(w)} \end{bmatrix} R(i\,w) \begin{bmatrix} e^{i\frac{\alpha}{2}h(w)} & 0 \\ 0 & e^{i\frac{\beta}{2}h(w)} \end{bmatrix},$$

where R is the 2×2 *rational* scattering matrix of the low-pass model of the filter and the exponential terms are due to the access lines used to perform the measurements. The transformation $h(w)$ maps normalized frequencies (low-pass model) to high frequencies (original system).

In order to cast the identification problem to a rational approximation problem we first need to identify the non-rational delay components, that is to evaluate α and β. We base our delay compensation procedure on analytic completion techniques [8], [7], [9]. The latter consist in extending partial frequency measurements performed on the broadband J to the whole imaginary axis under causality constraints. In the special case of our filter with rational response we will make the strong assumption that, if the delay components are properly compensated, the measurements can be extended on J_c (the complementary of J on the imaginary axis) by a polynomial of low order in the variable $1/iw$ such as to form a causal transfer function. In other words, if $S_{ij}(iw)$ denotes the measurements for which the delays have been compensated, we should be able to find a polynomial $p_{ij}(1/iw)$ of low degree such that the complemented elements $S_{ij}(iw) \vee p_{ij}(1/iw)$ have a "small" anti-causal component and have a smooth behavior at the boundaries of J. The simplicity of the extensions p_{ij} accounts for the absence of delay and the fact that the measurements on the broadband J already capture most of the complexity of the rational responses, that can therefore be represented on J_c by a short Taylor expansion p_{ij} at infinity. We expose in what follows the convex optimization problems that are considered to extract the delay components and extend the data on the whole imaginary axis.

To a given value τ of a delay compensation, we associate the polynomial which gives the "most causal" completion

$$p_\tau = \arg\min_{p\in\mathcal{P}} \|P_-(\tilde{S}_{11}(iw)e^{-i\tau h(w)} \vee p(1/iw))\|^2_\mu, \quad (2)$$

where $\mathcal{P} = \{p; \deg p \leq n_c, \sup_{w\in J_c} |p(1/iw)| \leq 1\}$. This modular bound on p is meaningful as our filter is passive. We then choose the delay α to be the value of the compensation τ that gives the smallest discontinuities at the concatenation points between the data and p_τ. To determine β we proceed in the same manner using the measurements of $\tilde{S}_{22}$ instead of those of $\tilde{S}_{11}$.

Now the delays are known, we improve the completion by relaxing (2) and imposing a better behavior near the concatenation points. We select a sub-collection of measurement indices $I = \{k, |w_k| > w_c\}$ where w_c is chosen sufficiently large (tail of the broadband J). We thus consider the optimization problem :

$$\min_{p\in\mathcal{P}} \sum_{k\in I} |p(1/iw_k) - S_{ij}(iw_k)|^2$$

under the additional constraint

$$\|P_-(S_{ij}(iw) \vee p(1/iw))\|^2_\mu \leq E.$$

This problem has a unique optimal solution unless its admissible set is empty. This will be the case provided that $E \geq E_{min}$ where E_{min} is the optimal criterion obtained from the preceding problem. In practice the values $w_c = 2.5$ (normalized frequencies) and $n_c = 4$ seem to give very good results when the broadband is three time bigger than the passband.

If p_{ij} are the polynomial completions computed by the later method we define

$$\tilde{F}_{ij} = P_+(S_{ij}(iw) \vee p_{ij}(1/iw)).$$

Those functions can be seen as the compensated, causal, stable projections of our initial data; note that, by construction, their $L^2(d\mu)$ distance to compensated data S on J is less than $\sqrt{E}$.

4 From continuous-time to discrete-time

To deal with rational approximation, we shift to the disk or discrete-time framework. A good reason to do this is that the Hardy spaces of the disk are simpler in some sense than that of the half-plane. The fact that the unit disk has a finite Lebesgue measure has some nice implications, as the inclusions $L^\infty(\mathbb{T}) \subset L^2(\mathbb{T}) \subset L^1(\mathbb{T})$.

Moreover, functions in $L^2(\mathbb{T})$ can be represented by their Fourier series and we don't have to cope with sampling.

The space $L^2(\mathbb{T})$ splits into two orthogonal subspaces

$$L^2(\mathbb{T}) = H^2 \oplus H^2_\perp,$$

where H^2 consists of functions whose Fourier coefficients of negative index are zero, while $H^2_\perp$ consists of functions whose Fourier coefficients of non-negative index are zero. Hardy spaces $H^2(\mathbb{D})$ and $H^2(\mathbb{E})$ of the disk and its exterior $\mathbb{E} = \mathbb{C} \setminus \mathbb{D}$ may be defined as those of the half-planes (integrals on vertical lines are then replaced by integrals over circles). Then, by analytic continuation, H^2 can be identified with $H^2(\mathbb{D})$ and $H^2_\perp$ with a strict subspace of $H^2(\mathbb{E})$ (functions vanishing at infinity). We denote by P_{H^2} and $P_{H^2_\perp}$ the orthogonal projections onto H^2 and $H^2_\perp$ respectively. Discrete-time stable and causal rational transfer functions naturally belong to $H^2(\mathbb{E})$.

There are many ways to transform a continuous-time function into a discrete-time one. We shall use either the usual bilinear transform or a variant of it, with the Möbius transformation from the z-plane to the s-plane

$$z \mapsto s = \frac{z+1}{z-1},$$

which sends $\mathbb{T}$ onto the imaginary axis and $\mathbb{C}^+$ onto $\mathbb{E}$.

4.1 Bilinear transform

The bilinear transform is the map

$$\tilde{F}(s) \mapsto F(z) = \tilde{F}\left(\frac{z+1}{z-1}\right).$$

This map is an isometry from $L^2(d\mu)$ onto $L^2(\mathbb{T})$ which preserves the McMillan degree. It sends the space $H^2_\mu(\mathbb{C}^+)$ onto $H^2(\mathbb{E})$.

4.2 $H^2(\mathbb{C}^+) \to H^2_\perp$ isometry

The map

$$\tilde{F}(s) \mapsto F(z) = \frac{\sqrt{2}}{z-1}\tilde{F}\left(\frac{z+1}{z-1}\right). \qquad (3)$$

is an isometry from $L^2(i\mathbb{R})$ onto $L^2(\mathbb{T})$. It also preserves the McMillan degree. It sends the space $H^2(\mathbb{C}^+)$ onto $H^2_\perp$. With this transformation, the formulas which allow us to derive a realization $(A, B, C, 0)$ of $F(z)$ from a realization $(\tilde{A}, \tilde{B}, \tilde{C}, 0)$ of $\tilde{F}(s)$ and reciprocally, are completely symmetric:

$$\begin{array}{ll} C = \tilde{C} & \tilde{C} = C \\ A = -(I-\tilde{A})^{-1}(I+\tilde{A}) & \tilde{A} = -(I-A)^{-1}(I+A) \\ B = \sqrt{2}(I-\tilde{A})^{-1}\tilde{B} & \tilde{B} = \sqrt{2}(I-A)^{-1}B \end{array}$$

Both methods are used to transport the function $\tilde{F}$ obtained in Section 3 from continuous-time to discrete-time. If the $H^2(\mathbb{C}^+) \to H^2_\perp$ isometry is used, then the value at infinity must be taken off. It will be reset after the approximation step. This method preserves the value at infinity obtained in the completion stage. If the bilinear transform is used, the value at infinity can thus be improved by the rational approximation step. This method usually gives better results.

5 Stable rational approximation of given McMillan degree

In the completion stage, we dealt with each entry separately. In rational approximation, the constraint on the McMillan degree involves the whole scattering matrix, and it is not possible to handle each entry independently. A specific approach has been developed at INRIA to cope with this problem, which is based on the following points

- the optimization range is reduced to a compact set,
- an atlas of charts is used to parametrize the optimization domain.

To simplify our writing, we keep denoting by $L^2(\mathbb{T})$, H^2 and $H^2_\perp$ the spaces of matrix-valued functions with entries in $L^2(\mathbb{T})$, H^2 and $H^2_\perp$ respectively. The proper dimension of the matrix will be understood from the context. The L^2-norm of a matrix-valued function derives from the scalar product

$$< F, G >= \frac{1}{2\pi}\mathrm{Tr}\int_0^{2\pi} F(e^{it})G(e^{it})^* dt.$$

The rational approximation problem we consider is, given a $p \times m$ matrix-valued function $F \in H^2(\mathbb{E})$, to minimize the L^2 distance to the set of rational stable functions of McMillan degree less than or equal to n. In our application $m = p = 2$.

Using the orthogonal decomposition

$$H^2(\mathbb{E}) = \mathbb{C} \oplus H^2_\perp,$$

we can see that any solution H must satisfy $H(\infty) = F(\infty)$. Thus, we may restrict our study to the case of

strictly proper transfer functions, that is to the space $H^2_\perp$.

A number of qualitative results are available in the literature which assert that the problem is well-posed and which pave the way to convergent algorithms. It was proved in [5] that *the global minimum of the L^2 criterion does exist*, as well as the *normality property*: if F is not of McMillan degree strictly less than n, then any local minimum of the criterion over the set of systems of order less than or equal to n has effective order n. The problem can thus be stated as:

Rational approximation problem. Given $F \in H^2_\perp$ of order $\geq n$, find $\hat{H}$ such that

$$\hat{H} = \text{argmin}_{H \in \mathcal{S}_n} \|F - H\|_2^2 \tag{4}$$

where

$$\mathcal{S}_n = \{H \in H^2_\perp, \deg H = n\}$$

is the set of rational strictly proper stable transfer functions of exact degree n.

The following *consistency* result must also be mentioned: if F has McMillan degree n, then the only critical point of the L^2 criterion is F itself [10].

The present approach was first proposed in the SISO case [6] and then in the MIMO case [14]. The first step is the reduction of the parameters space (see Section 6). The second step is to find a clever parametrization of the reduced optimization space namely the space of stable all-pass systems of fixed order. This parametrization, an atlas of charts, takes into account the precise structure of the space, namely a *non-trivial* differentiable manifold ([2], [13]). Several atlases have been proposed in the literature (see [2], [20]). The new atlas which is used in this paper particularly suits to state-space representations and has been preferred for computation facilities.

The question of the parametrization of LTI systems has been widely studied. The non-zero entries in classical canonical forms (companion, Hessenberg, tridiagonal forms) are often used as parameters because of their simplicity. See [37, chap.7] and [31] for an overview of these parametrization issues and their properties. Compared to these representations, our parametrization guarantees

- the non-redundancy in the parameters (injective mapping)
- restrictions on the system (stability, fixed order) automatically taken into account
- numerical robustness of the algorithm due to the use of unitary matrices

6 Reduction of the optimization set

In the SISO case, it is known that if $h = \frac{p}{q}$ is a best approximant of f then the numerator p can be easily computed from the denominator q by solving linear equations. The projection theorem in a Hilbert space asserts that h must be the projection of f onto the vector space $V_q = \{\frac{p}{q}; \deg p < n\}$. This space is the orthogonal complement of $B\,H^2_\perp$ in $H^2_\perp$, where $B = \prod_{i=1}^n \frac{1-\bar{a}_i z}{z-a_i}$ is the Blaschke product whose denominator is q,

$$H^2_\perp = B\,H^2_\perp \oplus V_q.$$

This approach was already developed in [33] and can be generalized to the MIMO case. The right generalization of the fraction description to the MIMO case is the Douglas-Shapiro-Shields factorization: a $p \times m$ rational matrix function $H \in H^2_\perp$ can be written in the form

$$H = G\,P, \tag{5}$$

where G is $p \times p$ lossless of McMillan degree n and $P \in H^2$. Recall that a rational lossless matrix is a matrix Blaschke product: $G(z)$ is contractive for $z \in \mathbb{D}$ and unitary for $z \in \mathbb{T}$. Multiplication by G lossless in $H^2_\perp$ is an isometry. In (5), G is called the lossless factor and brings the pole structure of H and thus its dynamics. It is unique up to a right unitary constant matrix $U \in \mathbb{U}_p$.

Now if $H = G\,P$ is a local approximant of F, thus H is completely determined from G as the orthogonal projection of F onto V_G

$$V_G = \{H \in H^2_\perp; \quad H = G\,P, \quad P \in H^2\}.$$

Equivalently, using multiplication by $G^\sharp(z) = G(z)^{-1}$ which is an isometry in $L^2(\mathbb{T})$, we get

$$< F - H, GH^2 > = < G^\sharp F - C, H^2 > = 0,$$

so that P is the orthogonal projection of $G^\sharp F$ onto H^2,

$$P = P_{H^2}(G^\sharp F).$$

The rational approximation problem is then to minimize the criterion

$$\psi_n : \; G \mapsto \|F - G P_{H^2}(G^\sharp F)\|^2 \tag{6}$$

over $\mathbb{L}^p_n/\mathbb{U}_p$ the right quotient of the set $\mathbb{L}^p_n$ of $p \times p$ lossless functions of McMillan degree n by unitary constant matrices. In least-square optimization and using a state-space formulation, this elimination step is classical and known under the name of Separable Least Square. It presents some important advantages: the dimension of the parameter space is reduced and mostly, lossless functions enter the picture.

Let us derive the state-space formulation which is implemented in our software RARL2 [26]. Let

$$F(z) = \mathcal{C}(zI_N - \mathcal{A})^{-1}\mathcal{B} + \mathcal{D}. \tag{7}$$

be a realization of F and let $\hat{H}(z) = \mathcal{D}+\hat{C}(zI-\hat{A})^{-1}\hat{B}$ be a local approximant of F. The error $F-\hat{H}$ has realization

$$\tilde{A} = \begin{bmatrix} \mathcal{A} & 0 \\ 0 & \hat{A} \end{bmatrix}, \tilde{B} = \begin{bmatrix} \mathcal{B} \\ \hat{B} \end{bmatrix}, \quad \tilde{C} = \begin{bmatrix} \mathcal{C} & -\hat{C} \end{bmatrix}, \tilde{D} = 0,$$

and the square of the L^2 error can be computed as

$$\|F - H\|_2^2 = \mathrm{Tr}\left(\tilde{B}^*\tilde{Q}\tilde{B}\right) = \mathrm{Tr}\left(\tilde{C}\tilde{P}\tilde{C}^*\right),$$

where $\tilde{P}$ and $\tilde{Q}$ are the reachability and observability gramians of the error. Necessary conditions for optimality may easily be found by computing the gradient of the square of the L^2-norm with respect to the state-space parameters. Partitioning the gramians in the same way as $\tilde{A}$,

$$\tilde{P} = \begin{bmatrix} \mathcal{P} & P_{12} \\ P_{12}^* & \hat{P} \end{bmatrix}; \quad \tilde{Q} = \begin{bmatrix} \mathcal{Q} & Q_{12} \\ Q_{12}^* & \hat{Q} \end{bmatrix},$$

we get

$$\begin{aligned} Q_{12}^*\mathcal{B} &= -\hat{Q}\hat{B} \\ \mathcal{C}P_{12} &= \hat{C}\hat{P} \\ Q_{12}^*\mathcal{A}P_{12} &= -\hat{Q}\hat{A}\hat{P}. \end{aligned} \tag{8}$$

These necessary conditions were first obtained by Wilson [38] and many model reduction algorithms were proposed in the literature based on these conditions (see [29], [3], [21], [36], [19], [16] and [11]). However, to our knowledge no other algorithm is available which guarantees the stability of the approximant and works in the MIMO case.

The reduction of the optimization set translates in this state-space setting as follows. First choose a realization of $\hat{H}$ such that the observable pair $(\hat{C}, \hat{A})$ is output normal

$$\hat{A}^*\hat{A} + \hat{C}^*\hat{C} = I,$$

which means that $\hat{Q}$ is the identity. Then, the first necessary condition yields $\hat{B} = -Q_{12}^*\mathcal{B}$. The rational approximation problem is thus to minimize the criterion

$$J_n(C, A) = \|F\|_2^2 - \mathrm{Tr}\left(\mathcal{B}^*Q_{12}Q_{12}^*\mathcal{B}\right) \tag{9}$$

over the set of output normal observable pairs (C, A). Note that Q_{12} is determined from A and C as the solution of the Stein equation

$$\mathcal{A}^*Q_{12}A + \mathcal{C}^*C = Q_{12}. \tag{10}$$

The derivative of the criterion with respect to some parameter λ can be computed as

$$\frac{dJ_n}{d\lambda} = 2\ \mathrm{Re\ Tr}\left(\frac{dQ_{12}}{d\lambda}B\mathcal{B}^*\right).$$

Using

$$AP_{12}^*\mathcal{A}^* + B\mathcal{B}^* = P_{12}^*,$$

we get

$$\frac{dJ_n}{d\lambda} = 2\ \mathrm{Re\ Tr}\left(P_{12}^*\left[\frac{dQ_{12}}{d\lambda} - \mathcal{A}^*\frac{dQ_{12}}{d\lambda}A\right]\right).$$

Now differentiating (10) with respect to A and C, we get the relations

$$\begin{aligned} &\frac{\partial Q_{12}}{\partial A} - \mathcal{A}^*\ \frac{\partial Q_{12}}{\partial A}\ A = \mathcal{A}^*\ Q_{12}, \\ &\frac{\partial Q_{12}}{\partial C} - \mathcal{A}^*\ \frac{\partial Q_{12}}{\partial C}\ A = \mathcal{C}^* \end{aligned}$$

so that we finally have

$$\frac{dJ_n}{d\lambda} = 2\ \ \mathrm{Re\ Tr}\left(P_{12}^*\left[\mathcal{A}^*\ Q_{12}\ \frac{\partial A}{\partial \lambda} + \mathcal{C}^*\ \frac{\partial C}{\partial \lambda}\right]\right). \tag{11}$$

The connection between observable pairs and lossless functions is stressed by the following result (see [25]).

Proposition 1 (Lossless embedding) *Given an observable pair (C, A) with A asymptotically stable, let Q be its observability gramian. Then, the rational matrix $G(z) = D + C(zI - A)^{-1}B$, with*

$$\begin{aligned} B &= -(A - \nu I)Q^{-1}(I - \nu A^*)^{-1}C^* \\ D &= I - CQ^{-1}(I - \nu A^*)^{-1}C^* \end{aligned}$$

is lossless for every ν such that $|\nu| = 1$.
The lossless function G satisfies $G(\nu) = I$. The map $(C, A) \mapsto G$ is a one-to-one correspondence between the set of observable pairs (C, A), A asymptotically stable, up to similarity, and the set of lossless functions up to a right unitary matrix.

This correspondence is in fact a diffeomorphism (see Cor.2.1 in [2]). If in addition the pair (C, A) is output normal, the matrix $[A\ C]^T$ has orthonormal columns and the lossless embedding consists in completing it into a unitary matrix

$$\begin{bmatrix} A \\ C \end{bmatrix} \mapsto \begin{bmatrix} A & B \\ C & D \end{bmatrix}.$$

We shall call *unitary realization*, a realization (A, B, C, D) such that A is asymptotically stable and the corresponding realization matrix unitary. Unitary realizations are precisely the balanced realizations of lossless functions

(see [20, prop.3.2]). In what follows, we shall use balanced realizations and the associated unitary matrix to represent lossless matrices. This is of course very advantageous from a numerical viewpoint.

7 Minimization over a manifold.

We said that an atlas of charts happens to be the right representation in order to use differential tools (as a gradient algorithm). A differential manifold is precisely a set endowed with a differential structure by means of an atlas of charts. An atlas is a collection of charts or (coordinate) maps $\phi_i : \mathcal{D}_i \to \mathbb{R}^d$, where $\mathcal{D}_i$ is an open subset of the manifold, which satisfy some compatibility conditions: the union of the $\mathcal{D}_i$ covers the manifold and the transition maps or changes of coordinates $\phi_i \circ \phi_j^{-1}$ are smooth. By means of the coordinate maps, differential calculus on $\mathbb{R}^d$ can be carried over to the manifold. The coordinate maps then become diffeomorphisms. The dimension of the manifold is d. An atlas of charts is thus the way to parametrize a non-trivial manifold in a local smooth manner. In the next section, we describe such an atlas of charts for the manifold $\mathbb{L}_n^p$ of dimension $2np$.

7.1 Lossless mutual encoding

Let $\Omega = (W, X, Y, Z)$ be a unitary realization. The idea is to attach to Ω a chart or coordinate map.

Let $G(z) = D + C(zI - A)^{-1}B$ be a balanced realization of G lossless. Note that such a realization is unique up to a state isometry. Let Λ be the unique solution to the Stein equation

$$\Lambda - A^*\Lambda W = C^*Y \tag{12}$$

and

$$V = D^*Y + B^*\Lambda W \tag{13}$$

Remark. In fact, the matrix V satisfies

$$\frac{1}{2i\pi}\int_{\mathbb{T}} G^\sharp(z)\, Y(zI - W)^{-1}\, dz = V, \tag{14}$$

which is known as a Nudelman interpolation condition for $G^\sharp$ (see [27]).

Formulas (12) and (13) can be rewritten in a matrix form

$$\begin{bmatrix} A & B \\ C & D \end{bmatrix}\begin{bmatrix} \Lambda \\ V \end{bmatrix} = \begin{bmatrix} SW \\ Y \end{bmatrix}. \tag{15}$$

Note that the realization of G being balanced by assumption, $P = \Lambda^*\Lambda$ satisfies

$$Y^*Y + W^*PW = V^*V + P. \tag{16}$$

We now assume that P is positive definite, a condition which allows for a parametrization of the set of solutions of the Nudelman problem (14) [4]. The matrix Λ is thus invertible and we may normalize the triple (Y, W, V) as

$$(\tilde{Y}, \tilde{W}, \tilde{V}) = (Y\Lambda^{-1}, \Lambda W\Lambda^{-1}, V\Lambda^{-1}). \tag{17}$$

We get from (15)

$$\begin{bmatrix} \tilde{W} \\ \tilde{Y} \end{bmatrix}^* \begin{bmatrix} A & B \\ C & D \end{bmatrix}\begin{bmatrix} I \\ \tilde{V} \end{bmatrix} = \tilde{W}^*\tilde{W} + \tilde{Y}^*\tilde{Y}.$$

Under the assumption $P > 0$,

$$K = \tilde{W}^*\tilde{W} + \tilde{Y}^*\tilde{Y} = \tilde{V}^*\tilde{V} + I \tag{18}$$

is positive definite and if $K^{1/2}$ denotes its Hermitian square root, then

$$\begin{bmatrix} \tilde{W}K^{-1/2} \\ \tilde{Y}K^{-1/2} \end{bmatrix}^* \begin{bmatrix} A & B \\ C & D \end{bmatrix}\begin{bmatrix} K^{-1/2} \\ \tilde{V}K^{-1/2} \end{bmatrix} = I.$$

We will now specify from V and Ω two unitary completions $\mathcal{U}$, $\mathcal{V}$ of these orthonormal columns,

$$\mathcal{U} = \begin{bmatrix} \tilde{W}K^{-1/2} & * \\ \tilde{Y}K^{-1/2} & * \end{bmatrix}, \quad \mathcal{V} = \begin{bmatrix} K^{-1/2} & * \\ \tilde{V}K^{-1/2} & * \end{bmatrix}$$

and define a map

$$\phi_\Omega : (A, B, C, D) \mapsto (V, D_0),$$

where D_0 is the unitary matrix such that

$$\mathcal{U}^*\begin{bmatrix} A & B \\ C & D \end{bmatrix}\mathcal{V} = \begin{bmatrix} I & 0 \\ 0 & D_0 \end{bmatrix}. \tag{19}$$

The map ϕ_Ω will be then invertible.

We must first fix the balanced realization (A, B, C, D) of G we start with. We will say that a realization of G is in canonical form with respect to Ω iff Λ given by (12) is positive definite and Hermitian. Then $\Lambda = P^{1/2}$ where P is the solution of (16). We denote by $\mathcal{D}_\Omega$ the set of unitary realizations in canonical form with respect to Ω.

The matrix $\mathcal{V}$ is chosen according to Proposition 1 (with $\nu = -1$)

$$\mathcal{V} = \begin{bmatrix} (I + \tilde{V}^*\tilde{V})^{-1/2} & -\tilde{V}^*(I + \tilde{V}\tilde{V}^*)^{-1/2} \\ \tilde{V}(I + \tilde{V}^*\tilde{V})^{-1/2} & (I + \tilde{V}\tilde{V}^*)^{-1/2} \end{bmatrix}. \tag{20}$$

The matrix $\mathcal{U}$ is computed from Ω as follows:

- perform the state isomorphism with matrix $\Lambda = P^{1/2}$, $(W, X, Y, Z) \longrightarrow (\tilde{W}, \tilde{X}, \tilde{Y}, Z)$:

$$(\tilde{W}, \tilde{X}, \tilde{Y}, Z) = (\Lambda W \Lambda^{-1}, \Lambda X, Y \Lambda^{-1}, Z).$$

- compute a Cholesky factorization of the matrix

$$\begin{bmatrix} K & L \\ L^* & N \end{bmatrix} = \begin{bmatrix} \tilde{W} & \tilde{X} \\ \tilde{Y} & Z \end{bmatrix}^* \begin{bmatrix} \tilde{W} & \tilde{X} \\ \tilde{Y} & Z \end{bmatrix}. \tag{21}$$

using the well-known formula [12, Sec. 0.2],

$$\begin{aligned} \begin{bmatrix} K & L \\ L^* & N \end{bmatrix} &= \begin{bmatrix} I & 0 \\ L^* K^{-1} & I \end{bmatrix} \begin{bmatrix} K & 0 \\ 0 & M^{-1} \end{bmatrix} \begin{bmatrix} I & K^{-1} L \\ 0 & I \end{bmatrix} \\ &= \begin{bmatrix} K^{1/2} & K^{-1/2} L \\ 0 & M^{-1/2} \end{bmatrix}^* \begin{bmatrix} K^{1/2} & K^{-1/2} L \\ 0 & M^{-1/2} \end{bmatrix} \end{aligned}$$

where $M^{-1} = N - L^* K^{-1} L$ can be computed by inverting the matrix (21) [12, Formula (0.8)].

We thus define

$$\mathcal{U} = \begin{bmatrix} \tilde{W} & \tilde{X} \\ \tilde{Y} & Z \end{bmatrix} \begin{bmatrix} K^{-1/2} & -K^{-1} L M^{1/2} \\ 0 & M^{1/2} \end{bmatrix}. \tag{22}$$

The matrices L and Z are given by

$$L = \tilde{Y}^* Z + \tilde{W}^* \tilde{X} \tag{23}$$
$$M = Z^* Z + X^* P^{-1} X. \tag{24}$$

Theorem 2 *Let $\Omega = (W, X, Y, Z)$ be a unitary realization and $\mathcal{D}_\Omega$ the set of unitary realizations in canonical form with respect to Ω (Λ in (12) is positive definite and Hermitian). The map*

$$\begin{aligned} \phi_\Omega : \quad \mathcal{D}_\Omega \quad &\to \mathbb{R}^{2np} \times \mathbb{U}_p, \\ (A, B, C, D) \quad &\mapsto \quad (V, D_0) \end{aligned}$$

where V, $\mathcal{V}$, $\mathcal{U}$ and D_0 are successively computed by (13), (20), (22) and (19), is a chart of $\mathbb{L}_n^p$.
The map $\phi_\Omega^{-1} : (V, D_0) \mapsto \mathcal{U} \operatorname{diag}(I, D_0) \mathcal{V}^*$ is a local canonical form.
The family $(\mathcal{D}_\Omega, \phi_\Omega)$, Ω unitary realization forms an atlas of $\mathbb{L}_n^p$.

Note that $\Omega \in \mathcal{D}_\Omega$ and has parameters $V = 0$ and $D_0 = I$. The chart is centered on G_Ω and is called an *adapted chart* for G_Ω. A parametrization of the quotient space $\mathbb{L}_n^p / \mathbb{U}_p$ is obtained by fixing D_0 within the chart and letting only V vary.

7.2 Illustration

Up to a right constant unitary matrix, any 2×2 lossless matrices of McMillan degree 1 with real coefficients can be written in the form

$$B_{a,\phi} = I + (\zeta_a - 1) \begin{bmatrix} \cos\frac{\phi}{2} \\ \sin\frac{\phi}{2} \end{bmatrix} \begin{bmatrix} \cos\frac{\phi}{2} \\ \sin\frac{\phi}{2} \end{bmatrix}^T, \phi \in [0, 2\pi[,$$

where ζ_a is a normalized Blaschke factor

$$\zeta_a(z) = \nu \frac{1 - \bar{a}\, z}{z - a}, \quad \begin{cases} \nu = -\frac{a}{|a|}, \ 0 < |a| < 1, \\ \nu = 1, a = 0. \end{cases} \tag{25}$$

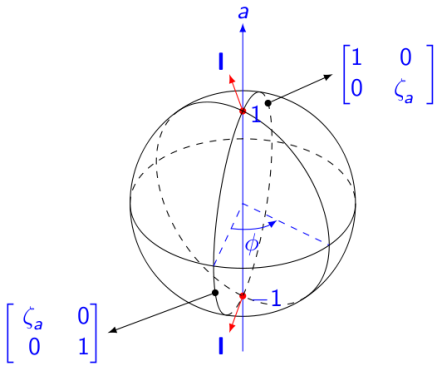


Fig. 3. The quotient space of 2×2, degree 1, real lossless functions

The dimension of this manifold is 2 and it can be represented by a sphere of which the poles have been excluded (see Picture 3). The poles of the sphere correspond to a drop of degree.

Consider the chart centered at $B_{0,0}(z) = \operatorname{diag}(1/z, 0)$. The parameter V is a 2-vector and $\Lambda = \sqrt{1 - \|V\|^2}$, so that the parameter domain is just the open unit disk ($\|V\| < 1$). The canonical form is

$$V = \begin{bmatrix} x \\ y \end{bmatrix} \mapsto \left[\begin{array}{c|cc} -x & 1 - \sigma x x^* & -\sigma x^* y \\ \hline \Lambda & x^* & y^* \\ -y & -\sigma y^* x & 1 - \sigma y y^* \end{array}\right].$$

The corresponding lossless function is of the form $B_{a,\phi}\, U$ with $a = -x$, $\cos\phi/2 = \Lambda$, $\sin\phi/2 = y$ and U a unitary matrix. Since Λ cannot be zero, the matrices on the meridian $\phi = \pi$ are not represented in this chart. They show up on the boundary of the chart except for $x = \pm 1$ and $y = 0$ which correspond to the poles of the sphere. As for the sphere, two charts at least are needed to represent the whole set. Adding the chart centered at $B_{0,\pi}$ we get an atlas.

7.3 Optimization in RARL2.

The software RARL2 is a Matlab based software which performs rational approximation following the principle we just described. It divides into two libraries

- `arl2lib` contains all the computations concerning the L^2 criterion and its gradient. The function can be given by a realization or Fourier coefficients as in our application. In this case, the matrix $\hat{B}$ (see Section 6) is computed by $\sum_{k\geq 0}^{N-1} F_{k+1}CA^k$, where the $(F_k)_{0\leq k\leq N}$ are the matrix Fourier coefficients.
- `boplib` is concerned with the parametrization of lossless functions (balanced output pairs) by means of the lossless mutual encoding method described in section 7.1. It also provides a minimization process which could handle **any criterion** defined over the manifold $\mathbb{L}_n^p/\mathbb{U}_p$.

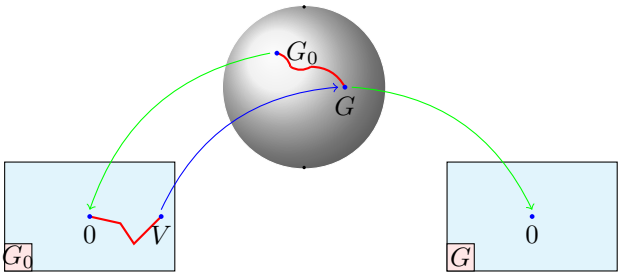


Fig. 4. Optimization over a manifold

The minimization process makes use of the Matlab solver `fmincon`. It starts at some initial point G_0 which is encoded in its adapted chart $\Omega = (A, B, C, D)$ (see section 7.1). Then `fmincon` performs the optimization of the criterion submitted to the nonlinear constraint $P > 0$, where P is the solution to

$$P - A^*PA = C^*C - V^*V$$

and V the parameter of the current point in the chart. This constraint ensures we remain within the domain of the chart. When a constraint violation occurs, a new adapted chart is computed and the optimization pursues within this new chart (see Figure 4) until a minimum is reached.

The convergence of the algorithm has been proved under mild assumptions in the SISO case [6] but never in the MIMO case. The main obstruction to the convergence is if the boundary of the manifold is reached, that is to say if the constraint violation (P singular) corresponds to the non-minimality of the canonical realization, that is to a drop of degree for the lossless functions. This would result in changing chart indefinitely.

7.4 Initialization

Since the criterion may possess many local minima, the choice of an initial point in the optimization process is essential. Projection based model reduction proposes a panel of low-cost computational methods to get a reasonable starting point. In particular, optimal Hankel norm model reduction [15] or balanced truncation [30] presents guarantees of quality and error bound on the result. In RARL2, the balanced truncation method of [23] has been implemented and is used as a starting point for the identification of microwave filters.

8 Results and conclusion.

A long-standing cooperation with the space agency CNES resulted in an original method to extract coupling parameters from frequency scattering measurements, and in two dedicated software programs which are now fully integrated in the design and tuning process. In this paper we have described in detail the identification step performed by the software PRESTO-HF [35] (which includes the rational approximation software RARL2). In Figures 5,6,7 the results of our procedure are shown at hand of a real-life example provided by the CNES. It consists of measurements of a microwave filter of 8th order in 800 frequency points. As shown by the error function in Figure 7, an excellent agreement is obtained between the final rational model and the measurements. The latter is obtained in less than 15 seconds on a Intel Core I7 processor, which makes our approach compatible with a real-time tuning procedure of the filter. The software PRESTO-HF [35] is currently used for this purpose by several of our industrial partners.

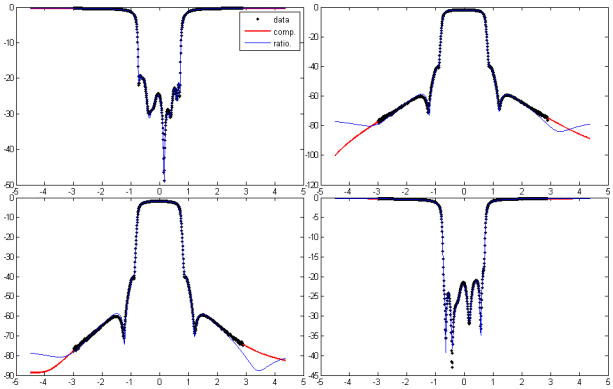


Fig. 5. CNES 2×2 hyperfrequency filter; Bode diagram of data (dots), completion (red) and approximant at order 8 (blue line).

In a second step we extract the coupling parameters from the identified model. This step is performed by the software DEDALE-HF [34]. Based on computer algebra methods, this software computes a realization of the form (1) in which the coupling geometry of M has been specified.

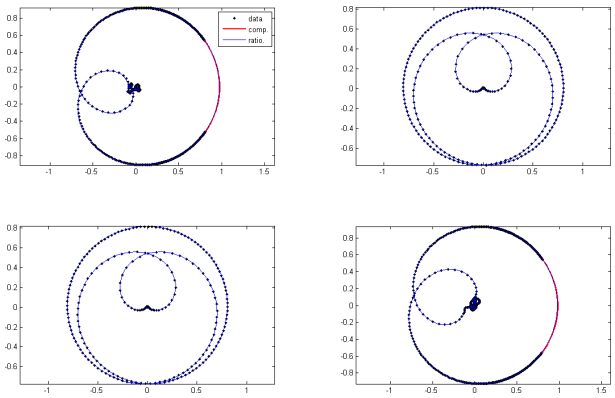


Fig. 6. CNES 2×2 hyperfrequency filter: data (dots), completion (red), and approximant (blue) at order 8 (Nyquist diagram).

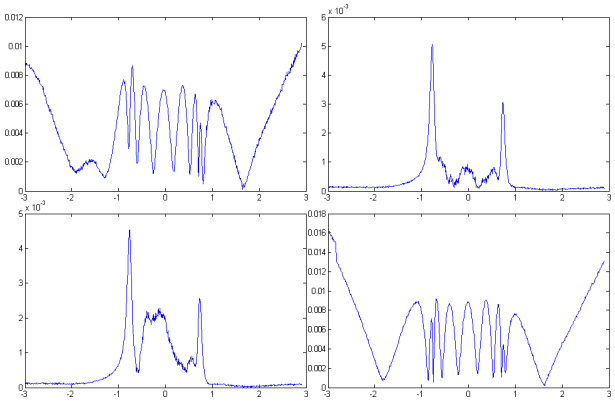


Fig. 7. CNES 2×2 hyperfrequency filter: Magnitude of the point-wise error between rational approximation (including delay components) and measurements for each entry.

The case of output multiplexers (OMUX) where several filters of the previous type are coupled on a common guide has also been considered. The model is obtained upon chaining the corresponding scattering matrices, and mixes up rational elements and complex exponentials (because of the delays). This makes the identification much more challenging. Nowadays, the new trend is to remove the waveguides that are an important part of the mass and bulk, in order to integrate more strongly these devices and to simplify their architecture. The devices obtained in this way (compact OMUX) differ from a single filter by the number of ports m greater than 2. They are described by a scattering matrix of order m. We are currently investigating the possibility to apply our methods to such devices. If the model reduction step directly applies to this case, the completion as well as the extraction of coupling parameters from the model require new developments.

References

[1] R.S. Adve, T.K. Sarkar, S.M. Rao, E.K. Miller, and R. Pflug. Application of the Cauchy method for extrapoling/interpolating narrow-band system responses. *IEEE Transactions on Microwave Theory and Techniques*, 45:837–845, 1997.

[2] D. Alpay, L. Baratchart, and A. Gombani. On the differential structure of matrix-valued rational inner functions. *Operator Theory: Advances and Applications*, 73:30–66, 1994.

[3] J. D. Aplevich. Gradient methods for optimal linear system reduction. *Int. J. Control*, 18, 1973.

[4] J.A. Ball, I. Gohberg, and L. Rodman. *Interpolation of rational matrix functions*, volume 45 of *Operator Theory: Advances and Applications*. Birkhäuser, 1990.

[5] L. Baratchart. Existence and generic properties for L^2 approximants of linear systems. *I.M.A. Journal of Math. Control and Identification*, 3:89–101, 1986.

[6] L. Baratchart, M. Cardelli, and M. Olivi. Identification and rational L^2 approximation: a gradient algorithm. *Automatica*, 27(2):413–418, 1991.

[7] L. Baratchart and J. Leblond. Hardy approximation to L^p functions on subsets of the circle with $1 \leq p < \infty$. *Constructive Approximation*, 14:41–56, 1998.

[8] L. Baratchart, J. Leblond, J. R. Partington, and N. Torkhani. Robust identification from band-limited data. *IEEE Trans. on Atom. Control*, 42(9):1318–1325, 1997.

[9] L. Baratchart, J. Leblond, and F. Seyfert. Extremal problems of mixed type in H^2 of the circle. Rapport de recherche RR-7087, INRIA, 2009.

[10] L. Baratchart and M. Olivi. Critical points and error rank in best H^2 matrix rational approximation. *Constructive Approximation*, 14:273–300, 1998.

[11] P. Van Dooren, K.A. Gallivan, and P.-A. Absil. H^2-optimal model reduction of MIMO systems. *Applied Mathematics Letters*, 21, 2008.

[12] H. Dym. *J-contractive matrix functions, reproducing kernel spaces and interpolation*, volume 71 of *CBMS lecture notes*. American Mathematical Society, Rhode Island, 1989.

[13] P. Fuhrmann and U. Helmke. Homeomorphism between observable pairs and conditioned invariant subspaces. *Systems and Control Letters*, 30:217–223, 1997.

[14] P. Fulcheri and M. Olivi. Matrix rational H^2-approximation: a gradient algorithm based on Schur analysis. *SIAM Journal on Control and Optimization*, 36(6):2103–2127, 1998.

[15] K. Glover. All optimal Hankel norm approximations of linear multivariable systems and their L^∞-error bounds. *Int. J. Contr.*, 39:1115–1193, 1984.

[16] S. Gugercin, A.C. Antoulas, and Beattie C. H^2 model reduction for large-scale linear dynamical systems. *SIAM J. Matrix Anal. Appli.*, 30, 2008.

[17] B. Gustavsen and A. Semlyen. Rational approximation of frequency domain responses by vector fitting. *IEEE Trans. Power Delivery*, 14(3):1052–1061, 1999.

[18] B. Gustavsen and A. Semlyen. A robust approach for system identification in the frequency domain. *IEEE Trans. Power Delivery*, 19(3):1167–1173, 2004.

[19] B. Hanzon and J.M. Maciejowski. Constructive algebra methods for the L^2-problem for stable linear systems. *Automatica*, 32(12), 1996.

[20] B. Hanzon, M. Olivi, and R.L.M. Peeters. Balanced realizations of discrete-time stable all-pass systems and the tangential Schur algorithm. *Linear Algebra and its Applications*, 2006.

[21] D.C. Hyland and D.S. Bernstein. The optimal projection equations for model reduction and the relationships among the method of Wilson, Skelton and Moore. *IEEE Trans. Aut. Control*, 29, 1985.

[22] I. Kollár. Frequency domain system identificaton toolbox for matlab. Gamax Ltd, Budapest, 2004.

[23] S.-Y. Kung and D.W. Lin. Optimal hankel-norm model reductions: Multivariable systems. *IEEE trans. on Automatic control*, AC-26(4):832–852, 1981.

[24] A.G. Lampérez, T.K. Sarkar, and M.S. Palma. Filter model generation from scattering parameters using the Cauchy method. In *European Microwave Conference*, 2002.

[25] H. Lev-Ari and T. Kailath. State-space approach to factorization of lossless transfer functions and structured matrices. *Linear Algebra and its Applications*, 162–164:273–295, 1992.

[26] J. P. Marmorat and M. Olivi. RARL2: a Matlab based software for H^2 rational approximation. http://www-sop.inria.fr/apics/RARL2/rarl2.html, 2004.

[27] J.-P. Marmorat and M. Olivi. Nudelman interpolation, parametrizations of lossless functions and balanced realizations. *Automatica*, 43:1329–1338, 2007.

[28] J.-P. Marmorat, M. Olivi, B. Hanzon, and R.L.M. Peeters. Matrix rational H^2 approximation: a state-space approach using Schur parameters. In *Proceedings of the CDC, Las-Vegas, USA*, 2002.

[29] L. Meier and D. G. Luenberger. Approximation of linear constant systems. *IEEE Trans. Aut. Control*, 12:585–587, 1967.

[30] B.C. Moore. Principal component analysis in linear systems: Controllability, observability, and model reduction. *IEEE Transactions on Automatic Control*, AC-26:17–32, 1981.

[31] P. Van Overschee, B. De Moor M. Gevers, and G. Li. *Parametrizations in control, estimation and filtering problems*. Springer-Verlag, 1993.

[32] J. Partington. *Interpolation, Identification and Sampling*. Oxford University Press, 1997.

[33] G. Ruckebusch. Sur l'approximation rationnelle des filtres. Technical report, CMA, Ecole polytechnique, 1978.

[34] F. Seyfert. DEDALE-HF: a Matlab toolbox dedicated to the equivalent network synthesis for microwave filters. http://www-sop.inria.fr/apics/Dedale/WebPages/.

[35] F. Seyfert. PRESTO-HF: a Matlab based toolbox dedicated to the identification problem of low pass coupling parameters of band pass microwave filters, 2004.

[36] J.T. Spanos, M.H. Milman, and D.L. Mingori. A new algorithm for l^2 optimal model reduction. *Automatica*, 28:897–909, 1992.

[37] M. Verhaegen and V. Verdult. *Filtering and system identification: a least-square approach*. Cambridge university press, 2007.

[38] D. Wilson and R. Mishra. Optimal reduction of multivariable systems. *Int. J. Control*, 29(2):267–278, 1979.

5.1.3 Detection of instabilities in power amplifiers using the decomposition $L^2 = H^2 \oplus \overline{H_0^2}$

Following paper is reproduced in this section:

- Adam Cooman, Fabien Seyfert, Martine Olivi, Sylvain Chevillard, and Laurent Baratchart. "Model-Free Closed-Loop Stability Analysis: A Linear Functional Approach". In: *IEEE Transactions on Microwave Theory and Techniques* (Sept. 2017). DOI: 10.1109/TMTT.2017.2749222. URL: https://hal.inria.fr/hal-01381731

Model-Free Closed-Loop Stability Analysis: A Linear Functional Approach

Adam Cooman, *Member, IEEE*, Fabien Seyfert, Martine Olivi, Sylvain Chevillard and Laurent Baratchart

arXiv:1610.03235v3 [cs.SY] 20 Feb 2018

***Abstract*—Performing a stability analysis during the design of any electronic circuit is critical to guarantee its correct operation. A closed-loop stability analysis can be performed by analysing the impedance presented by the circuit at a well-chosen node without internal access to the simulator. If any of the poles of this impedance lie in the complex right half-plane, the circuit is unstable. The classic way to detect unstable poles is to fit a rational model on the impedance.**

In this paper, a projection-based method is proposed which splits the impedance into a stable and an unstable part by projecting on an orthogonal basis of stable and unstable functions. When the unstable part lies significantly above the interpolation error of the method, the circuit is considered unstable. Working with a projection provides one, at small cost, with a first appraisal of the unstable part of the system.

Both small-signal and large-signal stability analysis can be performed with this projection-based method. In the small-signal case, a low-order rational approximation can be fitted on the unstable part to find the location of the unstable poles.

Frequency domain simulation methods, like Harmonic Balance (HB) or a DC analysis, impose a structure on the obtained solution of the circuit [1]: The DC analysis only allows for a fixed solution, while HB imposes a frequency grid. Any circuit solution that requires more than the imposed frequencies, e.g. an extra oscillation not on the imposed grid, cannot be represented in the constrained frequency grids of DC and HB. The simulator will still find a valid solution, but the obtained orbit will be locally unstable: it cannot recover from small perturbations and will be physically unobservable in the circuit [2]. It is therefore necessary to perform a local stability analysis on each of the circuit solutions obtained with a DC and HB analysis [1].

Over the years, several methods have been developed to determine the local stability of a circuit solution. Some techniques, like the analysis of the characteristic system [2], require access to the simulator. Open-loop techniques, like the analysis of the normalised determinant [1], require access to the intrinsic device models. These classic techniques are therefore hard to implement in commercial simulators.

Closed-loop stability analysis methods can easily be applied as a post-processing step without any internal knowledge of the circuit and can be used in commercial simulators. This is the reason why they have attracted a large interest lately [1], [3], [4]. A closed-loop local stability analysis performs linearisation of the circuit around the orbit to check the stability thereof: if the linearised circuit has at least one pole in the complex right half-plane, the orbit is unstable. It is moreover assumed that, conversely, the absence of unstable pole implies stability, although no published proof of this fact seems available yet. The question is more subtle than it looks: there exist delay systems which are unstable and still their transfer-function has no unstable pole [5], moreover has an example of an ideal circuit with this property. Nevertheless, it is claimed in [6] that a circuit whose elements are passive at arbitrary high frequencies must indeed have some unstable pole if it is unstable.

The poles of the linearisation around the circuit orbit cannot be obtained directly. Instead a FRF of the linearised circuit is obtained with small-signal simulations on a discrete set of frequencies. The closed-loop stability analysis then aims at determining whether the underlying FRF has a pole in the complex right half-plane. In a pole-zero stability analysis, a rational approximation is fitted on the FRFs. If the rational approximation contains poles in the complex right half-plane, the solution is declared unstable.

Note that the FRF of circuits with distributed elements, like transmission lines, is not rational. Therefore it must be argued that the poles of the computed rational approximant convey information on the poles of the true FRF. This is a delicate issue and a particular instance of a recurring question in approximation theory, namely: what do the singularities of an approximant tell us about the singularities of the approximated function? We observe that no such information can be drawn from the mere quality of approximation in a range of frequencies, since a famous theorem by Runge entails that a continuous function on a segment can be approximated arbitrary well by a proper rational function with prescribed pole location [7]. Thus, for singularity detection, the choice of the approximation algorithm (and not just the fit of the approximant) does matter.

For instance, methods based on linear interpolation, like Padé or multipoint Padé approximation, are famous for generating spurious poles that wander about the domain of analyticity of the approximated function. This phenomenon was intensively studied for meromorphic and branching functions [8]–[10], in particular the convergence in capacity of Padé approximants implies that spurious poles have a nearby zero when the order gets large, leading to so-called near pole-zero cancellations (also known as Froissart doublets). Modifications of Padé approximants were proposed to offset this issue [11], but they do not eliminate the problem [12]. Apparently, the theoretically less studied vector fitting method which is a least squares version of linear interpolation, popular today in system analysis, is also prone to producing spurious poles and near cancellations (see [13] for issues on convergence of this

Adam Cooman, Fabien Seyfert, Martine Olivi, Sylvain Chevillard and Laurent Baratchart are with APICS at INRIA, Sophia Antipolis. e-mail: adam.cooman@inria.fr

method).

In system identification, near cancellations are often ascribed to overmodelling. The terminology suggests an analogy with the stochastic identification paradigm: though measurements may not correspond to a rational transfer function, the basic assumption is that they arise from a well-defined rational system R with added noise. This point of view leads one to postulate the existence of a "correct order" to identify the Frequency Response Function (FRF), i.e. the degree of R, while using a higher degree results in approximating the noise term with inessential, nearly simplifying rational elements. However, if the transfer function is not rational, requirements to keep the degree small conflict with the need to make the approximation error small as well (not to incur undermodelling), thus calling for a compromise akin to the classical trade-off between bias and variance from parametric stochastic identification [14]. To quote [15][1]: "it is not always trivial to discriminate between overmodelling quasi-cancellations and physical quasi- cancellations that really reflect an unstable behaviour".

To resolve this issue, the approach proposed in [15] is to cut the frequency band into smaller intervals and use low-order local rational approximations to assess the stability of the FRF on each interval separately. On small enough intervals, rational approximation can be performed accurately in low degree, and if unstable poles occur their physical character is checked by re-modelling the FRF locally around each of them and verifying that the unstable pole remains present in the new model. This procedure is commercially available in the STAN tool [16], [17] and successful applications on several examples are reported in [18]–[21].

Still, justifying the above-described technique presently rests on heuristic arguments, and putting it to work is likely to require some know-how since several parameters need to be adjusted adequately (Appendix A, for example, shows that local models of a stable FRFs can become unstable). This is why the authors feel that it may be interesting to develop an alternative viewpoint, focusing more on estimating the unstable part of the glsFRF.

Below, we propose a closed-loop stability analysis method devoid of local models, in which the FRF is projected onto the orthogonal basis of stable and unstable functions. If a significant part of the FRF is projected onto the unstable basis functions, the circuit solution is unstable. Calculating the projection boils down to computing a Fourier transform once the FRF is mapped from the imaginary axis to the unit circle. Using the Fast Fourier Transform (FFT), this can be done fast and in a numerically robust way.

Functional projection onto a stable and unstable basis is a linear operation, simple to implement, and no optimisation step is required. No model-order or maximum approximation error needs to be specified. The parameters in the projection method are the frequency range on which the FRF is determined and the amount of simulation points. When the amount of simulated points is too low, an interpolation error is present in the result of the method. It is shown that the level of this error can easily be estimated and used to correctly choose the amount of needed simulation points.

[1]The rational approximation technique used in this reference is described as "frequency domain least squares identification"

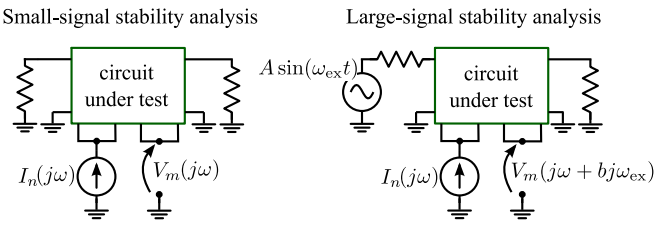


Fig. 1. A Small-signal current source is connected to a well-chosen node in the circuit under test to perform the local stability analysis.

Once the unstable part of the FRF has been obtained, it is compared to the level of the interpolation error to determine whether the unstable part is significant or not. This final step can be done visually, or a significance threshold can be chosen by the user, both will require some experience with the method.

A final benefit of the projection-based approach is that it may help exploiting the fact that the unstable part is rational in a small-signal stability analysis [6]. The unstable part can therefore be approximated by a rational function without influence of the distributed elements, which are projected onto the stable part of the FRF.

The following of the paper is structured as follows: First, the simulation set-up used to determine the FRF of the linearised circuit is discussed (Section I). Then, the details of the functional projection are provided (Section II). In Section III, the method is applied to four examples: First, an artificial example is considered. Then, the small-signal stability of two amplifiers is investigated and finally, the method is applied to investigate the large-signal stability of a circuit.

I. Determining the frequency responses

In this paper, the (trans)impedance presented by the circuit to a small-signal current source will be used as FRF (Fig. 1). In the remainder of this paper, it will be assumed that the unstable poles are observable in the FRF. To reduce the chance of missing an instability in the circuit due to a pole-zero cancellation, many different FRFs can be analysed one-by-one. Having a fast method to determine stability of a single FRF is therefore critical to a robust stability analysis.

The FRF of the linearised circuit is obtained by first placing the circuit in the required orbit, using either a DC or HB analysis and running a small-signal simulation around this orbit.

In a small-signal stability analysis, the stability of the DC solution of the circuit is investigated, so the FRF of the linearised circuit is obtained with an AC simulation. The impedance of the circuit is then obtained as:

$$Z_{mn}(j\omega) = \frac{V_m(j\omega)}{I_n(j\omega)} \tag{1}$$

where $I_n(j\omega)$ is the small-signal current injected into the selected node n and $V_m(j\omega)$ is the voltage response of the circuit measured at node m in the circuit.

In a large-signal stability analysis, the stability of a large-signal solution of the circuit is investigated. The circuit is driven by a periodic continuous-wave excitation at a pulsation ω_{ex} and the circuit solution is obtained with a HB simulation.

The FRF of the linearised system around the HB orbit is obtained with a mixer-like simulation[2]. As the small-signal will mix with the large signal, several transfer impedances with a different frequency translation are obtained:

$$Z_{mn}^{[b]}(j\omega) = \frac{V_m(j\omega + bj\omega_{\text{ex}})}{I_n(j\omega)} \quad b \in \mathbb{Z}$$

The stability analysis now needs to determine whether the obtained impedances have poles in the right half-plane. The stability analysis of a large-signal orbit doesn't differ much from the analysis of a DC solution [4]. The small-signal stability analysis can be considered a special case where only $Z_{mn}^{[0]}(j\omega)$ is analysed.

II. Stable/unstable projection

The projection described here has been used before to perform stable interpolation and extrapolation of FRF data [22]. With slight modifications, it can be turned into a full-blown stability analysis. We first start with a brief introduction to the notion of Hardy spaces.

The Hardy space $H^2(\mathbb{C}^+)$, is defined as the set of all functions g defined on $\mathbb{C}^+$ such that:

- $\forall z \in \mathbb{C}^+$, g is holomorphic at z
- $\sup_{x>0} \int_{-\infty}^{+\infty} |g(x+j\omega)|^2 d\omega < \infty$

A classical result [7], [23] states that every function $g \in H^2$ admits a limiting function $G(j\omega)$ defined on the imaginary axis $j\mathbb{R}$. The latter is obtained by taking the limit of $g(z)$ when z tends non tangentially toward $j\omega$. Moreover $\forall z \in \mathbb{C}^+$, $g(z)$ is equal to the poisson integral of G, that is:

$$g(z = x + jy) = \int_{-\infty}^{+\infty} G(j\omega) \frac{x}{x^2 + (y-\omega)^2} d\omega. \quad (2)$$

The holomorphic nature of g ensures that it is also the Cauchy integral of its boundary value G:

$$g(z = x + jy) = \frac{1}{2\pi} \int_{-\infty}^{+\infty} G(j\omega) \frac{1}{j\omega - z} d\omega. \quad (3)$$

There is a one to one linear correspondence between the Hardy functions g and its boundary value function G. Using this identification, $H^2(\mathbb{C}^+)$ becomes a subspace of the Hilbert space $L^2(j\mathbb{R})$ of square integrable functions on $j\mathbb{R}$. A direct consequence of (2) is that $H^2(\mathbb{C}^+)$ is closed in $L^2(j\mathbb{R})$ and therefore admits an orthogonal complement. An important result [23] asserts that, $\left(H^2(\mathbb{C}^+)\right)^\perp = H^2(\mathbb{C}^-)$ where $H^2(\mathbb{C}^-)$ is defined exactly as $H^2(\mathbb{C}^+)$ above by replacing $\mathbb{C}^+$ by $\mathbb{C}^-$ and taking the supremum over $x < 0$. We therefore have that

$$L^2(j\mathbb{R}) = H^2\big(\mathbb{C}^+\big) \oplus H^2\big(\mathbb{C}^-\big)$$

This decomposition asserts that any square integrable function on $j\mathbb{R}$ decomposes uniquely as the sum of the traces on the imaginary axis, of an analytic function in the right half-plane and a function analytic in the left half-plane. The projection on $H^2(\mathbb{C}^+)$ defines the stable part of the function. The projection onto $H^2(\mathbb{C}^-)$ is the unstable part. As an example, consider P/Q a strictly proper ($\deg(P) < \deg(Q)$) rational function devoid of poles on the imaginary axis. We write its partial fraction expansion as,

$$\frac{P(s)}{Q(s)} = \sum_{i \in I^+} \sum_{k=1}^{k_i} \frac{a_{i,k}}{(s-\lambda_i)^k} + \sum_{i \in I^-} \sum_{k=1}^{k_i} \frac{a_{i,k}}{(s-\lambda_i)^k}$$

where the $\lambda_i' s$ with $i \in I^-$ are poles belonging to $\mathbb{C}^-$, and the ones with $i \in I^+$ belong to $\mathbb{C}^+$. The strict properness of P/Q ensures its square integrability on $j\mathbb{R}$. By unicity its stable part obtained after projection on $H^2(\mathbb{C}^+)$ is found to be,

$$\sum_{i \in I^-} \sum_{k=1}^{k_i} \frac{a_{i,k}}{(s-\lambda_i)^k},$$

while its unstable part is

$$\sum_{i \in I^+} \sum_{k=1}^{k_i} \frac{a_{i,k}}{(s-\lambda_i)^k}.$$

In the general case the projection boils down to calculating certain inner products of $Z_{mn}^{[b]}(j\omega)$ with the basis functions B_k, which form an orthogonal basis of $L^2(j\mathbb{R})$

$$c_k = \left\langle Z_{mn}^{[b]}(j\omega), B_k \right\rangle = \int_{-\infty}^{\infty} Z_{mn}^{[b]}(j\omega) \, \overline{B_k(j\omega)} d\omega \quad (4)$$

$$B_k(s) = -\sqrt{\frac{\alpha}{\pi}} \frac{(s-\alpha)^k}{(s+\alpha)^{k+1}} \qquad k \in \mathbb{Z} \quad (5)$$

The overbar $\overline{\bullet}$ indicates the complex conjugate. α is a positive constant used for scaling. All B_k with $k \geq 0$ create a basis for the stable part, while the B_k with negative k form a basis for $H^2(\mathbb{C}^-)$. Once the c_k coefficients are calculated, the stable and unstable parts are easily recovered by calculating

$$Z_{\text{stable}}(j\omega) = \sum_{k=0}^{\infty} c_k B_k(j\omega) \quad (6)$$

$$Z_{\text{unstable}}(j\omega) = \sum_{k=1}^{\infty} c_{-k} B_{-k}(j\omega) \quad (7)$$

The inner product in (4) runs over all frequencies while the impedance function $Z_{mn}^{[b]}(j\omega)$ is only known over a frequency range $\mathbf{f} = [f_{\min}, f_{\max}]$. To impose the finite frequency band on the data, the impedance is filtered before the analysis

$$Z_{\text{f}}(j\omega) = Z_{mn}^{[b]}(j\omega) \, H(j\omega) \quad (8)$$

The filter $H(j\omega)$ is a high-order elliptic lowpass filter with its first transmission zero placed at $f_{\max}$ that imposes band limitation. This filter will stabilise poles close to its cutoff frequency, so $f_{\max}$ should be chosen well beyond the maximum frequency at which the circuit can become unstable. $f_{\min}$ should be placed very close to DC. If $f_{\min}$ can't be close to DC, a bandpass filter should be used for $H(j\omega)$. The filtering ensures that $Z_{\text{f}}(j\omega) \in L^2(j\mathbb{R})$ by suppressing anything outside of the frequency band of interest. The smooth

[2]In Keysight's Advanced Design System (ADS), this mixer-like simulation is called a Large-Signal Small-Signal (LSSS) analysis.

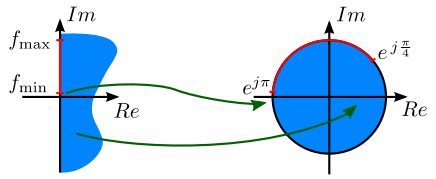


Fig. 2. The Möbius transform used in the analysis maps the complex right-half plane to the inside of the unit disc. DC is mapped to -1. When $\alpha = 2\pi f_{\max}\left(\sqrt{2}-1\right)$, $f_{\max}$ is mapped to $e^{j\frac{\pi}{4}}$.

decay to zero of $Z_{\mathrm{f}}(j\omega)$ at the edges of the frequency interval will avoid instabilities to pop up due to the discontinuity of $Z_{mn}^{[b]}(j\omega)$.

In this paper, we use an elliptic filter of order 10 to filter the data. The filter has one transmission zero which is placed exactly at $f_{\max}$.

A. Transforming to the unit circle

Working with the basis functions defined in (5) is troublesome from a numeric point of view. Performing the projection when working on unit disc yields better results [22]. The mapping from the complex plane to the unit circle is performed with the Möbius mapping visualised in Fig. 2:

$$Z_{\mathrm{f}}(s) \underset{\text{Möb}}{\longmapsto} Z_{\mathrm{f}}^{\mathrm{disc}} = \sqrt{\pi\alpha}\frac{2}{z-1}Z_{\mathrm{f}}\left(\alpha\frac{1+z}{1-z}\right) \tag{9}$$

Our mapping of choice converts square integrable functions on the frequency axis into square integrable functions of the same norm on the unit circle. Square integrable functions which are analytic on the right half-plane are mapped onto square integrable functions which are analytic inside the unit disc. Appendix B shows that the basis functions B_k map onto powers of z

$$B_k(s) \underset{\text{Möb}}{\longmapsto} B_k^{\mathrm{disc}}(z) = z^k \tag{10}$$

Projecting on this basis boils down to calculating the Fourier series of $Z_{\mathrm{f}}^{\mathrm{disc}}(z)$, with coefficients given by

$$c_k = \frac{1}{2\pi}\int_0^{2\pi} Z_{\mathrm{f}}^{\mathrm{disc}}\left(e^{j\theta}\right)e^{-jk\theta}d\theta \tag{11}$$

This Fourier series can be calculated in a numerically efficient way using the Fast Fourier Transform (FFT). The FFT requires that the θ-values are linearly spaced between 0 and 2π. Due to the mapping from the complex plane to the unit disc, the samples will not satisfy this constraint. A simple interpolation, can be used to obtain $Z_{\mathrm{f}}^{\mathrm{disc}}$ on the θ-values required to perform the FFT.

The interpolation can introduce artefacts in the unstable part if the FRF is not sampled on a sufficiently dense frequency grid, which will be shown on an example later. The level of this interpolation error can be estimated by interpolating the FRF using only the data at the even data points. The difference between the original and the interpolated data for the odd data points will give an indication of the interpolation error encountered in the stability analysis. When the unstable part lies significantly above the interpolation error, we can conclude that the original impedance is unstable.

The threshold to determine when the unstable part lies significantly above the level of the interpolation error will depend on the simulation set-up for the circuit. In the examples that follow, a level of 20 dB has been used as a threshold. A more strict threshold could be used at the cost of requiring a more dense frequency grid and an increased simulation time. When measured components are used in the simulation set-up, the noise level in those measurements should be taken into account to choose the correct threshold.

B. Summary of the projection method

The stable and unstable parts of a FRF are determined using the following steps:

1) Multiply the FRF with a high-order filter as in (8)
2) Transform the filtered FRF to the unit disc using (9)
3) Interpolate the transformed FRF to a linear grid and use the FFT to calculate the c_k coefficients
4) Reconstruct the stable and unstable part using (6-7)

C. Obtaining the unstable poles

A function is meromorphic on $\mathbb{C}$ if it is holomorphic on $\mathbb{C}$ but on a countable number of isolated poles. The function $\tanh(\omega)$ is for example meromorphic, having infinitely many isolated poles on the imaginary axis placed at $j\pi/2 + jk\pi$ $(\forall k \in \mathbb{Z})$ and being analytic elsewhere. In a small-signal stability analysis of a circuit composed of lumped elements, transmission lines and active devices modelled by negative resistors, the impedances can be shown to be meromorphic functions of the frequency. Under the additional realistic assumption that active elements can only deliver power over a finite bandwidth, the impedances are proven to possess only finitely many unstable poles in $\mathbb{C}^+$ [6]. Under the generic condition that $Z_{mn}^{[b]}(j\omega)$ is devoid of poles on the imaginary axis, and that the filtering function $H(j\omega)$ decays strongly enough in order to render $Z_{mn}^{[b]}(j\omega)\,H(j\omega)$ square integrable, we conclude that the unstable part of $Z_{mn}^{[b]}(j\omega)\,H(j\omega)$ is a rational function. Its poles coincide with the unstable poles of $Z_{mn}^{[b]}(j\omega)$: note here that the multiplication by $H(j\omega)$ does not add any unstable pole as $H(j\omega)$ is stable. This means that most of the complexity of the frequency response, like the delay, will be projected onto the stable part, while the unstable part can easily be approximated by a low-order rational model to recover the unstable poles[3].

Classic rational approximation tools can be used to approximate the unstable part and determine the unstable poles. When multiple frequency responses are analysed simultaneously, an approximation method suited for approximation of rational matrices is preferred [24]. In our current implementation, Kung's method [25], [26] is used to estimate the poles of the unstable part of a single FRF at a time. Alternatively, more sophisticated rational approximation engines like RARL2 [24] can be used to recover and track unstable poles.

[3]Interpolation error will be present in the obtained unstable part, but its influence can be minimised by weighting the rational estimator with the obtained interpolation error level.

Compared to working with a high-order rational approximation of the total impedance of the circuit, the 'split-first, approximate later' approach proposed here could be a faster and easier method to recover the unstable poles. The post-processing will be faster, but the amount of points required to obtain a sufficiently low interpolation error might be higher than the amount of points required for a tool based on rational approximation.

When the circuit is stable, designers often require information about critical stable poles, to determine how far the circuit is from instability and to track the location of the poles as the circuit varies. In its current form, the projection-based analysis does not simplify finding the location of the stable poles. To perform such an analysis, methods based on local modelling may still be required.

III. Examples

The stability analysis will now be applied to four different examples. The first is an artificial example generated in Matlab on which we can demonstrate that the unstable poles in the circuit are recovered perfectly. In the second example, an unstable balanced amplifier is analysed to show that the method works for RF circuits. The third example is a two-stage GaN Power Amplifier (PA). In the final example, a large-signal stability analysis is performed to verify the stability of a circuit orbit obtained in a HB simulation.

All simulations were performed in Keysight's Advanced Design System (ADS) and the post-processing was performed in Matlab.

A. Example 1: Random state space system

As a first example, the stability analysis is applied to a random system of order 202 generated with the rss function from Matlab[4]. The test system has an unstable pole pair at 1 GHz, as can be seen on its pole-zero map (Fig. 3). A zero is placed close to the unstable poles. This makes that the unstable poles are difficult to observe in the FRF. To introduce delay in the test system, a time delay of 2 ns is added to the system. The frequency response of the system is calculated on 5000 linearly spaced frequency points between 0 Hz and 5 GHz and is shown in green in Fig. 4. The obtained stable and unstable parts after projection are shown in red and blue on the same figure. The maximum interpolation error is very low in this example (-120 dB) . We will focus on the effect of the interpolation error in more detail in example 2.

The obtained unstable part peaks at 1 GHz, which matches the location of the unstable pole pair of the system. Note also that the obtained Z_{unstable} is very simple: it is clearly a second-order system. Most of the complexity of the frequency response, including the delay, is projected onto the stable part. This observation supports the proposed approach of estimating a rational model only after projection. A good fit was obtained with a rational model that consisted of two unstable poles and a single zero. The two poles obtained with a rational approximation of Z_{unstable} coincide exactly with the unstable poles in the circuit as is shown in Fig. 3.

[4]The rss function in Matlab returns models with poles and zeroes around 1Hz. The example here was scaled up in frequency to represent an RF circuit.

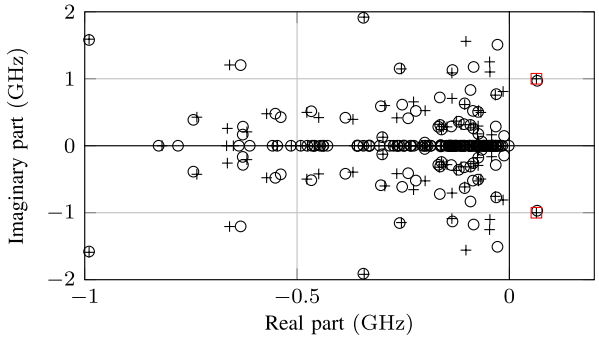


Fig. 3. Pole-zero map of the test system. There are 202 poles (+) and 200 zeroes (○). The two poles placed in the right half-plane are easily recovered after the projection by fitting a low-order model on the unstable part (□).

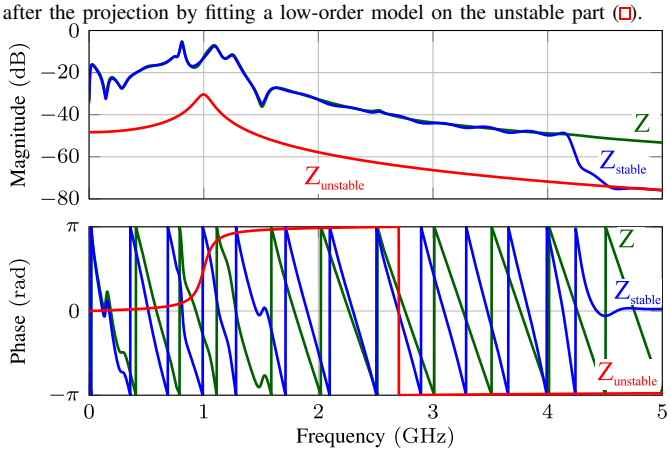


Fig. 4. The stable/unstable projection splits the original frequency response (–) in a stable part indicated with (–) and an unstable part indicated with (–)

B. Example 2: Balanced amplifier

The second example is a balanced amplifier built as a student project for operation around 3.4 GHz (Fig. 5). Two BFP520 transistors were used to construct the amplifier. During measurement, the design oscillated around 1.43 GHz when terminated with 50 Ω, so the circuit is a good candidate to verify the proposed method to find the instability in simulations.

The small-signal current source was connected to the collector of the top transistor. The BFP520 has a f_{T} of 45 GHz, so the maximum frequency for the simulation was set to 50 GHz. The impedance of the circuit was determined starting from DC in 10 MHz steps. The obtained impedance is shown in green in Fig. 6, the obtained stable and unstable parts are also shown in the same figure. The instability around 1.46 GHz is detected, but also some artefacts can be observed in the obtained unstable part at higher frequencies. The high level of the interpolation error at the frequency of these artefacts indicates that they are due to the interpolation in step 3 of the stable/unstable projection. At the frequency of the detected instability, the unstable part lies about 30 dB above the error level, which indicates that the instability is not an interpolation artefact.

To confirm that the artefacts are caused by the interpolation, a second simulation was run, but now 1 MHz steps were used instead of 10 MHz. The stable/unstable projection of

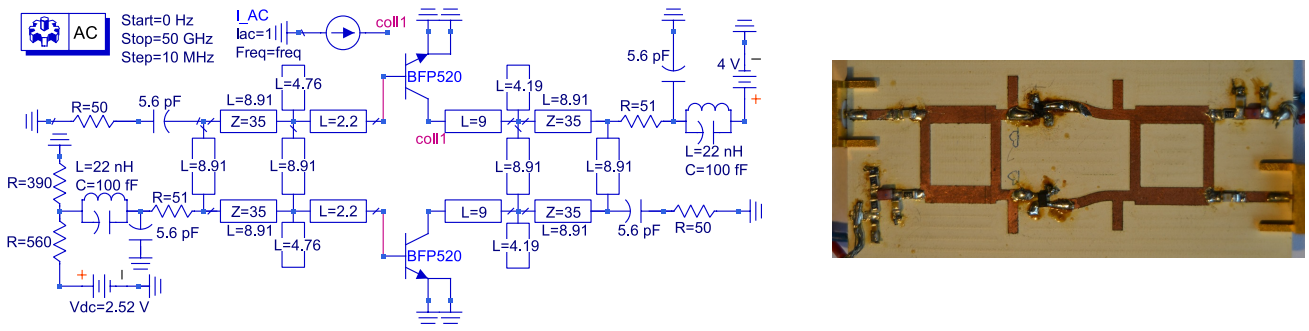


Fig. 5. Simulation set-up and photograph of the balanced amplifier. In the simulations for the stability analysis, the amplifier is excited at the collector of one of its transistors. All transmission lines in the circuit are 50Ω lines unless stated otherwise. The length of the transmission lines is given in millimetres. The TLINP model was used for the transmission lines with $\epsilon_r = 6.15$, $\tan(\delta) = 0.003$ and conductor losses $A = 2.5$ $^{\mathrm{dB}}/_{\mathrm{m}}$.

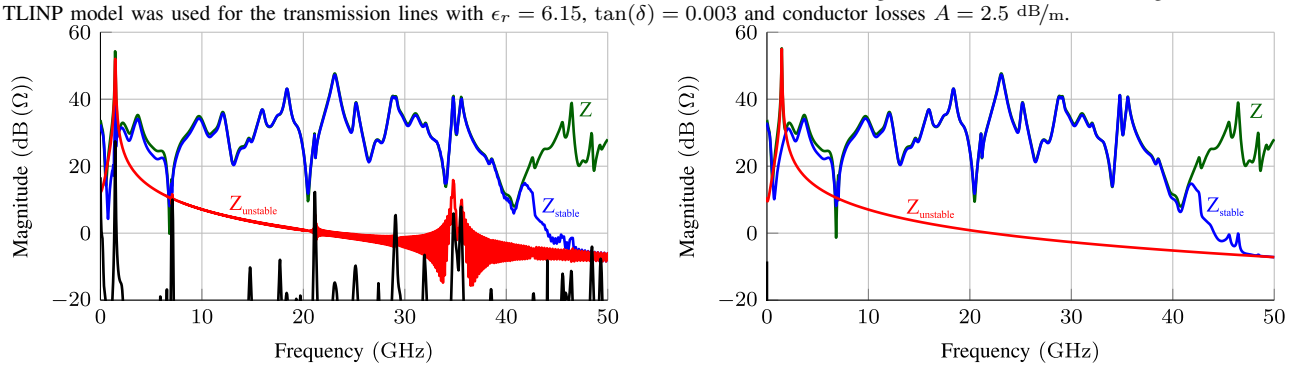


Fig. 6. The separation of the impedance (–) into the stable part (–) and unstable part (–) reveals the instability around 1.46 GHz. The interpolation error is shown with (–). In the plot on the left, there are artefacts present in the unstable part due to interpolation of the coarsely obtained impedance data. When the impedance of the balanced amplifier is simulated on a finer frequency grid, the artefacts disappear (right).

the denser frequency response data (Fig. 6) still predicts the instability around 1.46 GHz, but the artefacts in the unstable part at higher frequencies are gone. The maximum of the interpolation error went down to -80 dB(Ω).

C. Example 3: Two-stage Power Amplifier

As a third example, we consider the small-signal stability analysis of an X-band PA designed in the 0.25μm GaN HEMT technology GH25-10 of UMS [27]. The circuit and its design are described in great detail in [18]. The resulting MMIC is shown in Fig. 7.

The PA is a two-stage design where the second stage consists of two branches with each two transistors in parallel. In simulation, the second stage of the PA demonstrated an odd-mode instability [18], so a stabilisation resistor was added between the drains of the top and bottom halves of the second stage of the PA (as indicated in the Fig.).

The simulation of the complete PA was performed in ADS. The passive structures in the circuit were simulated with EM simulations in Momentum and combined with the non-linear transistor models afterwards. To verify the stability of the amplifier, the circuit impedance was determined at the gate of the top most transistor of the second stage.

The obtained impedance for $R_{\mathrm{stab}} = 500\Omega$ is shown in Fig. 8. The impedance is simulated on 945 logarithmically spaced points between 1MHz and 50GHz. Because 1MHz is not sufficiently close to DC, a bandpass filter was used in the stability analysis. Due to the low amount of data points in the resonances of the circuit, a Padé interpolation was used in the stability analysis. The resulting stable and unstable parts are shown in blue and red on the same figure. It is clear that the circuit is unstable for $R_{\mathrm{stab}} = 500\Omega$. The unstable part peaks around 9.5GHz and lies about 40 dB above the interpolation error level.

The odd-mode instability can be resolved by decreasing the resistance of R_{stab} [18]. In a second stability analysis, we determined the stability of the PA for $R_{\mathrm{stab}} = 25\Omega$. The results of this second analysis are shown in Fig. 9. The obtained unstable part coincides with the level of the interpolation error, which indicates that the circuit is now stable.

D. Example 4: R-L-diode circuit

The final example in this paper shows that the stability analysis can also be used to determine the stability of HB simulations of the R-L-diode circuit shown in Fig. 10. The circuit is based on [28], but a realistic diode model was used to represent the diode in the circuit instead of the three equations provided in the original paper.

The circuit is excited by a single-tone voltage source with an amplitude V_{in} and a frequency of 100 kHz. Because the diode has a transit-time of 4 μs, the circuit generates period-doubling solutions starting from sufficiently high amplitudes V_{in}. For even higher V_{in}, the circuit will create chaotic solutions.

To visualise this behaviour, a bifurcation diagram is constructed using time-domain simulations in the same way as is described in [28]: For every value of V_{in}, 1030 periods of 100 kHz are simulated and the final 30 periods are sampled every $^1/_{100\ \mathrm{kHz}}$. If the circuit solution is periodic with the same period as the input source, all 30 sampled points will fall on

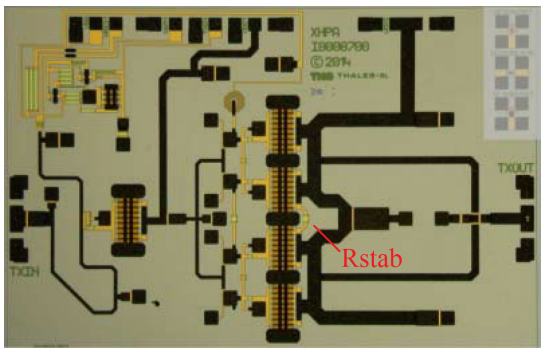


Fig. 7. Microphotograph of the MMIC. The stabilisation resistor R_{stab} is indicated in red.

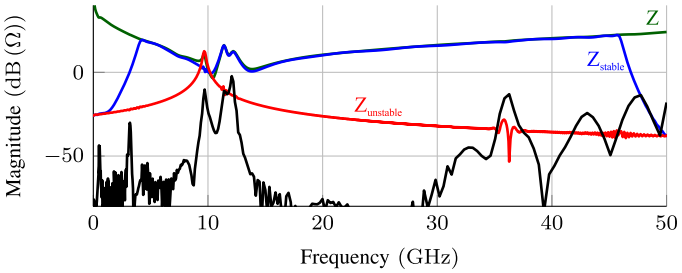


Fig. 8. Impedance seen at the gate of the first stage of the PA for $R_{\text{stab}} = 500\Omega$. Its obtained stable and unstable parts clearly indicate that the DC solution of the amplifier is unstable. The interpolation error (–) is quite high due to the low amount of simulation points.

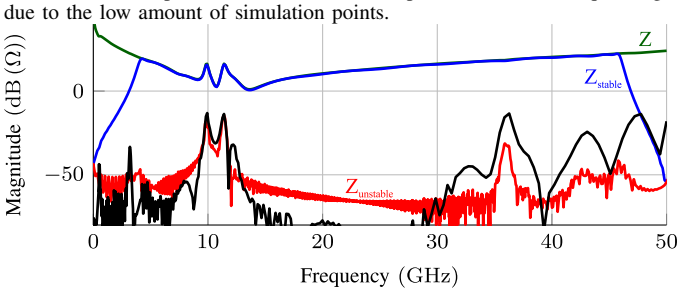


Fig. 9. Impedance presented by the PA at the gate of the transistor in the first stage for $R_{\text{stab}} = 25\Omega$. The obtained stable and unstable parts indicate that the DC solution is now stable. The interpolation error is shown with (–).

top of each-other. If a period-doubling occurs in the circuit, two different values will be obtained.

The obtained bifurcation diagram for our R-L-diode example is shown in Fig. 11. It is clear that a period-doubling occurs for V_{in} higher than 0.8 V. Starting from 1.8 V the period quadruples. For the highest input amplitudes, a chaotic solution is obtained.

If this R-L-diode circuit is simulated with HB, the circuit solution is constrained to harmonics of 100 kHz. For input amplitudes higher than 0.8 V, where the circuit wants to go to a period-doubling solution, the constrained HB solution will be locally unstable.

We run two HB simulations on this circuit. Both HB simulations have a base frequency of 100 kHz and an order of 10. In the first simulation, V_{in} is set to 0.5 V, which will result in a stable orbit. The second simulation has a V_{in} of 1.5 V, which will cause the orbit to be unstable.

The frequency response of the circuit around the HB solution is obtained with a mixer-like simulation, as explained in the introduction of this paper. The small-signal excitation was swept in both cases on a linear frequency grid starting from $(1\text{ kHz} + 1\text{ Hz})$ up to $(2\text{ MHz} + 1\text{ Hz})$ in 1 kHz steps. The 1 Hz was added to the start and stop values of the sweep to avoid overlap with the tones of the HB simulation.

The mixer-like simulation in ADS uses Single-Sideband (SSB) current excitations $i(t) = e^{j\omega t}$, which causes the obtained frequency responses $Z'^{[b]}_{mn}(j\omega)$ with $b \neq 0$ to be non-Hermitian:

$$Z'^{[b]}_{mn}(j\omega) \neq \overline{Z'^{[b]}_{mn}(-j\omega)}$$

An alternative representation can make $Z'^{[b]}_{mn}(j\omega)$ Hermitian by transferring to a sine and cosine basis from the exponential basis [29], [30]

$$Z^{[b]}_{mn}(j\omega) = \frac{1}{2}\left[Z'^{[b]}_{mn}(j\omega) + Z'^{[-b]}_{mn}(j\omega)\right]$$
$$Z^{[-b]}_{mn}(j\omega) = \frac{j}{2}\left[Z'^{[b]}_{mn}(j\omega) - Z'^{[-b]}_{mn}(j\omega)\right]$$

$Z^{[-1]}_{mn}(j\omega)$ $Z^{[0]}_{mn}(j\omega)$ and $Z^{[+1]}_{mn}(j\omega)$ are then analysed with the stable/unstable projection method. The results are shown in Fig. 12. The HB solution obtained for $V_{\text{in}} = 0.5$ V is clearly stable: its unstable part is more than 70 dB smaller than its stable part.

In the case for $V_{\text{in}} = 1.5$ V, the solution is clearly unstable as the unstable part lies far above the stable part of the frequency response. Note that the lowest-frequency peak in the unstable part is located around 50 kHz and that copies of the resonance are found at 150 kHz, 250 kHz,... This behaviour is to be expected and indicates that the circuit wants to go to a period-doubling solution.

During the stability analysis of a periodic orbit, the unstable part will contain both the unstable base pole and all its higher-order copies. The unstable part will be simple, just like in the small-signal case, and it will be possible to approximate it by a finite set of base poles. Due to the infinite amount of higher-order copies however, it will not be possible to approximate it by a low-order rational approximation as is the case in the stability analysis of a DC solution.

IV. Conclusion

This paper introduces a closed-loop local stability analysis without using a rational approximation. Instead, the impedance functions are split into a stable and unstable part by projecting onto an orthogonal basis. Transforming the problem to the unit disc allows to calculate this projection with the FFT which makes the projection-based stability analysis very fast. In a small-signal stability analysis, once the unstable part is obtained, a low-order rational model can be used to find the unstable poles in the circuit.

Due to the model-free nature of the proposed method, it is a very simple method to use: no choice of model order or approximation error needs to be made. The only requirements of the projection-based stability analysis are that the frequency responses are sampled on a sufficiently dense frequency grid and that the maximum frequency of the simulations is large enough. When the circuit impedance is simulated on a too

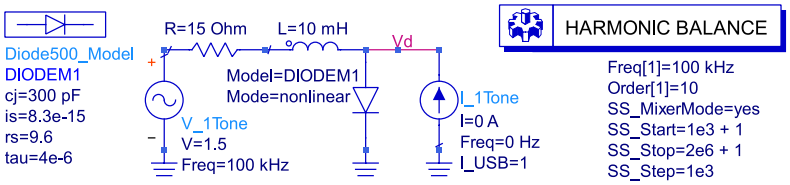


Fig. 10. The circuit R-L-diode circuit is excited with a small-signal current source at the diode.

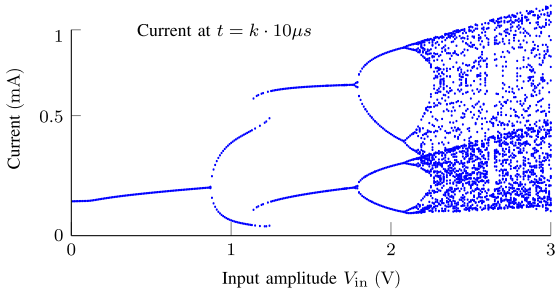


Fig. 11. The bifurcation diagram of the R-L-diode circuit shows that a period-doubling occurs for input amplitudes higher than 0.8 V.

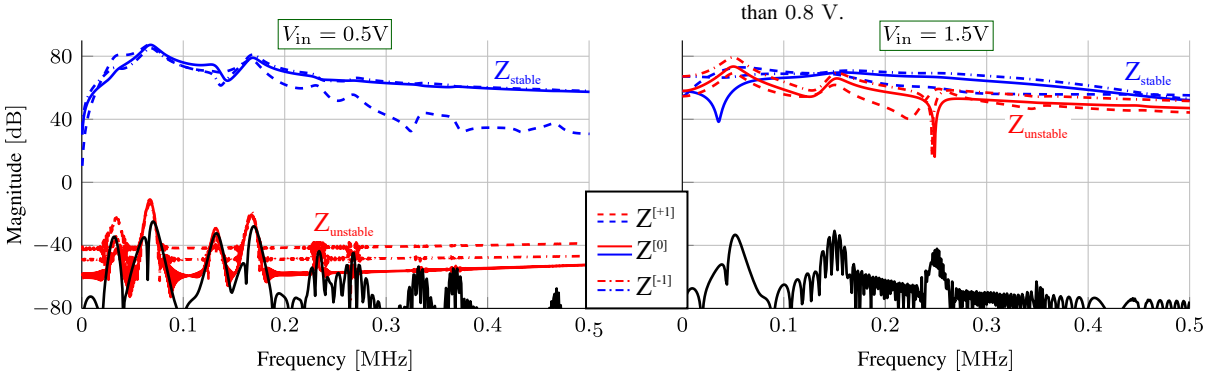


Fig. 12. The results of the stability analysis of $Z_{mn}^{[-1]}(j\omega)$ $Z_{mn}^{[0]}(j\omega)$ and $Z_{mn}^{[+1]}(j\omega)$ in the two HB simulations show that, for $V_{\text{in}} = 0.5$ V, the orbit is stable, but for $V_{\text{in}} = 1.5$ V, the orbit is unstable. The interpolation error in shown with (–).

coarse frequency grid, a large interpolation error is introduced in the results. The level of this interpolation error can easily be determined and used to improve the accuracy of the method.

Once the stable and unstable parts of the impedance are obtained with a sufficiently low interpolation error, the obtained unstable part can be compared to the interpolation error to determine whether it is significant or not. From experience, we found that, when the unstable part lies more than 20 dB above the interpolation error level, the circuit can be considered unstable. Further work is to be done towards automated decision making regarding stability.

The stable/unstable projection has been successfully applied to both the stability analysis of DC and large-signal solutions of RF circuits.

Appendix A
On Local Rational Approximation

Due to the presence of distributed elements in microwave circuits it is impossible to obtain a rational model of the impedances of the circuit over the full simulated frequency range. In a small frequency band however, it is possible to obtain a good low-order rational approximation of the impedance. This is the basis of pole-zero based stability analysis.

It is however not guaranteed that the poles of the local model correspond to the poles of the global model. We will demonstrate this in this appendix using an simple example without distributed elements.

Consider the artificial example in equation (12). All poles of this rational function lie in $s = -5$, so $Z(s)$ is stable. In the frequency band $[-1, 1]$ however, the FRF of this stable rational function closely resembles an unstable FRF (Figure 13)

To verify whether the global poles are obtained with a local model, we estimate a local rational model on 1000 frequency points on the interval $\omega \in [-0.6, 0.6]$ using vector fitting [31]. The model order for the local model is not known in advance, so it is swept from 2 to 11. The maximum of the phase error of the obtained fit is shown in Figure 14. Starting from model order 8, the phase error lies below 10^{-3} degrees so the fit can be considered sufficiently good for stability analysis [17].

The obtained local model is unstable however. In fact, an unstable local model is obtained for all model, which shows that the poles of a local model do not necessarily correspond to the poles of the underlying impedance. This is an artificial example of course, but the example indicates that the use of local lower-order models to determine the stability could lead to misleading results.

Appendix B
Mapping of the Basis Functions onto the Unit Disc

Applying transform (9) to the basis functions of the complex plane (5) yields the following:

$$B_k^{\text{disc}}(z) = \sqrt{\pi\alpha}\frac{2}{z-1}B_k\left(\alpha\frac{1+z}{1-z}\right)$$
$$= -\sqrt{\pi\alpha}\frac{2}{z-1}\sqrt{\frac{\alpha}{\pi}}\frac{\left(\alpha\frac{1+z}{1-z}-\alpha\right)^k}{\left(\alpha\frac{1+z}{1-z}+\alpha\right)^{k+1}} = z^k$$

$$Z(s) = \frac{-228.5s^{14} - 153.7s^{13} - 875.2s^{12} - 550.2s^{11} - 1364.9s^{10} - 789.9s^{9} \\ -1118.6s^{8} - 584.2s^{7} - 520.9s^{6} - 239.2s^{5} - 140.7s^{4} - 54.7s^{3} - 21.4s^{2} - 5.9s - 1}{(s+5)^{15}/_{9.9281 \cdot 10^{10}}} \tag{12}$$

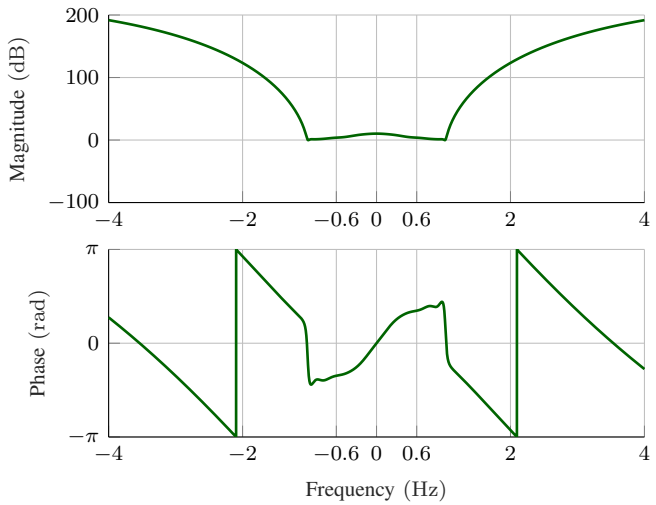


Fig. 13. FRF used in Appendix A. The FRF is stable, but resembles an unstable FRF in the interval [-1,1]. Using a local model of this FRF might result in false positives during stability analysis.

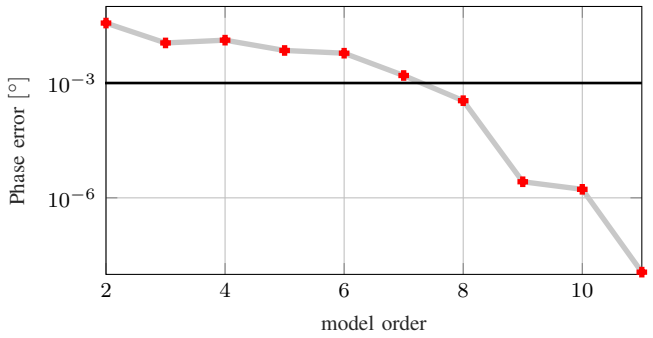


Fig. 14. Maximum phase error as a function of the model order obtained when estimating a local model in the range $\omega \in [-0.6, 0.6]$ of the stable impedance (12).

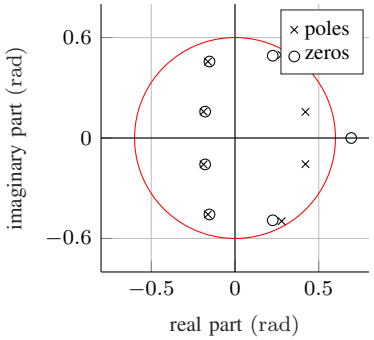


Fig. 15. The obtained poles and zeroes for model order 8. The model was estimated using data in the interval $\omega \in [-0.6, 0.6]$. The red circle indicates all points s in the complex plane for which $|s| < 0.6$.

ACKNOWLEDGEMENT

This research was partly supported by the French space agency CNES, partly by the Flemish Agency for Innovation by Science and Technology (IWT-Vlaanderen) and partly by the Strategic Research Program of the VUB (SRP-19). We are also thankful to Juan-Marie Collantes (UPV) for fruitful discussions on the topic of closed loop stability analysis. We would like to thank Kurt Homan, Johan Nguyen and Dries Peumans for the design and measurement of the balanced amplifier. Finally, we would like to thank Marc van Heijningen for providing the data of the MMIC PA.

REFERENCES

[1] A. Suarez, "Check the stability: Stability analysis methods for microwave circuits," *Microwave Magazine, IEEE*, vol. 16, no. 5, pp. 69–90, June 2015.
[2] A. Suarez and R. Quere, *Stability analysis of nonlinear microwave circuits*. Artech House, 2002.
[3] J. Jugo, J. Portilla, A. Anakabe, A. Suarez, and J. Collantes, "Closed-loop stability analysis of microwave amplifiers," *Electronics Letters*, vol. 37, no. 4, pp. 226–228, Feb 2001.
[4] J. Collantes, I. Lizarraga, A. Anakabe, and J. Jugo, "Stability verification of microwave circuits through floquet multiplier analysis," in *Circuits and Systems, 2004. Proceedings. The 2004 IEEE Asia-Pacific Conference on*, vol. 2, Dec 2004, pp. 997–1000 vol.2.
[5] J. Partington, *Linear operators and linear systems*, ser. Student texts. London Math. Soc., 2004, no. 60.
[6] L. Baratchart, S. Chevillard, and F. Seyfert. (2014) On transfer functions realizable with active electronic components. hal-01098616. Inria Sophia Antipolis. [Research Report] RR-8659.
[7] W. Rudin, *Real and Complex analysis*. Mc Graw-Hill, 1982.
[8] G. A. Baker and P. Graves-Morris, *Padé Approximants*, ser. Encyclopedia of Mathematics and its Applications. Cambridge University Press, 1996, vol. 59.
[9] A. A. Gonchar and E. Saff, Eds., *Spurious Poles in Diagonal Rational Approximation*, ser. Series in Computational Mathematics, vol. 19. New York: Springer, 1992, in Progress in Approximation Theory.
[10] H. Stahl, "Spurious poles in padé approximation," *Journal of Computational and Applied Mathematics*, no. 99, 1998.
[11] S. G. P. Gonnet and L. N. Trefethen, "Robust padé approximation via svd," *SIAM Rev.*, vol. 55, pp. 101–117, 2013.
[12] B. Beckermann and A. Matos, "Algebraic properties of robust padé approximants," *Jour. Approx. Theory*, vol. 190, pp. 91–115, 2015.
[13] A. Lefteriu and A. C. Antoulas, "On the convergence of the vector fitting algorithm," *IEEE Transactions on Microwave Theory and Techniques*, vol. 61, no. 4, 2013.
[14] L. Ljung, *System identification: Theory for the user*. Prentice–Hall, 1987.
[15] A. Anakabe, N. Aylloon, J. Collantes, A. Mallet, G. Soubercaze-Pun, and K. Narendra, "Automatic pole-zero identification for multivariable large-signal stability analysis of rf and microwave circuits," in *Microwave Conference (EuMC), 2010 European*, Sept 2010, pp. 477–480.
[16] AMCAD. engineering, *STAN tool - Stability Analysis*.
[17] I. Mallet, A. Ankabe, G. Soubercaze-Pun, and J.-M. Collantes, "Automation of the zero-pole identification methods for the stab analysis of microwave active circuits," U.S. Patent 8,407,637 B2, 2013.
[18] M. van Heijningen, A. P. de Hek, F. E. van Vliet, and S. Dellier, "Stability analysis and demonstration of an x-band gan power amplifier mmic," in *2016 11th European Microwave Integrated Circuits Conference (EuMIC)*, Oct 2016, pp. 221–224.
[19] N. Otegi, A. Anakabe, J. Pelaz, J. Collantes, and G. Soubercaze-Pun, "Experimental characterization of stability margins in microwave amplifiers," *Microwave Theory and Techniques, IEEE Transactions on*, vol. 60, no. 12, pp. 4145–4156, Dec 2012.
[20] J. Collantes, N. Otegi, A. Anakabe, N. Aylloon, A. Mallet, and G. Soubercaze-Pun, "Monte-carlo stability analysis of microwave amplifiers," in *Wireless and Microwave Technology Conference (WAMICON), 2011 IEEE 12th Annual*, April 2011, pp. 1–6.
[21] N. Aylloon, J. M. Collantes, A. Anakabe, I. Lizarraga, G. Soubercaze-Pun, and S. Forestier, "Systematic approach to the stabilization of multitransistor circuits," *IEEE Tran. on Microwave Theory and Techniques*, vol. 59, no. 8, pp. 2073–2082, 2011.

[22] M. Olivi, F. Seyfert, and J.-P. Marmorat, "Identification of microwave filters by analytic and rational h2 approximation," *Automatica*, vol. 10, Februari 2013.
[23] K. Hoffman, *Banach spaces of analytic functions*, ser. Prentice-Hall series in modern analysis. Prentice-Hall, 1962.
[24] J. P. Marmorat and M. Olivi, "RARL2: a Matlab based software for H^2 rational approximation," http://www-sop.inria.fr/apics/RARL2/rarl2.html, 2004.
[25] S. Kung, "A new identification and model reduction algorithm via singular value decomposition," in *Proceedings of the 12th Asilomar Conference on Circuits, Systems and Computers*, 1978, pp. 705–714.
[26] I. Markovsky, *Low Rank Approximation: Algorithms, Implementation, Applications*. Springer, 2012. [Online]. Available: http://homepages.vub.ac.be/~imarkovs/book.html
[27] D. Floriot, H. Blanck, D. Bouw, F. Bourgeois, M. Camiade, L. Favede, M. Hosch, H. Jung, B. Lambert, A. Nguyen, K. Riepe, J. Splettstosse, H. Stieglauer, J. Thorpe, and U. Meiners, "New qualified industrial algan/gan hemt process: Power performances amp; reliability figures of merit," in *2012 7th European Microwave Integrated Circuit Conference*, Oct 2012, pp. 317–320.
[28] A. Azzouz, R. Duhr, and M. Hasler, "Transition to chaos in a simple nonlinear circuit driven by a sinusoidal voltage source," *IEEE Transactions on Circuits and Systems*, vol. 30, no. 12, pp. 913–914, Dec 1983.
[29] E. Louarroudi, "Frequency domain measurement and identification of weakly nonlinear time-periodic systems," Ph.D. dissertation, Vrije Universiteit Brussel (VUB), 2014.
[30] H. Sandberg, E. Mollerstedt, and Bernhardsson, "Frequency-domain analysis of linear time-periodic systems," *Automatic Control, IEEE Transactions on*, vol. 50, no. 12, pp. 1971 – 1983, dec. 2005.
[31] B. Gustavsen and A. Semlyen, "Rational approximation of frequency domain responses by vector fitting," *Power Delivery, IEEE Transactions on*, vol. 14, no. 3, pp. 1052–1061, Jul 1999.

Martine Olivi was born in France, in 1958. She got the engineer's degree from Ecole des Mines de St-Etienne, France, and the PhD degree in Mathematics from Université de Provence, Marseille, France, in 1983 and 1987 respectively. Since 1988, she is with the Institut National de Recherche en Informatique et Automatique (INRIA), Sophia Antipolis, France. Her research interests include: rational approximation, parametrization of linear multivariable systems, Schur analysis, identification and design of resonant systems. Detailed information and publications are available at www-sop.inria.fr/members/Martine.Olivi

Adam Cooman was born in Belgium in 1989. He graduated as an Electrical Engineer in Electronics and Information Processing in 2012 at Vrije Universiteit Brussel (VUB) and obtained his Ph.D at the Department ELEC of the VUB in December 2016. Now, Adam is part of the APICS team at INRIA, Sophia Antipolis, France. His main interests are the design of Electronic circuits, from low frequencies up to the microwave frequencies.

Sylvain Chevillard was born in Paris, France, in 1983. He received his Ph.D. in computer science from Université de Lyon - École Normale Supérieure de Lyon in 2009. He is Chargé de recherche (junior researcher) at Inria Sophia Antipolis, France. Some of his research interests are reliable computing, approximation theory, computer algebra, inverse problems.

Fabien Seyfert graduated from the "Ecole superieure des Mines" (Engineering School) in St Etienne (France) in 1993 and received his Ph.D in mathematics in 1998. From 1998 to 2001 he was with Siemens (Munich, Germany) as a researcher specialized in discrete and continuous optimization methods. Since 2002 he occupies a full research position at INRIA (French agency for computer science and control, Nice, France). His research interest focuses on the conception of effective mathematical procedures and associated software for problems from signal processing including computer aided techniques for the design and tuning of microwave devices.

Laurent Baratchart received his Docteur Ingénieur degree from Ecole des Mines de Paris in 1982 (advisor: Y. Rouchaleau) and his Thèse d'état in Mathematics from the University of Nice in 1987 (advisor: A. Galligo). He was the head of INRIA's project team MIAOU (Mathematics and Informatics in Automatic control and Optimization for the User) from 1988 to 2003 and is currently the head of the project team APICS (Analysis of Problems of Inverse type in Control and Signal processing) at INRIA Sophia-Antipolis since 2004. His main interests lie with Complex and Harmonic Analysis, Inverse Problems, as well as System and Circuit Theory.

5.2 Multi-band frequency design

Following paper is reproduced in this section:

V. Lunot, F. Seyfert, S. Bila, and A. Nasser. "Certified computation of optimal multiband filtering functions". In: *IEEE Transactions on Microwave Theory and Techniques* 56.1 (2008), pp. 105–112

Certified Computation of Optimal Multiband Filtering Functions

Vincent Lunot, Fabien Seyfert, Stéphane Bila, and Abdallah Nasser

***Abstract*—In this paper, we focus on the problem of computing multiband filtering characteristics with a guarantee on their global optimality with respect to a Zolotarev-like criterion. An iterative algorithm based on linear programming is presented. This algorithm ensures quadratic convergence to the optimal solution. We also provide an equiripple-like criterion that allows one to check in a very simple manner whether a computed filtering function is optimal or not. The latter is used to analyze two practical design examples based on asymmetric dual-band specifications. In particular, it is shown that the selectivity of the dual-band response does not necessarily increase with the filter's order. This study yields some striking results when compared to the usual single-band situation, and introduces the idea that for certain asymmetric specifications some of the filter order values are more suitable than others. Finally, the practical implementations of the filtering devices in inline dual-mode cavities and stacked single-mode cavities are detailed.**

***Index Terms*—Differential correction algorithm, filter synthesis, multiband filter, Zolotarev problem.**

I. Introduction

In BOTH space and terrestrial communications, advanced filtering characteristics have become a major way of improving and simplifying the architecture of systems. In particular, a dual-band filtering characteristic allows one to incorporate the two passbands within the single filter structure. The latter is an interesting replacement for the doubly multiplexed solution that combines single-band filters with a junction at the input/output. The main advantage is to save mass and volume, but also to simplify both the manufacturing and the tuning of the hardware since the multiband filter architecture can be realized with topologies and technologies commonly used for single-band filter design.

Some recent studies [1], [2] exposed methods using frequency transformations to design multiband filters. However, these lack generality. Indeed, the response is limited by symmetric specifications or by the position of the transmission zeros.

For general specifications, some optimization methods are known [3], [4]. However, they do not guarantee the optimality of the response.

In [5], an iterative algorithm based on linear programming was proposed in order to compute the "best" filtering function in the Zolotarev sense. Unlike general optimization methods that might end up at a local, but nonglobal optimum, the latter method guarantees the global optimality of the response. A drawback, however, is its poor rate of convergence, yielding a high number of necessary iterations to obtain a reasonable convergence. In this paper, modifications are introduced in the main computation loop of the algorithm that ensure quadratic convergence to the optimal solution. We also give a simple characterization of the optimal filtering function in terms of an alternation property, whereas in the single-band case, this characterization reduces to the classical equiripple property this is no longer true for the multiband situation: we give an example of an optimal, but nonequiripple response.

Two examples are constructed of the design of dual-band filters when beginning from asymmetrical frequency specifications. The ability to certify the optimality of the computed responses appears to be crucial and leads to some striking results when compared to the usual single-band situation. In the latter, the selectivity of the response increases with the filter's order. This is no longer true for dual-band specifications for which specific degrees appear to be more adapted than others. A trivial example of this is given by symmetric dual-band specifications: in this case, one can show that the optimal filtering function is always of even order. Adding an extra reflection zero to an even-order characteristic will only deteriorate the response. This type of phenomenon also occurs in a less predictable manner when dealing with asymmetric frequency specifications. To illustrate this, we analyze frequency specifications for which the best characteristic with, at most, ten reflection zeros and three transmission zeros is shown to be of 9–3 type.

Finally, we detail the practical implementation of the filtering devices respectively in aligned dual-mode cavities and stacked single-mode cavities.

II. Statement of the Synthesis Problem

A. Polynomial Structure of the Scattering Matrix

The scattering matrix associated with the classical low-pass lossless circuit prototype [6] has the following structure:

$$S = \frac{1}{E}\begin{bmatrix} F & P \\ P & (-1)^n F^* \end{bmatrix} \tag{1}$$

where n is the number of resonators. The polynomial P is of degree $m < n-1$ and satisfies the condition $P = (-1)^{n+1}P^*$ (which implies that the set of transmission zeros is symmetric with respect to the imaginary axe, i.e., paraconjugated). F is

Manuscript received April 15, 2007.

V. Lunot and F. Seyfert are with the Institut National de Recherche en Informatique et en Automatique (INRIA), 06902 Sophia Antipolis, France (e-mail: Vincent.Lunot@sophia.inria.fr; fseyfert@sophia.inria.fr).

S. Bila and A. Nasser are with XLIM, University of Limoges, 87060 Limoges, France (e-mail: bila@ircom.unilim.fr; abdallah.nasser@gmail.com).

Digital Object Identifier 10.1109/TMTT.2007.912234

0018-9480/$25.00 © 2007 IEEE

of degree n and monic and the denominator E is the unique Hurwitz polynomial satisfying the following spectral equation:

$$EE^* = FF^* + (-1)^{n+1}P^2. \tag{2}$$

Using the latter equations, the squared modulus of the transmission parameter is expressed as

$$|S_{21}(j\omega)|^2 = \frac{1}{1 + \left|\dfrac{F(j\omega)}{P(j\omega)}\right|^2} \tag{3}$$

where $D = F/P$ is known as the filtering or characteristic function.

In the case of a single passband and for given transmission zeros (i.e., P is fixed), the classical formula using the $\mathrm{Arccosh}$ function [6] allows the computation of a polynomial F that yields an equiripple quasi-elliptic filtering characteristic. 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 multiband situation, explicit formulas no longer exist for D, 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.

B. General Zolotarev Problem

Given a set of passbands (see Fig. 1), i.e., a collection of intervals on the real axes $I_{1,\ldots,}I_r$ (which union we call I), as well as a collection of stopbands $J_{1,\ldots,}J_p$ (which union we call J), the "best" multiband response is such that the modulus of the transmission is as big as possible on the I intervals (passbands) and as small as possible on the J intervals (stopbands). The latter translates in a straightforward manner to the following normalized optimization problem specifying what the best filtering function is:

$$T_m^n = \left\{(F,P) \in H^n \times H^m, \left\|\frac{F}{P}\right\|_I \leq 1\right\} \tag{4}$$

$$\text{solve}: \max_{(F,P)\in T_m^n} \min_{\omega \in J} \left|\frac{F(\omega)}{P(\omega)}\right| \tag{5}$$

where H^k is the set of polynomials of degree less than k and $\|.\|_I$ is the sup norm over the set I (see Fig. 1). Indeed, if F'/P' is an optimal solution of (5) with F' monic, one may verify that setting $F(s) = F'(s/j)$ and $P(s) = j^{(n+1)}P'(s/j)/\varepsilon$ yields a scattering matrix with structure (1), with the lowest possible transmission on all the stopbands J_i, provided $|S_{21}|^2 \geq 1/(1+\varepsilon^2)$ on the passbands I_i.

C. Real Zolotarev Problem

In this study, we consider solving (5) under the additional condition that F and P are polynomials with real coefficients. This in particular implies that the synthesized scattering matrix satisfies $S_{11} = S_{22}$ and that the reflection zeros are paraconjugated, 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 it is less restricted).

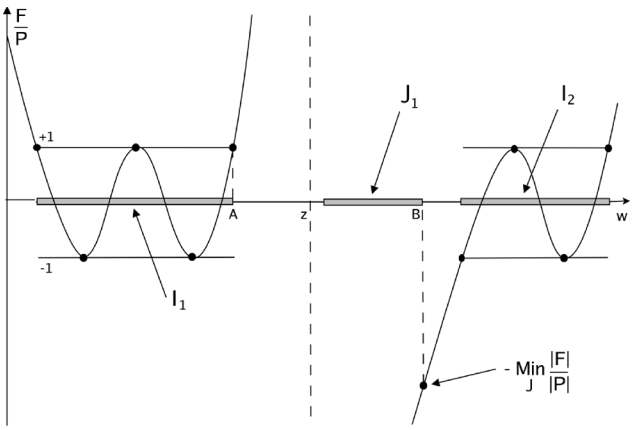


Fig. 1. Graph of function F/P with one transmission zero in z, two passbands I_1 and I_2, and one stopband J_1.

III. 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 stopband $J = [1.1, 2]$ and one passband $I = [-1, 1]$. We are interested in the all-pole filter of order 2 that solves the related Zolotarev problem, i.e., among all polynomials of degree less than 2 that are bounded by 1 on I, find the one with the fastest growth on J. The solution to this problem is, of course, known to be the Chebyshev polynomial $F(x) = \cos(2 \cdot \mathrm{Arcos}(x)) = 2x^2 - 1$. We will now see that this result can be recovered from a numerical algorithm in a guaranteed manner. The advantage of this procedure is that it will extend to multiband situations for which closed-form formulas are not known. Once a sign has been chosen for the polynomial $F = ax^2 + bx + c$ on J (say, positive), the original Zolotarev problem can be formulated as the following optimization problem:

$$\text{solve}: \max h$$
$$\text{subject to} \begin{cases} \forall x \in J, & h \leq ax^2 + bx + c & \text{(i)} \\ \forall x \in I, & 1 \geq ax^2 + bx + c & \text{(ii)} \\ \forall x \in I, & -1 \leq ax^2 + bx + c & \text{(iii)}. \end{cases}$$

Here, h is an auxiliary variable, which expresses the minimum of the polynomial over J. Evaluating inequalities (i) at sample points in the interval J and inequalities (ii) and (iii) at sample points in the interval I yields a set of linear inequalities in the variables (a, b, c, h). In this manner, the original Zolotarev problem is cast into a linear optimization problem under linear constraints: a linear program (LP). These types of problems have been widely studied and efficient software to solve them in a guaranteed manner exist (e.g., Cplex, MATLAB, Maple, LpSolve). Using the LP solver of MATLAB and taking 100 sample points over the intervals I and J yields the solution $a = 2.0002$, $b = 1e-12$, and $c = -1.0002$.

The advantage of this method as compared to closed-form formulas is that it generalizes to any number and any arrangement of the intervals I and J. We strongly encourage the reader to derive from what precedes a simple algorithm to solve the multiband synthesis problem in the special case of all-pole filters.

In the following sections, 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.

IV. Sign Combinations and Characterization of the Solution

A. Sign Combinations

Our goal is now to eliminate the absolute value in (5) to get a "linear" version of the problem. If F'/P' is an optimal solution of (5) and is irreducible ($gcd(F', P') = 1$) then, as the value of the $\max\min$ in (5) is positive, F' has no zero in J and, as the absolute value of F'/P' is bounded by 1 over I, P' has no zero in I. Therefore, F' has a constant sign on every interval J_j and P' has a constant sign on every interval I_i. Thus there exists a sign function σ (such that $\sigma(\omega) = \pm 1$) that is constant on every interval I_i and J_j such that F'/P' has a representative in the convex set

$$A_m^n = \left\{ \begin{array}{l} (F,P) \in H^n \times H^m, \forall \omega \in J : F(\omega)\sigma(\omega) \geq 0 \\ lc(P) = 1, \left\| \dfrac{F}{P} \right\|_I \leq 1, \forall \omega \in I : P(\omega)\sigma(\omega) \geq 0 \end{array} \right\} \tag{6}$$

where $lc(P) = 1$ signifies that the coefficient of x^m in P is equal to 1.

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 (5) by

$$\text{solve} : \max_{(F,P)\in A_m^n} \min_{\omega \in J} \frac{\sigma(\omega)F(\omega)}{|P(\omega)|}. \tag{7}$$

Solving (7) for all possible sign combinations and retaining the overall best solution yields an optimal solution of (5).

B. Characterization of the Solution

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 solution of (7). The latter is based on an alternation property.

Let λ be the value of the minimum of $|F/P|$ on J. We call J^+ (respectively, J^-) the union of intervals J_i such that $\sigma(J_i) = 1$ (respectively, $\sigma(J_i) = -1$). We define the following sets of "extreme" points:

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

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

In Fig. 1, ten "extreme" points are plotted.

A sequence of consecutive points $(\omega_1 < \omega_2 < \cdots < \omega_k)$ is called "alternant" if its points belong alternatively to the sets $E^+(F', P')$ and $E^-(F', P')$. In Fig. 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 to determine whether a function is the solution of (7) or not. The following indeed holds.

- A concave maximization problem, i.e., (7), admits a unique solution.
- F'/P' is an optimal solution of "full rank" 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 $E^+(F', P')$ and $E^-(F', P')$ with $N = m + n$.

The latter 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 passband and stopbands. However, in the multiband case, this is no longer true in general. Fig. 2(a) shows the optimal 6–4 function for the stopbands $[-2; -1.3]$, $[-0.6; 0]$, $[1.3; 2]$, and for the passbands $[-1; -0.8]$, $[0.6; 1]$. The attenuation level attained in the stopbands is of -32.2 dB, whereas the return loss is set to -20 dB. The 12 "extreme" points certify that this 6–4 nonequiripple function is the optimal solution with respect to the specifications. As pointed out by some reviewers, one might enlarge a bit the passbands and try to obtain an equiripple response with different return-loss levels in the passbands. This was done by solving the problem with following passbands $[-1; -0.75]$ and $[0.5; 1]$ and return loss levels of, respectively, -25 and -20 dB. As shown in Fig. 2(b), the optimal frequency response for these new specifications is equirriple. These new specifications are harder to meet than 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, 12 "extreme" points certify the optimality of the response.

In the following, we present a quadratic convergent algorithm for (7).

V. Algorithm

A. Geometry of the Sub-Problem

We will now study (7) from a geometric point of view. If we denote by M the value of the criterion $\min$ in (7) for a given (F, P) (M can be seen as the rejection level of F/P in the stopbands), then the convex set $E(M)$ defined by

$$E(M) = \{(F,P) \in A_m^n, \forall \omega \in J : \sigma(\omega)F(\omega) - M|P(\omega)| \geq 0\} \tag{8}$$

is, in a way, the set containing all the functions that have at least a rejection level M in the stopbands. Let M' be the value of the criterion $\max\min$ in (7) (M' is the best possible rejection). By definition of the $\max$, $E(M')$ is then the set of representatives of the optimal function F'/P'.

The key point for computing the solution of (7) is that, for $M_1 < M_2 < M' < M_3$, the following holds (see Fig. 3).

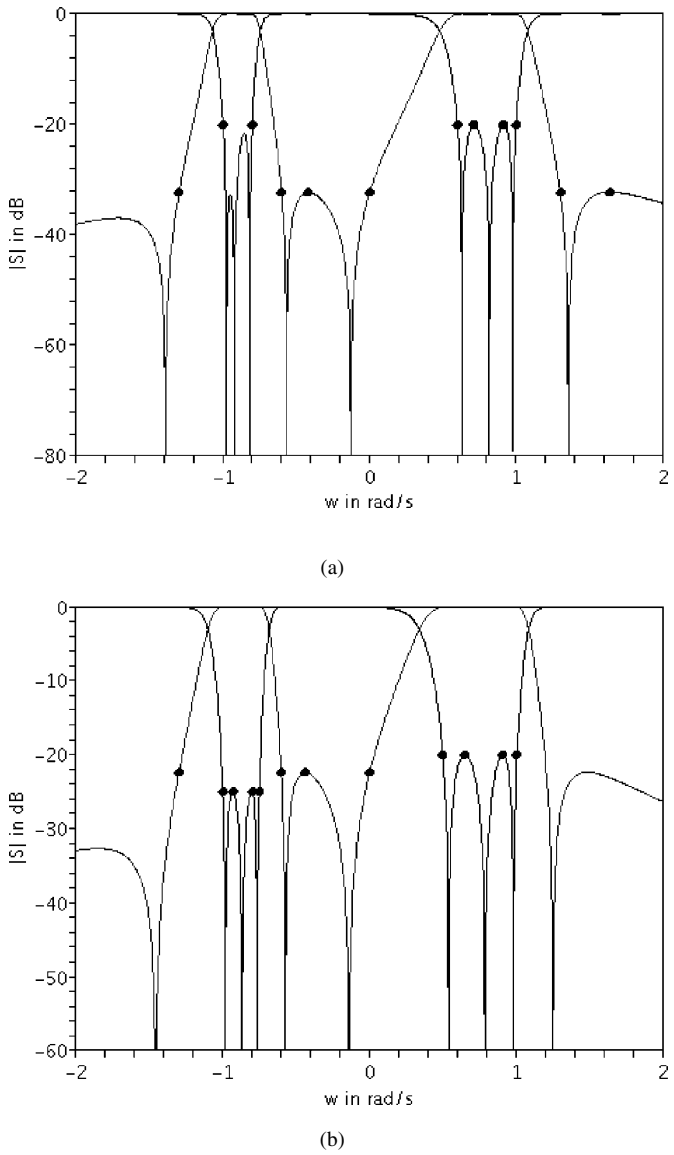


Fig. 2. (a) Optimal, but nonequiripple filtering function with six poles and four zeros. (b) Optimal 6–4 response with enlarged passbands and unequal return-loss levels in the passbands.

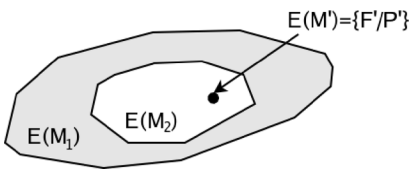


Fig. 3. Sets $E(M)$ for $M_1 < M_2 < M'$.

- $E(M_3) = \varnothing$.
- $E(M') \cong \{F'/P'\}$.
- $E(M') \subset E(M_2) \subset E(M_1)$.

Indeed, by making a hypothesis on the possible rejection level M and by checking the emptiness of $E(M)$, the following information on M' is known.

- If $E(M)$ is empty, $M' < M$.
- If $E(M)$ is nonempty, $M' \geq M$.

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 $E(M)$ allows to check nonemptiness using linear programming.

B. Detailed Equations for Checking Emptiness

One way of checking the emptiness of $E(M)$ is to find (F, P) in A_m^n, which maximizes the following function f_M:

$$f_M(F,P) = \min_{\omega \in J} \left(\sigma(\omega)F(\omega) - M\,|P(\omega)|\right). \tag{9}$$

Computation of (F, P) can be done by discretizing the I and J intervals. Indeed, in this way, the equations of the constraints in A_m^n become linear in the coefficients of F and P. Thus, the problem of finding (F, P) is done by solving the LP problem

$$\begin{aligned}&\text{solve}: \max h\\ &\text{subject to}\\ &\begin{cases}\sigma(y_j)F(y_j) - MP(y_j) \geq h, & \text{for all } y_j\\ \sigma(y_j)F(y_j) + MP(y_j) \geq h, & \text{for all} y_j\\ \sigma(x_j)P(x_j) \geq 0, & \text{for all } x_j\\ -\sigma(x_j)P(x_j) \leq F(x_j) \leq \sigma(x_j)P(x_j), & \text{for all } x_j\end{cases}\end{aligned} \tag{10}$$

where (x_j) [respectively, (y_j)] are a discretization of I (respectively, J).

If the maximum h is positive, then (F, P) in A_m^n, which maximizes f_M, has been computed, therefore, the set $E(M)$ is nonempty. Else, if $h \leq 0$, the set $E(M)$ is empty.

Accuracy depends, of course, on the number and placement of chosen points. On each interval, taking approximately 20 Chebyshev points gives satisfactory results.

C. Differential Correction-Like Algorithm

Instead of using dichotomy as previously suggested, we now come to an algorithm that adjusts M in a more efficient way by using the information gained from solving (10). The latter is an adaptation of the differential correction algorithm introduced by Cheney [7] for rational approximation.

Step 0: Initialization.

Choose polynomials (F_0, P_0) in A_m^n. Compute

$$M_0 = \min_{\omega \in J} \left| \frac{F_0}{P_0}(\omega) \right|. \tag{11}$$

Step k: Compute (F_k, P_k), which solves the LP problem (10) for $M = M_{k-1}$

$$f_{M_{k-1}}(F_k, P_k) = \max_{(F,P) \in A_m^n} f_{M_{k-1}}(F, P). \tag{12}$$

If $f_{M_{k-1}}(F_k, P_k) \leq 0$, return (F_{k-1}, P_{k-1}) else compute

$$M_k = \min_{\omega \in J} \left| \frac{F_k}{P_k}(\omega) \right|. \tag{13}$$

This iterative algorithm is proven to converge to the solution of (7) in the nondegenerated case (degree $P = m$). Instead of

taking the leading coefficient of P equal to 1, another normalization can be taken to ensure convergence in the general case.

For initialization, the choice of (F_0, P_0) in A_m^n does not influence the solution. One way of computing possible (F_0, P_0) is to solve (12) with M_{-1} properly chosen (e.g., $M_{-1} = 1$ generally works).

D. Improvements: Specifications and Quadratic Convergence

This algorithm can be easily generalized to compute the best filtering function subject to arbitrary specifications Ψ (Ψ is a positive function) in passbands and stopbands: in f_M, replace M by $M + \Psi(\omega)$ and in A_m^n, replace bound 1 over I by $\Psi(\omega)$.

Furthermore, the rate of convergence can be greatly increased by a slight modification of function f_M. Indeed, for a fast computation of the solution with respect to specifications Ψ, replace $f_{M_{k-1}}$ at step k by

$$f_{M_{k-1}}(F,P) = \min_{\omega \in J} \left(\frac{\sigma(\omega)F(\omega) - (\psi(\omega) + M_{k-1})\,|P(\omega)|}{|F_{k-1}(\omega)|} \right). \tag{14}$$

The associated LP problem is defined by

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

$$\text{subject to } \begin{cases} \sigma(y_j)F(y_j) - (\psi(y_j)+M)P(y_j) \geq h\,|F_{k-1}(y_j)|, \\ \quad \text{for all } y_j \\ \sigma(y_j)F(y_j) + (\psi(y_j)+M)P(y_j) \geq h\,|F_{k-1}(y_j)|, \\ \quad \text{for all } y_j \\ \sigma(x_j)P(x_j) \geq 0, \quad \text{for all } x_j, \\ -\sigma(x_j)\psi(x_j)P(x_j) \leq F(x_j) \leq \sigma(x_j)\psi(x_j)P(x_j), \\ \quad \text{for all } x_j. \end{cases}$$

An adaptation of the proof for rational approximation (see [8]) shows that the convergence is quadratic whenever the solution is of "full rank."

VI. Examples

We consider two design examples based on asymmetric frequency specifications. Here we discuss the computation of the filtering functions, whereas their practical implementations are considered in Section VI-A.

Example 1

A first example is taken from [5] with the following electrical specifications:

- return loss at 20 dB in the passbands ($I_1 = [-1, -0.625]$ and $I_2 = [0.25, 1]$ on the ω-axis, normalized frequency);
- 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. Using the previously presented algorithm and working for practical reasons with finite intervals (so the two chosen "outside" stopbands are set to $[-10, -1.188]$ and [1.212, 10]), we obtain the filtering function plotted in Fig. 4. Only nine transmission zeroes and 14 "extreme" points appear on the graph, which seems at first glance to contradict

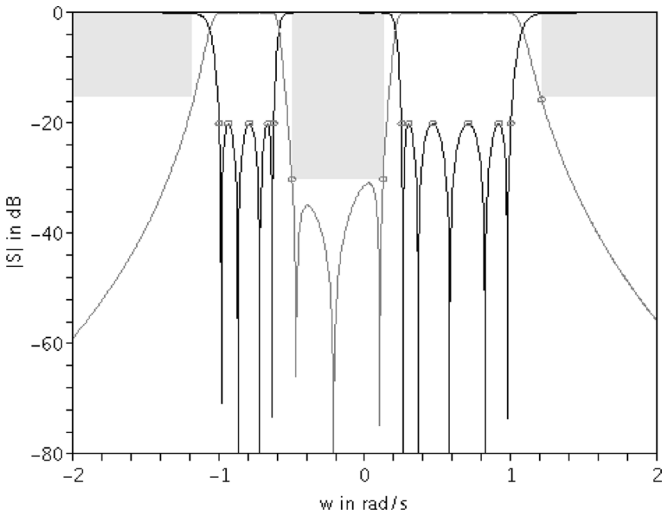


Fig. 4. Optimal transmission and reflection parameters (example 1).

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., in $\omega = -10$, together with a reflection zero that was rejected at $\omega = -100$. 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, ten reflection zeros (respectively, at most, three 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 certify the optimality of the computed filtering function is crucial. Someone using a generic optimizer may insist on 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.

Example 2

A second example from [9] is taken whose electrical specifications are defined by the following:

- return loss at 23 dB in the passbands ($I_1 = [-1, -0.383]$ and $I_2 = [0.383, 1]$ on the ω-axis);
- in the lower stopband ($J_1 =]-\infty, -1.864]$), rejection is set at 10 dB on $]-\infty, -1.987]$ and 15 dB on $[-1{,}987, -1{,}864]$. 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 [11]. However, the same phenomenon as in example 1 occurs, and the optimal solution appears to be of type 7–3. The slight difference between the filtering function in Fig. 5 and the one in [9] is due to the fact that reflection zeros were originally laboriously optimized "by hand." Here, transmission and reflection zeros are optimized simultaneously. In the upper stopband, the rejection level is lower than in [9], but is improved in the other stopbands and, contrary

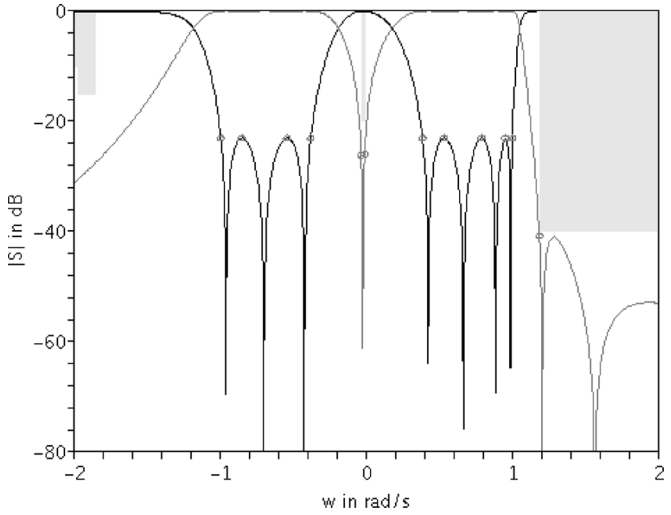


Fig. 5. Optimal transmission and reflection parameters (example 2).

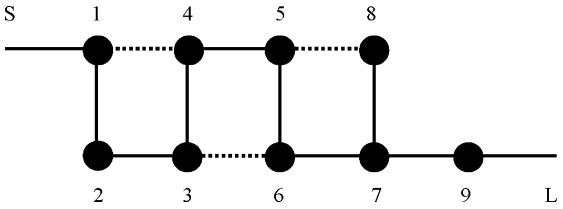


Fig. 6. Extended-box coupled resonator network for the realization of the ideal 9–3 dual-band response in Fig. 4.

to the original solution, the return loss is equiripple in the two passbands.

VII. Design and Implementation at Microwave Frequencies

Example 1: Nine-Pole Three-Zero Dual-Band Filter Implemented in Inline Dual-Mode Cavities

The low-pass specifications given in Fig. 4 correspond to the following passbands and stopbands at microwave frequencies: the two passbands are, respectively, $I_1 = [8.28, 8.31]$ 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. 6) 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 on their dual-mode TE_{111} and coupled by rectangular irises, as shown in Fig. 7.

Applying an exhaustive coupling matrix synthesis [11], 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 [10]. However, in this case, an exhaustive computation of all the solutions to the coupling matrix

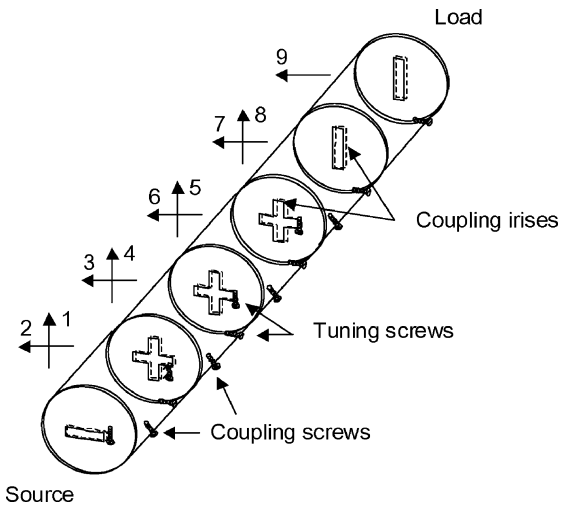


Fig. 7. Implementation of the nine-pole three-zero dual-band filter with inline dual-mode cylindrical cavities, network topology illustrated in Fig. 6.

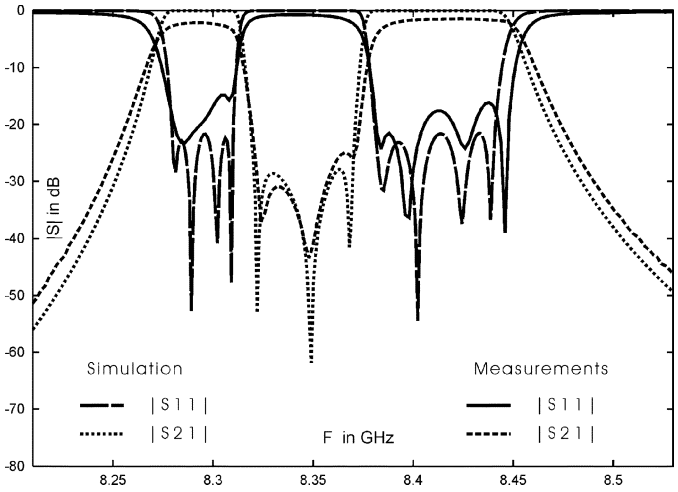


Fig. 8. Measurements and simulation of the nine-pole three-zero dual-band filter physically illustrated in Fig. 7.

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 coherency on the solution modifications (corresponding coupling values). The CAD model is a finite-element model. Metallic losses are not considered during CAD tuning for facilitating 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 approximately 4000, but can be improved using silver-plated cavities. However, measured and simulated results are in good agreement, as shown in Fig. 8. Insertion losses are 2.15 dB in the first passband and 1.45 dB in the second one.

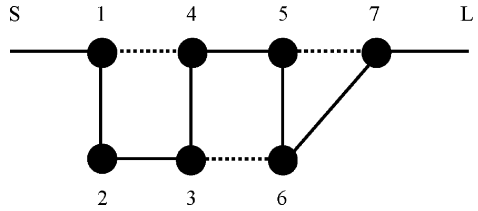


Fig. 9. Pseudo extended-box coupled resonator network for the realization of the ideal 7–3 dual-band response in Fig. 5.

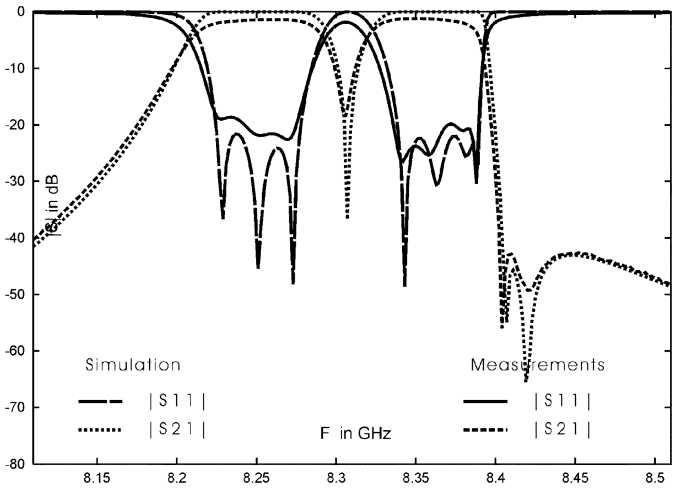


Fig. 10. Measurements and simulation of the 7–3 dual-band filter, network topology illustrated in Fig. 9.

Example 2: Seven-Pole Three-Zero Filter Implemented in Stacked Single-Mode Cavities

At microwave frequencies, the low-pass specifications shown in Fig. 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, respectively, at $J_1 =]-\infty, 8.158]$ 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. 9.

This configuration of the coupled-resonator network leads to three real solutions for realizing the ideal filtering characteristic.

A solution is chosen for being implemented in stacked single-mode rectangular cavities, as described in [9]. 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. 10. Insertion losses are, respectively, 1.4 and 1.25 dB in the passbands.

VIII. Conclusion and Perspectives

We have presented an iterative algorithm with a quadratic convergence rate to perform the optimal synthesis of multiband filtering functions. We also provided a simple equiripple-like optimality criterion that allows one to check for the optimality of any filtering function. The main advantages of this approach, as opposed to direct optimization schemes, are to provide the user with some optimal placements for the reflection and transmission zeros while validating the result. Based on the latter, we showed that, for some dual-band specifications, an increase of the filter's order does not necessarily yield an improvement of its selectivity: this is a major difference as compared to the usual single-band situation and, to the best of our knowledge, the analysis of this phenomenon is new to the filtering community. Finally, the process was validated by the design of two asymmetric dual-band bandpass filters.

References

[1] R. J. Cameron, M. Yu, and Y. Wang, "Direct-coupled microwave filters with single and dual stopbands," *IEEE Trans. Microw. Theory Tech.*, vol. 53, no. 11, pp. 3288–3297, Nov. 2005.

[2] G. Macchiarella and S. Tamiazzo, "Design techniques for dual-passband filters," *IEEE Trans. Microw. Theory Tech.*, vol. 53, no. 11, pp. 3265–3271, Nov. 2005.

[3] S. Amari, "Synthesis of cross-coupled resonator filters using an analytical gradient-based optimization technique," *IEEE Trans. Microw. Theory Tech.*, vol. 48, no. 9, pp. 1559–1563, Sep. 2000.

[4] M. Mokhtaari, J. Bornemann, K. Rambabu, and S. Amari, "Coupling-matrix design of dual and triple passband filters," *IEEE Trans. Microw. Theory Tech.*, vol. 54, no. 11, pp. 3940–3946, Nov. 2006.

[5] V. Lunot, S. Bila, and F. Seyfert, "Optimal synthesis for multi-band microwave filters," in *IEEE MTT-S Int. Microw. Symp. Dig.*, Jun. 2007, pp. 115–118.

[6] R. J. Cameron, "General coupling matrix synthesis methods for Chebyshev filtering functions," *IEEE Trans. Microw. Theory Tech.*, vol. 47, no. 4, pp. 433–442, Apr. 1999.

[7] E. W. Cheney, *Approximation Theory*. New York: Chelsea Press, 1982.

[8] D. Braess, *Nonlinear Approximation Theory*. Berlin, Germany: Springer-Verlag, 1986.

[9] S. Bila, R. J. Cameron, P. Lenoir, V. Lunot, and F. Seyfert, "Chebyshev synthesis for multi-band microwave filters," in *IEEE MTT-S Int. Microw. Symp. Dig.*, Jun. 2006, pp. 1221–1224.

[10] S. Bila, D. Baillargeat, S. Verdeyme, M. Aubourg, P. Guillon, F. Seyfert, J. Grimm, L. Baratchart, C. Zanchi, and J. Sombrin, "Direct electromagnetic optimization of microwave filters," *IEEE Micro*, vol. 2, no. 1, pp. 46–51, Mar. 2001.

[11] R. J. Cameron, J. C. Faugère, and F. Seyfert, "Coupling matrix synthesis for a new class of microwave filter configuration," in *IEEE MTT-S Int. Microw. Symp. Dig.*, Jun. 2005, pp. 119–122.

Vincent Lunot was born in Chambray-lès-Tours, France, in June 1978. He received the Master's degree in cryptography and coding theory from the University of Limoges, Limoges, France, in 2002, and is currently working toward the Ph.D. degree in applied mathematics at the Institut National de Recherche en Informatique et en Automatique (INRIA), Sophia Antipolis, France.

His research interests include optimization, rational approximation, and algorithmics and their applications to the synthesis of microwave filters.

Fabien Seyfert received the Engineering degree and Ph.D. degree in mathematics from the Ecole superieure des Mines (Engineering School), St. Etienne, France, in 1993 and 1998, respectively.

From 1998 to 2001, he was with Siemens, Munich, Germany, where he was a Researcher specializing in discrete and continuous optimization methods. Since 2002, he has been a Full Researcher with the Institut National de Recherche en Informatique et en Automatique (INRIA), Sophia Antipolis, France. His research interest focuses on the conception of effective mathematical procedures and associated software for problems from signal processing including computer-aided techniques for the design and tuning of microwave devices.

Stéphane Bila was born in Paris, France, in September 1973. He received the Ph.D. degree from the University of Limoges, Limoges, France, in 1999.

He then held a one-year post-doctoral position with the Centre National d'Etudes Spatiales (CNES), Toulouse, France. In 2000, he became a Researcher with the Centre National de la Recherche Scientifique (CNRS), and joined IRCOM (now XLIM), Limoges, France. His research interests include numerical modeling, optimization, and computer-aided techniques for the advanced synthesis of microwave components and circuits.

Abdallah Nasser was born in Baalbek, Lebanon, in 1980. He received the Engineer degree from the Lebanese University, Beyrouth, Lebanon, in 2004, and is currently working toward the Ph.D. degree in high-frequency electronics and opto-electronics at XLIM, University of Limoges, Limoges, France.

His research interests include synthesis methods based on computer-aided techniques for the design and optimization of microwave components and circuits for space applications.

5.3 Coupling matrix synthesis problem

5.3.1 An algebraic framework for coupling matrix synthesis problem

Following paper is reproduced in this section:

- Richard J. Cameron, Jean-Charles Faugère, Fabrice Rouillier, and Fabien Seyfert. "Exhaustive approach to the coupling matrix synthesis problem and application to the design of high degree asymmetric filters". In: *International Journal of RF and Microwave Computer-Aided Engineering* 17.1 (Jan. 2007), pp. 4–12. DOI: 10.1002/mmce.20190. URL: https://hal.inria.fr/hal-00663777

Exhaustive Approach to the Coupling Matrix Synthesis Problem and Application to the Design of High Degree Asymmetric Filters

Richard J. Cameron,[1] Jean-Charles Faugere,[2] Fabrice Rouillier,[3] Fabien Seyfert[4]

[1] Com Dev Space, Aylesbury, Bucks, UK
[2] Univ. Paris VI, 75252 Paris Cedex 05, France
[3] INRIA, 78153 Rocquencourt, France
[4] INRIA, 06902 Sophia Antipolis, France

Received 5 September 2005; accepted 21 June 2006

ABSTRACT: In this paper a new approach to the synthesis of coupling matrices for microwave filters is presented. The new approach represents an advance on existing direct and optimization methods for coupling matrix synthesis, in that it will exhaustively discover all possible coupling matrix solutions for a network if more than one exists. This enables a selection to be made of the set of coupling values, resonator frequency offsets, parasitic coupling tolerance, etc. that will be best suited to the technology it is intended to realize the microwave filter with. To demonstrate the use of the method, the case of the recently introduced "extended box" coupling matrix configuration is taken. The extended box is a new class of filter configuration adapted to the synthesis of asymmetric filtering characteristics of any degree. For this configuration the number of solutions to the coupling matrix synthesis problem appears to be high and offers therefore some flexibility that can be used during the design phase. We illustrate this by carrying out the synthesis process of two asymmetric filters of 8th and 10th degree. In the first example a ranking criterion is defined in anticipation of a dual mode realization and allows the selection of a "best" coupling matrix out of 16 possible ones. For the 10th degree filter a new technique of approximate synthesis is presented, yielding some simplifications of the practical realization of the filter as well as of its computer aided tuning phase. © 2006 Wiley Periodicals, Inc. Int J RF and Microwave CAE 17: 4–12, 2007.

Keywords: coupling matrix; filter synthesis; bandpass filter; Groebner basis; inverted characteristic; multiple solutions

I. INTRODUCTION

In Ref. 1, a synthesis method for the "Box Section" configuration for microwave filters is introduced. Box sections are able to realize a single transmission zero (TZ) each and have an important advantage that no "diagonal" inter-resonator couplings are required to realize the asymmetric zero, as would the equivalent trisection. Also the frequency characteristics are reversible by retuning the resonators alone [2], retaining the same values and topology of the inter-resonator couplings.

The first feature leads to particularly simple coupling topologies, and is suitable for realization in the very compact waveguide or dielectric dual-mode res-

Correspondence to: F. Seyfert; e-mail: Fabien.Seyfert@sophia.inria.fr.

DOI 10.1002/mmce.20190
Published online 1 December 2006 in Wiley InterScience (www.interscience.wiley.com).

© 2006 Wiley Periodicals, Inc.

WILEY InterScience® DISCOVER SOMETHING GREAT

onator cavity, while the ability to reverse the characteristics by retuning makes the box-filter useful for diplexer applications, the same structure being usable for the complementary characteristics of the two channel filters.

Ref. 1 continued on to introduce the extended box configuration for filter degrees $N > 4$, able to realize a maximum of $(N - 2)/2$ (N even) or $(N - 3)/2$ (N odd) symmetric or asymmetric TZs. Figure 1 gives extended box networks of even degree 4 (basic box section), 6, 8, and 10, showing the particularly simple ladder network form of the extended box configuration. In each case, the input and output are from opposite corners of the ladder network. The extended box network also retains the property of giving lateral inversion of the frequency characteristics by retuning of the resonators alone.

The prototype coupling matrix for the extended box network may be easily synthesized in the folded or "arrow" forms. However, it appears that there is no simple closed form equation or procedure that may be used to transform the folded or arrow coupling matrix to the extended box form. In Ref. 1 a method is described which is essentially the reverse of the general sequence that reduces any coupling matrix to the folded form, for which a regular sequence of rotation pivots and angles does exist. Using this method means that some of the rotation angles cannot be determined by calculation from the pretransform coupling matrix (as can be done from the "forward" method) and so they have to be determined by optimization. Other methods (e.g. [3, 4]) are also known to produce a solution.

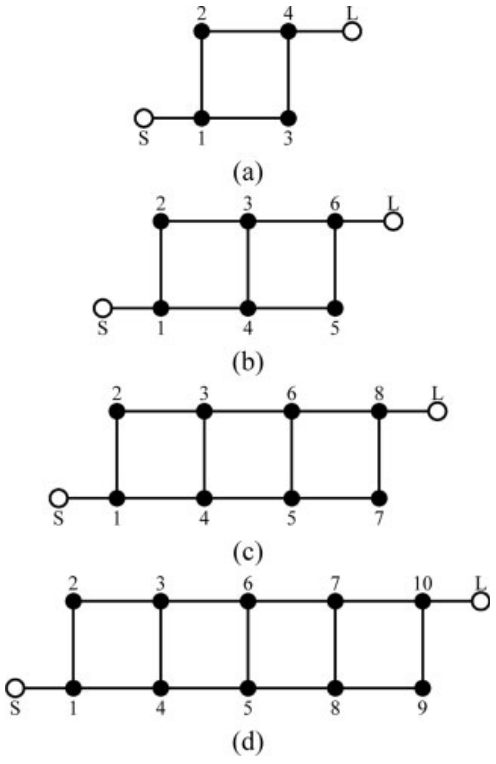


Figure 1. Coupling and routing diagrams for extended box section networks: (a) 4th degree (basic box section), (b) 6th degree, (c) 8th degree, and (d) 10th degree.

Although most target coupling matrix configurations (eg propagating in-line) have one or two unique solutions, the extended box configuration is distinct in having multiple solutions, all returning exactly the same performance characteristics under analysis as the original prototype folded or arrow configuration. The solutions converged upon by existing optimization methods tend to be dependent upon the starting values given to the coupling values or rotation angles, and it can never be guaranteed that all possible solutions have been found. In Ref. 2 an approach based on computer algebra was outlined that allows to compute all the solutions for a given coupling matrix topology, including those with complex values (which of course are discarded from the solutions considered for the realization of the hardware). In this paper we detail the latter procedure as well as a modification in the choice of the set of algebraic equations to solve that leads to an important improvement of the algorithm's efficiency in practice.

Having a range of solutions enables a choice to be made of the coupling value set most suited to the technology it is intended to realize the filter with. Considerations influencing the choice include ease of the design of the coupling elements, minimization of parasitic couplings, or resonator frequency offsets. Some of the coupling matrix solutions may contain coupling elements with values small enough to be ignored without damage to the overall electrical performance of the filter, and so simplifying the manufacture and tuning processes.

In the following section we describe the multi-solution synthesis method, applicable to the extended box network and others that support multiple solutions. Finally we apply our procedure to the synthesis of filtering characteristics of degree 8 and 10. We demonstrate how the ability to choose among several coupling matrices simplifies the practical realization of the filter in dual-mode waveguide or dielectric resonator cavities. In particular an approximate synthesis technique based on a post-processing optimization step is presented and improves the approach in Ref. 2.

II. GENERAL FRAMEWORK FOR THE COUPLING MATRIX SYNTHESIS PROBLEM

In this section we work with a fixed coupling topology, that is we are given a set of independent nonzero couplings associated to a low pass prototype of some filter with N resonators. Starting with numerical values for the couplings (coupling matrix M) and the input/output (i/o) loads (R_1, R_2) one can easily compute the admittance matrix using following formula:

$$Y(s) = C(sI - jM)^{-1}C^t = \sum_{k=0}^{\infty} \frac{Cj^k M^k C^t}{s^{k+1}} \qquad (1)$$

with

$$C = \begin{bmatrix} \sqrt{R_1} & \dots 0 \dots & 0 \\ 0 & \dots 0 \dots & \sqrt{R_N} \end{bmatrix}$$

The coupling matrix synthesis problem is actually about inverting the latter procedure: given an admittance matrix we want to find values for the i/o loads and couplings that realize it. To formalize this we give a name to the mapping that builds the admittance matrix from the free electrical parameters and we define

$$T : p = (\sqrt{R_1}, \sqrt{R_N} \dots M_{i,j}) \rightarrow (CC^t, \dots CM^k C^t, \dots CM^{2N-1}C^{\mathrm{t}})$$

The above definition is justified by the fact that the admittance matrix is entirely determined by the first $2N$ coefficients of its power expansion at infinity [5].

Now suppose that each of the electrical parameters move around in the complex plane: what about the corresponding set of admittance matrices? The latter can be identified with the image by T of C^r (C is here the field of complex numbers) where r is the number of free electrical parameters. We call this set V ($=T(C^r)$) and refer to it as the set of admissible admittance matrices with respect to the coupling topology.

In this setting the coupling matrix synthesis problem is the following: given an element w in V compute the solution set of

$$T(p) = w \qquad (2)$$

Now from the definition of T it follows that eq. (2) is a nonlinear polynomial system with r unknowns, namely, the square roots of the i/o loads and the free couplings of the topology. From the polynomial structure of the latter system we can deduce following mathematical properties (we will take them here for granted):

- Equation (2) has a finite number of solutions for all generic w in V (generic means for almost all w in V) if and only if the differential of T is generically of rank r. In this case we will say that the coupling topology is nonredundant.
- The number of complex solutions of the eq. (2) is generically constant with regard to w in V. Because of the sign symmetries this number is a multiple of 2^N and can therefore be written as $m2^N$. The number m is the number of complex solutions up to sign symmetries and we will call it the "reduced order" of the coupling geometry.

Remarks: The nonredundancy property ensures that a coupling geometry is not over-parameterized, which would yield a continuum of solutions to our synthesis problem. We illustrate this with the 6th degree topology of Figure 2.

- If no diagonal couplings are present (as suggested by the gray dots in Fig. 2), the topology is redundant, i.e. the synthesis problem admits an infinite number of solutions.
- If, for example, the coupling (1,4) is removed, the topology becomes nonredundant and is adapted to a 6-2 symmetric filtering characteristic. In this case the resulting coupling topology is the so called arrow form for which the coupling matrix synthesis problem is known to have only one solution. The reduced order of the latter topology is therefore 1.
- Finally, if diagonal couplings are allowed, the topology becomes nonredundant, and is actually the 6th degree extended box topology of Figure 1 and is adapted to a 6-2 asymmetric filtering characteristic. We will see in the following section that its reduced order is 8.

The use of the adjective "generic" in the latter statements is necessary for their mathematical cor-

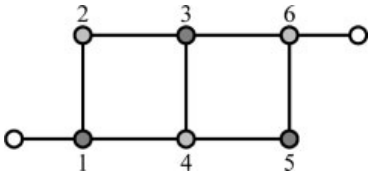


Figure 2. Redundant topology.

TABLE I. Reduced Order and Observed Number of Real Solutions

Topology	Max. No. of TZs	Reduced Order	Observed No. of Real Solutions
Figure 1(a)	1	2	2
Figure 1(b)	2	8	6
Figure 1(c)	3	48	16
Figure 1(d)	4	384	36, 58
Figure 3	8	3	1

rectness. In fact properties concerning parameterized algebraic systems are often true for all possible values of the parameters but an exceptional set. An example of this is given by following polynomial:

$$p(x) = ax^2 + 1.$$

The latter polynomial has two distinct roots for almost all complex values of the parameter a: the exceptional parameter set where the latter property does not hold is characterized by the equation $a = 0$ and is very "thin" (or non-generic) as a subset of the complex plan.

The constructive nature of our framework for the synthesis problem depends strongly on our ability to invert numerically the mapping T, i.e. compute the solution set of eq. (2). In the next section we briefly explain how this can be done using Groebner basis computations.

III. GROEBNER BASIS

As an example of the use of Groebner basis, suppose we are given the following system:

$$\begin{cases} x^2 + 2xy + 1 = 0 & \text{(a)} \\ x^2 + 3xy + y + 1 = 0 & \text{(b)} \end{cases}$$

By combining equations we get the following polynomial consequences:

$$\begin{aligned} &\text{(b)} - \text{(a)}: && xy + y + 1 = 0 && \text{(c)} \\ &\text{(c)}x - \text{(b)}y: && 3xy^2 - yx - x + y^2 + 2y = 0 && \text{(d)} \\ &\text{(d)} - \text{(c)}y: && -yx - x - 2y^2 - y = 0 && \text{(e)} \\ &\text{(e)} + \text{(c)}: && -x - 2y^2 + 1 = 0 && \text{(f)} \\ &\text{(f)}y + \text{(c)}: && -2y^3 + 2y + 1 = 0 && \text{(g)} \end{aligned}$$

Note that eq. (g) is a univariate polynomial in the unknown y. Solving the latter numerically yields the following 3-digit approximations for y: $\{-0.56 + 0.25j, -0.56 - 0.25j, 1.19\}$ and from eq. (f) we get the corresponding values for $x = \{0.42 - 0.61j, 0.42 + 0.61j, -1.84\}$. Now we can verify that the latter three pairs of values for (x,y) are also solutions of eqs. (a) and (b) and therefore the only three solutions of our original system. Equations (f) and (g) are what is called a Groebner basis [6] of our original system and allows us to reduce the resolution of a multivariate polynomial system to the one of a polynomial in a single unknown.

The technique that we have presented is a simple example is called "elimination" and can be thought as the nonlinear version of the classical Gaussian elimination technique for linear systems. The fact that the process of variables elimination by means of combinations of equations always ends up with a polynomial in a single variable is equivalent to the property that the original system has only isolated solutions [7]. In the case of our synthesis problem this is ensured by the nonredundancy of the considered coupling topology.

In practice, computing a Groebner basis can be computationally very costly: the number of necessary combinations of equations can be very large and strongly grows with the total number of variables of the system. Therefore, the use of specialized algorithms and their effective software implementation is strongly recommended. In this work we have used the tool Fgb [8].

Table I summarizes the reduced order and the number of real solutions observed for a particular filtering characteristic for each of the extended box networks of Figure 1. The synthesis method is not limited to the case of extended box topologies: Table I also mentions the case of a 10th degree topology (see Fig. 3) adapted to 10-8 symmetric characteristics.

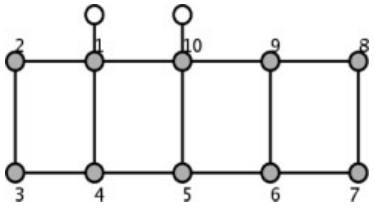


Figure 3. Coupling topology adapted to 10-8 symmetric characteristics.

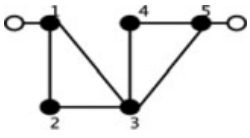


Figure 4. Academic example of a 5th degree coupling topology adapted to 5-2 asymmetric characteristics.

0	0.4	0	0	0
0.4	0.3	0.1	0	0.1
0	0.1	.2	0.2	0.2
0	0	0.2	0.2	1
0	0.1	0.2	1	0.1

Figure 5. Canonical coupling matrix in arrow form of a 5-2 filtering function, admitting only complex coupling matrices when using the topology of Figure 4.

The reduced order of the latter is equal to 3 and is therefore much smaller than the reduced order of 384 of its 10th degree extended box analogue. This is something we observed empirically by testing our method on various networks: topologies adapted to asymmetric characteristics seem to have a much higher reduced order than those adapted to symmetric ones.

Although the reduced order depends only on the coupling geometry, the number of real solutions depends on the prototype characteristic the network is realizing (position of TZs, return loss, etc…) and is, by definition, bounded from above by the reduced order. One can even construct some coupling topologies and some filtering characteristics for which the synthesis problem admits only complex solutions. An academic example of this is given by the topology of Figure 4 and the filtering characteristic, the canonical coupling matrix in arrow form of which is given on Figure 5. In this latter case the reduced order of the coupling topology is 2 but both solutions to the synthesis problem are complex and equal to the matrix of Figure 6 and to its conjugate.

IV. PRACTICAL IMPLEMENTATION OF THE SYNTHESIS PROCEDURE AND EXAMPLES

A. 8th Degree Extended Box Filter

As an application we will consider the synthesis of an 8th degree filter in extended box configuration (see Fig. 1c). Using a computer algebra system (e.g. Maple), we check that this topology is nonredundant and from the application of the minimum path rule

0	0.41-0.001j	0.006+0.074j	0	0
0.41-0.001j	0.3-0.035j	0.079+0.031j	0	0
0.006+0.074j	0.079+0.031j	.099-0.2j	0.3-0.075j	0.043-0.54j
0	0	0.3-0.075j	0.3+0.23j	1.2+0.02j
0	0	0.043-0.54j	1.2+0.02j	0.1

Figure 6. Complex solution to the synthesis problem with coupling topology of Figure 4 and coupling matrix in canonical arrow form of Figure 5.

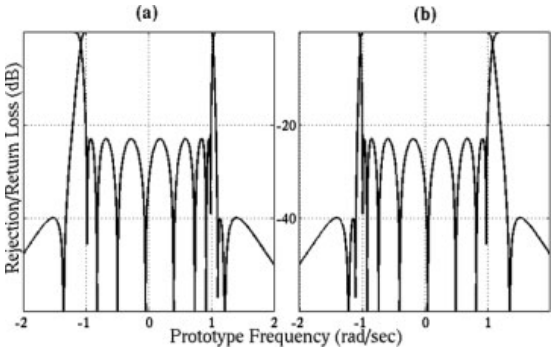


Figure 7. (a) Original and (b) inverted rejection and return loss performance of an 8-3 asymmetric characteristic in extended box configuration.

we conclude that the set of admissible admittances consists of rational reciprocal matrices of degree 8 with at most 3 TZs. Using classical quasi-elliptic synthesis techniques an 8th degree filtering characteristic is designed with a 23 dB return loss and three prescribed TZs, producing one rejection lobe level of 40 dB on the lower side and two at 40 dB on the upper side (see Fig. 7a).

Now computing the $2N$ first terms of the power expansion of the admittance matrix yields the left hand term of eq. (2) which in turn could be solved using Groebner basis computations. At this point it is important to mention that the complexity of the Groebner basis computations of a system increases with its total number of complex solutions. The natural sign symmetries of the system derived from

0.0107	-0.2904	0	-0.8119	0	0	0	0
-0.2904	-0.9804	0.1081	0	0	0	0	0
0	0.1081	0.0605	0.5475	0	0.5984	0	0
0.8119	0	0.5475	0.1384	-0.0663	0	0	0
0	0	0	-0.0663	0.0152	0.5334	0.6782	0
0	0	0.5984	0	0.5334	0.0226	0	-0.1260
0	0	0	0	0.6782	0	0.0113	0.8530
0	0	0	0	0	-0.1260	0.8530	0.0107

(a)

0.0107	0.0001	0	-0.2464	0	0	0	0
0.0001	-0.9590	0.2094	0	0	0	0	0
0	0.2094	0.0498	0.4681	0	-0.4681	0	0
-0.2464	0	0.4681	0.0115	0.3744	0	0	0
0	0	0	0.3744	-0.0439	0.3744	0.8165	0
0	0	-0.4681	0	0.3744	0.0115	0	0.8623
0	0	0	0	0.8165	0	0.1975	0.0001
0	0	0	0	0	0.8623	0.0001	0.0107

(b)

Figure 8. "$N \times N$" coupling matrices for an 8-3 asymmetric prototype: (a) extended box configuration, (b) "cul-de-sac" configuration. $R_1 = R_N = 1.0878$.

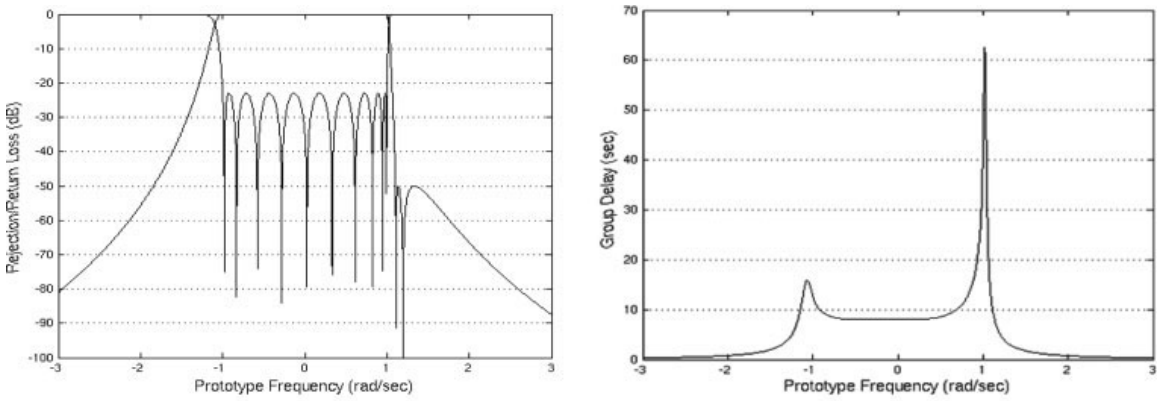


Figure 9. 10-2-2 asymmetric characteristic: (a) rejection and return loss (b) group delay.

eq. (2) tend to artificially increase the latter (total number of solutions $= m2^N$) and may dramatically increase the computation time of the corresponding Groebner basis. Before continuing on with the synthesis we therefore explain how a rewriting of eq. (2) allows us to get rid of these unwanted sign symmetries.

An alternative to eq. (2) to invert the mapping T is to use an algebraic version of the approach presented in Ref. 9 that is based on similarity transforms. If M is a coupling matrix in canonical form realizing the admittance matrix, then eq. (2) is "equivalent" to the following matrix equation where the unknown is a similarity transform P.

$$P = \begin{pmatrix} 1 & \dots 0 \dots & 0 \\ \vdots & H & \vdots \\ 0 & \dots 0 \dots & 1 \end{pmatrix} \quad \text{(a)}$$
$$H^{\mathrm{t}}H = Id \quad \text{(b)} \qquad (3)$$
$$\forall (i,j) \in I \ (P^{\mathrm{t}}MP)_{i,j} = 0 \quad \text{(c)}$$

In the latter, I is the set of indices corresponding to the couplings that must be zero in the target topology (in our example $I = \{[(1,3), (1,5), (1,6).\dots\dots]\}$). If P is a solution of eq. (3); it is readily seen that all the similarity transforms that are obtained from P by inverting some of the columns vectors of the submatrix H are also solutions of eq. (3). To break these symmetries the "trick" is to slightly modify eq. (3b). We denote by h_i the ith column vector of H. Some of the equations of eq. (3b) indicate that the vectors h_i are unitary with regard to the Euclidean norm. We replace these normalizing equations by

$$u_i^{\mathrm{t}} h_i = 1 \qquad (4)$$

where u_i is a randomly-chosen vector. We call eq. (3′) the resulting system. It can be verified that for a generic choice of the u_i's, all the solutions of eq. (3) that are equivalent up to sign changes of their column vectors correspond to a single solution of eq. (3′). More precisely to every set of solutions of eq. (3) of the form

$$H = (\pm h_1, \pm h_2 \cdots \pm h_i \cdots) \qquad (5)$$

there corresponds a unique solution $G = (g_1 \cdots g_i \cdots)$ of eq. (3′) where the column vectors g_i are given by

0.0145	0.7712	0	0.3879	0	0	0	0	0	0
0.7712	0.2493	-0.5232	0	0	0	0	0	0	0
0	-0.5232	0.0554	0.1925	0	-0.5393	0	0	0	0
0.3879	0	0.1925	-0.9071	-0.0010	0	0	0	0	0
0	0	0	-0.0010	-0.7492	-0.2683	0	0.3110	0	0
0	0	-0.5393	0	-0.2683	0.0437	-0.4668	0	0	0
0	0	0	0	0	-0.4668	0.3195	-0.4934	0	-0.2040
0	0	0	0	0.3110	0	-0.4934	-0.1000	0.4827	0
0	0	0	0	0	0	0	0.4827	-0.0021	0.8388
0	0	0	0	0	0	-0.2040	0	0.8388	0.0145

Figure 10. Coupling matrix of the 10-2-2 characteristic of Figure 9 with the extended box topology and a "small" M_{45} coupling, $R_1 = R_N = 1.04326$.

0.0161	0.7655	0	0.4053	0	0	0	0	0	0
0.7655	0.2705	-0.5173	0	0	0	0	0	0	0
0	-0.5173	0.0560	0.2057	0	-0.5386	0	0	0	0
0.4053	0	0.2057	-0.8923	0	0	0	0	0	0
0	0	0	0	-0.7810	-0.2512	0	0.2968	0	0
0	0	-0.5386	0	-0.2512	0.0445	-0.4761	0	0	0
0	0	0	0	0	-0.4761	0.2867	-0.5041	0	-0.1984
0	0	0	0	0.2968	0	-0.5041	-0.0850	0.4851	0
0	0	0	0	0	0	0	0.4851	0.0016	0.8427
0	0	0	0	0	0	-0.1984	0	0.8427	0.0173

Figure 11. Coupling matrix of the 10-2-2 characteristic of Figure 9 with a simplified topology, (i.e. $M_{45} = 0$), $R_1 = 1.0969$, $R_N = 1.0963$.

$$g_i = \frac{h_i}{u_i^t h_i} \tag{6}$$

With regard to the Groebner basis computation system, eq. (3′) has shown to be much more tractable than the algebraic system derived from eq. (2).

Getting back to our 8th degree example, we compute M the associated coupling matrix in arrow form and set up eq. (3′). The latter is an algebraic system of linear and quadratic equations in the entries of H. The computation of its Groebner basis leads to the following result:

- The reduced order of the topology is 48.
- For this particular filtering characteristic, 16 of the 48 solutions are real-valued.

Only the real solutions have a physical interpretation and are therefore of practical interest.

The criterion used to choose the best coupling matrix out of the 16 realizable ones will depend on the hardware implementation of the filter. Having in mind a realization with dual mode cavities, we choose to select solutions where the asymmetry between the two "arms" of each cross-iris is maximized in order to minimize parasitic couplings. The best ratios between couplings of the relevant pairs (M_{14}, M_{23}), (M_{36}, M_{45}), and (M_{57}, M_{68}) are found for the solution shown in Figure 8a, where each cross-iris has one of its coupling values at least five times larger than the other one.

Figure 8b illustrates that sometimes solutions emerge which have very small values for certain couplings (M_{12} and M_{78} in this case), which may be safely omitted for the implementation without damaging the final response of the network. In this case a quasi cul-de-sac network is produced, similar to the 8-3 example given in Ref. 1. In fact one can show that with some renumbering, the cul-de-sac network of Ref. 1 is a sub-topology of the extended box where the couplings M_{12} and M_{78} are set to zero. The cul-de-sac topology is more restrictive than the extended box one in the sense that it is only adapted for the synthesis of auto-reciprocal characteristics, such that $S_{11} = S_{22}$ holds. However, our current filtering characteristic is, up to numerical errors, auto-reciprocal and this explains why in this example a quasi cul-de-sac network is found among all possible coupling matrices.

Finally it is shown that only the resonators need to be retuned in order to obtain an inverted characteristic. Figure 7b shows the rejection and return loss obtained from the coupling matrices of Figure 8 when the signs of their diagonal elements $M_{i,i}$ are changed (see Ref. 4 for details).

B. 10th Degree Extended Box Filter and Approximate Synthesis Technique

We consider the synthesis of a 10th degree filter in the extended box topology of Figure 1d. Using our procedure we check that this topology is nonredundant and that it is adapted to asymmetric characteristics with up to 4 TZs. A filtering characteristics is designed with a 23 dB return loss, 2 TZs at $+j1.10929$ and $+j1.19518$ to give two 50 dB rejection lobes on the upper side and 2 more complex zeros at $\pm 0.75877 - j0.13761$ for group delay equalization purposes (see Fig. 9).

The corresponding coupling matrix in arrow form is determined and the computation of a Groebner basis of system (2) yields the following:

- The reduced order of the topology is 384.
- For our specific filtering characteristic 36 real and therefore realizable solutions are found.

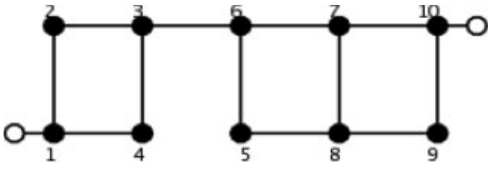


Figure 12. Simplified 10th degree topology.

0.0161	0.4053	0	0.7655	0	0	0	0	0	0
0.4053	-0.8923	-0.2057	0	0	0	0	0	0	0
0	-0.2057	0.0560	0.5173	0	-0.5386	0	0	0	0
0.7655	0	0.5173	0.2705	0	0	0	0	0	0
0	0	0	0	-0.7810	-0.2512	0	0.2968	0	0
0	0	-0.5386	0	-0.2512	0.0445	-0.4761	0	0	0
0	0	0	0	0	-0.4761	0.2867	-0.5041	0	0.1984
0	0	0	0	0.2968	0	-0.5041	-0.0850	-0.4851	0
0	0	0	0	0	0	0	-0.4851	0.0016	0.8427
0	0	0	0	0	0	0.1984	0	0.8427	0.0173

Figure 13. Coupling matrix with a simplified topology and the most asymmetric irises, $R_1 = 1.0969$, $R_N = 1.0963$.

When realized with dual mode cavities this topology requires four cross- irises. Our aim is to demonstrate how our exhaustive approach may allow the "replacement" of a cross-iris by an iris with a single arm as well as to simplify the future computer-aided tuning process of the filter.

Among all the possible coupling matrices the one with the smallest coupling corresponding to an iris is selected, which leads to the matrix of Figure 10 where M_{45} is equal to -0.001. Setting M_{45} to zero yields a small but undesirable variation of the return loss as well as of the upper-band rejection lobes. The remaining couplings are therefore retuned, thanks to an optimization step that minimizes the discrepancy between the original response and the one obtained by imposing that M_{45} be zero (see Fig. 11 for the resulting coupling matrix). A quasi perfect fit is obtained between the two responses: the least square error between the two return losses on the normalized broadband $[-3,3]$ equals 8.83×10^{-5} (on the Bode plot there is visually no difference).

Finally the simplified coupling topology of Figure 12 is considered as a new topology in its own right. Using our procedure its reduced order is found to be equal to 2 and a second equivalent coupling matrix with the same coupling topology is computed (see Fig. 13). With regard to the "iris asymmetry criterion" of the last section the latter matrix is the best one.

Note that besides the removal of a cross-iris we have also lowered the reduced order of our target topology from 384 to 2. This is important if one wants to use a computer-aided tuning process [10] that typically identifies a coupling matrix from measured data. In the cases of topologies with multiple solutions, such a tool will return a set of equivalent coupling matrices and leave to the user the "expert" task of choosing the "right" one. This can be done by using some extra information concerning the physical device, like for example an a priori estimation of the coupling value realizable by some irises. Nevertheless, the latter task is of course much easier to carry out with a short list of equivalent coupling matrices than with a huge one.

V. CONCLUSION

In this paper, a new method for the synthesis of the full range of coupling matrices for networks that support multiple solutions is presented. This procedure yields an exhaustive list of all the solutions to the synthesis problem. Based on the latter, an approximate synthesis technique is derived which allows the reduction of the constructional complexity of high-degree asymmetric filters in dual-mode technologies. In addition it has been shown that a knowledge of which solutions are possible is important when reconstructing the coupling matrix from measured data, during development or computer-aided tuning (CAT) processes.

A software called Dedale-HF and dedicated to the presented exhaustive synthesis technique has recently been released and is accessible under: http://www.sop.inria.fr/apics/Dedale

REFERENCES

1. R.J. Cameron, A.R. Harish, and C.J. Radcliffe, Synthesis of advanced microwave filters without diagonal cross-couplings, IEEE Trans Microwave Theory Tech MTT-50 (2002), 2862–2872.
2. G. Macchiarella, A powerful tool for the synthesis of prototype filters with arbitrary topology, 2003 IEEE MTT-S Int Microwave Symp Dig 3 (2003), 1721–1724.
3. S. Amari, Synthesis of cross-coupled resonator filters using an analytical gradient-based optimization technique, IEEE Trans Microwave Theory Tech MTT-48 (2000), 1559–1564.

4. R.J. Cameron, J.C. Faugere, and F. Seyfert, Coupling matrix synthesis for a new class of microwave filter configuration, 2005 IEEE MTT-S Int Microwave Symp Dig (2005), 119–122.
5. T. Kailath, Linear systems, Prentice Hall, Upper Saddle River, NJ, 1980.
6. D. Cox, J. Little, and D. O'Shea, Ideals, varieties, and algorithms, Springer, Berlin, 1997.
7. L. Gonzalez-Vega, F. Rouillier, and M.F. Roy, Some tapas of computer algebra, In: Algorithms and computation in mathematics, Vol. 4, Springer, Berlin, 1999, pp. 34–65.
8. J.C. Faugere, A new efficient algorithm for computing Groebner bases without reduction to zero (F5), Proc Int Symp on Symbolic and Algebraic Comp, New York, 2002, pp. 75–83.
9. R.J. Cameron, General coupling matrix synthesis method for Chebyshev filtering functions, IEEE Trans Microwave Theory Tech MTT-47 (1999), 433–442.
10. F. Seyfert, L. Baratchart, J.P. Marmorat, S. Bila, and J. Sombrin, Extraction of coupling parameters for microwave filters: Determination of a stable rational model from scattering data, 2003 IEEE MTT-S Int Microwave Symp Dig 1 (2003), 25–28.

BIOGRAPHIES

Richard J. Cameron (M'83–SM′94–F′02) received the B.Sc. degree in telecommunications and electronic engineering from Loughborough University, UK, in 1969. In 1969, he joined Marconi Space and Defence Systems, Stanmore, UK. His activities there included small earth-station design, telecommunication satellite system analysis, and computer-aided RF circuit and component design. In 1975, he joined the European Space Agency's technical establishment (ESTEC, The Netherlands), where he was involved in the research and development of advanced microwave active and passive components and circuits, with applications in telecommunications, scientific and earth observation spacecraft. Since joining Com Dev Ltd., Aylesbury, Bucks, UK, in 1984, he has been involved in the software and methods for the design of high-performance components and sub-systems for both space and terrestrial application. Prof. Cameron is a Fellow of the Institution of Electrical Engineers (IEE), UK, and of the Institute of Electrical and Electronic Engineers (IEEE), USA. He has recently taken up an appointment as a Visiting Professor to the University of Leeds, England.

Jean-Charles Faugère was born in Normandie, France, in 1966. He graduated from the Ecole Normale Supérieure and he received the Ph.D. degree from the University of Paris 6 (France) in 1994 . Presently he is a CNRS researcher in Université Pierre et Marie Currie (France). His research interest are in Computer Algebra, polynomial solving, Groebner basis, efficient software and applications of Computer Algebra. He is currently the scientific leader of the SPIRAL team (University Paris 6) whose goal is mainly to solve algebraic systems of equations.

Fabrice Rouillier received his Ph.D. degree in mathematics from the University of Rennes, France, in 1996. Since October 1996, he is research scientist at INRIA. He visited the LORIA unit (1996–2000), the team of Pr. D. Lazard (2000–2003), at the University of Paris VI before joining the Rocquencourt unit in June 2003. Since January 2006, he is the scientific leader of the SALSA (Software for Algebraic Systems and Applications) project.

Fabien Seyfert graduated from the ''Ecole Supérieure des Mines'' (Engineering School) in StEtienne (France) in 1993 and received his Ph.D. in mathematics in 1998. From 1998 to 2001 he joined Siemens (Munich, Germany) as a researcher specialized in discrete and continuous optimization methods. Since 2002 he has a full research position at INRIA in Sophia-Antipolis. His research interest focuses on the conception of effective mathematical procedures and associated software for problems from signal processing including computer-aided techniques for the design and tuning of microwave devices.

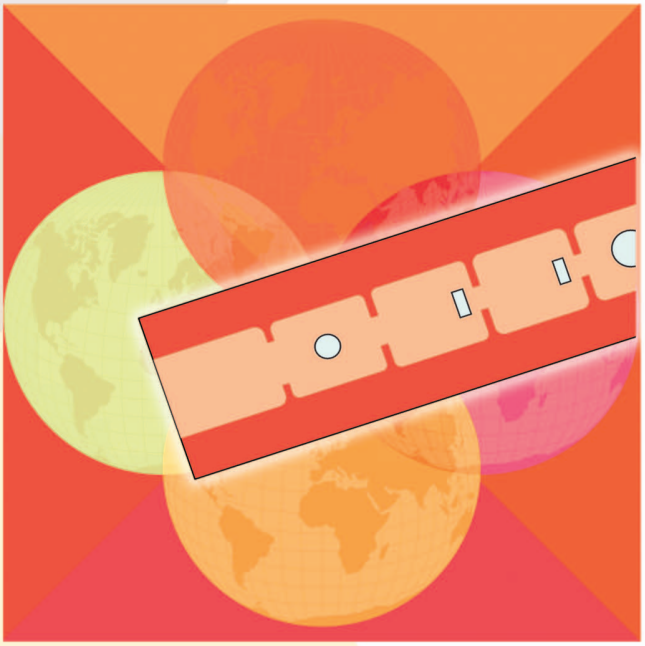
5.3.2 Classification of coupling topologies: a survey

Following paper is reproduced in this section:

- Fabien Seyfert and Stéphane Bila. "General synthesis techniques for coupled resonator networks". In: *IEEE Microwave Magazine* 8.5 (2007), pp. 98–104. DOI: 10.1109/MMW.2007.4383440. URL: https://hal.inria.fr/hal-00663533

General Synthesis Techniques for Coupled Resonator Networks

Fabien Seyfert and Stéphane Bila



© ARTVILLE

With modern communication systems, the allocated frequency spectrum has become more crowded and the demand for high-performance microwave filters complying with stringent specifications has considerably increased. Telecommunication systems require high selectivity to prevent interference, together with flat in-band group-delay and amplitude to minimize signal degradation. The design of microwave filters is usually a tradeoff between various electrical performances (selectivity, insertion loss, group delay) along with minimization of mass and volume, development time, and manufacturing cost [1]. For particular applications, additional constraints such as power handling, thermal stability, or mechanical stability must also be analyzed carefully [2].

Several types and implementation technologies of distributed microwave filters [3] are available, and the choice is driven by the application. However, the design is generally based on the same scheme [4]. The first step consists of synthesizing a lumped-element network from a polynomial filtering function that fulfills the electrical specifications. The second step then converts the lumped-element network into a practical microwave filter.

Applying this scheme, a designer has to face two major problems: the derivation of the lumped-element network which has to be compatible with the polynomial filtering function to be realized, and the dimensioning of the distributed microwave filter. This article details previous points, focusing on the design of coupled resonator filters, i.e., filters that comply with the coupling matrix representation.

Fabien Seyfert (fseyfert@sophia.inria.fr) is with INRIA, 06902 Sophia Antipolis, France
Stéphane Bila (stephane.bila@xlim.fr) is with Xlim, 87060 Limoges, France.

Digital Object Identifier 10.1109/MMM.2007.904720

 1527-3342/07/$25.00©2007 IEEE

Compatibility of Coupling Topologies with Specific Classes of Filtering Functions

As addressed in [5] and [6], the lowpass prototype circuit (Figure 1) is widely used as a coarse model for the synthesis of coupled resonator filters. The coupling topology—or, in other words, the way resonators are coupled to each other—is imposed by realizability issues that depend on the technology that is intended for the filter implementation. For example, in dual-mode waveguide technology [7], the presence of diagonal cross couplings yields severe complications in the manufacturing process, and efforts have been made to derive topologies that are "diagonal cross-coupling free" [8]. For planar technologies, elementary space constraints also yield some restrictions on the coupling topology, and every designer inevitably faces the following question: what kind of frequency responses can I possibly adjust given the constraints I have on my coupling topology? In the following, we give some guidelines to answer this question.

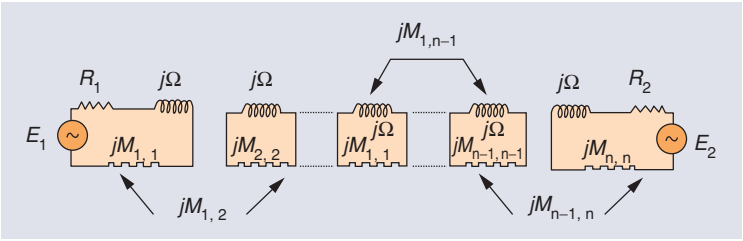


Figure 1. *Lowpass circuit prototype, where* j Ω *symbolizes a unit inductance.*

The nondissipative passive nature of the circuit (Figure 1) and its reciprocity ($S_{12} = S_{21}$) implies mechanically the general polynomial form of its associated scattering matrix

$$S = \begin{bmatrix} S_{11} & S_{12} \\ S_{21} & S_{22} \end{bmatrix} = \frac{1}{E} \begin{bmatrix} F & P \\ P & (-1)^n F^* \end{bmatrix}, \qquad (1)$$

where n is the number of resonators and F, P, and E are polynomials with complex coefficients of the complex variable $s = \sigma + j\omega$ where ω is the normalized frequency.

The polynomial P is of degree $m < n - 1$ and satisfies the condition $P = (-1)^{n+1} P^*$ (which implies that the set of transmission zeros is symmetric with respect to the imaginary axis, i.e., paraconjugated). F is of degree n and monic, and the denominator E is the unique Hurwitz polynomial satisfying the following spectral equation

$$EE^* = FF^* + (-1)^{n+1} P^2. \qquad (2)$$

These properties indicate that the scattering parameters are entirely governed by the two numerator polynomials F and P, in terms of which the squared modulus of the transmission S-parameter is expressed simply as

$$\left|S_{21}(j\omega)\right|^2 = \frac{1}{1 + \left|\frac{F(j\omega)}{P(j\omega)}\right|^2}, \qquad (3)$$

where $D = F/P$ is known as the filtering or characteristic function.

This formula is the starting point of efficient frequency synthesis techniques and formulas that exist, for example, for F (given P) in order to obtain very selective quasi-elliptic filtering characteristics [9], [10]. Techniques based on the predistortion of the filter response to compensate for the losses in the final device also make heavy use of the general polynomial structure (1); in particular, the reflexion zeros (zeros of F) are shifted into the right complex plan by this method [11]. More recently, methods were developed to determine F and P in an optimal manner with respect to some general multiband specifications [12]. In all these techniques, advantage is taken from the fact that F and P can be chosen freely up to limitations on their degrees and the paraconjugated nature of P. It is, therefore, natural to ask if these limitations are sufficient to ensure the realizability of a general polynomial scattering matrix of the form (1) by a low-pass prototype circuit (Figure 1) with a specific coupling topology.

When no constraint is given on the topology of the coupling matrix, the answer to this question is yes. A constructive demonstration of this is given in [9], where the author starts from a polynomial model and derives a full coupling matrix that realizes the model (see also [13] for mathematical details). Reduction steps, involving the use of analytically computed similarity transforms (see [14]), allow reducing the full coupling matrix to matrices with well-known canonical topologies like the arrow form (Figure 2) and the folded form (Figure 3). To tackle more general coupling topologies, we first list necessary conditions relevant to the compatibility question between filtering characteristics and topologies.

Shortest Path Rule

For a given topology, let l be the length of the shortest path in the coupling graph from the input to the output resonator. Then $n - l - 1$ is the maximum number of transmission zeros this topology can accommodate. This rule is an algebraic consequence of the structure of the lowpass prototype, and a proof of it can be found in [15].

Degrees of Freedom of a Class of Filtering Characteristics

For specific classes of filtering characteristics, we can evaluate the number of free parameters that define the polynomials F and P. This number is called the dimension of the class. If m is the number of allowed transmission zeros, we have:

- *General Asymmetric Functions*: n complex transmission zeros can be chosen independently, while $m + 1$ real parameters define the polynomial P (its coefficients are alternatively real and pure imaginary). This yields a total of $2n + m + 1$ free real parameters.

From the synthesized lowpass prototype circuit, normalized couplings can be used for a preliminary dimensioning of the distributed filter.

- *Symmetric Functions:* For this kind of response, F has real coefficients and P is restricted to be even (and, therefore, m as well). This yields a total of $n + m/2 + 1$ free real parameters.

This little counting exercise leads to the following useful rule: in order to accommodate a class of responses [such as (n, m) asymmetric] characterized by a given number of free parameters, a coupling topology must possess at least the same number of free electrical parameters. If these two numbers are equal, then the realization problem has a finite number of solutions (but possibly none).

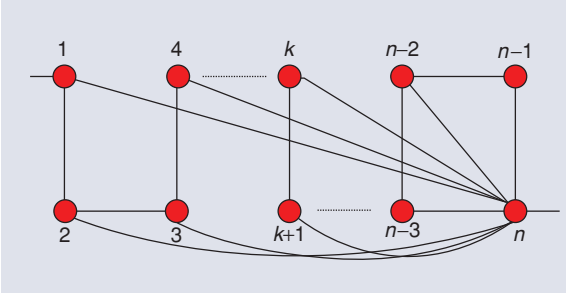


Figure 2. *General "arrow" form.*

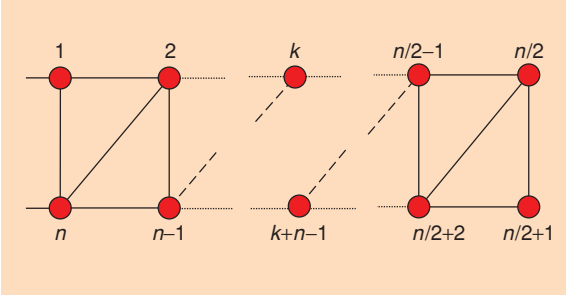


Figure 3. *General folded form.*

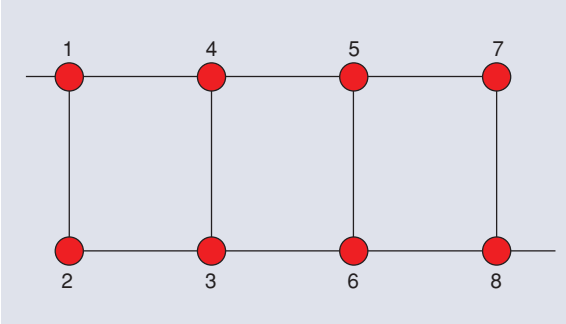


Figure 4. *Eighth-degree extended box topology.*

Canonical Coupling Topologies: Example of the Arrow Form

The general arrow form (Figure 2) entails the following free electrical parameters: n self couplings $M_{i,i}$, $n-1$ couplings between adjacent resonators $(M_{i,i+1})$, $n-2$ additional couplings between the last resonator and all others, and two source/load couplings which yield a total of $3n-1$ free electrical parameters. Using the minimum path rule, the maximum number of transmission zeros is computed to be $(n-1)-2+1=n-2$. On the other hand, the number of free parameters for the $(n, n-2)$ asymmetric class is, according to our preceding remark, $2n+n-2+1=3n-1$, which is consistent with the fact that the arrow form is a canonical form as mentioned earlier. Moreover, we may try here to give a precise definition of the intuitive notion of canonical form; if C is a class of responses of dimension k, then a form is called canonical if it entails exactly k nonzero independent electrical parameters and if the associated realization problem is guaranteed to have a single solution (up to the usual sign changes) for each element of C.

Canonical forms adapted to responses with less transmission zeros can be obtained by enlarging the shortest path, i.e., by canceling progressively the $M_{k,n}$ couplings. The limiting form obtained by this procedure is the classical all-pole topology, where resonators are coupled in a line. The latter is compatible with purely Chebyshev characteristics [$(n, 0)$ type].

For symmetric characteristics, the use of topologies where all couplings $M_{i,j}$ are zero if $i+j$ is even are commonly used; as a matter of fact, the responses of such circuits are structurally symmetric [14], [16] so that no additional relations between couplings are necessary to ensure the symmetry of the response (i.e., the electrical parameters are free). This yields a general arrow form adapted to symmetric responses where all $M_{i,i}$ are set to 0 as well as every second coupling of the form $M_{k,n}$. We leave to the reader's curiosity the care of verifying that the total number of free parameters in this form is equal to $n+(n-2)/2+1$ (for even n), which is also the dimension of the class of $(n, n-2)$ symmetric characteristics.

General Coupling Topologies

For general topologies, one may ask if our necessary conditions of compatibility between a topology and a class of functions are also sufficient. Do they guarantee the existence of a solution to the coupling matrix synthesis problem? The answer to this question is, roughly, yes for the two classes we defined previously, but additional material is needed (mathematical definition of nonredundancy) for a proper formulation. Interested readers will find the complete statement of this compatibility condition in [16]. For practical matters, it is of course crucial to derive a general method that performs the realization step for filtering functions and topologies

where the compatibility rules are fulfilled. The lack of an explicit reduction process for general topologies led the filtering community to derive various approaches based on optimization to solve the underlying nonlinear multivariate problem [10], [17]. Even if algorithms perform relatively well in practice, no guarantee exists about the derivation of a solution, or all solutions, to the coupling matrix synthesis problem. A notable exception to this is made by [18], where a certified process is derived for special topologies made of cascaded triplets or quadruplets. Recently, a procedure [16] based on the use of Groebner basis and homotopy techniques tackled the problem of solving exhaustively the related nonlinear system of equations and finally led to a complete solution of the synthesis problem for all relevant topologies (at least for the time being). This technique has been made accessible to the filtering community through the software Dedale-HF [19], which is available on the Web and free for any academic usage.

A typical application of this is made with the recently introduced extended box topologies [8], which are especially convenient for dual-mode cavities filters with asymmetric characteristics. Consider for example the eight-degree extended box topology in Figure 4. The shortest path rule indicates that, at most, three transmission zeros are supported by this topology. Counting the parameters yields eight self-couplings, ten couplings, and two source/load couplings for a total of 20 free electrical parameters. On the other hand, the dimension of the class of (8, 3) asymmetric characteristics is, according to our formula, $2 \times 8 + 3 + 1 = 20$. The topology and the filtering characteristics class (8, 3) are therefore compatible (see [16] for a rigorous proof of this). Using Dedale-HF, a strongly asymmetric (8, 3) characteristic is computed (see Figure 5), and all 16 possible coupling matrices with the prescribed topology are derived. It is now up to the designer to decide which coupling matrix is most convenient for the application. For more details about this example, see Dedale-HF's tutorial [19] and [20] for applications to equivalent network simplification methods.

Another interesting class of characteristics is autoreciprocal ones, which are characterized by the additional condition $S11 = S22$. Topologies that admit a symmetry plan across the center of the circuit—i.e., that have a coupling matrix which is symmetric across both of its diagonals—are especially suited for this kind of response as their scattering matrix is structurally autoreciprocal. Such topologies are called symmetric, and our previous counting exercise can be repeated to derive necessary realizability conditions. For single-band characteristics, the latter condition happens to be sufficient, but, unfortunately, there exists autoreciprocal dual-band characteristics that admit no symmetric circuit realization. This technical point is beyond the scope of this article, but details on this will be given in forthcoming publications.

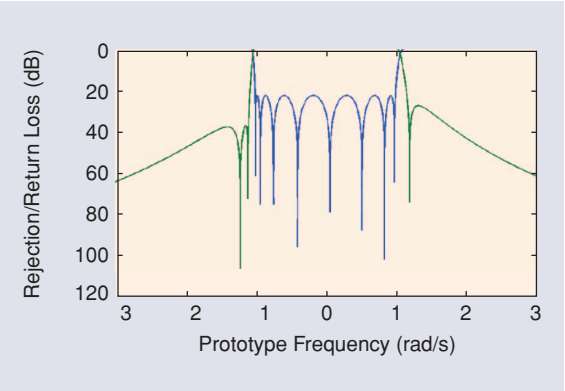


Figure 5. *(8, 3) asymmetric filtering characteristic.*

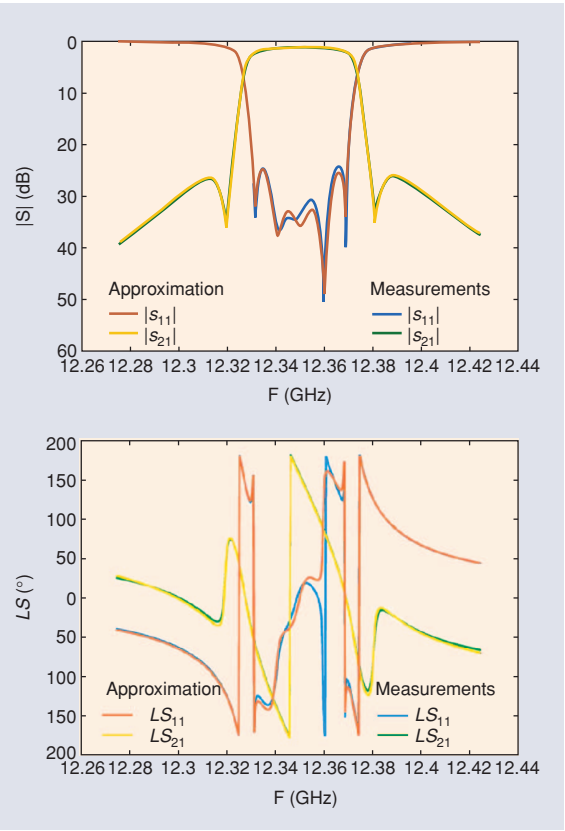


Figure 6. *Rational approximation of measured scattering parameters.*

Computer-Aided Design and Tuning

From the synthesized lowpass prototype circuit, normalized couplings can be used for a preliminary dimensioning of the distributed filter. This first-order dimensioning is generally not sufficient for a precise tuning, especially for narrow-band filters, even in the presence of the tuning element within the hardware. A more accurate dimensioning step, generally involving an electromagnetic (EM) model together with an elaborated process for tuning its dimensions, is then necessary. Moreover, computer-aided tuning is also necessary in some cases to guide the designer while adjust-

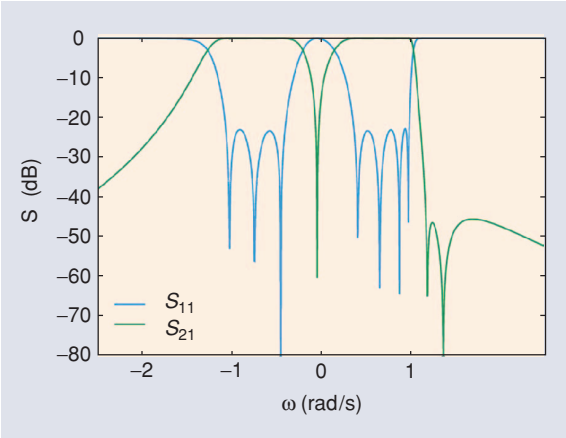


Figure 7. *Normalized (lowpass) scattering parameters (seven-pole, three-zero asymmetrical characteristic).*

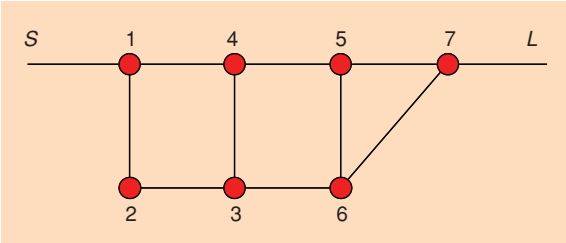


Figure 8. *Coupled resonator network: generalized extended box providing a seven-pole, three-zero asymmetrical characteristic.*

ing the tuning elements (typically tuning screws) of a manufactured prototype.

Extracting coupling parameters from measured or simulated scattering data is an effective approach for tuning, step by step, an EM model or a hardware including tuning elements. Indeed, the comparison between identified parameters and synthesized ones provides an accurate diagnosis of tuning deviation as well as a direction for a better adjustment.

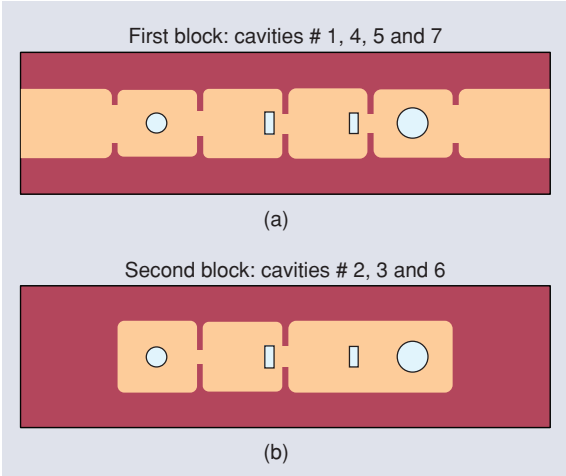


Figure 9. *Seven-pole, three-zero dual-band filter implemented with stacked mono-mode rectangular cavities.*

Pioneering works on computer-aided tuning of microwave filters [21] were based on optimizing the coupling parameters of an equivalent lumped-element model by fitting the measured scattering parameters. However, the efficiency of such a straight approach depends on a favorable initial guess of the coupling parameters, and substantial effort has been spent up to now to propose more robust methods.

Currently, most parameter extraction techniques [22]–[25] consist in, first, deriving a rational approximation of the simulated or measured scattering parameters and, second, synthesizing the resulting lowpass coupled resonator network. A cornerstone of these techniques is clearly the determination of a stable rational model of scattering parameters that coincides with the number of poles and zeros of the polynomial characteristic function [26]. The fundamental problem is to map the simulated or measured scattering parameters, which integrate delays due inherently to in/out coupling systems, with the polynomial formulation that is required for synthesizing the coupled resonator network. A strategy consists of estimating and then removing these delays by adjusting input/output reference planes [26], [27] to reduce the problem to a pure rational approximation problem. Figure 6 compares measured scattering parameters with their rational approximation.

Once a good rational approximation of the scattering parameters is found, the problem becomes once again how to synthesize the lowpass coupled resonator network. In the case of a coupled resonator network leading to a unique coupling matrix-for instance, a canonical network-the synthesis always delivers a single coupling matrix that can be exploited for tuning iteratively the CAD model or the hardware. However, when several coupling matrices result from the synthesis, identifying the proper one is not always obvious, especially when the filter is substantially detuned.

A preliminary selection can then be completed by eliminating coupling matrices whose coupling signs are not consistent with the realized filter. Undeniably, coupling signs are controlled by the arrangement of coupling elements between resonators, and all coupling matrices that cannot correspond to this arrangement can be removed. A further step consists of tracking the evolution of remaining coupling matrices between close tuning steps. In this case, a tuning element is slightly modified to perturb the filter response and, consequently, the coupling matrices. Since the selected tuning element is related to a particular coupling parameter, the proper solution can be recognized by seeking coherency between the tuning element modification and the evolution of coupling parameters within each coupling matrix. This step is done naturally while tuning the filter, but the number of tuning elements that are adjusted at the same time must be limited in this case to follow the proper coupling matrix without ambiguity.

Design Example

Here we use as an example the design of a seven-pole, three-zero dual-band bandpass filter. The two passbands are 50-MHz wide and centered at 8.253 and 8.265 GHz, respectively. The generation of characteristic polynomials from electrical specifications is detailed in [12]. The resultant scattering parameters, normalized in the lowpass frequency domain, are shown in Figure 7.

The topology of the coupled resonator network chosen for realizing the previous characteristic is a generalized extended-box topology presented in Figure 8. As can be verified using the guidelines of the preceding section, this network is compatible with the class of (7, 3) asymmetrical characteristics. Using Dedale-HF, three possible realizations (4)–(6), shown at the bottom of the page, of the ideal response are computed. The solution in (4) is selected since it has the most homogeneous coupling values.

The filter could be constructed with dual-mode resonators (cavities), but this requires a complex coupling system, such as offsetting coupling and resonator elements [28], for controlling both couplings M_{57} and M_{67}. The filter is, therefore, chosen to be implemented using mono-mode rectangular cavities as shown in Figure 9. The structure consists of two stacked blocks,

Several types and implementation technologies of distributed microwave filters are available, and the choice is driven by the application.

each block gathering several cavities and separated by a metallic plate with several coupling apertures. All cavities are excited on their TE_{111} mode, except the sixth cavity, which is excited on its TE_{112} mode for facilitating the coupling with both cavities 5 and 7. Rectangular windows couple the cavities within each block, whereas rectangular or circular apertures are used in the metallic plate for realizing either a magnetic or electric coupling.

The computer-aided design is performed using an EM model of the filter. A preliminary dimensioning stage, using simplified structures, is applied for initializing, respectively, the width of each cavity, the width of each coupling window, and the width or the radius of each coupling aperture with respect to the ideal coupling parameters specified in (4). The dimensions of the EM model are then adjusted more precisely, identifying, at each step, the proper coupling

$$\begin{pmatrix} 0 & 0.8990 & 0 & 0 & 0 & 0 & 0 & 0 & \\ 0.899 & 0.076 & 0.265 & 0 & -0.823 & 0 & 0 & 0 & 0 \\ 0 & 0.265 & -0.961 & 0.121 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0.121 & -0.512 & 0.256 & 0 & 0.366 & 0 & 0 \\ 0 & -0.823 & 0 & 0.256 & 0.151 & 0.434 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0.434 & 0.568 & 0.193 & 0.346 & 0 \\ 0 & 0 & 0 & 0.366 & 0 & 0.193 & -0.220 & 0.793 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0.346 & 0.793 & 0.076 & 0.899 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0.899 & 0 \end{pmatrix} \quad (4)$$

$$\begin{pmatrix} 0 & 0.899 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0.899 & 0.076 & 0.498 & 0 & -0.708 & 0 & 0 & 0 & 0 \\ 0 & 0.498 & -0.018 & 0.098 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0.098 & -0.916 & 0.242 & 0 & -0.161 & 0 & 0 \\ 0 & -0.708 & 0 & 0.242 & 0.078 & 0.666 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0.666 & 0.145 & 0.473 & 0.262 & 0 \\ 0 & 0 & 0 & -0.161 & 0 & 0.473 & -0.264 & 0.824 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0.262 & 0.824 & 0.076 & 0.899 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0.899 & 0 \end{pmatrix} \quad (5)$$

$$\begin{pmatrix} 0 & 0.899 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0.899 & 0.076 & 0.570 & 0 & -0.651 & 0 & 0 & 0 & 0 \\ 0 & 0.570 & -0.292 & 0.442 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0.442 & -0.685 & 0.153 & 0 & 0.072 & 0 & 0 \\ 0 & -0.651 & 0 & 0.153 & 0.305 & 0.595 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0.595 & 0.058 & 0.511 & 0.319 & 0 \\ 0 & 0 & 0 & 0.072 & 0 & 0.511 & -0.361 & 0.804 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0.319 & 0.804 & 0.076 & 0.899 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0.899 & 0 \end{pmatrix} \quad (6)$$

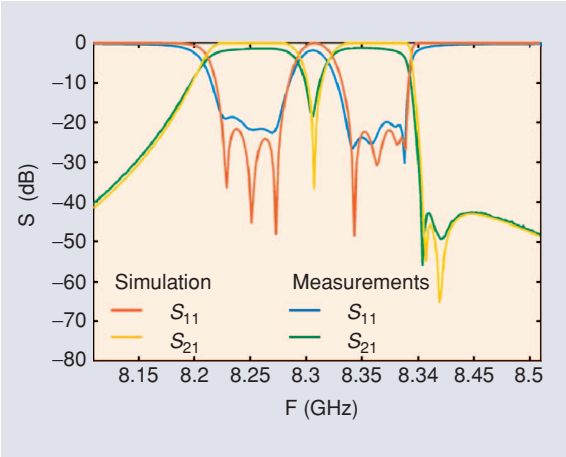


Figure 10. *Simulated and measured scattering parameters of the seven-pole, three-zero dual-band filter.*

parameters from the exhaustive set of solutions as explained in the previous section. One can note that during tuning iterations, the number of extracted coupling matrices fluctuates since the number of real solutions depends on the coefficients of the characteristic polynomials.

The hardware prototype is also tuned using coupling parameter extraction for adjusting tuning screws in each cavity and coupling window. The scattering parameters obtained from the EM model and from the hardware prototype are compared in Figure 10.

Conclusions

This article presented general techniques for the synthesis and design of coupled resonator filters. The synthesis of the prototype circuit focuses on the compatibility between the coupling topology and the filtering function to be realized, providing some guidelines to select a proper coupling topology and to solve the coupling matrix synthesis problem. The dimensioning of the distributed filter is centered on parameter extraction techniques.

References

[1] R.V. Snyder, "Practical aspects of microwave filter development," *IEEE Microwave Mag.*, vol. 8, no 2, pp. 42–54, Apr. 2007.

[2] C. Kudsia, R. Cameron, and W.C. Tang, "Innovation in microwave filters and multiplexing networks for communications satellite systems," *IEEE Trans. Microwave Theory Tech.*, vol. 40, no. 6, pp. 1133–1149, Jun. 1992.

[3] R. Levy, R.V. Snyder, and G.L. Matthaei, "Design of microwave filters," *IEEE Trans. Microwave Theory Tech.*, vol. 50, no. 3, pp. 783–793, Mar. 2002.

[4] G.L. Matthaei, L. Young, and E.M.T. Jones, *Microwave Filters, Impedance-Matching Networks and Coupling Structures.* New York: McGraw-Hill, 1964.

[5] D. Swanson, G. Macchiarella, "Microwave filter design by synthesis and optimization," *IEEE Microwave Mag.*, vol. 8, no. 2, pp. 55–69, Apr. 2007.

[6] H.C. Bell, "The coupling matrix in low-pass prototype filters," *IEEE Microwave Mag.*, vol. 8, no. 2, pp. 70–76, Apr. 2007.

[7] A.E. Atia and A.E. Williams, "Narrow-bandpass waveguide filters," *IEEE Trans. Microwave Theory Tech.*, vol. 20, no. 4, pp. 258–265, Apr. 1972.

[8] R. Cameron, "Synthesis of advanced microwave filters without diagonal cross-couplings," *IEEE Trans. Microwave Theory Tech.*, vol. 50, no. 12, pp. 2862–2871, Dec. 2002.

[9] R. Cameron, "General coupling matrix synthesis methods for Chebyshev filtering functions," *IEEE Trans. Microwave Theory Tech.*, vol. 47, no. 4, pp. 433–442, Apr. 1999.

[10] S. Amari, "Synthesis of cross-coupled resonator filters using an analytical gradient-based optimization technique," *IEEE Trans. Microwave Theory Tech.*, vol. 48, no. 9, pp. 1559–1564, Sept. 2000.

[11] M. Yu, W.-C. Tang, A. Malarky, V. Dokas, R. Cameron, and Y. Wang, "Predistortion technique for cross-coupled filters and its application to satellite communication systems," *IEEE Trans. Microwave Theory Tech.*, vol. 51, no. 12, pp. 2505–2515, Dec. 2003.

[12] V. Lunot, S. Bila, and F. Seyfert, "Optimal synthesis for multiband microwave filters," in *2007 IEEE MTT-S Int. Microwave Symp. Dig.*, pp. 115–118, Jun. 2007.

[13] F. Seyfert "From S to M" tutorial. [Online]. Available: http://www-sop.inria.fr/apics/Dedale/Doc/S2M.html

[14] H.C. Bell, "Canonical asymmetric coupled-resonator filters," *IEEE Trans. Microwave Theory Tech.*, vol. 30, pp. 1335–1340, Sept. 1982.

[15] S. Amari, "On the maximum number of finite transmission zeros of coupled resonator filters with a given topology," *IEEE Microwave Guided Wave Lett.*, vol. 9, no. 9, pp. 354–356, Sept. 1999.

[16] R. Cameron, J.C. Faug`ere, F. Roullier, and F. Seyfert, "Exhaustive approach to the coupling matrix synthesis problem and application to the design of high degree asymmetric filters," *Int. J. RF Microwave Computer Aided Eng.*, vol. 17, no. 1, pp. 4–12, Jan. 2007.

[17] W.A. Atia, K.A Zaki, and A.E Atia, "Synthesis of general topology multiple coupled resonators filters by optimization," in *IEEE MTT-S Int. Symp. Dig.*, Sept. 1998, vol. 2, pp 1693–1698.

[18] S. Tamiazzo and G. Macchiarella, "An analytical technique for the synthesis of cascaded N-tuplets cross-coupled resonators microwave filters using matrix rotations," *IEEE Trans. Microwave Theory Tech.*, vol. 53, no. 5, pp. 1693–1698, May 2005.

[19] Dedale-HF's page. [Online]. Available: http://www.sop-inria.fr/apics/Dedale

[20] P. Lenoir, S.Bila, F.Seyfert, D. Baillargeat, and S.Verdeyme, "Synthesis of assymetrical dual-band bandpass filter based on equivalent network simplification," *IEEE Trans. Microwave Theory Tech.*, vol. 54, no. 7, pp. 3090–3097, Jul. 2006.

[21] H.L. Thal, "Computer-aided filter alignment and diagnosis," *IEEE Trans. Microwave Theory Tech.*, vol. 26, no. 12, pp. 958–963, Dec. 1978.

[22] A. Garcia-Lamperez, S. Llorente-Romano, M. Salazar-Palma, and T.K. Sarkar, "Efficient electromagnetic optimization of microwave filters and multiplexers using rational models," *IEEE Trans. Microwave Theory Tech.*, vol. 52, no. 2, pp. 508–521, Feb. 2004.

[23] P. Kozakowski and M. Mrozowski, "Quadratic programming approach to coupled resonator filter CAD," *IEEE Trans. Microwave Theory Tech.*, vol. 54, no. 11, pp. 3906–3913, Nov. 2006.

[24] S. Bila, D. Baillargeat, M. Aubourg, S. Verdeyme, P. Guillon, F. Seyfert, J. Grimm, L. Baratchart, C. Zanchi, and J. Sombrin, "Direct electromagnetic optimization of microwave filters," *IEEE Microwave Mag.*, vol. 2, no. 1, pp. 46–51, Mar. 2001.

[25] P. Harscher, R. Vahldieck, and S. Amari, "Automated filter tuning using generalized low-pass prototype networks and gradient-based parameter extraction," *IEEE Trans. Microwave Theory Tech.*, vol. 49, no. 12, pp. 2532–2538, Dec. 2001.

[26] F. Seyfert, L. Baratchart, J.P. Marmorat, S. Bila, and J. Sombrin, "Extraction of coupling parameters for microwave filters: Determination of a stable rational model from scattering data," in *2003 IEEE MTT-S Int. Microwave Symp. Dig.*, pp. 25–28, June 2003.

[27] F.T. Hsu, Z. Zhang, K.A. Zaki, and A.E. Atia, "Parameter extraction for symmetric coupled-resonator filters," *IEEE Trans. Microwave Theory Tech.*, vol. 50, no. 12, pp. 2971–2978, Dec. 2002.

[28] R.J. Cameron, "Dual-mode realisation for asymmetric filter characteristics," *ESA J.*, vol. 6, no. 3, pp. 339–356, 1982.

5.3.3 Design simplification via exhaustive solving of the CM synthesis problem

Following paper is reproduced in this section:

- Philippe Lenoir, Stéphane Bila, Fabien Seyfert, Dominique Baillargeat, and Serge Verdeyme. "Synthesis and design of asymmetrical dual-band bandpass filters based on equivalent network simplification". In: *IEEE Transactions on Microwave Theory and Techniques* 54.7 (2006), pp. 3090–3097. DOI: 10.1109/TMTT.2006.877037. URL: https://hal.inria.fr/hal-00663496

Synthesis and Design of Asymmetrical Dual-Band Bandpass Filters Based on Equivalent Network Simplification

Philippe Lenoir, Stéphane Bila, Fabien Seyfert, Dominique Baillargeat, *Member, IEEE*, and Serge Verdeyme, *Member, IEEE*

***Abstract*—Although the synthesis of symmetrical dual-band bandpass filters has been studied, little seems to be known about the general asymmetrical case. In this paper, a procedure for the synthesis of general asymmetrical dual-band bandpass filters implemented with inline dual-mode cavities is proposed. The inline architecture, as well as the asymmetrical nature of the response, lead naturally to the choice of an extended box topology or generalizations of the latter. It was recently shown that these topologies possess the property of multiple solutions, meaning that the related coupling matrix synthesis problem admits several solutions. On one hand, this multiplicity offers some flexibility to the designer, but on the other hand, working with multiple solutions may lead to ambiguities during the tuning process. Our procedure takes the best of both worlds: using the list of equivalent coupling matrices, simplifications of the original topology can be obtained by canceling some particular couplings. The locations of cancelled couplings are chosen so as to preserve the electrical response and to provide some hardware simplifications. It is also shown that the resulting simplified topology no longer has the multiple solution property, therefore solving ambiguity problems. The procedure is detailed and demonstrated on two examples.**

***Index Terms*—Circuit synthesis, coupling matrix, dual-band filters, microwave filters.**

I. Introduction

THE DEMAND for advanced filtering functions has considerably increased with the development of space telecommunications. For example, in satellite communication systems, highly selective transfer functions with self-equalized group delays are required for input multiplexer (IMUX) channels. Another emerging application in this domain is the design of dual-band bandpass filters used to transmit noncontiguous channels to the same geographical region through one beam [1]. In this case, a single high-power amplifier (HPA) can be used together with the dual-band bandpass filter, dramatically simplifying the system architecture.

An approach for implementing such a circuit consists of designing two classical single-band bandpass filters, one for each passband. Their input/output ports are then connected together through waveguide junctions. However, this approach leads to a complex design procedure since waveguide junctions and filters have to be optimized together to comply with the mechanical constraints. Indeed, each channel must have the same length, and input/output waveguide ports must have the same orientation. Another approach consists of designing a single circuit realizing the dual-band characteristic. This straightforward approach requires the synthesis of an advanced filtering function, but makes the hardware implementation easier since a classical filter architecture can be used.

Narrowband filters, dedicated to space applications, are generally implemented using cavities or resonators since they offer better performances in terms of losses and power handling, typically up to 150 W in output multiplexer (OMUX) channel filters. For reducing mass and volume, these resonant elements are often excited on dual modes. Furthermore, nonadjacent couplings between resonant elements are generally required in order to add transmission zeroes to the transfer function for improving the selectivity and/or flattening the group delay. A practical way to implement a filter with dual-mode cavities, while permitting nonadjacent couplings, is the inline architecture, which consists of connecting dual-mode cavities in a row. The latter architecture is used in this study for implementing general asymmetrical dual-band bandpass filters with no restriction on the placement of transmission zeroes on the frequency axis, as opposed, for example, to [3].

The synthesis of a microwave filter starts with the selection of a transfer function that fulfills the electrical specifications. For single-band filtering characteristics, quasi-elliptic polynomial functions given by explicit formulas are widely employed [2]. For symmetrical dual-band characteristics of even degree and with an even number of transmission zeroes, the latter formulas may be adapted by means of frequency transformations [3]. This is no longer the case for more general situations and gets designers to use direct optimization methods [1], [4]–[6]. In this study, a local optimization method detailed in [6] is applied where the starting point is computed from the quasi-elliptic synthesis of each individual channel.

In a second step, an equivalent lumped-element network is synthesized in order to realize the selected transfer function. The equivalent network is characterized by its coupling topology, specifying the distribution of zero and nonzero couplings between resonators. The latter has to obviously be consistent with the filter architecture, as well as with the transfer function it is supposed to realize. The main difficulty comes here from the conflicting requirements of the design; on one hand, a topology able to realize several asymmetric transmission zeroes, while on the other hand, the latter should remain simple enough to admit a classical hardware implementation. In our case, the hardware

Manuscript received October 19, 2005; revised February 7, 2006.

P. Lenoir, S. Bila, D. Baillargeat, and S. Verdeyme are with the Institut de Recherche en Communications Optiques et Microondes, 87060 Limoges, France (e-mail: bila@ircom.unilim.fr).

F. Seyfert is with the Institut National de Recherche en Informatique et Automatique, 06902 Sophia Antipolis, France (e-mail: seyfert@sophia.inria.fr).

Digital Object Identifier 10.1109/TMTT.2006.877037

0018-9480/$20.00 © 2006 IEEE

implementation is an inline dual-mode cavity architecture ideally with no cross-couplings. This will lead us to consider the use of extended box coupling topologies [7] that are consistent with general asymmetrical characteristics. Generalizations of latter topologies are also proposed in this study in order to synthesize some extra transmission zeroes. For all these topologies, it was shown recently that the number of solutions to the coupling matrix synthesis problem is high [8]. This differs from [6] where the symmetrical nature of the filtering function allowed the use of a canonical topology with a single solution to the coupling matrix synthesis problem. In the current study, an exact and exhaustive synthesis method [8] is used to determine all the solutions to the coupling matrix synthesis problem. Furthermore, this study demonstrates how to use this extra flexibility by providing rules to be applied in order to obtain significant simplifications of the original topology while keeping the electrical response nearly unchanged. The latter translates into hardware simplifications like transformation of cross-irises into single-arm irises or realignment of irises with respect to the cavities. Such a simplification approach has been employed in [9] for implementing a symmetrical filter architecture while realizing an asymmetrical single-band transfer function.

Finally, the proposed approach is also consistent with numerical modeling techniques, which are often used by designers. These methods are used along with a coupling matrix extraction algorithm [10]–[13]: this allows driving the tuning process, so as to converge towards a device implementing the ideal coupling matrix. Nevertheless, as opposed to the classical situation dealing with canonical topologies [6], [9]–[13], when working with topologies, which admit multiple solutions, the latter coupling matrix extraction step returns a list of several equivalent coupling matrices. This leaves the difficult task of choosing the right one to the designer and, thus, represents the main drawback of topologies with multiple solutions. The current study shows how hardware simplifications obtained in the preceding step of our procedure solve the addressed ambiguity and allow us to use a tuning process based on a well-posed coupling matrix extraction problem.

The approach is illustrated by the design of two asymmetrical dual-band filters at Ka-band. In Section II, an 11-pole dual-band bandpass filter with four transmission zeroes is designed. The synthesis procedure is presented from the determination of the characteristic function, up to the simplified network construction. The latter network allows the simplification of cross-irises into single-arm irises without any notable effect on the circuit behavior. A numerical model and an experimental model are also investigated in order to demonstrate the efficiency of the proposed approach. In Section III, the approach is repeated, synthesizing an 11-pole dual-band bandpass filter with five zeroes. Here, the simplified network allows realignment of all the distributed elements for an easier hardware implementation of the inline dual-mode cavity filter.

II. Design of an 11-Pole Four-Zero Dual-Band Filter

The electrical specifications of the dual-band bandpass filter to be designed are: 1) a first passband centred at 18.362 GHz with a 39-MHz bandwidth and 2) a second passband centred at 18.508 GHz with a 78.5-MHz bandwidth.

A 20-dB return loss in each passband and a 10-dB insertion loss in the intermediate stopband are required. An insertion loss greater than 25 dB is also desired in the lower and upper stopbands. Starting from these electrical specifications, the transfer and reflection functions of the dual-band filter are calculated.

A. Characteristic Function Selection

The characteristic function $D(s)$ can be written as a polynomial rational function

$$D(s) = \frac{\prod_{i=1}^{N}(s - s_{ri})}{\prod_{i=1}^{Nz}(s - s_{pi})} = \frac{R(s)}{P(s)} \tag{1}$$

where s_{ri} and s_{pi} are, respectively, the normalized reflection and transmission zeroes, N is the number of reflection zeroes, i.e., the order of the filtering function, and Nz is the number of transmission zeroes.

The modulus of the transfer function admits the following simple expression in terms of its characteristic function:

$$|S_{21}(s)|^2 = \frac{1}{1 + \varepsilon^2 |D(s)|^2} \tag{2}$$

where ε is an adjustable real parameter.

Applying the procedure described in [6], the initial dual-band characteristic function is constructed from two single-band functions. In our case, the lower passband is realized with a fifth-order quasi-elliptic function and the upper passband is realized with a sixth-order quasi-elliptic function. Each single-band characteristic function presents two transmission zeroes for improving the selectivity in the stopbands.

As a result, the dual-band characteristic function is initialized with four transmission zeroes and 11 reflection zeroes. The initial reflection and transmission zeroes are then slightly retuned in order to improve the transmission feature (2) within the two passbands. Since the attenuation level of each single-band transfer function is high in the other passband, the local optimization method converges rapidly, leading to the following normalized values:

$$\begin{aligned}
s_{p1} &= -j1.063 \\
s_{p2} &= -j0.565 \\
s_{p3} &= j0.205 \\
s_{p4} &= j1.063 \\
s_{r1} &= -j0.996 \\
s_{r2} &= -j0.943 \\
s_{r3} &= -j0.823 \\
s_{r4} &= -j0.682 \\
s_{r5} &= -j0.610 \\
s_{r6} &= j0.261 \\
s_{r7} &= j0.354 \\
s_{r8} &= j0.557 \\
s_{r9} &= j0.774 \\
s_{r10} &= j0.922 \\
s_{r11} &= j0.985
\end{aligned} \tag{3}$$

with $\varepsilon = 47$.

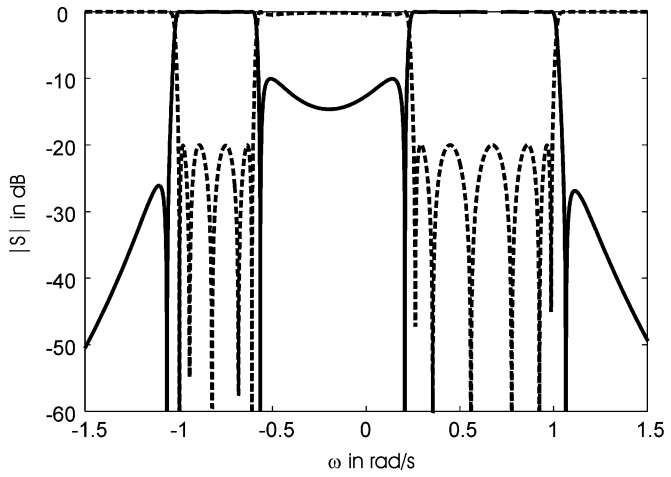


Fig. 1. Ideal 11-pole four-zero transfer (—) and reflection (- - -) functions.

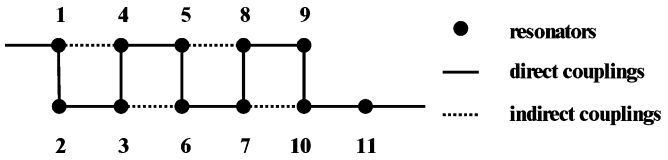


Fig. 2. Coupling topology realizing 11-pole four-zero transfer functions.

The transfer and reflection functions corresponding to these values are presented in Fig. 1.

B. Exact Synthesis

An equivalent lumped-element network that realizes the previous transfer function is now synthesized. The related coupling topology has to be adapted to the exact synthesis of the desired asymmetrical characteristic, as well as be compatible with the inline dual-mode cavity architecture. The extended box topology presented in Fig. 2 meets our requirements since this coupling topology allows us to realize any asymmetrical transfer function of order 11 with four transmission zeroes.

The exhaustive synthesis method presented in [8] is applied in order to determine all the coupling matrices that correspond to the previous coupling topology. The method is based on computations that exhaustively solve an algebraic system of equations related to the synthesis problem.

Generically, an 11-pole four-zero transfer function can be realized in 384 manners with the above coupling topology. This theoretical number, called the reduced order, is the number of complex solutions to the synthesis problem and does not depend on the considered filtering characteristic (only on the coupling topology). Nevertheless, the number of real solutions, i.e., the only ones of physical interest, depends on the numerical values of the characteristic polynomials. For our particular filtering characteristic, 66 real solutions are found.

Theoretically, any solution among these 66 could be chosen to design the filter. Our goal is now to use this extra flexibility in order to simplify the initial coupling topology by cancelling one or several couplings without severely affecting the electrical response. To this end, some rules have to be observed when seeking to simplify the coupling topology, which are as follows.

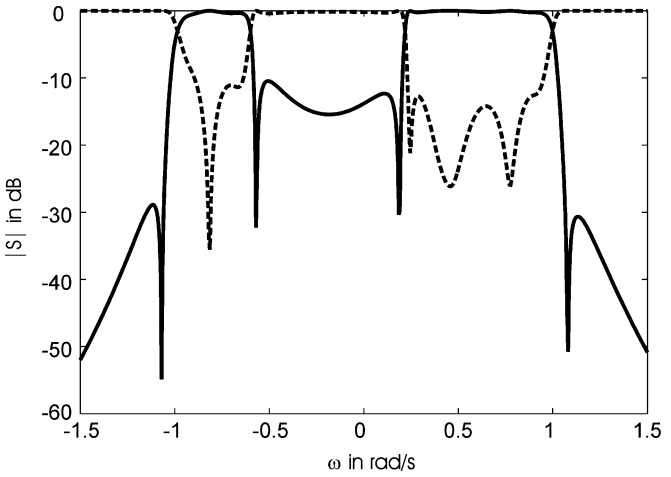


Fig. 3. 11-pole four-zero transfer (—) and reflection (- - -) functions when couplings M_{14} and M_{58} are neglected (no compensation).

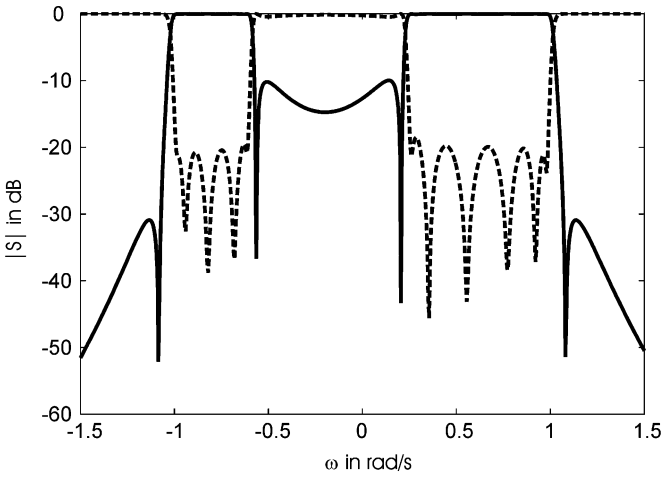


Fig. 4. 11-pole four-zero transfer (—) and reflection (- - -) functions when couplings M_{14} and M_{58} are compensated.

- The number of couplings in the shortest coupling path, between source and load, has to be preserved in order to keep the number of transmission zeroes constant.
- Couplings corresponding to irises are cancelled in priority since couplings realized with screws can hardly be completely set to zero in practice because of remaining residual couplings.

The latter rule indicates that, starting from the coupling diagram in Fig. 1, our simplification will apply only to cancel one or several horizontal couplings. The shortest path rule imposes some conditions on cancelable horizontal couplings. For example, if coupling M_{14} (between resonators 1–4) is cancelled, all the couplings in the inferior path ($M_{23}, M_{36}, M_{67}, M_{710}$) needs to remain nonzero.

C. Approximate Synthesis With a Simplified Network

Following latter rules, solutions with low cross couplings M_{14} and M_{58} are explored. A good candidate, out of all 66 matrices, is the first matrix shown at the bottom of the following page. The cancellation of M_{14} and M_{58} modifies the resulting transfer and reflection functions, as shown in Fig. 3, but compensating this effect with the remaining couplings, the original transfer function is almost recovered, as shown in Fig. 4. The

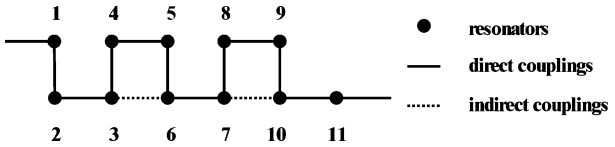


Fig. 5. Simplified coupling topology leading to the 11-pole four-zero approximate transfer function in Fig. 4.

coupling topology is the one presented in Fig. 5 and the final simplified coupling matrix is the second matrix shown at the bottom of this page.

Our approach leads to the simplification of two cross-coupling irises into single-arm irises. Moreover, the tuning process through coupling matrix extraction will be simplified since the reduced order of the simplified topology is one as computed with methods detailed in [8]. More precisely, the previous simplified topology does not allow to realize all the transfer functions of eleventh order with four transmission zeroes, but when the latter is a realizable one, there corresponds only one coupling matrix. In other words, the original transfer function in Fig. 1 cannot be realized exactly with the simplified coupling topology, but the approximate transfer function in Fig. 4 can only be realized with the final simplified coupling matrix. To summarize the approach, the approximate synthesis yields some important hardware simplifications, while at the same time solving identifiability problems inherent to the use of topologies with multiple solutions.

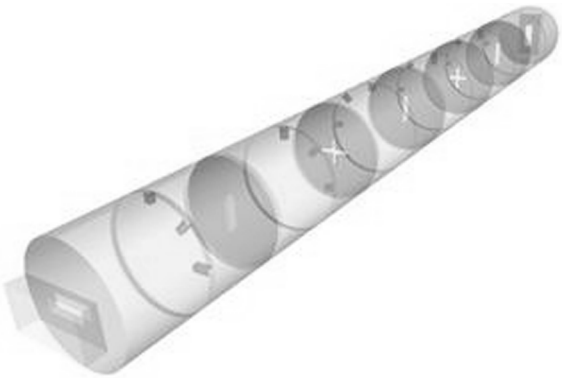


Fig. 6. Electromagnetic model of the 11-pole four-zero asymmetrical dual-band bandpass filter.

D. Electromagnetic optimization

The electromagnetic model of the inline dual-mode cavity filter is presented in Fig. 6. The filter is designed for the TE_{113} mode, applying the electromagnetic optimization procedure presented in [10]. Each electromagnetic analysis is followed by a coupling matrix extraction step yielding some corrections on the modeled geometrical dimensions.

The transfer and reflection functions obtained from the electromagnetic model are given in Fig. 7. The numerical model behavior is slightly different from the ideal one since parasitic couplings between resonant elements have been compensated [14].

$$R\mathrm{in} = R\mathrm{out} = 0.563$$

$$\begin{pmatrix}
-0.115 & 0.773 & 0 & -0.079 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0.773 & 0.142 & 0.521 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0.521 & -0.221 & 0.514 & 0 & -0.264 & 0 & 0 & 0 & 0 & 0 \\
-0.079 & 0 & 0.514 & 0.296 & 0.697 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0.697 & -0.305 & 0.425 & 0 & 0.094 & 0 & 0 & 0 \\
0 & 0 & -0.264 & 0 & 0.425 & 0.318 & 0.507 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0.507 & -0.323 & 0.189 & 0 & 0.390 & 0 \\
0 & 0 & 0 & 0 & 0.094 & 0 & 0.189 & 0.482 & 0.314 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0.314 & -0.101 & 0.267 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0.390 & 0 & 0.267 & 0.144 & 0.777 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0.777 & -0.115
\end{pmatrix}$$

$$R\mathrm{in} = R\mathrm{out} = 0.571$$

$$\begin{pmatrix}
-0.106 & 0.781 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0.781 & 0.147 & 0.476 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0.476 & -0.243 & 0.478 & 0 & -0.267 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0.478 & 0.315 & 0.725 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0.725 & -0.313 & 0.422 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & -0.267 & 0 & 0.422 & 0.320 & 0.555 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0.555 & -0.297 & 0.149 & 0 & 0.399 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0.149 & 0.472 & 0.300 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0.300 & -0.113 & 0.245 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0.399 & 0 & 0.245 & 0.130 & 0.787 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0.787 & -0.106
\end{pmatrix}$$

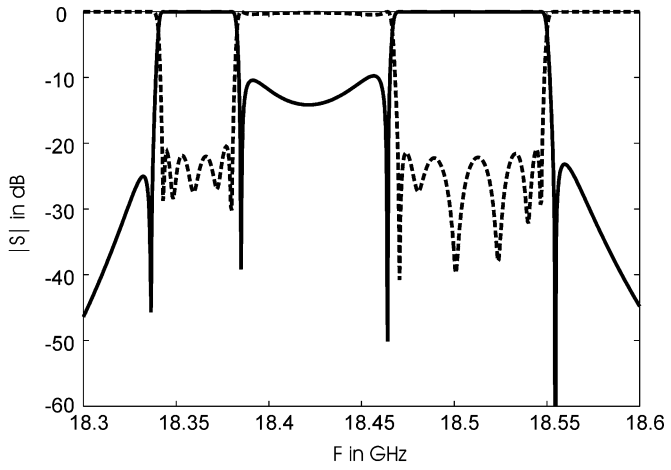


Fig. 7. 11-pole four-zero transfer (—) and reflection (- - -) functions obtained with the electromagnetic model.

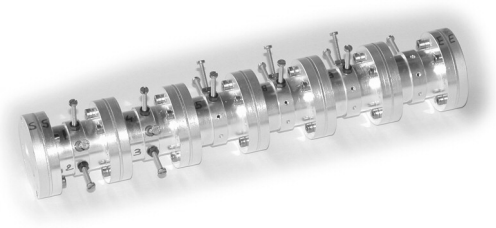


Fig. 8. Realized 11-pole four-zero asymmetrical dual-band bandpass filter.

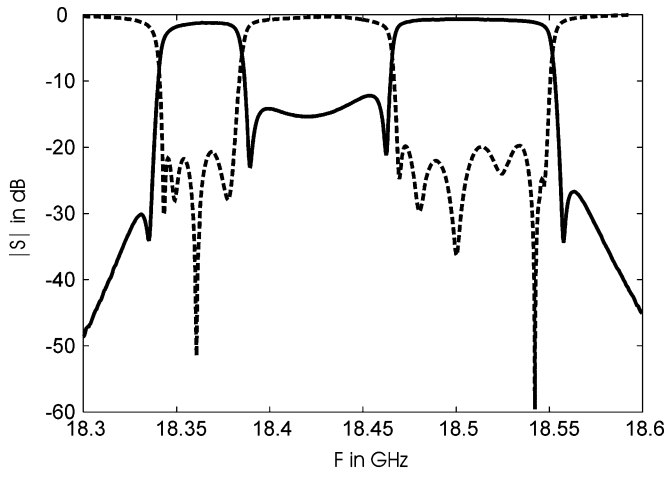


Fig. 9. Experimental 11-pole four-zero transfer (—) and reflection (- - -) functions.

The critical parameters governing the behavior of the structure are the dimensions of coupling irises and cavities. The sensitivities to the latter parameters are found to be consistent with standard manufacturing tolerances (around 10 μm).

E. Measurements

The filter has been built and tested. A photograph of the realized prototype is presented in Fig. 8. The measured transfer and reflection functions are presented in Fig. 9. The first passband is 39-MHz wide and is centred at 18.362 GHz. The second passband is 81-MHz wide and is centered at 18.508 GHz. The insertion losses are 1.2 and 0.6 dB, respectively, in the first and second passbands. Therefore, a good agreement is achieved between theory and measurement, validating our design approach with simplified coupling topologies.

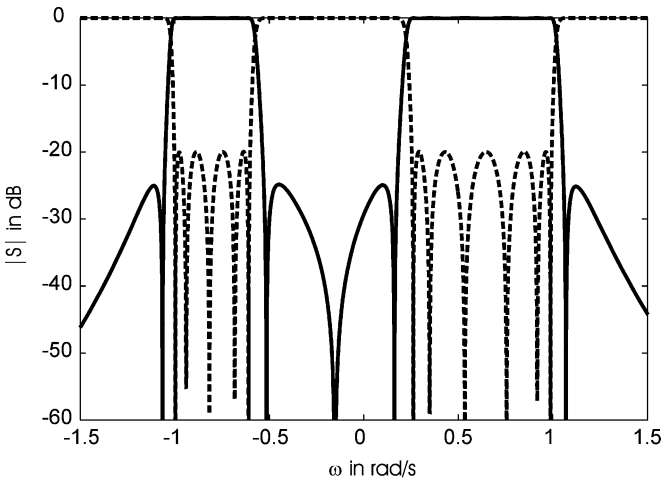


Fig. 10. Ideal 11-pole five-zero transfer (—) and reflection (- - -) functions.

III. Synthesis of an 11-Pole Five-Zero Dual-Band Filter

In order to deepen the intermediate stopband, a fifth transmission zero is added to the transfer function. The insertion loss in the intermediate stopband is then specified to be 25 dB. The previous synthesis procedure is repeated.

A. Characteristic Function Selection

Applying the same method, the following transmission and reflection zeros are computed:

$$\begin{aligned}
s_{p1} &= -j\,1.065\\
s_{p2} &= -j\,0.515\\
s_{p3} &= -j\,0.153\\
s_{p4} &= j\,0.161\\
s_{p5} &= j\,1.067\\
s_{r1} &= -j\,0.996\\
s_{r2} &= -j\,0.940\\
s_{r3} &= -j\,0.818\\
s_{r4} &= -j\,0.683\\
s_{r5} &= -j\,0.610\\
s_{r6} &= j\,0.261\\
s_{r7} &= j\,0.346\\
s_{r8} &= j\,0.532\\
s_{r9} &= j\,0.753\\
s_{r10} &= j\,0.915\\
s_{r11} &= j\,0.985
\end{aligned} \tag{4}$$

with $\varepsilon = 40$.

The transfer and reflection functions corresponding to these values are presented in Fig. 10.

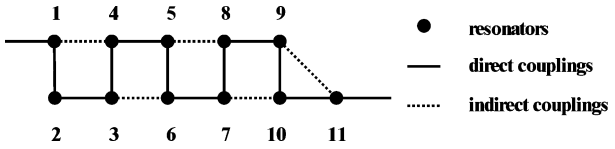


Fig. 11. Coupling topology realizing 11-pole five-zero transfer functions.

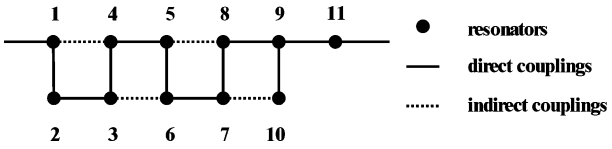


Fig. 12. Simplified coupling topology proposed for realizing the 11-pole five-zero approximate transfer function.

B. Exact Synthesis

In order to realize and adjust an additional transmission zero, the extended-box topology needs to be modified while keeping in mind the following facts.

- One degree of freedom, i.e., one extra coupling, must be added to the actual extended box topology in order to enable to adjust the position of the new transmission zero.
- The shortest path between resonators 1 and 11 in the new coupling topology must be of length 5 to satisfy the minimum path rule.

The latter requirements are met by the coupling topology in Fig. 11 by adding cross-coupling M_{911} to the original extended box (Fig. 2). This coupling topology allows us to realize any transfer function of order 11 with five transmission zeroes.

The reduced order of this topology is found to be 963, and applying an exhaustive synthesis from the selected transfer function leads to 81 real coupling matrices.

The cross-coupling M_{911} is the main problem of the above coupling topology, as an angle is necessary between the last coupling iris and the last cavity in order to realize it. Our approach will, therefore, focus on simplifications of the coupling topology that allow a realization with aligned irises and cavities.

Obviously, the rules given for the first example still hold valid. However, one can note that now the shortest coupling path is unique. Consequently, none of the following couplings $M_{14}, M_{45}, M_{58}, M_{89}$, and M_{911} can be cancelled.

In order to recover an aligned architecture, a possible way is to suppress coupling M_{1011}. Indeed, the latter cancellation will lead to the simplified coupling topology presented in Fig. 12.

C. Approximate Synthesis With a Simplified Network

Following our approach based on approximate synthesis, solutions with low cross-coupling M_{1011} are investigated. The coupling matrix shown at the bottom of this page is then a good candidate.

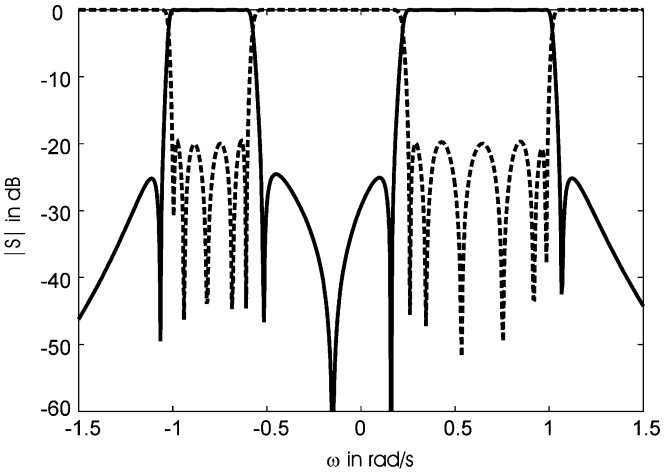


Fig. 13. Approximate 11-pole five-zero transfer (—) and reflection (- - -) functions (neglected coupling M_{1011} is compensated).

Considering this late matrix, couplings M_{34} and M_{78} also have weak values and should also be cancelled applying the approximate synthesis; but since these couplings are implemented with coupling screws, they are preserved in the simplified coupling topology.

Neglecting the coupling M_{1011}, the resulting transfer and reflection functions are only slightly modified, and by compensating with the remaining couplings, the original transfer function is recovered as shown in Fig. 13. The final coupling matrix, which is consistent with the coupling topology presented in Fig. 12, is then as shown at the top of the following page.

Applying this approach, the hardware implementation is highly simplified since all the irises and cavities are aligned.

$$R\text{in} = R\text{out} = 0.563$$

$$\begin{pmatrix}
-0.116 & 0.352 & 0 & -0.694 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0.352 & -0.588 & 0.309 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0.309 & -0.612 & 0.046 & 0 & 0.262 & 0 & 0 & 0 & 0 & 0 \\
-0.694 & 0 & 0.046 & 0.332 & 0.308 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0.308 & -0.038 & 0.653 & 0 & -0.311 & 0 & 0 & 0 \\
0 & 0 & 0.262 & 0 & 0.653 & 0.166 & 0.174 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0.174 & 0.813 & 0.074 & 0 & 0.152 & 0 \\
0 & 0 & 0 & 0 & -0.311 & 0 & 0.074 & -0.461 & 0.430 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0.430 & 0.146 & 0.198 & 0.778 \\
0 & 0 & 0 & 0 & 0 & 0 & 0.152 & 0 & 0.198 & 0.726 & -0.004 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0.778 & -0.004 & -0.116
\end{pmatrix}$$

$$R\text{in} = R\text{out} = 0.563$$

$$\begin{pmatrix} -0.116 & 0.352 & 0 & -0.694 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0.352 & -0.588 & 0.309 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0.309 & -0.612 & 0.046 & 0 & 0.262 & 0 & 0 & 0 & 0 & 0 \\ -0.694 & 0 & 0.046 & 0.331 & 0.307 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0.307 & -0.038 & 0.653 & 0 & -0.311 & 0 & 0 & 0 \\ 0 & 0 & 0.262 & 0 & 0.653 & 0.166 & 0.174 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0.174 & 0.813 & 0.075 & 0 & 0.152 & 0 \\ 0 & 0 & 0 & 0 & -0.311 & 0 & 0.075 & -0.460 & 0.430 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0.430 & 0.145 & 0.196 & 0.778 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0.152 & 0 & 0.196 & 0.727 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0.778 & 0 & -0.116 \end{pmatrix}$$

Moreover, the tuning process through coupling matrix extraction will be also simplified since, here again, the reduced order of our simplified topology is found to be one.

IV. Conclusion

This paper has presented an approach based on equivalent network simplification for the synthesis and the design of asymmetrical dual-band bandpass filters implemented with inline dual-mode cavities. The first step involves an exact and exhaustive synthesis yielding a list of equivalent coupling matrices that are consistent with the extended box coupling topology or variations of it. In a second step, the proposed approach takes advantage of the multiple solution property of these coupling topologies by providing some rules for selecting a coupling matrix to be used as the starting point for an approximate synthesis procedure. The approximate synthesis allows then some substantial simplifications of the initial coupling topology by cancelling one or several weak couplings between resonators. The simplified coupling topology makes the hardware implementation easier and also solves ambiguity problems that may occur during the tuning phase by restoring the well posedness of the coupling matrix extraction step.

The proposed approach is applied to synthesize and design two asymmetrical dual-bandpass filters implemented with inline dual-mode cavities. When applied to an 11-pole four-zero microwave filter, the proposed approach allows to replace two cross irises by single-arm irises when compared with an exact synthesis. A numerical model and an experimental prototype of this filter have been fabricated in order to validate the theoretical results. The approach is repeated with an 11-pole five-zero microwave filter and the approximate synthesis allows realignment of all of the distributed elements compared with an exact synthesis.

Acknowledgment

The authors would like to acknowledge J. Puech, C. Zanchi, and J. Sombrin, all with the Centre National d'Etudes Spatiales (CNES), Toulouse, France, for supporting this study.

References

[1] S. Holme, "Multiple passband filters for satellite applications," in *Proc. 20th AIAA Int. Commun. Satellite Syst. Conf. Exhibit*, 2002, pp. 1993–1996.

[2] R. J. Cameron, "General coupling matrix synthesis methods for Chebyshev filtering functions," *IEEE Trans. Microw. Theory Tech.*, vol. 47, no. 4, pp. 433–442, Apr. 1999.

[3] G. Macchiarella and S. Tamiazzo, "A design technique for symmetric dualband filters," in *IEEE MTT-S Int. Microw. Symp. Dig.*, Long Beach, CA, Jun. 2005, 4 pp.

[4] J. Lee, M. S. Uhm, and I. B. Yom, "A dual-passband filter of canonical structure for satellite applications," *IEEE Microw. Wireless Compon. Lett.*, vol. 14, no. 6, pp. 271–273, Jun. 2004.

[5] J. Lee, M. S. Uhm, and J. S. Park, "Synthesis of self-equalized dual-passband filter," *IEEE Microw. Wireless Compon. Lett.*, vol. 15, no. 4, pp. 256–258, Apr. 2005.

[6] P. Lenoir, S. Bila, D. Baillargeat, and S. Verdeyme, "Design of dual-band bandpass filters for space applications," presented at the Proc. Eur. Microw. Assoc., Sep. 2005, accepted for publication.

[7] R. J. Cameron and A. R. Harish and C. J. Radcliffe, "Synthesis of advanced microwave filters without diagonal cross-couplings," *IEEE Trans. Microw. Theory Tech.*, vol. 50, no. 12, pp. 2862–2872, Dec. 2002.

[8] F. Seyfert, R. Cameron, and J. C. Faugère, "Coupling matrix synthesis for a new class of microwave filter configuration," in *IEEE MTT-S Int. Microw. Symp. Dig.*, Long Beach, CA, Jun. 2005, 4 pps.

[9] S. Bila, D. Baillargeat, S. Verdeyme, F. Seyfert, L. Baratchart, C. Zanchi, and J. Sombrin, "Simplified design of microwave filters with asymmetric transfer functions," in *Eur. Microw. Conf.*, Munich, Oct. 2003, pp. 1357–1360.

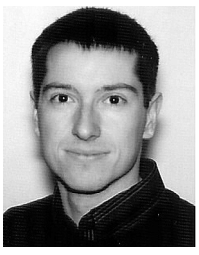
[10] S. Bila, D. Baillargeat, S. Verdeyme, M. Aubourg, P. Guillon, F. Seyfert, J. Grimm, L. Baratchart, C. Zanchi, and J. Sombrin, "Direct electromagnetic optimization of microwave filters," *IEEE Micro*, vol. 2, no. 1, pp. 46–51, Mar. 2001.

[11] P. Harsher, R. Vahldieck, and S. Amari, "Automated filter tuning using generalized low-pass prototype networks and gradient-based parameter extraction," *IEEE Trans. Microw. Theory Tech.*, vol. 49, no. 12, pp. 2532–2538, Dec. 2001.

[12] M. Kahrizi, S. Safavi-Naeini, S. K. Chaudhuri, and R. Sabry, "Computer diagnosis and tuning of RF and microwave filters using model-based parameter estimation," *IEEE Trans. Circuits Syst. I, Fundam. Theory Appl.*, vol. 49, no. 9, pp. 1263–1270, Sep. 2002.

[13] A. Garcia-Lamperez, S. Llorente-Romano, M. Salazar-Palma, and T. K. Sarkar, "Efficient electromagnetic optimization of microwave filters and multiplexers using rational models," *IEEE Trans. Microw. Theory Tech.*, vol. 52, no. 2, pp. 508–521, Feb. 2004.

[14] S. Bila, D. Baillargeat, M. Aubourg, S. Verdeyme, F. Seyfert, L. Baratchart, C. Boichon, F. Thevenon, J. Puech, C. Zanchi, L. Lapierre, and J. Sombrin, "Finite element modelling for the design optimization of microwave filters," *IEEE Trans. Magn.*, vol. 40, no. 2, pp. 472–475, Mar. 2004.



Philippe Lenoir was born in Limoges, France, in September 1979. He received the Ph.D. degree in high-frequency electronics and opto-electronics from the University of Limoges, Limoges, France.

He is currently with the Institut National de Recherche en Informatique et Automatique (INRIA), Sophia Antipolis, France, within a post-doctoral position. His research interests include the synthesis method based on computer-aided techniques for the design and optimization of microwave components and circuits for space applications.

Stéphane Bila was born in Paris, France, in September 1973. He received the Ph.D. degree from the University of Limoges, Limoges, France, in 1999.

He then held a post-doctoral position for one year with the Centre National d'Etudes Spatiales (CNES), Toulouse, France. In 2000, he became a Researcher with the Centre National de la Recherche Scientifique (CNRS). He then joined the Microwave Circuits and Devices Team with the Institut de Recherche en Communications Optiques et Microondes (IRCOM), Limoges, France. His research interests include numerical modeling and computer-aided techniques for the advanced synthesis and design of microwave components and circuits.

Fabien Seyfert received the Engineer degree from the Ecole Superieure des Mines, St. Etienne, France, in 1993, and the Ph.D. degree in mathematics from the Ecole Superieure des Mines, Paris, France, in 1998.

From 1998 to 2001, he was with Siemens, Munich, Germany, where he was a Researcher specializing in discrete and continuous optimization methods. Since 2002, he has had a full research position with the Institut National de Recherche en Informatique et Automatique (INRIA), Sophia Antipolis, France. His research interest focuses on the conception of effective mathematical procedures and associated software for problems from signal processing, including computer-aided techniques for the design and tuning of microwave devices.

Dominique Baillargeat (M'04) was born in Le Blanc, France, in 1967. He received the Ph.D. degree from the Institut de Recherche en Communications Optiques et Microondes (IRCOM), University of Limoges, Limoges, France, in 1995.

From 1995 to 2005, he was an Associate Professor with the Microwave Circuits and Devices Team, IRCOM Laboratory. He is currently a Professor. His fields of research concern the development of methods of design for microwave devices. These methods include computer-aided design (CAD) techniques based on hybrid approach coupling electromagnetic, circuits and thermal analysis, synthesis and electromagnetic optimization techniques, etc. He is mainly dedicated to the packaging of millimeter wave and opto-electronics modules and to the design of millimeter original filters based on new topologies, concepts (electromagnetic bandgap (EBG), etc.) and/or technologies (silicon, low-temperature co-fired ceramic (LTCC), etc.).

Serge Verdeyme (M'99) was born in Meilhards, France, in June 1963. He received the Doctorat degree from the University of Limoges, Limoges, France, in 1989.

He is currently a Professor with the Institut de Recherche en Communications Optiques et Microondes (IRCOM), University of Limoges, and Head of the Microwave Circuits and Devices Team. His main area of interest concerns the design and optimization of microwave devices.

5.4 Matching problems

5.4.1 Nevanlinna-Pick interpolation and matching problems

Following paper is reproduced in this section:

- Laurent Baratchart, Martine Olivi, and Fabien Seyfert. "Boundary Nevanlinna-Pick interpolation with prescribed peak points. Application to impedance matching". In: *SIAM Journal on Mathematical Analysis* (2017). DOI: 10.1137/16M1085577. URL: https://hal.inria.fr/hal-01377782

BOUNDARY NEVANLINNA-PICK INTERPOLATION WITH PRESCRIBED PEAK POINTS. APPLICATION TO IMPEDANCE MATCHING.

LAURENT BARATCHART [*], MARTINE OLIVI [†], AND FABIEN SEYFERT [‡]

Abstract

We study a generalized Nevanlinna Pick interpolation problem on the half-plane for rational functions of prescribed degree, where peak points are imposed and interpolation conditions may lie on the real axis. This generalizes previous work by T. Georgiou, C. Byrnes, A Lindquist and A. Megretski. The problem is motivated by the issue of broadband matching in electronics and microwave system design. We prove existence and uniqueness of a solution by differential-topological techniques. The approach is put to work numerically on a real example, using a continuation method.

1. Introduction. Nevanlinna-Pick interpolation is a classical topic from function theory that has undergone several generalizations and enjoys deep connections with circuits and systems theory. In its original form, the problem consists in finding a Schur function to meet a finite set of interpolation conditions on the disk or the half-plane; here and below, a Schur function is a complex analytic function bounded by 1 in modulus. This kind of interpolation owns attractive necessary and sufficient conditions for a solution to exist (the non-negativity of the so-called Pick matrix), along with a parametrization of all solutions by Schur functions (the so-called Nevanlinna parametrization) [21, Ch. I, sec. 2, Ch. IV, sec. 6]. Composing with a conformal map, the problem can equivalently be stated in terms of Carathéodory functions, that is, analytic functions with non-negative real part. The theory has been extended in various directions including meromorphic, multiply connected, multivariable, operator-valued and non-commutative settings, as well as boundary interpolation, see *e.g.* [1, 2, 4, 7, 6, 5, 17, 18, 35, 3, 39]. Meantime, the links of such interpolation problems to sensitivity minimization and model matching, initially stressed in [37, 19], started a success story in robust control of linear systems, see *e.g.* [28, 29, 30, 31, 41] and the survey in [8].

Still, the relevance of Nevanlinna-Pick interpolation to Engineering problems had been pointed at earlier in a circuit-theoretic context, in relation with oscillator design, Darlington synthesis and broadband matching of dissipative devices [40, 13, 33]. In particular, the two issues of describing rational solutions of given degree and determining those of minimal degree were raised in [40]. Both turn out to be rather subtle. The second is still fairly open, but the first made substantial progress through the works [22, 14, 15, 23]. These show that if there are N interpolation conditions and the Nevanlinna-Pick matrix is positive definite, then rational Schur interpolants f of degree at most $N-1$ are essentially parametrized by the zeros of $1-ff^*$, a rational function of degree at most $2(N-1)$ which is positive on the boundary of the analyticity domain of f (the disk or the half-plane); here, f^* stands for the paraconjugate function (see definition in section 3). The stable zeros of $1-ff^*$ (so-called spectral zeros) may in turn be regarded (except in degenerate cases where cancellation occurs) as extra design parameters, see for example [26] where they are used to shape a robust feedback loop while bounding the degree of the controller. Since $f^*=\bar{f}$ on the boundary of the domain of analyticity, we note that if the spectral zeros lie on that boundary then they are maximum places for $|f|$ (*i.e.* places where $|f|=1$), hereafter called *peak points* of f.

Motivated by the broadband matching problem for filters, we present in this work a still more general result where some or all interpolation points may lie *on* the boundary of the analyticity domain, and still the zeros of $1-ff^*$ essentially parametrize the interpolants. For example, given N interpolation points inside or on the boundary of the domain, and a polynomial r of degree at most $N-1$ which is nonzero at every interpolation point, then there is a continuously invertible correspondence between sets of admissible interpolation values and polynomials p of degree at most $N-1$, the correspondence being that p/q meets the interpolation conditions with q the (normalized) stable polynomial such that $qq^*=pp^*+rr^*$. Moreover, along "most" paths between two sets of interpolation conditions, this correspondence is smoothly invertible. This allows us to tackle the problem numerically using continuation methods. Even when the interpolation points lie interior to the domain, this procedure is more efficient than minimizing entropy-like criteria as proposed in [14] (which are nevertheless interesting for themselves, see *e.g.* [26] for an application to model reduction).

The gist of our application to broadband matching for filters is that peak points at the ends of the band-

[*]EPI-APICS, Inria, BP 93, Sophia-Antipolis cedex (Laurent.Baratchart@inria.fr, https://www-sop.inria.fr/members/Laurent.Baratchart/).

[†]EPI-APICS, Inria, BP 93, Sophia-Antipolis cedex (Martine.Olivi@inria.fr, https://www-sop.inria.fr/members/Martine.Olivi/).

[‡]EPI-APICS, Inria, BP 93, Sophia-Antipolis cedex (Fabien.Seyfert@inria.fr, https://www-sop.inria.fr/members/Fabien.Seyfert/).

width (*i.e.* zeros of r) will ensure selectivity of the filter, while appropriate interpolation conditions in the bandwidth will guarantee perfect match at designated frequencies. Once the location of the zeros and the interpolation points is chosen, the results of the present paper allow one to compute the transmission parameter of the filter (the unique Schur rational function meeting the interpolation conditions and having zeros as prescribed, of degree the number of interpolation conditions minus 1) from which the whole scattering matrix is easily deduced. A scattering matrix corresponding to a physical RLC network (*i.e.* one with with real elements R, L, C) is obtained upon using appropriate conjugate-symmetric interpolation conditions. Let us stress, however, that complex elements are common when modelling microwave devices, due to the use of a low-pass transformation. The degree constraint on the filter is here essential, for it should be kept as small as possible, while meeting given specifications, in order to contain unmodelled losses and keep the physical size small. In the experiments presented in Section 5, a numerical search is performed on the location of the interpolation points so as to minimize the maximum of the reflection over the bandwidth. This way the paper offers new avenues in broadband matching, which is today a critical step in circuit design as passbands grow larger and efficiency concerns more stringent.

In the above-mentioned problem, interpolation values are admissible if those corresponding to boundary interpolation points have modulus strictly less than 1 while those corresponding to interior interpolation points satisfy Pick's criterion. We also consider another interpolation problem where an additional interpolation condition is imposed on the boundary, whose value has modulus 1 (we do not prescribe the angular derivative, though); in this case solutions are sought in degree $N+1$. Both problems are relevant to filter design, in which a unimodular normalization of the filtering function at infinity is sometimes necessary (the setting here is the half-plane). Note, however, that no interpolation value of modulus 1 is ever needed in the bandwidth since perfect match cannot take place at frequencies where the load is fully reflective. Still, it would be interesting to know which features of the interpolant can still be injectively and continuously specified when some unimodular interpolation values are imposed, that is, when interpolation points may at the same time be peak points. It is also natural to ask if in the case of matrix-valued interpolation (tangential or higher dimensional), spectral factors of $I-FF^*$ can again be used to parametrize solutions to the Nevanlinna-Pick problem whose McMillan degree is less than some prescribed bound. This question has been addressed in [36, **?**] in the standard setting where the interpolation points are in the interior. Both issues are left here for future research.

The paper is organized as follows. In Section 2, we discuss broadband matching which motivates the two interpolation problems raised in Section 3, called $\mathscr{P}$ and $\hat{\mathscr{P}}$ respectively. In Section 4 we state and prove our main results concerning existence, uniqueness, and generic smoothness of a solution with respect to the interpolation data. Though different from those in [22, 14, 23], our proofs likewise have a differential-topological flavour. Injectivity of the evaluation map at interpolation points (*cf.* equation (12)) is the most difficult issue, and is handled using ideas from orthogonal polynomials theory. The two interpolation problems are treated in parallel, but the authors were not able to reduce one of them to the other; in particular, Nevanlinna's iteration does not seem to be effective to do this. Finally, some numerical illustrations are given in Section 5. Our computational scheme uses continuation techniques, justified by the generic smoothness previously established. For the convenience of the reader, we provide him with an index of notation at the end of the paper.

2. Broadband matching. Communication devices such as multiplexers, routers, power dividers, couplers or antenna receptor chains, are realized by connecting together elementary components among which filters and N-port junctions are most common. For example, multiplexers are realized by plugging $N-1$ filters (one per channel) to a N-port junction. In fact, filters are typical two-port components which are present in almost every telecommunication device.

Now, when connecting a filter to some existing system, a recurring issue is to determine which frequencies will carry energy to the system across the filter, and which frequencies will bounce back. In this respect, the system L shown in Figure 1 (to be seen as the load of the filter S) is characterized by its reflection coefficient L_{11}, which is a complex-valued function of the frequency ω as the latter ranges over real numbers. We stress that the loads we consider may vary with frequency, *i.e.* they need not be purely resistive. Hereafter, we abbreviate real and complex numbers by $\mathbb{R}$ and $\mathbb{C}$, respectively. The effect of the filter is described by a 2×2 scattering matrix S, whose entries are again $\mathbb{C}$-valued functions of $\omega\in\mathbb{R}$. We assume in our discussion that the filter is lossless, meaning that S is unitary at all frequencies:

$$S(\omega)^*S(\omega)=Id,\qquad \omega\in\mathbb{R}, \tag{1}$$

where superscript "*" stands for "transpose-conjugate". Then, the reflection coefficient G_{11} at port 1 of the

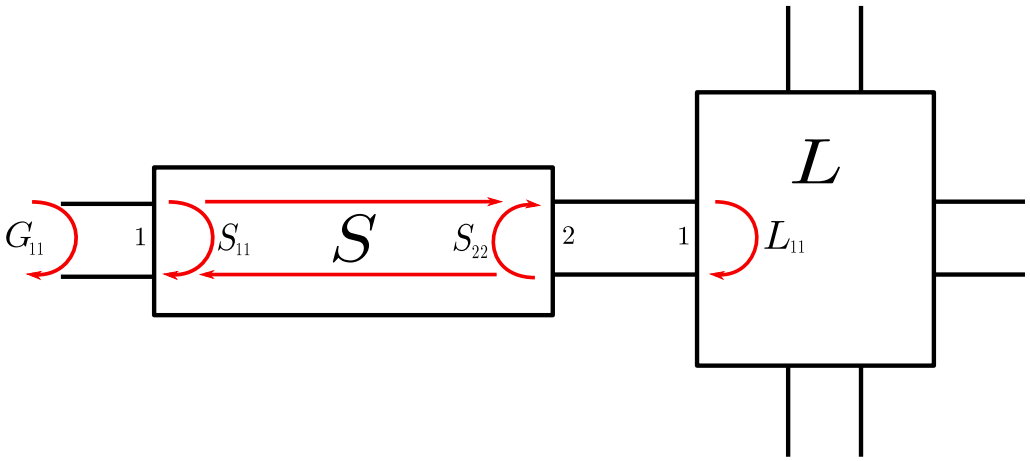


FIG. 1. *Filter plugged to a load L with reflexion coefficient* L_{11}

global system, consisting of the connected pair (S,L), is easily computed at the frequency ω to be

$$\begin{aligned} G_{11}(\omega) &= S_{11}(\omega) + \frac{S_{12}(\omega)S_{21}(\omega)L_{11}(\omega)}{1-S_{22}(\omega)L_{11}(\omega)} \\ &= \frac{S_{11}(\omega) - L_{11}(\omega)\det(S(\omega))}{1-S_{22}(\omega)L_{11}(\omega)} \\ &= \det(S(\omega))\frac{\overline{S_{22}(\omega)} - L_{11}(\omega)}{1-S_{22}(\omega)L_{11}(\omega)}. \end{aligned} \tag{2}$$

By definition, a matching frequency is some $\omega \in \mathbb{R}$ for which $G_{11}(\omega) = 0$. This means that the filter transmits to the load all the energy carried by the signal entering port 1 at frequency ω. If $|L_{11}(\omega)| < 1$, hence also $|S_{22}(\omega)L_{11}(\omega)| < 1$, it follows from (2) that ω is matching if and only if

$$S_{22}(\omega) = \overline{L_{11}(\omega)}. \tag{3}$$

In contrast, a stopping frequency is defined by the property that $|G_{11}(\omega)| = 1$. This means that all the energy carried by the signal entering port 1 at frequency ω bounces back and does not feed the load. If $|L_{11}(\omega)| < 1$, this amounts in view of (2) to say that $|S_{22}(\omega)| = 1$, which is in turn equivalent by (1) to

$$S_{12}(\omega) = S_{21}(\omega) = 0. \tag{4}$$

The problem of synthesizing the filter S, or the matching network L, so that $|G_{11}|$ is smallest possible on a given frequency band is a very old one. When the filter is finite-dimensional, this issue gave rise to the matching theory of Fano and Youla [16]. Specifically, if the model for the load is rational with 2 ports, this theory provides one with a parametrization of all responses G of those global systems that can be realized by plugging a filter S of given degree to a given load L. Such G are characterized in terms of their transmission zeros, which account for the fact that the load L can be "extracted" from the global response. However, it is unknown even today how to deduce filtering characteristics from this parametrization when the load has degree greater that one. This may contribute to explain why the Fano-Youla theory had little impact in practice. Also, the need to derive a rational model for the load, and to estimate its transmission zeros, might have impeded its dissemination in the engineering community. Instead, system manufacturers often use blackbox “optimization” in spite of usual drawbacks and uncertainties pertaining to this approach. Another method was proposed by J. Helton [25] in the infinite dimensional setting, where the matching problem gets reformulated as a H^∞ approximation problem of Nehari type whose solution is elegantly formulated in terms of the norm and maximizing vectors of a Hankel operator. This technique amounts to convexify the problem and it yields hard bounds on the achievable matching error, no matter the degree of the filter, along with an optimal (non-rational) solution to match this bound. However, this optimal filter has infinite degree which makes it hardly realizable or even computable in practice. In the present paper, we propose an intermediate approach where a finite-dimensional filter response of prescribed degree is being synthesized by imposing matching and stopping frequencies when the load is given.

3. Two interpolation problems. Below, we regard the scattering matrix of a filter as a function of a *real* variable, namely the frequency. This differs from the more usual convention where the transfer function is defined on the imaginary axis rather than the real line. In the present framework, a stable finite dimensional filter is one whose scattering matrix is rational with poles in the open upper half-plane $\mathbb{C}^+$ and entries with numerator's degree not exceeding the degree of the denominator. Equivalently, the scattering matrix belongs to the Hardy space $H^\infty(\mathbb{C}^-)$ of bounded holomorphic functions in the open lower half-plane $\mathbb{C}^-$. Also, a physical RLC network has a scattering matrix whose entries are ratios of polynomials whose coefficients in even degree are real and those in odd degree are pure imaginary (so that the function is real on the imaginary axis).

A polynomial with no root in $\mathbb{C}^-$ is said to be *stable in the broad sense*. A polynomial is called *stable* if it has no root in $\overline{\mathbb{C}^-}$, the *closed* lower half-plane.

The scattering matrix S of a lossless filter is termed *inner* in $H^\infty(\mathbb{C}^-)$, meaning that it satisfies (1). By the maximum principle, this entails that S is contractive in $\mathbb{C}^-$:

$$\|S(z)\xi\| \leq \|\xi\|, \quad \forall z \in \mathbb{C}^-,\ \xi \in \mathbb{C}^2,$$

where "$\|.\|$" indicates the Euclidean norm and the inequality is strict unless $S(z)\xi$ is constant.

We denote by $\mathbb{D}$ the open unit disk. For a rational matrix valued function $F(s)$, we define its *paraconjugate* $F^*(s)$ by

$$F^*(s) = \big(F(\bar{s})\big)^*, \quad s \in \mathbb{C}.$$

When F is constant, this notation agrees with the one introduced previously for the transpose conjugate of a complex matrix. Note that $F^*(s)$ indeed takes values on $\mathbb{R}$ which are transpose conjugate to those of F. Clearly "$*$" is an involution: $(F^*)^* = F$. If p is a polynomial, then its paraconjugate p^* is the polynomial obtained by conjugating the coefficients, in particular it has the same degree as p and roots conjugate to those of p.

Recall (see *e.g.* [27, 20]) that the McMillan degree of a $\ell_1 \times \ell_2$ rational matrix R is the smallest non-negative integer ℓ for which one can write $R(s) = C(sI_\ell - A)^{-1}B + D$, where A, B, C and D are complex matrices of size $\ell \times \ell$, $\ell \times \ell_2$, $\ell_1 \times \ell$ and $\ell_1 \times \ell_2$ respectively, and I_ℓ is the identity matrix of size $\ell \times \ell$. Equivalently, the McMillan degree is the smallest possible degree for the determinant of an invertible polynomial matrix P such that PR is also a polynomial matrix. A function of the form q^*/q where q is a stable polynomial of degree d is called a *Blaschke product* of degree d. When a rational matrix R is inner, then its determinant is a Blaschke product whose degree is equal to the McMillan degree of R [10].

Every 2×2 rational inner matrix S of McMillan degree N such that $\lim_{s\to\infty} S(s) = I_2$ admits the following representation (Belevitch form [12]):

$$S = \frac{1}{q}\begin{bmatrix} p^* & -r \\ r^* & p \end{bmatrix}, \tag{5}$$

where p, q are monic complex polynomials of degree N while r is a complex polynomial of degree at most $N-1$ having no common real root with p and q is computed from p and r as the unique monic stable polynomial satisfying the Feldtkeller equation:

$$qq^* = pp^* + rr^*. \tag{6}$$

Sometimes we say that q satisfying (6) is a stable *spectral factor* of $pp^* + rr^*$, see Proposition 2 for more details about existence and uniqueness of q.

By (3), a finite set $\{x_1 \dots x_m\} \subset \mathbb{R}$ consists of matching frequencies for the filter (5) with respect to a load having reflection coefficient L_{11} if, and only if

$$\frac{p}{q}(x_k) = \overline{L_{11}(x_k)} \stackrel{def}{=} \gamma_k, \qquad 1 \leq k \leq m. \tag{7}$$

We shall assume throughout that $|\gamma_k| < 1$, because the matching problem at fully reflecting frequencies for the load is ill-defined: indeed expression (2) is either of modulus 1 or indeterminate of the form $0/0$ when $|L_{11}(\omega)| = 1$, which makes it impossible to meet $G_{11}(\omega) = 0$.

In addition to the interpolation conditions (7) which take place on $\mathbb{R}$, it is often desirable (*e.g.* to prevent oscillations of the response or to account for unmodelled resistive effects) to meet additional interpolation

conditions inside the stability domain $\mathbb{C}^-$. To accomodate this case as well, we consider points $\{z_1 \dots z_l\}$ in $\mathbb{C}^-$ where following "complex" matching condition should hold:

$$\frac{p}{q}(z_k) = \overline{L_{11}(\Re(z_k))} \stackrel{def}{=} \beta_k, \tag{8}$$

where "$\Re$" indicates the real part. Let us abbreviate $(z_1, \cdots, z_l)^T$ and $(\beta_1, \cdots, \beta_l)^T$ by Z and β respectively (hereafter M^T denote the transpose of a matrix M). We define $P(Z,\beta)$ to be the so-called Pick matrix associated with the interpolation data (z_k, β_k), namely the Hermitian $l \times l$ matrix defined by:

$$P_{k,j}(Z,\beta) = \frac{1-\beta_k\overline{\beta_j}}{i(z_k - \overline{z_j})}. \tag{9}$$

It is classical that $P(Z,\beta)$ is positive semi-definite if and only if there is a Schur function f on $\mathbb{C}^-$ (*i.e.* holomorphic and such that $|f| \leq 1$) to meet the interpolation conditions $f(z_k) = \beta_k$ for $1 \leq k \leq l$, see *e.g.* [21, Ch. I, Cor. 2.3] for a version on the disk which immediately implies the present one using the conformal map $z \mapsto (i+z)/(i-z)$ from $\mathbb{C}^-$ onto $\mathbb{D}$. When such a function exists, $P(Z,\beta)$ is actually positive definite unless the solution to this constrained interpolation problem is unique, in which case the unique solution is a Blaschke product of degree equal to the rank of $P(Z,\beta)$. Conversely, if there is a solution to the interpolation problem which is a Blaschke product of degree $\delta < l$ then $P(Z,\beta)$ has rank δ. Functions in $H^\infty(\mathbb{C}^-)$, in particular Schur functions, have nontangential limits at a.e. point of $\mathbb{R}$ which allow one to speak of their boundary values [21, Ch. I, Thm. 5.3]. Moreover, knowing the boundary values on an arbitrary subset of positive measure of $\mathbb{R}$ determines the function uniquely [21, Ch. II, Cor. 4.2]. This implies that positive definiteness of $P(Z,\beta)$ is equivalent to the existence of a Schur solution to the interpolation problem whose trace on $\mathbb{R}$ has modulus strictly less than 1 on a set of positive measure. Indeed, if two distinct solutions have modulus 1 a.e. on $\mathbb{R}$, then any convex combination yields another solution having modulus strictly less than 1 at every point where the initial solutions take on distinct values.

We call $\mathbb{P}_Z^+$ the set of those $\beta \in \mathbb{C}^l$ such that $P(Z,\beta)$ is positive definite. Clearly $\mathbb{P}_Z^+$ is open in $\mathbb{C}^l$, and it is also convex as follows easily from the equivalence of positive definiteness with the existence of Schur solutions to the interpolation problem having modulus strictly less than 1 on a subset of $\mathbb{R}$ of positive measure. In particular, $\mathbb{P}_Z^+$ is connected. For simplicity, we often drop the subscript Z when the interpolation points are understood, and write $\mathbb{P}^+$ instead $\mathbb{P}_Z^+$.

Next, if we want to impose $N-1$ stopping frequencies for S in $\mathbb{R} \cup \{\infty\}$ which are distinct from the x_k, it is equivalent in view of (4) to prescribes the roots of the transmission polynomial r in (5) (or of r^* since it has the same real zeros as r). Here, we count multiplicities by repetition and a zero at infinity means a drop in degree. We shall first consider the situation where the leading coefficient of r is imposed as well, so that r itself is prescribed. This leads us to raise the following matching problem.

Problem $\mathscr{P}$: *Given* m distinct real frequencies $(x_1, x_2 \dots x_m)$, m interpolation conditions $(\gamma_1, \gamma_2 \dots \gamma_m)$ in $\mathbb{D}^m$, l distinct "complex frequencies" $(z_1, z_2 \dots z_l)$ associated to l interpolation values $(\beta_1, \beta_2 \dots \beta_l)$ in $\mathbb{P}^+$ and $r \not\equiv 0$ a complex polynomial of degree at most $m+l-1$, such that $r(x_k) \neq 0$, $k = 1, \dots, m$, *to find* (p,q) a pair of monic complex polynomials of degree $N = m+l$ such that,

$$\left\{ \begin{array}{rcll} \frac{p}{q}(x_k) & = & \gamma_k, & \text{for } k = 1,..,m \\ \frac{p}{q}(z_\ell) & = & \beta_\ell, & \text{for } \ell = 1,..,l \\ qq^* - pp^* & = & rr^* & \end{array} \right. \tag{10}$$

and q has no root in the open lower half-plane $\mathbb{C}^-$ (*i.e.* q is stable in the broad sense).

Nothing in the formulation of Problem $\mathscr{P}$ prevents the denominator polynomial q from vanishing at real points. If this happens, then the McMillan degree of S will drop since a real zero of q is a zero both of p and r with the same multiplicity (because $|p|^2 + |r|^2 = |q|^2$ on $\mathbb{R}$ by (10)), *cf* (5). Observe in this case, by the assumption in Problem $\mathscr{P}$, that a common zero to p and r cannot be one of the x_k, hence $\frac{p}{q}(x_k)$ in (10) is still equal to $p(x_k)/q(x_k)$. Remark also that the real roots of r are peak points for the modulus of p/q, *i.e.* points where the maximum value $|p/q| = 1$ is attained.

Problem $\mathscr{P}$ may be viewed as a generalization of the Nevanlinna-Pick interpolation problem with degree constraint studied in [22], in which the interpolation points are now allowed to lie *on* the real axis, whereas in [22] they are confined to the stability domain $\mathbb{C}^-$. We also extend the results announced in [11], in that the polynomial r can have real roots provided these are not interpolation points. Both generalizations

are *crucial* to approach the matching problem with interpolation techniques as described in Section 2, which was the main incentive for the authors to undertake the present study.

Problem $\mathscr{P}$ imposes the condition $(p/q)(\infty) = 1$ since p, q are monic of degree N. In other words there is an implicit extra interpolation node on the real line (namely ∞) with interpolation value equal to 1. In connection with the matching problem discussed in Section 1, where p/q is thought of as the entry S_{22} of the scattering matrix S of a filter (*cf.* (5)), this is the right normalization in that the low-pass equivalent model to a LC-resonant filter behaves like an open circuit at infinite frequency, which results into the condition $S(\infty) = Id$. However, if we add for example a transmission line in front of the filter, the line can be modeled in the narrow band approximation by a reflexion coefficient which is unimodular (with free phase) at infinity. This extra design parameter can be used to meet an additional interpolation condition or, dually, to reduce the degree of p, q while keeping the interpolation properties of p/q. This leads us naturally to the following "non-normalized" version of problem $\mathscr{P}$.

Problem $\hat{\mathscr{P}}$: *Given* m distinct real frequencies $(x_1, x_2 \dots x_m)$, m interpolation conditions $(\gamma_1, \gamma_2 \dots \gamma_m)$ in $\mathbb{D}^m$, l distinct "complex frequencies" $(z_1, z_2 \dots z_l) \in (\mathbb{C}^-)^l$ associated to l interpolation values

$$(\beta_1, \beta_2 \dots, \beta_l)$$

in $\mathbb{P}^+$ and $r \not\equiv 0$ a complex polynomial of degree at most $m+l-1$, such that $r(x_k) \neq 0$, $k = 1, \dots, m$, *to find* (p, q) a couple of complex polynomials of degree at most $\hat{N} = m+l-1$ such that,

$$\left\{ \begin{array}{rcl} \frac{p}{q}(x_k) & = & \gamma_k, \quad \text{for } k = 1, .., m \\ \frac{p}{q}(z_\ell) & = & \beta_\ell, \quad \text{for } \ell = 1, .., l \\ qq^* - pp^* & = & rr^* \end{array} \right. \tag{11}$$

where q is stable in the broad sense and normalized so that $q(x_1) > 0$ if $m > 0$, $q(z_1) > 0$ otherwise.

Although problems $\mathscr{P}$ and $\hat{\mathscr{P}}$ will be treated in a similar way, the authors were not able to reduce one of them to the other.

4. Solution to $\mathscr{P}$ and $\hat{\mathscr{P}}$: two matching theorems. We begin with the analysis of problem $\mathscr{P}$. It relies on the study of a specific evaluation map to be defined presently. According to the statement of the problem, we fix $(x_1, x_2 \dots x_m)^T \in \mathbb{R}^m$, $(z_1, z_2 \dots z_l)^T \in (\mathbb{C}^-)^l$, and a polynomial r of degree at most $m+l-1$ such that $r(x_k) \neq 0$ for all k. We let $\mathbf{PM}_N$ designate the set of monic polynomial of degree $N = m+l$ with complex coefficients. This set is topologized as $\mathbb{C}^N \sim \mathbb{R}^{2N}$, using coefficients as coordinates except for the leading one which is equal to 1 by definition. Specifically, we identify $p(z) = z^N + p_{N-1}z^{N-1} + \dots + p_0$ with the vector $(p_0, p_1, \dots, p_{N-1})^T \in \mathbb{C}^N$. Hereafter, the degree of a polynomial p is abbreviated as $\deg p$.

As r is fixed with $\deg r < N$, equation (6) associates to each $p \in \mathbf{PM}_N$ a unique polynomial $q = q(p) \in \mathbf{PM}_N$ which is stable in the broad sense, *cf.* Proposition 2 to come. Since $|p|^2 \leq |p|^2 + |r|^2 = |q|^2$ on $\mathbb{R}$, the rational function p/q has modulus at most 1 there. In particular it has no real pole, and no pole in $\mathbb{C}^-$ either since q is stable in the broad sense. Thus, by the maximum principle, we conclude that $|p/q| \leq 1$ on $\overline{\mathbb{C}^-}$. In addition, since no x_k is a root of r by assumption, we have that $|p(x_k)/q(x_k)| < 1$ hence p/q is not a Blaschke product. Therefore the Pick matrix associated with the interpolation data $(z_k, p(z_k)/q(z_k))$ is positive definite, and we can define an evaluation map $\psi : \mathbf{PM}_N \to \mathbb{D}^m \times \mathbb{P}^+_Z$ by the formula

$$\psi(p) = \begin{pmatrix} p(x_1)/q(x_1) \\ \vdots \\ p(x_m)/q(x_m) \\ p(z_1)/q(z_1) \\ \vdots \\ p(z_l)/q(z_l) \end{pmatrix} \in \mathbb{D}^m \times \mathbb{P}^+_Z. \tag{12}$$

The result which yields existence and uniqueness of a solution to Problem $\mathscr{P}$, along with generic differentiability thereof, may now be stated as follows.

THEOREM 1. *ψ is a homeomorphism from $\mathbf{PM}_N$ onto the product space $\mathbb{D}^m \times \mathbb{P}^+$. Moreover, the restriction of ψ to those $p \in \mathbf{PM}_N$ having no common real root with r is a diffeomorphism onto its image.*

REMARK 4.1. *From the uniqueness part of Theorem 1, it follows that if the set of interpolation points x_k, z_ℓ is stable under the map $z \mapsto -\bar{z}$, and if the interpolation values at x_k and $-x_k$ (resp. z_ℓ and $-\bar{z}_\ell$) are*

conjugate, then p and q have pure imaginary coefficients in odd degree and real coefficients in even degree. Equivalently, p/q is real-valued on the imaginary axis.

The proof of Theorem 1 will be given in sections 4.5 and 4.6, after some preparatory work.

4.1. Continuity and differentiability of ψ. For $k \geq 0$ an integer, we let $\mathbf{P}_k$ be the space of complex polynomials of degree at most k and $\mathbf{PE}_k$ the subset comprising polynomials of exact degree k. We occasionally write $\mathbf{P}_{\mathbb{R},k}$ for the real subspace of polynomials with real coefficients. The space $\mathbf{P}_k$ identifies with $\mathbb{C}^{k+1} \sim \mathbb{R}^{2k+2}$, using coefficients as coordinates. Thus, $p(z) = p_k z^k + p_{k-1} z^{k-1} + \cdots + p_0$ is regarded as $(p_0, p_1, \cdots, p_k)^T \in \mathbb{C}^{k+1}$. With this definition, $\mathbf{PM}_N \subset \mathbf{P}_N$ is the hyperplane $\{p_N = 1\}$ which in turn identifies with $\mathbb{C}^N$ as pointed out earlier. We further denote by $\mathbf{SB}_N$ the set of polynomials of degree at most N which are stable in the broad sense, and by $\mathbf{SBM}_N$ the subset of monic polynomials of degree N stable in the broad sense. The set of stable polynomials of degree at most N will likewise be denoted by $\mathbf{S}_N$, the subset of stable polynomial of exact degree N by $\mathbf{SE}_N$, and the subset of stable monic polynomials of degree N by $\mathbf{SM}_N$.

We write $\mathbf{P}^+_{2N}$ for the set of polynomials of degree at most $2N$ which are non-negative on $\mathbb{R}$. Such a polynomial must have real coefficients, even degree, positive dominant coefficient, and its real roots have even multiplicity. Moreover, it is equal to its para-conjugate. We put $\mathbf{PE}^+_{2N}$ for the subset of non-negative polynomials of exact degree $2N$ and $\mathbf{PM}^+_{2N}$ for the subset of non-negative monic polynomials of degree $2N$. The sets $\mathbf{P}^+_{2N}$ and $\mathbf{PE}^+_{2N}$ will be regarded as embedded in $\mathbb{R}^{2N+1}$, and again $\mathbf{PM}^+_{2N}$ will be seen as a subset of $\mathbb{R}^{2N}$ for it is the intersection of $\mathbf{P}^+_{2N} \subset \mathbb{R}^{2N+1}$ with the hyperplane $\{p_{2N+1} = 1\}$.

The interior $\overset{\circ}{\mathbf{PM}}{}^+_{2N}$ of $\mathbf{PM}^+_{2N} \subset \mathbb{R}^{2N}$ consists of monic polynomials of degree $2N$ which are strictly positive on $\mathbb{R}$. Indeed, if $p \in \mathbf{PM}^+_{2N}$ is such that $p(x_0) = 0$, then adding a small negative constant to p will destroy positivity at x_0 and therefore p cannot lie interior to $\mathbf{PM}^+_{2N}$ in $\mathbb{R}^{2N}$. Conversely, let $p \in \mathbf{PM}^+_{2N}$ have no zero on $\mathbb{R}$. Then, there is $\varepsilon > 0$ such that $|p(x)| > \varepsilon$ for $x \in \mathbb{R}$. Write $p(x) = x^{2N+1} + p_{2N} x^{2N} + \cdots + p_0$ and put $a := \max\{1, \varepsilon + 2\Sigma_{j=0}^{2N} |p_j|\}$. If we let $(\delta_0, \cdots, \delta_{2N-1})^T \in \mathbb{R}^{2N}$ be such that $\Sigma |\delta_j| a^j < \varepsilon/2$, we easily get upon setting $\delta p(x) := \Sigma_{j=0}^{2N-1} \delta_j x^j$ that $|p + \delta p| > \varepsilon/2$ on $[-a, a]$ and that

$$|p(x) + \delta p(x)| \geq |x|^{2N} \left(|x| - \varepsilon/2 - \Sigma_{j=1}^{2N-1} |p_j| \right) > \frac{|x|^{2N+1}}{2}, \qquad |x| > a.$$

Hence p lies interior to $\mathbf{PM}^+_{2N}$. Likewise, the interior $\overset{\circ}{\mathbf{PE}}{}^+_{2N}$ of $\mathbf{PE}^+_{2N} \subset \mathbb{R}^{2N+1}$ consists of polynomials of exact degree $2N$ which are strictly positive on $\mathbb{R}$. In another connection, the interior of $\mathbf{SBM}_N \subset \mathbf{PM}_N$ is $\mathbf{SM}_N$. Indeed, if $p \in \mathbf{SBM}_N \setminus \mathbf{SM}_N$, then p must have a real root x, and replacing the latter with $x - i\varepsilon$ for small $\varepsilon > 0$ produces a nearby polynomial which is unstable. Hence p is not an interior point of $\mathbf{SBM}_N$. Conversely if $p \in \mathbf{SM}_N$, then it has N-roots in $\mathbb{C}^-$ and we can pick a smooth curve $\Gamma \subset \mathbb{C}^-$ encompassing them. Since Γ is compact and p does not vanish on Γ, we have that $|p| > \eta > 0$ on Γ and if $q \in \mathbf{PM}_N$ is close enough to p then $|q| > \eta$ on Γ as well. By the argument principle, we have that $\int_\Gamma q'/q dz = 2i\pi n$ where n is the number of roots of q inside Γ, and if q is sufficiently close to p this integral is arbitrary close to $\int_\Gamma p'/p dz = 2i\pi N$ so that $n = N$, implying that $q \in \mathbf{SM}_N$. Thus, p lies interior to $\mathbf{SBM}_N$. As well, the interior of $\mathbf{SB}_N$ is $\mathbf{SE}_N$. Indeed, if $p(z) \in \mathbf{SBM}_N$ has degree strictly less than N, multiplying it by $(1 - i\varepsilon z)$ for small $\varepsilon > 0$ yields a nearby polynomial (in the topology of $\mathbf{SBM}_N$) which is unstable, and if p has a real root x then replacing x by $x - i\varepsilon$ again produces an unstable nearby polynomial. Thus, the interior of $\mathbf{SB}_N$ is included in $\mathbf{SE}_N$, and the converse inclusion follows from an application of the argument principle similar to the one already used to show that $\mathbf{SM}_N$ is the interior of $\mathbf{SBM}_N$.

After these rather mechanical preliminaries, we are in position to prove our first result:

PROPOSITION 2. *To any non zero $P \in \mathbf{P}^+_{2N}$, one can associate $q \in \mathbf{SB}_N$ such that*

$$P(t) = |q(t)|^2 = q(t) q^*(t), \quad t \in \mathbb{R}. \tag{13}$$

The polynomial $q(s)$ is unique up to a multiplicative unimodular constant, and if P has exact degree $2N$ then q has exact degree N. For fixed $z \in \mathbb{C}^-$ and $x \in \mathbb{R}$, define three maps φ_z, φ_x and φ_N by the formulas:

a) $\varphi_z : \mathbf{P}^+_{2N} \setminus \{0\} \to \mathbf{SB}_N$, with $\varphi_z(P)$ the unique solution to (13) meeting $q(z) > 0$,

b) $\varphi_x : \mathbf{P}^+_{2N} \setminus \{p \in \mathbf{P}^+_{2N}, p(x) = 0\} \to \mathbf{SB}_N$, with $\varphi_x(P)$ the unique solution to (13) meeting $q(x) > 0$,

c) $\varphi_N : \mathbf{PM}^+_{2N} \to \mathbf{SBM}_N$ with $\varphi_N(P)$ the unique monic solution to (13).

The maps φ_z, φ_x, φ_N are continuous and define homeomorphisms $\mathbf{P}^+_{2N} \setminus \{0\} \to \{p \in \mathbf{SB}_N,\ p(z) > 0\}$, $\mathbf{P}^+_{2N} \setminus \{p \in \mathbf{P}^+_{2N}, p(x) = 0\} \to \{p \in \mathbf{SB}_N,\ p(x) > 0\}$,and $\mathbf{PM}^+_{2N} \to \mathbf{SBM}_N$ respectively.

Moreover, the restriction of φ_N to $\overset{\circ}{\mathbf{PM}}{}^+_{2N}$ is a diffeomorphism onto $\mathbf{SM}_N$, and the restriction of φ_z (resp. φ_x) to $\overset{\circ}{\mathbf{PE}}{}^+_{2N}$ is a diffeomorphism onto the open subset $\{p \in \mathbf{SE}_N,\ p(z) > 0\}$ (resp. $\{p \in \mathbf{SE}_N,\ p(x) > 0\}$) of the linear subspace $\mathfrak{V}_z$ (resp. $\mathfrak{V}_x$) of $\mathbf{P}_N$ consisting of polynomials of degree at most N whose value at z (resp. x) is real. Specifically, the derivatives of φ_z, φ_x, and φ_N are given by:

- *if $P \in \overset{\circ}{\mathbf{PE}}{}^+_{2N}$ and δP is a real polynomial of degree at most $2N$, then*

$$D\varphi_z(P)[\delta P] = u$$

where u is the unique polynomial such that

$$u^*\varphi_z(P) + u\varphi_z^*(P) = \delta P, \qquad u \in \mathbf{P}_N, \quad u(z) \in \mathbb{R}; \tag{14}$$

- *if $P \in \overset{\circ}{\mathbf{PE}}{}^+_{2N}$ and δP is a real polynomial of degree at most $2N$, then*

$$D\varphi_x(P)[\delta P] = u$$

where u is the unique polynomial such that

$$u^*\varphi_x(P) + u\varphi_x^*(P) = \delta P, \qquad u \in \mathbf{P}_N, \quad u(x) \in \mathbb{R}; \tag{15}$$

- *if $P \in \overset{\circ}{\mathbf{PM}}{}^+_{2N}$ and δP is a real polynomial of degree at most $2N-1$, then*

$$D\varphi_N(P)[\delta P] = u$$

where u is the unique polynomial such that

$$u^*\varphi_N(P) + u\varphi_N^*(P) = \delta P, \qquad u \in \mathbf{P}_{N-1}. \tag{16}$$

Proof. It is elementary to check that $q \in \mathbf{SB}_N$ satisfies (13) if and only if its roots are the real roots of P with half their multiplicity and the non-real roots of P having strictly positive imaginary part with their multiplicity, while its dominant coefficient has square modulus equal to the dominant coefficient of P. This shows the existence of q and its uniqueness up to a multiplicative unimodular constant. Alternatively, the result also follows upon applying to $P(i(e^{i\theta}+1)/e^{i\theta}-1))$ a classical result by Fejèr and Riesz asserting that non-negative trigonometric polynomials are square moduli of algebraic polynomials on the unit circle [34, sec. 53].

Next, we prove that φ_z is continuous. Let (P_k) be a sequence in $\mathbf{P}^+_{2N} \setminus \{0\}$ converging to $P \in \mathbf{P}^+_{2N} \setminus \{0\}$. We must show that $q_k := \varphi_z(P_k)$ converges to $\varphi_z(P)$. As a basis of $\mathbf{P}_N$, pick the Lagrange interpolation polynomials L_n, $n = 0, 1 \ldots, N$, associated with the integer points $x = 0, 1, \ldots, N$. In other words, to each $n \in \{0, \cdots, N\}$, we have for $0 \leq j \leq N$ that $L_n(j) = \delta_{n,j}$, the Kronecker delta function. The coordinates of q_k in this basis are $(q_k(0), q_k(1), \ldots, q_k(N))$. As $|q_k(j)| = \sqrt{P_k(j)}$ is bounded since (P_k) converges, the sequence (q_k) is in turn bounded in $\mathbf{P}_N$. Thus we may extract a convergent sub-sequence from any subsequence, and *we claim* that the limit is $\varphi_z(P)$; this will prove the announced continuity. Assume indeed that a subsequence, again denoted by (q_k) for simplicity, converges to $q \in \mathbf{P}_N$. Since taking products and conjugates of polynomials is continuous $\mathbf{P}_N \times \mathbf{P}_N \to \mathbf{P}_{2N}$ and $\mathbf{P}_N \to \mathbf{P}_N$ respectively, we get in the limit from the relation $P_k = q_k q_k^*$ that $P = qq^*$. In particular $q \not\equiv 0$. Moreover $q(z) \geq 0$ because pointwise evaluation is also continuous. In order to prove the claim, it remains to show that $q \in \mathbf{SB}_N$. Suppose for a contradiction that q has some unstable root $s_0 \in \mathbb{C}^-$ with multiplicity μ. As q is not identically zero, s_0 is an isolated root so there exists $R > 0$ such that the disk $D = \{s, |s - s_0| \leq R\}$ is included in $\mathbb{C}^-$ and the circle $\partial D = \{s, |s - s_0| = R\}$ contains no root of q. As the sequence (q_k) converges uniformly to q on every compact subset of $\mathbb{C}$, the argument principle implies that q_k has μ roots in D counting multiplicities, as soon as k is large enough, which yields the desired contradiction.

Next, consider the map $\tilde{\varphi} : \mathbf{P}_N \to \mathbf{P}^+_{2N}$ defined by $\tilde{\varphi}(q) = qq^*$. Clearly, the restriction of $\tilde{\varphi}$ to the subset $\{p \in \mathbf{SB}_N,\ p(z) > 0\}$ (resp. $\{p \in \mathbf{SB}_N,\ p(x) > 0\}$, $\mathbf{SBM}_N$) is a continuous inverse to φ_z (resp. φ_x, φ_N). In addition $\tilde{\varphi}$ is C^∞-differentiable, and its differential $D\tilde{\varphi}(q)$ at q acts on $dq \in \mathbf{P}_N$ by the formula

$$D\tilde{\varphi}(q)[dq] \ = \ dqq^* + qdq^*. \tag{17}$$

Let us prove that the restriction of φ_z to $\overset{\circ}{\mathbf{P}}{}^+_{2N}$ is a diffeomorphism onto $H_z \overset{def}{=} \{p \in \mathbf{SE}_N,\ p(z) > 0\}$. By definition, if $P \in \overset{\circ}{\mathbf{PE}}{}^+_{2N}$ then $q = \varphi_z(P)$ lies in H_z which is obviously an open subset of $\mathfrak{V}_z$. Being a linear subspace

of $\mathbf{P}_N$ of codimension 1, $\mathfrak{V}_z$ identifies with $\mathbb{R}^{2N+1}$ and the restriction of $\tilde{\varphi}$ to $\mathfrak{V}_z$ is in turn C^∞-differentiable. Further, the restriction $\tilde{\varphi}_1$ of $\tilde{\varphi}$ to H_z is inverse to the restriction of φ_z to $\overset{\circ}{\mathbf{PE}}{}^+_{2N}$, and it is differentiable with derivative given by (17) restricted to $dq \in \mathfrak{V}_z$. *We claim* that this derivative is injective. Assume indeed that $D\tilde{\varphi}_1(q)[dq] = 0$. Then, since q and q^* are coprime polynomials (for their roots respectively lie in $\mathbb{C}^+$ and $\mathbb{C}^-$), we get from (17) that q divides dq so that $dq = \lambda q$ for some $\lambda \in \mathbb{C}$, because the degree of dq cannot exceed N which is the degree of q. In view of (17), we conclude from $D\tilde{\varphi}_1(q)[dq] = 0$ that $(\lambda + \bar{\lambda})qq^* = 0$, and since $qq^* \neq 0$ (for it has exact degree $2N$) we see that λ is pure imaginary. As $q(z) > 0$, this implies that $dq(z) = \lambda q(z)$ is pure imaginary, and since it is also real because $dq \in \mathfrak{V}_z$ we necessarily have that $\lambda = 0$ whence $dq = 0$. *This proves the claim.* As $D\tilde{\varphi}_1(q)$ maps $\mathfrak{V}_z$ injectively into reals polynomials of degree at most $2N$ and both spaces have dimension $2N+1$, we conclude that it is invertible. Now, the inverse function theorem asserts that $\tilde{\varphi}_1$ is a local diffeomorphism $H_z \to \overset{\circ}{\mathbf{PE}}{}^+_{2N}$. But we saw that $\tilde{\varphi}$ is a homeomorphism $\{p \in \mathbf{SB}_N,\ p(z) > 0\} \to \mathbf{P}^+_{2N} \setminus \{0\}$ under which the image of H_z is evidently $\overset{\circ}{\mathbf{PE}}{}^+_{2N}$, hence $\tilde{\varphi}_1$ is a global diffeomorphism $H_z \to \overset{\circ}{\mathbf{PE}}{}^+_{2N}$ This concludes the proof for φ_z.
The case of φ_x is similar, and the case of φ_N even simpler for dq in (17) will have degree at most $N-1$, making obvious that it must vanish if it is divisible by q. □

REMARK 4.2. *Continuity of spectral factorization can be given other, more analytic proofs based on the Poisson representation of* log*-moduli of outer functions, see e.g. [9, Lemma 1] for an alternative argument on the disk that easily carries over to the half-plane.*

Keeping in mind notation from Proposition 2, we may now represent the map ψ introduced in (12) as the composition of two functions, namely the map from $\mathbf{PM}_N$ into $\mathbf{PM}_N \times \mathbf{SBM}_N$ given by

$$p \to (p, \varphi_N(pp^* + rr^*))$$

followed by the evaluation map

$$(p,q) \to \left(\frac{p}{q}(x_1), \ldots, \frac{p}{q}(x_m), \frac{p}{q}(z_1), \ldots, \frac{p}{q}(z_l)\right)^T$$

from $\mathbf{PM}_N \times \mathbf{SBM}_N$ into $\mathbb{D}^m \times \mathbb{P}^+_Z$. Proposition 2 immediately yields:

COROLLARY 3. *The map ψ is continuous at every $p \in \mathbf{PM}_N$, and if p has no real root in common with r, then ψ is C^∞-smooth around p.*

4.2. An excursion into positive real functions. Recall that a holomorphic function f on $\mathbb{C}^-$ is a Schur function if $|f| \leq 1$, and a Carathéodory function if $\Re f \geq 0$. The map $f \mapsto (1-f)/(1+f)$ is an involution from Schur functions to Carathéodory functions and back. Like Schur functions, Carathéodory functions have non tangential limits a.e. on $\mathbb{R}$ from $\mathbb{C}^-$, allowing us to speak of their boundary values. Unlike Schur functions, though, rational Carathéodory functions may well have poles on $\mathbb{R}$. The function

$$z \mapsto -i/(z - x_0), \qquad x_0 \in \mathbb{R}, \tag{18}$$

is an example. This difference stems from the fact that the real part of a Schur function is the Poisson integral of a function on $\mathbb{R}$, wheras that of a Carathéodory function is generally the Poisson integral of a measure [21, Ch. I, Thm. 3.5]. In the previous example, the measure is a Dirac delta at x_0.

If (p,q) is a solution to Problem $\mathscr{P}$, then p/q is a Schur function as explained before (12). Our proof of Theorem 1 rests in part on the link, to be stressed momentarily, between problem $\mathscr{P}$ and its analog for Carathéodory functions.

For $p \in \mathbf{PM}_N$ and $q = \varphi_N(pp^* + rr^*)$, we put $\Sigma = \Sigma(p) := p/q$. By construction, this is a Schur rational function satisfying

$$1 - \Sigma^*\Sigma = \frac{rr^*}{qq^*}. \tag{19}$$

We now define a Carathéodory function Y (the so-called Cayley transform of Σ) by the formula

$$Y := \frac{1-\Sigma}{1+\Sigma} = \frac{q-p}{q+p}. \tag{20}$$

Then, a straightforward computation shows that

$$Y+Y^* = \frac{(q-p)(q^*+p^*)+(q^*-p^*)(q+p)}{(q+p)(q^*+p^*)} = \frac{2rr^*}{(q+p)(q^*+p^*)}. \tag{21}$$

By definition, the dissipation polynomial of a rational Carathéodory function is the numerator of the fraction $Y+Y^*$ when the latter is written in irreducible form. To us, given a rational Carathéodory function π/χ with π, χ polynomials, it is more convenient to define the *dissipation polynomial of the pair* (π,χ) to be the polynomial $\pi\chi^*+\pi^*\chi$. Thus, by (21), $2rr^*$ is the dissipation polynomial of the pair $(q-p,q+p)$.

In view of (20)-(21), Problem $\mathscr{P}$ is equivalent to an interpolation problem for rational Carathédory functions of the form π/χ where $\pi\in\mathbf{P}_{N-1}$ and $\chi\in\mathbf{SBM}_N$, with prescribed dissipation polynomial rr^* for the pair (π,χ). The corresponding interpolation conditions are $(\pi/\chi)(x_k)=(1-\gamma_k)/(1+\gamma_k)$ and $(\pi/\chi)(z_\ell)=(1-\beta_\ell)/(1+\beta_\ell)$. For this equivalent problem, the analog of equation (6) is

$$\pi\chi^*+\pi^*\chi = rr^*, \tag{22}$$

which entails that π/χ is a Carathéodory function when $\chi\in\mathbf{SBM}_N$. If r has no real root and (π,χ) is a solution to the Carathéodory analog of Problem $\mathscr{P}$, then χ is stable by (22), *i.e.* it lies in $\mathbf{SM}_N$ and not just in $\mathbf{SBM}_N$. This entails that χ, χ^* are coprime so that π is uniquely determined by r and χ through (22). In this case the Carathéodory analog to $\mathscr{P}$ is easier to handle than $\mathscr{P}$ itself, essentially because (22) is inear in χ and π whereas (6) is quadratic in q and p. Things change when r has a real root, say x_0. For if $\chi\in\mathbf{SBM}_N$ satisfies $\chi(x_0)=0$ and $\pi\in\mathbf{P}_{N-1}$ is a solution to (22) then for each $a>0$ the polynomial $\pi_a(s);=\pi(s)-ia\chi(s)/(s-x_0)$ is again a solution. So, when r and χ happen to have a common real root, they fall short of determining π *via* (22). This discrepancy arises because the dissipation polynomial of the pair $(-ia,(z-x_0))$ is identically zero and still $z\mapsto -ia/(z-x_0)$ is a non-zero Carathéodory function, see example (18).

Applications of Problem $\mathscr{P}$ to filter design discussed in Section 1 typically involve a transmission polynomial r having real zeros near the endpoints of the bandwidth of the filter, because these ensure stiffness of the response there. Thus, we find ourselves in the difficult case of the Carathéodory analog to $\mathscr{P}$. Nevertheless, the latter plays an important role in our proof of Theorem 1, when showing that ψ and its derivative are injective.

It will be convenient to introduce the Hardy space $H^2(\mathbb{C}^-)$ consisting of those holomorphic functions f in $\mathbb{C}^-$ satisfying

$$\sup_{y<0}\left(\int_{-\infty}^{+\infty}|f(x+iy)|^2\,dx\right)^{1/2}<+\infty. \tag{23}$$

Such a function has a nontangential limit at almost every $x\in\mathbb{R}$ that we denote again with $f(x)$, the argument being now in $\mathbb{R}$ and not in $\mathbb{C}^-$. This nontangential limit lies in the Lebesgue space $L^2(\mathbb{R})$, and in fact $\|f\|_{L^2(\mathbb{R})}$ is equal to the supremum in (23) [21, Ch. I, Thm. 5.3]. Moreover, for $z\in\mathbb{C}^-$, $f(z)$ can be recovered from f on $\mathbb{R}$ through a Cauchy as well as a Poisson integral [21, Ch. II, sec. 3]. In particular, a rational function π/χ with $\pi\in\mathbf{PE}_k$ and $\chi\in\mathbf{PE}_N$ does lie in $H^2(\mathbb{C}^-)$ if and only if $k<N$ and it has no pole in $\overline{\mathbb{C}^-}$, in other words if it vanishes at infinity and if every zero of χ in $\overline{\mathbb{C}^-}$ is cancelled by a corresponding zero of π. It follows easily that a rational Carathéodory function lies in $H^2(\mathbb{C}^-)$ if and only if its restriction to the real line lies in $L^2(\mathbb{R})$. Every $f\in H^2(\mathbb{C}^-)$ is the Cauchy integral of the non-tangential limit of its real part:

$$f(z) = -\frac{1}{i\pi}\int_{\mathbb{R}}\frac{\Re f(t)}{t-z}\,dt, \qquad z\in\mathbb{C}^-, \tag{24}$$

and the non-tangential limit of its imaginary part is the Hilbert transform of the nontangential limit of its real part [21, Ch. III, sec. 2]:

$$\Im f(x) = \frac{1}{\pi}\lim_{\varepsilon\to 0^+}\int_{|x-t|>\varepsilon}\frac{\Re f(t)}{t-x}\,dt, \qquad \text{a.e. } x\in\mathbb{R}. \tag{25}$$

Consequently, the nontangential limit of f can be recovered from its real part as

$$f(x) = \Re f(x) + \frac{i}{\pi}\lim_{\varepsilon\to 0^+}\int_{|x-t|>\varepsilon}\frac{\Re f(t)}{t-x}\,dt, \qquad \text{a.e. } x\in\mathbb{R}. \tag{26}$$

When f is smooth on $\mathbb{R}$, in particular if it is rational, then the last formula is valid for all $x\in\mathbb{R}$ and not just almost every x.

THEOREM 4. *Let $g \in \mathbf{PM}_{2N}^+$ and $d \in \mathbf{P}_{2K}^+$, with $K < N$ and $\frac{d}{g} \in L^2(\mathbb{R})$. Let further $(x_1, \cdots, x_m)^T \in \mathbb{R}^m$ and $(z_1, \cdots, z_l)^T \in (\mathbb{C}^-)^l$ with $m+l = N$, and assume that $d(x_k) \neq 0$ for all $k \in \{1, \cdots, m\}$. Then, the following three properties hold.*

1. *There exists a unique pair of polynomials $\chi_g \in \mathbf{SBM}_N$ and $\pi_{d,g} \in \mathbf{P}_{N-1}$, such that the rational function $Y_{d,g} = \frac{\pi_{d,g}}{\chi_g}$ satisfies:*
   (a) $Y_{d,g} \in H^2(\mathbb{C}^-)$
   (b) $\pi_{d,g}\chi_g^* + \pi_{d,g}^*\chi_g = d$
   (c) $Y_{d,g} + Y_{d,g}^* = \frac{d}{g}$
2. *Let g_1, g_2 in $\mathbf{PM}_{2N}^+$ such that $\frac{d}{g_1}$ and $\frac{d}{g_2}$ are in $L^2(\mathbb{R})$. If*
   (a) $\forall k \in \{1..m\}\ Y_{d,g_1}(x_k) = Y_{d,g_2}(x_k)$,
   (b) $\forall k \in \{1..l\}\ Y_{d,g_1}(z_k) = Y_{d,g_2}(z_k)$,
   then $g_1 = g_2$ so that $\pi_{d,g_1} = \pi_{d,g_2}$ and $\chi_{g_1} = \chi_{g_2}$ by property 1.
3. *For fixed $d \in \mathbf{P}_{2K}^+$, the evaluation map $\theta : \mathbf{PM}_{2N}^+ \to \mathbb{C}^N$ given by*

$$\theta(g) = \begin{pmatrix} Y_{d,g}(x_1) \\ \vdots \\ Y_{d,g}(x_m) \\ Y_{d,g}(z_1) \\ \vdots \\ Y_{d,g}(z_l) \end{pmatrix} \tag{27}$$

is well-defined and a diffeomorphism onto its image.

Proof. Let $u(z) = \Pi_{j=1}^{\ell}(z-x_j)^{2\kappa_j}$ be the monic divisor of g comprising all its real roots (if g has no real roots, then $\ell = 0$ and $u \equiv 1$). If $\pi_{d,g}$, χ_g satisfy (1b) and (1c), a short computation yields that $\chi_g = \varphi_N(g)$, where φ_N was defined in Proposition 2. In particular χ_g is uniquely determined by g and of necessity $u^{1/2} := \Pi_{j=1}^{\ell}(z-x_j)^{\kappa_j}$ divides χ_g. Then, condition (1a) implies that $u^{1/2}$ also divides $\pi_{d,g}$. Moreover, u divides d since $d/g \in L^2(\mathbb{R})$. After cancellation of the factor $u = u^{1/2}(u^{1/2})^*$ on both sides of (1b), we find that $\pi_{d,g}/u^{1/2}$ is uniquely determined in $\mathbf{P}_{N-1-\Sigma_j\kappa_j}$ by an equation of the Bezout type since the polynomials $\chi_g/u^{1/2}$ and $(\chi_g/u^{1/2})^*$ are coprime (for if χ_g had more real roots than those in $u^{1/2}$, counting multiplicities, they would also appear in χ_g^* and thus in g, contradicting the definition of u). This establishes the uniqueness part of property 1 and the existence part follows easily by reverting the computations.

Let $Y_{d,g}$ be as in property 1. It is a rational function in $H^2(\mathbb{C}^-)$ whose real part on $\mathbb{R}$ is $d/(2g)$ by (1c), therefore (26) implies for $k \in \{1 \dots m\}$ that

$$Y_{d,g}(x_k) = \frac{d}{2g}(x_k) + \frac{i}{2\pi}\lim_{\varepsilon\to 0}\int_{\varepsilon<|t-x_k|}\frac{d(t)}{g(t)}\frac{dt}{t-x_k} \tag{28}$$

and (24) entails that $\forall k \in \{1 \dots l\}$

$$Y_{d,g}(z_k) = \frac{i}{2\pi}\int_{-\infty}^{\infty}\frac{d(t)}{g(t)}\frac{dt}{t-z_k}. \tag{29}$$

Suppose now that g_1, g_2 are as in property 2. Note that $g_j(x_k) \neq 0$ for $j = 1,2$ and $1 \leq k \leq m$, since $d(x_k) \neq 0$ and $d/g_j \in L^2(\mathbb{R})$ by assumption. Separating real and imaginary parts in (28), we see from (2a) that $g_2(x_k) = g_1(x_k)$ for $1 \leq k \leq m$ and also that

$$\begin{aligned} J(x_k) &:= \lim_{\varepsilon\to 0}\int_{\varepsilon<|t-x_k|} d(t)\frac{g_2(t)-g_1(t)}{g_1(t)g_2(t)}\frac{dt}{t-x_k} \\ &= \int_{-\infty}^{\infty} d(t)\frac{g_2(t)-g_1(t)}{g_1(t)g_2(t)}\frac{dt}{t-x_k} = 0, \end{aligned} \tag{30}$$

where we omitted the principal value in the integral because *we claim* that the integrand is in fact non-singular. Indeed, even though g_1 and g_2 may have real zeros (some of which may be common to g_1 and g_2), the fraction $d(g_2-g_1)/g_1g_2$ has no pole on $\mathbb{R}$; for if λ is a zero of g_j with multiplicity μ_j and, say, $\mu_1 \geq \mu_2$, then λ is a zero of d with multiplicity at least μ_1 (as $d/g_1 \in L^2(\mathbb{R})$) and it is a zero of (g_2-g_1) of

multiplicity at least μ_2. Moreover, λ cannot coincide with x_k by our assumption that $d(x_k)\neq 0$, while we observed already that g_2-g_1 vanishes at x_k. *This proves the claim*.

Similarly we get from (29) and 2b that $\forall k\in\{1\dots l\}$

$$I(z_k)=\int_{-\infty}^{\infty} d(t)\frac{g_2(t)-g_1t)}{g_1(t)g_2(t)}\,\frac{dt}{t-z_k}=0, \tag{31}$$

and taking conjugates

$$\overline{I(z_k)}=\int_{-\infty}^{\infty} d(t)\frac{g_2(t)-g_1(t)}{g_1(t)g_2(t)}\,\frac{dt}{t-\overline{z_k}}=0. \tag{32}$$

We combine linearly equations (31), (32) and (30) using arbitrary complex coefficients $a=(a_1,\dots,a_l)^T$, $b=(b_1,\dots,b_l)^T$ and $c=(c_1,\dots,c_m)^T$ to obtain

$$\sum_{k=1}^{l}(a_kI(z_k))+\sum_{k=1}^{l}(b_k\overline{I(z_k)})+\sum_{k=1}^{m}c_kJ(x_k)=0.$$

Putting everything over a common denominator yields

$$\int_{-\infty}^{\infty} d(t)\frac{g_2(t)-g_1(t)}{g_1(t)g_2(t)\prod_{k=1}^{l}|t-z_k|^2}\,\frac{P_{a,b,c}(t)dt}{\prod_{k=1}^{m}(t-x_k)}=0, \tag{33}$$

where $P_{a,b,c}$ is the polynomial defined by

$$\begin{aligned}P_{a,b,c}(z)=&\sum_{k=1}^{l}a_k\prod_{j=1\dots l,j\neq k}(z-z_j)\prod_{j=1\dots l}(z-\overline{z_j})\prod_{j=1\dots m}(z-x_j)+\\ &\sum_{k=1}^{l}b_k\prod_{j=1\dots l}(z-z_j)\prod_{j=1\dots l,j\neq k}(z-\overline{z_j})\prod_{j=1\dots m}(z-x_j)+\\ &\sum_{k=1}^{m}c_k\prod_{j=1\dots l}(z-z_j)\prod_{j=1\dots l}(z-\overline{z_j})\prod_{j=1\dots m,j\neq k}(z-x_j).\end{aligned} \tag{34}$$

The $2l+m$ polynomials obtained by setting a_k, b_k, and c_k to 0 except for one of them which is set to 1 forms the Lagrange interpolating basis of $\mathbf{P}_{2l+m-1}$ at the points $\{x_j,z_k,\bar{z}_k\}$. Therefore $P_{a,b,c}$ ranges over $\mathbf{P}_{2l+m-1}$ as (a,b,c) ranges over $\mathbb{C}^m\times\mathbb{C}^l\times\mathbb{C}^l$. Observing now that g_2-g_1 vanishes at the x_k, (33) can be rewritten as

$$\int_{-\infty}^{\infty} d(t)\frac{P(t)P_{a,b,c}(t)}{g_1(t)g_2(t)\prod_{k=1}^{l}|t-z_k|^2}\,dt=0, \tag{35}$$

where $P(t)$ is the polynomial $(g_2(t)-g_1(t))/\prod_1^m(t-x_k)$. Note that P has degree at most $2N-1=2l+2m-1$ (for g_1, g_2 are monic of degree N). Hence, we can choose (a,b,c) so that $P_{a,b,c}=P$, and then we conclude from (35) that

$$\frac{d(t)P^2(t)}{g_1(t)g_2(t)\Pi_{k=1}^{l}|t-z_k|^2}=0,\qquad t\in\mathbb{R},$$

because it is everywhere non-negative and its integral is zero. Since d is not identically zero, we get that $P=0$ and consequently that $g_2=g_1$. This proves property 2.

As to property 3, observe that if $g\in\overset{\circ}{\mathbf{PE}}{}^{+}_{2N}$ (*i.e.* if $g\in\mathbf{PM}^{+}_{2N}$ has no real root, see discussion before Proposition 2), then d/g lies in in $L^2(\mathbb{R})$ hence also in $H^2(\mathbb{C}^-)$, and θ is well-defined by (27). Next, we compute the derivatives of $Y_{d,g}(x_k)$, $Y_{d,g}(z_k)$ with respect to the coefficients of g. Put

$$g(x)=x^{2N}+g_{2N-1}x^{2N-1}+\cdots+g_0.$$

Since $\chi_g=\varphi_N(g)$, we get from (16) that $\partial\chi_g/\partial g_j$ exists in $\mathbf{P}_{N-1}$ for $0\leq j\leq 2N-1$, and that

$$\chi_g^*(x)\frac{\partial\chi_g}{\partial g_j}(x)+\chi_g(x)\frac{\partial\chi_g^*}{\partial g_j}(x)=x^j \tag{36}$$

(note that $(\partial\chi_g/\partial g_j)^* = \partial\chi_g^*/\partial g_j$ since $*$ is a linear operation). Moreover, by property 1 already proved, $\pi_{d,g}$ is the solution to (1b) which is a nonsingular linear equation (for χ_g and χ_g^* are now coprime since they have no real root) whose coefficients depend linearly on the coefficients of χ_g. Hence $\partial\pi_{d,g}/\partial g_j$ also exists in $\mathbf{P}_{N-1}$, $0 \le j \le 2N-1$, and by the Leibnitz rule we have that

$$\chi_g^* \frac{\partial \pi_{d,g}}{\partial g_j} + \pi_{d,g} \frac{\partial \chi_g^*}{\partial g_j} + \chi_g \frac{\partial \pi_{d,g}^*}{\partial g_j} + \pi_{d,g}^* \frac{\partial \chi_g}{\partial g_j} = 0. \tag{37}$$

From the differentiability of χ_g, $\pi_{d,g}$ just pointed out, we get since evaluation at x_k is a linear operation and because $\chi_g(x_k) \neq 0$ that

$$\frac{\partial}{\partial g_j}\big(Y_{d,g}(x_k)\big) = F_{d,g,j}(x_k), \tag{38}$$

where

$$F_{d,g,j} = \frac{(\partial\pi_{d,g}/\partial g_j)\chi_g - \pi_{d,g}(\partial\chi_g/\partial g_j)}{\chi_g^2} \tag{39}$$

is a rational function in $H^2(\mathbb{C}^-)$ as it is the ratio of a polynomial of degree at most $2N-1$ by a stable polynomial of degree $2N$ (namely χ_g^2). Using (39), (37), (1b), (36) and the fact that $\chi_g = \varphi_N(g)$, we compute

$$\begin{aligned}
F_{d,g,j}(x) + F_{d,g,j}^*(x) &= \frac{\left(\frac{\partial\pi_{d,g}}{\partial g_j}\chi_g - \pi_{d,g}\frac{\partial\chi_g}{\partial g_j}\right)(\chi_g^*)^2 + \left(\frac{\partial\pi_{d,g}^*}{\partial g_j}\chi_g^* - \pi_{d,g}^*\frac{\partial\chi_g^*}{\partial g_j}\right)\chi_g^2}{\chi_g^2(\chi_g^*)^2}(x) \\
&= \frac{\left(\frac{\partial\pi_{d,g}}{\partial g_j}\chi_g^* + \frac{\partial\pi_{d,g}^*}{\partial g_j}\chi_g\right)\chi_g\chi_g^* - \left(\pi_{d,g}\frac{\partial\chi_g}{\partial g_j}(\chi_g^*)^2 + \pi_{d,g}^*\frac{\partial\chi_g^*}{\partial g_j}\chi_g^2\right)}{g^2}(x) \\
&= \frac{-\left(\frac{\partial\chi_g}{\partial g_j}\pi_{d,g}^* + \frac{\partial\chi_g^*}{\partial g_j}\pi_{d,g}\right)\chi_g\chi_g^* - \left(\pi_{d,g}\frac{\partial\chi_g}{\partial g_j}(\chi_g^*)^2 + \pi_{d,g}^*\frac{\partial\chi_g^*}{\partial g_j}\chi_g^2\right)}{g^2}(x) \\
&= \frac{-\frac{\partial\chi_g}{\partial g_j}\left(\pi_{d,g}^*\chi_g + \pi_{d,g}\chi_g^*\right)\chi_g^* - \frac{\partial\chi_g^*}{\partial g_j}\left(\pi_{d,g}\chi_g^* + \pi_{d,g}^*\chi_g\right)\chi_g}{g^2}(x) \\
&= -\frac{\left(\frac{\partial\chi_g}{\partial g_j}\chi_g^* + \frac{\partial\chi_g^*}{\partial g_j}\chi_g\right)d}{g^2}(x) = -\frac{d(x)x^j}{g^2(x)}.
\end{aligned} \tag{40}$$

Since $F_{d,g} + F_{d,g}^* = 2\Re F_{d,g}$ on $\mathbb{R}$, we obtain from (26), (38) and the previous computation:

$$\frac{\partial Y_{d,g}(x_k)}{\partial g_j} = -\frac{d(x_k)x_k^j}{2g^2(x_k)} - \frac{i}{2\pi}\lim_{\varepsilon\to 0^+}\int_{|x_k-t|>\varepsilon}\frac{d(t)t^j}{g^2(t)(t-x_k)}\,dt, \tag{41}$$

and combining linearly these partial derivatives leads us to the formula

$$D\big(Y_{d,g}(x_k)\big)[\delta g] = \frac{-d(x_k)\delta g(x_k)}{2g^2(x_k)} - \frac{i}{2\pi}\lim_{\varepsilon\to 0}\int_{\varepsilon<|t-x_k|}\frac{d(t)\delta g(t)}{g(t)^2}\,\frac{dt}{t-x_k}, \quad \forall \delta g \in \mathbf{P}_{\mathbb{R},2N-1}. \tag{42}$$

The companion formula

$$D\big(Y_{d,g}(z_\ell)\big)[\delta g] = \frac{-i}{2\pi}\int_{-\infty}^{\infty}\frac{d(t)\delta g(t)}{g(t)^2}\,\frac{dt}{t-z_\ell}, \quad \forall\ \delta g \in \mathbf{P}_{\mathbb{R},2N-1} \tag{43}$$

is obtained in the same manner, appealing to (24) rather than (26). Hereafter, we drop the dependence on d, g and we write for simplicity Y_{x_k} (resp Y_{z_ℓ}) instead of $Y_{d,g}(x_k)$ (resp. $Y_{d,g}(z_\ell)$). Then, we find that the application θ is differentiable with derivative

$$D\theta(g) : \delta g \in \mathbf{P}_{\mathbb{R},2N-1} \to \begin{pmatrix} DY_{x_1}(\delta g)) \\ \vdots \\ DY_{x_1}(\delta g) \\ DY_{z_1}(\delta g) \\ \vdots \\ DY_{z_l}(\delta g) \end{pmatrix} \in \mathbb{C}^N, \tag{44}$$

where $DY_{x_k}(\delta g)$ is given by (42) and $DY_{z_\ell}(\delta g)$ by (43).

Now, suppose that $\delta g \in \ker(D\theta)$. Separating real and imaginary parts in (42), we see that δg vanishes at every x_k. Consequently the principal value of the integral in (42) can be omitted, and this integral is zero for all x_k. Moreover, the integrals in (43) vanish for all z_ℓ. Thus, equating to zero an arbitrary linear combination of the integrals in (42) and those in (43) together with their conjugates, as x_k ranges over $\{x_1,\cdots,m\}$ and z_ℓ ranges over $\{z_1,\cdots,z_l\}$, we get in the same manner as we got (35) that

$$\forall P_{a,b,c} \in \mathbf{P}_{2l+m-1},\ \int_{-\infty}^{\infty} d(t)\frac{\hat{\delta}g(t)P_{a,b,c}(t)}{g^2(t)\prod_{\ell=1}^{l}|t-z_\ell|^2}\,dt = 0, \tag{45}$$

where $\hat{\delta}g$ is the real polynomial $\delta g/\prod_1^m(t-x_k)$. Picking $P_{a,b,c} = \hat{\delta}g$ in (45), we conclude since the integrand is nonnegative that $\hat{\delta}g = 0$, hence also $\delta g = 0$. Therefore $D\theta(g)$ is injective, thus it is invertible and θ is a local diffeomorphism. Finally, we know from property 2 that θ is injective, therefore it is a diffeomorphism from $\overset{\circ}{\mathbf{PM}}{}^{+}_{2N}$ onto its image. □

4.3. Injectivity of ψ. We can now establish that the map ψ introduced in (12) is one-to-one.

PROPOSITION 5. *The map ψ is injective.*

Proof. Let $\nu = (\gamma_1,\ldots,\gamma_m,\beta_1,\ldots,\beta_m) \in \mathbb{D}^m \times \mathbb{P}_Z^+$ and assume that there exist distinct polynomials $p_1(z)$ and $p_2(z)$ in $\mathbf{PM}_N$ such that $\psi(p_1) = \psi(p_2)$. Put $q_j = \varphi_N(p_j p_j^* + rr^*)$ for $j = \{1,2\}$, so that our assumption means:

$$\frac{p_1}{q_1}(x_k) = \frac{p_2}{q_2}(x_k),\ 1 \le k \le m, \quad \text{and} \quad \frac{p_1}{q_1}(z_\ell) = \frac{p_2}{q_2}(z_\ell),\ 1 \le l \le l. \tag{46}$$

By the Feldtkeller equation (6), $|p_j(t)/q_j(t)| \le 1$ for $t \in \mathbb{R}$, and $|p_j(t)/q_j(t)| = 1$ exactly when t is a real zero of r with multiplicity $\mu \ge 1$ which is not a zero of p_j of multiplicity greater than, or equal to μ; here, when p_j and q_j both vanish at t, the value $p_j(t)/q_j(t)$ is understood as the limit of $p_j(\tau)/q_j(\tau)$ when $\tau \to t$. In particular, there are at most $\deg r$ real numbers t for which $|p_j(t)/q_j(t)| = 1$, hence we can find a complex number ξ of modulus 1, distinct from -1, such that $1+\xi p_j/q_j$ is never zero on $\mathbb{R}$ for $j = \{1,2\}$. Consider the rational functions G_j, Y_j defined by

$$\begin{aligned} G_j(z) &\overset{def}{=} \frac{1-\xi\frac{p_j(z)}{q_j(z)}}{1+\xi\frac{p_j(z)}{q_j(z)}} = \frac{1-\xi}{1+\xi} + \left(\frac{2\xi}{1+\xi}\right)\frac{q_j(z)-p_j(z)}{q_j(z)+\xi p_j(z)} \\ &\overset{def}{=} \frac{1-\xi}{1+\xi} + Y_j(z). \end{aligned} \tag{47}$$

Being the Cayley transform of the Schur function $\xi p_j/q_j$, the function G_j is a Carathédory function and so is Y_j as it differs from G_j by the pure imaginary constant $(1-\xi)/(1+\xi)$. Now, our choice of ξ ensures the continuity of G_j, hence of Y_j, on the real axis. Moreover Y_j vanishes at infinity, since $\deg(p_j - q_j) \le N-1$ while $\deg(q_j+\xi p_j) = N$, therefore Y_j lies in $H^2(\mathbb{C}^-)$. A computation similar to (21) then yields that

$$Y_j + Y_j^* = G_j + G_j^* = \frac{2rr^*}{(q_j+\xi p_j)(q_j+\xi p_j)^*}. \tag{48}$$

We can apply Theorem 4 to $d = 2rr^*/|1+\xi|^2$ and $g_j = (q_j+\xi p_j)(q_j+\xi p_j)^*/|1+\xi|^2$, because on $\mathbb{R}$ we have that $d/g_j = \Re Y_j$ is square summable. So, if we set

$$\chi_j = \frac{q_j+\xi p_j}{1+\xi} \quad \text{and} \quad \pi_j = \left(\frac{2\xi}{(1+\xi)^2}\right)(q_j - p_j), \tag{49}$$

we see from (48), since $Y_j = \pi_j/\chi_j$, that the pair of polynomials χ_j, π_j satisfies assertions (1a), (1b), (1c) of that theorem. Therefore $\chi_j = \chi_{g_j}$ and $\pi_j = \pi_{d,g_j}$, hence property 2 of Theorem 4 implies that $\pi_1 = \pi_2$ and $\chi_1 = \chi_2$, consequently $p_1 = p_2$. □

4.4. Properness of ψ. Recall that a map is called *proper* if the preimage of a compact set is compact.

PROPOSITION 6. *The map $\psi : \mathbf{PM}_N \to \mathbb{D}^m \times \mathbb{P}_Z^+$ defined in (12) is proper.*

Proof. Let $K \subset \mathbb{D}^m \times \mathbb{P}^+$ be compact and put $W = \psi^{-1}(K)$. By the continuity of ψ, W is closed. Thus, it remains to prove that W is bounded in $\mathbf{PM}_N$.

Assume for a contradiction that there is an unbounded sequence p_n in $\psi^{-1}(K)$, and let us write $\psi(p_n) = (\gamma_1^{\{n\}}, \cdots, \gamma_m^{\{n\}}, \beta_1^{\{n\}}, \cdots \beta_l^{\{n\}})$. By definition $\gamma_j^{\{n\}} = p_n(x_k)/q_n(x_k)$ and $\beta_\ell^{\{n\}} = p_n(z_\ell)/q_n(z_\ell)$ with $q_n = \varphi_N(p_n p_n^* + rr^*)$, *cf.* Proposition 2, item *c*). Extracting a subsequence if necessary, we may assume that $\psi(p_n)$ converges to some $(\gamma_1, \cdots, \gamma_m, \beta_1, \cdots, \beta_l) \in K$ in $\mathbb{D}^m \times \mathbb{P}_Z^+$. For each n, by Euclidean division of $p_n(t)$ by $L(t) := \Pi_{k=1}^m (t - x_k)$, we can write

$$p_n(t) = \sum_{k=1}^{m} p_n(x_k) L_{x_k}(t) + L(t) h_n(t) \tag{50}$$

where $L_{x_k}(t) = \prod_{\substack{1 \le j \le m \\ j \neq k}} \frac{t - x_j}{x_k - x_j}$ is the k-th Lagrange interpolation polynomial of the set $\{x_1, \ldots x_m\}$ and h_n is a monic polynomial of degree $N - m = l$. It may of course happen that $m = 0$ (if there is no x_k), in which case we set $L \equiv 1$ and $L_{x_k} \equiv 0$; then $h_n = p_n$. To the opposite, it may be that $l = 0$ (if there is no z_ℓ) in which case $h_n = 1$.

Let $\|p_n\|$ indicate the norm of p_n in $\mathbf{PM}_N \sim \mathbb{C}^N$. The precise norm that we use is irrelevant for they are all equivalent. Since $p_n/\|p_n\|$ is bounded whereas $\|p_n\|$ is not, we may assume upon taking another subsequence if necessary that $\|p_n\| \to +\infty$ and $p_n/\|p_n\| \to g$ where $g \in \mathbf{P}_N$ is such that $\|g\| = 1$. In another connection, using (6), one easily checks that

$$\forall k \in \{1 \ldots m\} \; |p_n(x_k)|^2 = \frac{\left|\gamma_k^{\{n\}}\right|^2}{1 - \left|\gamma_k^{\{n\}}\right|^2} |r(x_k)|^2, \tag{51}$$

and since $\gamma_k^{\{n\}} \to \gamma_k \in \mathbb{D}$ we conclude that $p_n(x_k)$ is bounded independently of n. Then, dividing (50) by $\|p_n\|$ and letting $n \to \infty$, we get that $g = Lh$ where h is the limit of $h_n/\|p_n\|$. Observe that $h \in \mathbf{P}_{l-1}$ for $h_n/\|p_n\|$ has leading coefficient $1/\|p_n\| \to 0$ as $n \to \infty$. If $l = 0$ we are done, because then $h = 0$, contradicting the fact that $\|g\| = 1$.

Suppose next that $l > 0$ and rewrite the Feldtkeller equation after division by $\|p_n\|^2$ as

$$\frac{p_n p_n^*}{\|p_n\|^2} + \frac{rr^*}{\|p_n\|^2} = \varphi_{z_1}\left(\frac{p_n p_n^*}{\|p_n\|^2} + \frac{rr^*}{\|p_n\|^2}\right)\left(\varphi_{z_1}\left(\frac{p_n p_n^*}{\|p_n\|^2} + \frac{rr^*}{\|p_n\|^2}\right)\right)^*, \tag{52}$$

where the map φ_{z_1} defined in Proposition 2 item *a*) has been used since the polynomial $p_n p_n^*/\|p_n\|^2 + rr^*/\|p_n\|^2$ fails to be monic. Because φ_{z_1} is continuous except at 0, as shown in that proposition, and since $rr^*/\|p_n\|^2 \to 0$ while $p_n p_n^*/\|p_n\|^2 \to gg^* \neq 0$, we get from (52) that $\varphi_{z_1}((p_n p_n^* + rr^*)/\|p_n\|^2)$ converges to $\varphi_{z_1}(gg^*)$ in $\mathbf{P}_N$ as $n \to \infty$. Moreover, as $q_n/\|p_n\| = a_n \varphi_{z_1}\left((p_n p_n^* + rr^*)/\|p_n\|^2\right)$ for some $a_n \in \mathbb{C}$ with $|a_n| = 1$ by Proposition 2, we may assume upon extracting another subsequence that $a_n \to a$ with $|a| = 1$ and therefore that $q_n/\|p_n\| \to a\varphi_{z_1}(gg^*)$. In addition, since $L = L^*$ has only real roots, it holds that $\varphi_{z_1}(gg^*) = bL\varphi_{z_1}(hh^*)$ for some $b \in \mathbb{C}$ with $|b| = 1$. Therefore, because convergence in $\mathbf{P}_N$ implies pointwise convergence on $\mathbb{C}$ and since $\varphi_{z_1}(gg^*)$ has no zeros in $\mathbb{C}^-$ by definition of φ_{z_1}, we have that

$$\beta_\ell = \lim_{n\to\infty} \beta_\ell^{\{n\}} = \lim_{n\to\infty} \frac{p_n(z_\ell)}{q_n(z_\ell)} = \lim_{n\to\infty} \frac{p_n(z_\ell)/\|p_n\|}{q_n(z_\ell)/\|p_n\|} = \frac{g(z_\ell)}{a\varphi_{z_1}(gg^*)(z_\ell)} = \frac{h(z_\ell)}{ab\varphi_{z_1}(hh^*)(z_\ell)}.$$

Hence the $l \times l$ matrix $P(Z, \beta)$ defined by (9) is the Pick matrix corresponding to the interpolation data $(z_\ell, (h/(ab\varphi_{z_1}(hh^*)))(z_\ell))$, and since $h/(ab\varphi_{z_1}(hh^*))$ is a Blaschke product of degree at most $l - 1$ it cannot have full rank, see discussion after (9). This, however, contradicts the fact that $P(Z, \beta)$ is nonsingular by definition of $\mathbb{P}_Z^+$. ∎

4.5. ψ is a homeomorphism. We are now in position to prove the first claim of Theorem1. It will be convenient to invoke a famous result by Brouwer, known as *invariance of the domain* [32, chap. 10, sect. 62]: if $\Omega \subset \mathbb{R}^n$ is open and $f : \Omega \to \mathbb{R}^n$ is continuous and injective, then f is an open map; this means that f maps open sets to open sets. Hence $f(\Omega)$ is open and the inverse map $f^{-1} : f(\Omega) \to \Omega$ is continuous, that is: f is a homeomorphism onto its image.

PROPOSITION 7. *ψ defined in (12) is a homeomorphism from $\mathbf{PM}_N$ onto $\mathbb{D}^m \times \mathbb{P}^+$.*

Proof. We may regard ψ as a map from $\mathbb{R}^{2N}$ into $\mathbb{R}^{2N}$. By Corollary 3 and Proposition 5 it is continuous and injective, hence the image $\psi(\mathbf{PM}_N)$ is open and ψ is a homeomorphism onto this image, by invariance of the domain. In another connection, the properness of ψ implies that $\psi(\mathbf{PM}_N)$ is closed in $\mathbb{D}^m \times \mathbb{P}^+$. Indeed, suppose that $\psi(p_n)$ is a sequence in $\psi(\mathbf{PM}_N)$ that converges to some $\nu \in \mathbb{D}^m \times \mathbb{P}^+$. Because the union of a convergent sequence and its limit is compact, properness entails that we can extract a subsequence (p_{n_k}) converging to some $p \in \mathbf{PM}_N$, and then $\psi(p) = \nu$ by continuity. Hence $\psi(\mathbf{PM}_N)$ contains its limit point ν, thereby showing that it is closed.

Now, being the product of two connected topological spaces, $\mathbb{D}^m \times \mathbb{P}^+_Z$ is connected. Consequently $\psi(\mathbf{PM}_N)$, which is both open and closed in $\mathbb{D}^m \times \mathbb{P}^+_Z$, is either empty or the whole space. As it is certainly not empty ψ is surjective, as desired.

4.6. ψ is a diffeomorphism where differentiable. We established through Proposition 7 and Corollary 3 that $p \mapsto \psi(p)$ is a homeomorphism $\mathbf{PM}_N \to \mathbb{D}^m \times \mathbb{P}^+_Z$ which is differentiable at every p having no common real root with r. Clearly, such p form an open subset $\mathbf{PM}_N(r) \subset \mathbf{PM}_N$. To complete the proof of Theorem 1, it remains to prove:

PROPOSITION 8. *The map ψ is a diffeomorphism from $\mathbf{PM}_N(r)$ onto its image.*

Proof. We show that, locally, ψ restricted to $\mathbf{PM}_N(r)$ is a composition of diffeomorphisms involving the map θ defined in Theorem 4. This will ensure that ψ is a local diffeomorphism and, since it is a homeomorphism $\mathbf{PM}_N \to \mathbb{D}^m \times \mathbb{P}^+$ by Proposition 7, the proof will be complete.
If $p_0 \in \mathbf{PM}_N(r)$, then the polynomial $q_0 := \varphi_N(p_0 p_0^* + rr^*) \in \mathbf{SBM}_N$ is devoid of real roots. Argueing as we did before (47), there is $\xi \in \mathbb{C}$ of unit modulus, $\xi \neq -1$, such that $1 + \xi p_0/q_0$ is never zero on $\mathbb{R}$, hence $\xi p_0 + q_0$ has no real root and since $|p_0/q_0| < 1$ on $\mathbb{C}^-$ we conclude that $(\xi p_0 + q_0)/(1+\xi) \in \mathbf{SM}_N$.

Since $p_0 p_0^* + rr^* \in \mathbf{P}^{+\circ}_{2N}$ and $\mathbf{SM}_N$ is open in $\mathbf{PM}_N$ (*cf.* discussion before Proposition 2), the smoothness of φ_N around $p_0 p_0^* + rr^*$ and the continuity of $p \mapsto pp^* + rr^*$ ensures the existence of a neighborhood V of p_0 in $\mathbf{PM}_N(r)$ such that the map $\eta(p) := (\xi p + \varphi_N(pp^* + rr^*))/(1+\xi)$ is defined and differentiable on V with $\eta(V) \subset \mathbf{SM}_N$. *We claim* that its differential $D\eta$ is invertible at every $p \in V$. Indeed, it is enough to show that $D\eta$ is injective. Set for simplicity $q = \varphi_N(pp^* + rr^*)$ and observe that the kernel of $D\eta(p)$ consists of those $dp \in \mathbf{P}_{N-1}$ for which

$$\xi\, dp + dq = 0 \tag{53}$$

where $dq = D\varphi_N(pdp^* + p^* dp)$ satisfies (*cf.* (16))

$$q^* dq + q dq^* = pdp^* + p^* dp. \tag{54}$$

Combining the last two equations yields

$$\overline{\xi}(\xi p + q) dp^* + \xi(\xi p + q)^* dp = 0. \tag{55}$$

The polynomial $(\xi p + q)$ is strictly stable and therefore it is coprime with its paraconjugate, hence it must divide dp by (55). Since dp has degree at most $N-1$ while $(\xi p + q)$ has degree N (remember $\xi \neq -1$), this yields $dp = 0$ which *proves the claim*. Thus, η is a diffeormophism when restricted to V, in particular, $\eta(V)$ is open in $\mathbf{SM}_N$.

Next, consider the map $\mathfrak{m} : \eta(V) \to \mathbf{PM}^{+\circ}_{2N}$ given by $\mathfrak{m}(\nu) = \nu\nu^*$; to check that $\mathfrak{m}$ indeed maps $\eta(V)$ into the interior of $\mathbf{PM}^+_{2N}$, simply observe that $(\xi p + q)(\xi p + q)^*$ has no real root because so does $(\xi p + q)$ as it is strictly stable. Shrinking V if necessary, we get from Proposition 2 that $\mathfrak{m}$ is the restriction to $\eta(V)$ of φ_N^{-1} and therefore a diffeomorphism onto its image.

Then, putting $g = \eta(p)\eta(p)^*$ and $d = 2rr^*/|1+\xi|^2$, we see from Theorem 4 that the map θ defined in (27) allows us to evaluate at the interpolation points $(x_1, \cdots, x_m, z_1, \cdots, z_l)$ the positive real function

$$Y_{d,g} = (q - \xi p)/(q + \xi p) - \frac{1-\xi}{1+\xi}$$

in a diffeomorphic manner with respect to $m(\eta(p))$.

Eventually we need to come back to the "scattering domain", that is, we must compute the values $p(x_k)/q(x_k)$ and $p(z_\ell)/q(z_\ell)$, for $1 \le k \le m$ and $1 \le \ell \le l$, in terms of the $Y_{d,g}(x_k)$ and the $Y_{d,g}(z_\ell)$ in a

diffeomorphic manner. This is easily accomplished by smoothly inverting the correspondence $p_j/q_j \mapsto Y_j$ in equation (47). Specifically, upon defining $\tau : \mathbb{C}^+ \to \mathbb{D}$ by

$$\tau(z) = \overline{\xi}\frac{1-\left(\frac{1-\xi}{1+\xi}+z\right)}{1+\left(\frac{1-\xi}{1+\xi}+z\right)}, \qquad z \in \mathbb{C}^+, \tag{56}$$

we find that $\tau(Y_{d,g}) = p/q$. So, letting $\tau_N : (\mathbb{C}^+)^N \to \mathbb{D}^N$ act componentwise as τ, we find that on V

$$\psi = \tau_N \circ \theta \circ \varphi_N^{-1} \circ \eta \tag{57}$$

which expresses ψ locally as a composition of diffeomorphisms. □

REMARK 4.3. *In the decomposition (57), the maps τ_N and η depend on ξ and therefore on the point p_0 around which we carry out the local analysis of ψ. In fact, there is no global decomposition of ψ in terms of θ, but merely a collection of local ones, taylored so as to associate a non singular Carathéodory function Y (i.e. one having no pole on $\mathbb{R}$) to the initial Schur function (i.e. scattering element) p_0/q_0.*

Proposition 8 is of practical importance to solve Problem $\mathscr{P}$ numerically, because computationally efficient algorithms for the numerical inversion of ψ can be based on continuation techniques which themselves rely on the differentiability of ψ^{-1}, see Section 5. In this connection, we give below a genericity result that warrants the use of such techniques in the present context.

PROPOSITION 9. *$\psi(\mathbf{PM}_N(r))$ is an open, dense and connected subset of $\mathbb{D}^m \times \mathbb{P}^+$. Suppose that ν_0, ν_1 both lie in $\psi(\mathbf{PM}_N(r))$, and that γ is a continuous path from ν_0 to ν_1 in $\mathbb{D}^m \times \mathbb{P}^+$. Then, for every $\varepsilon > 0$ there exists a continuous path $\hat{\gamma}$ from ν_0 to ν_1 in $\psi(\mathbf{PM}_N(r))$ such that*

$$\sup_{t\in[0,1]} \|\hat{\gamma}(t) - \gamma(t)\| \leq \varepsilon,$$

where $\|.\|$ designates an arbitrary but fixed norm on $\mathbb{R}^{2N} \sim \mathbb{C}^N \supset \mathbb{D}^m \times \mathbb{P}^+$.

Proof. By Proposition 7 ψ is a homeomorphism $\mathbf{PM}_N \to \mathbb{D}^m \times \mathbb{P}^+$. Openness, density and connectedness of $\psi(\mathbf{PM}_N(r))$ in $\mathbb{D}^m \times \mathbb{P}^+$ will thus follow from the corresponding properties of $\mathbf{PM}_N(r)$ in $\mathbf{PM}_N$. These are easily verified, for if $\{\zeta_1, \cdots, \zeta_\mu\}$ are the real roots of r then $\mathbf{PM}_N(r)$ consists of those monic polynomials no root of which coincides with a ζ_j. This is clearly an open condition. Moreover, given any $p(z) = \Pi_{k=1}^N (z - \xi_k)$ in $\mathbf{PM}_N$, we can find ξ_k' arbitrary close to ξ_k which is not a ζ_j, thereby showing the density of $\mathbf{PM}_N(r)$. In addition, two polynomials $\Pi_{j=1}^N (z - \xi_k^{(1)})$ and $\Pi_{j=1}^N (z - \xi_k^{(2)})$ such that neither $\xi_k^{(1)}$ nor $\xi_k^{(2)}$ is a ζ_j can be deformed into each other within $\mathbf{PM}_N(r)$ by a map $t \mapsto \Pi_{j=1}^N (z - \xi_k(t))$ where $t \mapsto \xi_k(t)$, $t \in [0,1]$, is a continuous path from $\xi_k^{(1)}$ to $\xi_k^{(2)}$ in $\mathbb{C}$ which does not meet any ζ_j; hence $\mathbf{PM}_N(r)$ is connected.

Next, pick $\nu_0, \nu_1 \in \psi(\mathbf{PM}_N(r))$ and let $\gamma : [0,1] \to \mathbb{D}^m \times \mathbb{P}^+$ be a continuous map such that $\gamma(0) = \nu_0$ and $\gamma(1) = \nu_1$. Set $F : [0,1] \to \mathbf{PM}_N$ to be $F(t) = \psi^{-1}(\gamma(t))$. Thanks to the Stone-Weiestrass theorem, there is a sequence of polynomial maps $G_n : [0,1] \to \mathbf{PM}_N$ converging uniformly to F; here, by a polynomial map, we mean that each component is a polynomial in t. *We claim* that $\psi(G_n)$ converges uniformly to γ in the space of continuous maps $[0,1] \to \mathbb{D}^m \times \mathbb{P}^+$. To see this, we can select a compact neighborhood K of the compact set $F([0,1])$ in $\mathbf{PM}_N$ and observe, by Heine's theorem, that ψ is uniformly continuous on K. In particular, to each $\varepsilon > 0$, there exists $\delta > 0$ such that, for all $t \in [0,1]$ and $p \in \mathbf{PM}_N$, $||p - F(t)|| \leq \delta \Rightarrow ||\psi(p) - \gamma(t)|| \leq \varepsilon$. Letting $p = G_n(t)$ for n large enough that $\|G_n(t) - F(t)\| \leq \delta$ for all $t \in [0,1]$, we get that $\|\psi(G_n(t)) - \gamma(t)\| \leq \varepsilon$, *thereby proving the claim*.

We now show that γ can be uniformly approximated by paths contained in $\psi(\mathbf{PM}_N(r))$. Given $\varepsilon > 0$, let n_0 be so large that $||\psi(G_{n_0}(t)) - \gamma(t)|| \leq \varepsilon/2$ for all $t \in [0,1]$. For each root ζ_k of r, define a smooth map $\eta_k : [0,1] \to \mathbb{C}$ by $\eta_k(t) = -G_{n_0}(t)[\zeta_k]$, that is, evaluation of the polynomial $-G_{n_0}(t)$ at ζ_k. Sard's theorem [24, App.1] implies that the image of a smooth map from $\mathbb{R}$ into $\mathbb{R}^2$ has Lebesgue measure zero in $\mathbb{R}^2$, hence $\eta_k([0,1])$ has measure zero in $\mathbb{C}$. Therefore we can pick z of arbitrary small modulus in the set $\mathbb{C} \setminus \bigcup_{k=1}^\mu \eta_k([0,1])$. In particular, invoking Heine's theorem again, we can select $|z|$ so small that $||\psi(G_{n_0}(t) + z) - \psi(G_{n_0}(t)|| \leq \varepsilon/2$ for $t \in [0,1]$, which means that $||\psi(G_{n_0}(t) + z) - \gamma(t)|| \leq \varepsilon$. By construction, the polynomial $G_{n_0}(t) + z$ vanishes at no ζ_k which indicates that the path γ_1 defined by $\gamma_1(t) = \psi(G(t) + z)$ lies in $\psi(\mathbf{PM}_N(r))$ and uniformly approximates γ within a distance of ε.

Still, γ_1 does not meet our needs because its origin and endpoint need not be equal to ν_0 and ν_1 (although they lie within ε of them). To remedy this, we concatenate γ_1 with small line segments joining ν_0

to $\gamma_1(0)$ and v_1 to $\gamma_1(1)$ within $\psi(\mathbf{PM}_N(r))$. More precisely, as v_0 and v_1 both lie in the open set $\psi(\mathbf{PM}_N(r))$, we can find an open ball therein, centered at v_0 (resp. v_1) of radius $\varepsilon_0 < \varepsilon$. Let γ_2 be a path in $\psi(\mathbf{PM}_N(r))$ that uniformly approximates γ within $\varepsilon_0/3$. Such a path exists by the previous part of the proof. By uniform continuity of γ, there exists $\delta > 0$ such that $|t_0 - t_1| \leq \delta \Rightarrow ||\gamma(t_0) - \gamma(t_1)|| \leq \varepsilon_0/3$. We define the path $\hat{\gamma}$ by:

$$\hat{\gamma}(t) = \begin{cases} (1-\frac{t}{\delta})v_0 + \frac{t}{\delta}\gamma_2(\delta) & \text{if } t \in [0,\delta] \\ \gamma_2(t) & \text{if } t \in [\delta, 1-\delta] \\ (1-\frac{1-t}{\delta})\gamma_2(1-\delta) + \frac{1-t}{\delta}v_1 & \text{if } t \in [1-\delta, 1] \end{cases}$$

The triangular inequality yields that $||\gamma_2(\delta) - v_0|| \leq ||\gamma_2(\delta) - \gamma(\delta)|| + ||\gamma(\delta) - v_0|| \leq \frac{2\varepsilon_0}{3}$, which shows that the line segment between v_0 and $\gamma_2(\sigma)$ lies in $\psi(\mathbf{PM}_N(r))$. The same holds for the segment between $\gamma_2(1-\delta)$ and v_1. Besides, for $t \in [0,\delta]$, we have that

$$||\hat{\gamma}(t) - \gamma(t)|| \leq ||\hat{\gamma}(t) - v_0|| + ||v_0 - \gamma(t)|| \leq \frac{t}{\delta}||\gamma_2(\delta) - v_0|| + \varepsilon_0/3 \leq \varepsilon_0 < \varepsilon.$$

The same inequality holds for $t \in [1-\delta, 1]$, while for $t \in [\delta, 1-\delta]$ the equality $\hat{\gamma}(t) = \gamma_2(t)$ yields $||\hat{\gamma}(t) - \gamma(t)|| \leq \varepsilon_0/3 < \varepsilon$. This concludes the proof. □

4.7. Solution to $\hat{\mathscr{P}}$. Much like Problem $\mathscr{P}$, Problem $\hat{\mathscr{P}}$ can be studied *via* the evaluation map:

$$\hat{\psi}: \quad p \in \mathbf{P}_{\hat{N}} \to \begin{pmatrix} p(x_1)/q(x_1) \\ \vdots \\ p(x_m)/q(x_m) \\ p(z_1)/q(z_1) \\ \vdots \\ p(z_l)/q(z_l) \end{pmatrix}, \tag{58}$$

where, this time, q is computed from p using the maps defined in point b) or c) of Proposition 2 :

$$q = \begin{cases} \varphi_{x_1}(rr^* + pp^*) & \text{if } m > 0, \\ \varphi_{z_1}(rr^* + pp^*) & \text{if } m = 0. \end{cases} \tag{59}$$

Note that definition (59) is always legitimate, for $pp^* + rr^*$ is not the zero polynomial since $r \not\equiv 0$, and if $m > 0$ then $pp^* + rr^*$ cannot vanish at x_1 because $r(x_1) \neq 0$ by assumption.

Hereafter, we say that a polynomial $p \in \mathbf{P}_{\hat{N}}$ has n zeros at infinity if p has degree $\hat{N} - n$. Zeros at infinity are considered to lie on the real line.

The exact analog of Theorem 1 holds, namely:

THEOREM 10. *$\hat{\psi}$ is a homeomorphism from $\mathbf{P}_{\hat{N}}$ onto $\mathbb{D}^m \times \mathbb{P}^+$. The restriction of $\hat{\psi}$ to those $p \in \mathbf{P}_{\hat{N}}$ having no common real root with r (including at infinity) is a diffeomorphism onto its image.*

Remark 4.1 applies to Theorem 10 as well as to Theorem 1. It is worth emphasizing that the condition that p and r have no common zero at infinity, which is required in Theorem 10 for $\hat{\psi}$ to be a local diffeomorphism at p, means that one of them at least has exact degree $\hat{N}$.

The proof closely follows the path to Theorem 1 but with one significant difference, namely the analog of Y_j in (47), though still bounded, may no longer vanish at infinity. Thus, it needs not belong to $L^2(\mathbb{R})$ and Theorem 4 does not apply. Below, we state and prove a modified version of that theorem which is valid when d/g is merely bounded on $\mathbb{R}$. Subsequently, we outline a proof of Theorem 10 which runs parallel to that of Theorem 1. The statement refers to the notion of a smooth embedded manifold of dimension n_1 in $\mathbb{R}^{n_2}$, namely a subset of $\mathbb{R}^{n_2}$ which is locally the image of a C^∞-map $\Upsilon : U \to \mathbb{R}^{n_2}$, with $U \subset \mathbb{R}^{n_1}$ an open set, such that Υ is injective together with its derivative. Beyond this basic terminology, we use only two elementary facts from differential geometry, namely that the preimage of a manifold under a submersion (*i.e.* a map with surjective derivative) is a manifold with the same codimension, and that the image of a manifold under an immersion (*i.e.* a map with injective derivative) is locally a manifold of the same dimension, see *e.g.* [24, Ch. 1] or [38, Ch. 1]. In what follows, depending on whether $m > 0$ or $m = 0$, the normalization induced by (59) is either $q(x_1) > 0$ or $q(z_1) > 0$. We shall detail the proofs when $m > 0$, and indicate briefly the changes when $m = 0$.

THEOREM 11. *Let $d \in \mathbf{P}^+_{2\hat{N}}$ and $(x_1,\cdots,x_m)^T \in \mathbb{R}^m$, $(z_1,\cdots,z_l)^T \in (\mathbb{C}^-)^l$, with $m+l=\hat{N}+1$. Assume that $d(x_k)\neq 0$ for $k\in\{1,\cdots,m\}$. Then, the following three properties hold.*

1. *For each $g \in \mathbf{P}^+_{2\hat{N}}$ such that $\frac{d}{g}\in L^\infty(\mathbb{R})$, there uniquely exist polynomials $\chi_g \in \mathbf{SB}_{\hat{N}}$ and $\pi_{d,g}\in\mathbf{P}_{\hat{N}}$, with $\chi_g(x_1)>0$ (resp. $\chi_g(z_1)>0$ if $m=0$), such that the rational function $Y_{d,g}=\frac{\pi_{d,g}}{\chi_g}$ satisfies:*
   (a) $Y_{d,g}\in H^\infty(\mathbb{C}^-)$,
   (b) $\pi_{d,g}\chi_g^* + \pi_{d,g}^*\chi_g = d$,
   (c) $\Im(Y_{d,g}(x_1))=0$ (resp. $\Im(Y_{d,g}(z_1))=0$ if $m=0$),
   (d) $Y_{d,g}+Y_{d,g}^* = \frac{d}{g}$.
2. *Let g_1,g_2 in $\mathbf{P}^+_{2\hat{N}}$ be such that $\frac{d}{g_1}$ and $\frac{d}{g_2}$ are in $L^\infty(\mathbb{R})$. If*
   (a) $\forall k\in\{1..m\}\ Y_{d,g_1}(x_k)=Y_{d,g_2}(x_k)$,
   (b) $\forall k\in\{1..l\}\ Y_{d,g_1}(z_k)=Y_{d,g_2}(z_k)$,
   then $g_1=g_2$ whence $\pi_{d,g_1}=\pi_{d,g_2}$ and $\chi_{g_1}=\chi_{g_2}$, by 1.
3. *For fixed d, the evaluation map $\hat{\theta}:\overset{\circ}{\mathbf{PE}}{}^+_{2\hat{N}}\to\mathbb{R}\times\mathbb{C}^{\hat{N}}$ given by*

$$\hat{\theta}(g)=\begin{pmatrix} Y_{d,g}(x_1)\\ \vdots \\ Y_{d,g}(x_m)\\ Y_{d,g}(z_1)\\ \vdots\\ Y_{d,g}(z_l)\end{pmatrix} \tag{60}$$

   is well-defined and a diffeomorphism onto its image (observe that if $m>0$ then $Y_{d,g}(x_1)$ is real-valued and all other components of $\hat{\theta}$ are complex valued, whereas if $m=0$ then there are no x_k and $Y_{d,g}(z_1)$ is real valued while other components are complex valued).
4. *The set $\mathscr{M}_{2\hat{N}}(d)=\hat{\theta}^{-1}(\{1\}\times\mathbb{C}^{\hat{N}})$ is a smooth embedded submanifold of of $\overset{\circ}{\mathbf{PE}}{}^+_{2\hat{N}}$ of dimension $2\hat{N}$. For G the canonical projection from $\mathbb{R}\times\mathbb{C}^{\hat{N}}$ onto $\mathbb{C}^{\hat{N}}$ given by $(x,y_1\dots y_{\hat{N}})^t\to(y_1,\dots y_{\hat{N}})^t$, the map $\hat{\theta}_{red}\overset{def}{=}G\circ\hat{\theta}:\mathscr{M}_{2\hat{N}}(d)\to\mathbb{C}^{\hat{N}}$ is a diffeomorphism onto its image. Moreover, it holds that*

$$\mathscr{M}_{2\hat{N}}(d)=\left\{g\in\overset{\circ}{\mathbf{PE}}{}^+_{2\hat{N}},\ g(x_1)=d(x_1)/2\right\} \tag{61}$$

$$\left(\text{resp. } \mathscr{M}_{2\hat{N}}(d)=\left\{g\in\overset{\circ}{\mathbf{PE}}{}^+_{2\hat{N}},\ \int_{-\infty}^{\infty}\frac{d(t)}{g(t)}\frac{dt}{|t-z_1|^2}=-\frac{2\pi}{\Im(z_1)}\right\}\ \text{if } m=0\right). \tag{62}$$

Proof. As to property 1, observe from (1b) and (1d) that necessarily $\chi_g=\varphi_{x_1}(g)$ (resp. $\varphi_{z_1}(g)$ if $m=0$) . Let us check that equation (1b) is then solvable with respect to $\pi_{d,g}\in\mathbf{P}_{\hat{N}}$. In doing so, we may as well assume that $2\hat{N}$ is the exact degree of g, and therefore that $\deg d\leq 2\hat{N}$ (since $d/g\in L^\infty(\mathbb{R})$) as well as $\deg\chi_g=\hat{N}$. Let Δ be the monic g.c.d. of χ_g and χ_g^*. Clearly all roots of Δ are real, and $\Delta=\Delta^*$. Of necessity, Δ^2 divide g, therefore also d since $d/g\in L^\infty(\mathbb{R})$. Since χ/Δ and χ^*/Δ are coprime, we can certainly solve the Bezout-type equation $A\chi_g^*/\Delta+B\chi_g/\Delta=d/\Delta^2$ with $A,B\in\mathbf{P}_{\hat{N}-\deg\Delta}$. Since $(d/\Delta^2)^*=d/\Delta^2$, we may replace A with $A_1=(A+B^*)/2$ and B with $B_1=(A^*+B)/2$. Then, $\pi_{d,g}=\Delta A_1$ solves for (1b) and (1d) is satisfied by construction. Equation (1b) characterizes $\pi_{d,g}$ up to the addition of a pure imaginary multiple of χ_g only, but the latter is determined by condition (1c). Clearly $Y_{d,g}$ just constructed belongs to $H^\infty(\mathbb{C}^-)$, because it is a rational Carathéodory function with no pole on $\mathbb{R}$ since Δ divides $\pi_{d,g}$. This shows both existence and uniqueness of the pair $\chi_g,\pi_{d,g}$).

We turn to property 2. Note that the vanishing at infinity of $Y_{d,g}$ in Theorem 4 (induced by the condition $\deg\chi_g=N>N-1\geq\deg\pi_{d,g}$) is replaced here by the normalization condition (1c) at some interpolation point. This is to the effect that $Y_{d,g}$ (which belongs to $H^\infty(\mathbb{C}^-)$) may not belong to $H^2(\mathbb{C}^-)$ because it may not vanish at infinity. For that reason, slightly different kernels than those in (28) and (29) are required to represent $Y_{d,g}$ in terms of its real part on $\mathbb{R}$. Below, we discuss the case where $m>0$ so that (1c) bears on x_1.

First, let d, g, χ_g and $Y_{d,g}$ be as before. Then, it holds by (1c), (1d) that

$$Y_{d,g}(x_1)=\frac{d}{2g}(x_1). \tag{63}$$

By Euclidean division, we can write $Y_{d,g}(z) = C + H(z)$ where $H \in H^2(\mathbb{C}^-)$ and $C = Y_{d,g}(\infty)$ is a complex constant. Since $C = Y_{d,g}(x_1) - H(x_1)$, we get from (1c) and (25) that

$$\Im(C) = -\frac{i}{\pi} \lim_{\varepsilon\to 0^+} \int_{|x_1-t|>\varepsilon} \frac{\Re(H(t))}{t-x_1}\, dt, \tag{64}$$

ensuing by (1d) and (25) again that

$$\begin{aligned} Y_{d,g}(x_k) &= \frac{d}{2g}(x_k) + \frac{i}{\pi} \lim_{\varepsilon\to 0} \int_{\substack{\varepsilon<|t-x_k| \\ \varepsilon<|t-x_1|}} \Re(H(t)) \left(\frac{dt}{t-x_k} - \frac{dt}{t-x_1} \right) \\ &= \frac{d}{2g}(x_k) + \frac{i}{\pi} \lim_{\varepsilon\to 0} \int_{\substack{\varepsilon<|t-x_k| \\ \varepsilon<|t-x_1|}} \Re(H(t)) \frac{(x_k-x_1)dt}{(t-x_k)(t-x_1)}. \end{aligned} \tag{65}$$

In another connection, it is elementary to check that

$$\lim_{\varepsilon\to 0} \int_{\substack{\varepsilon<|t-x_k| \\ \varepsilon<|t-x_1|}} \frac{dt}{(t-x_k)(t-x_1)} = 0, \tag{66}$$

therefore $\Re(H)$ may be replaced by $\Re(Y_{d,g})$ under the integral sign in (65) to yield

$$Y_{d,g}(x_k) = \frac{d}{2g}(x_k) + \frac{i}{2\pi} \lim_{\varepsilon\to 0} \int_{\substack{\varepsilon<|t-x_k| \\ \varepsilon<|t-x_1|}} \frac{d(t)}{g(t)} \frac{(x_k-x_1)dt}{(t-x_k)(t-x_1)}, \qquad k \in \{2\dots m\}, \tag{67}$$

where we used (1d) again. Note that the kernel in (67) decays like $|t|^{-2}$ for large $|t|$, hence this singular integral makes sense even though d/g may not vanish at infinity. In the same manner, we obtain using (24) instead of (25) that

$$Y_{d,g}(z_\ell) = \frac{i}{2\pi} \int_{\varepsilon<|t-x_1|} \frac{d(t)}{g(t)} \frac{(z_\ell-x_1)dt}{(t-z_\ell)(t-x_1)}, \qquad \ell \in \{1\dots l\}. \tag{68}$$

Now, let g_1, g_2 be as in property 2. Then, in view of (63) and (67) where we separate real and imaginary parts, we get from property (2a) and (1d) that $g_1(x_k) = g_2(x_k)$ for $1 \le k \le m$. Next, writing by (2a) again that $Y_{d,g_1} - Y_{d,g_2}$ vanishes at x_k, z_ℓ for $2 \le k \le m$, $1 \le \ell \le l$, and adjoining the equations conjugate to those at z_ℓ while using representations (67) and (68), we get $m+2l-1$ equations that we can linearly combine together so as to get (33), where this time $P_{a,b}$ ranges over $\mathbf{P}_{2l+m-2}$ and again $d(g_1-g_2)/g_1g_2$ has no real pole while g_1-g_2 vanishes at x_k for $1 \le k \le m$. Since $(g_1-g_2)/\Pi_{k=1}^m(z-x_k)$ has degree at most $2\hat{N}-m = m+2l-2$, we can pick $P_{a,b}$ to be that polynomial thereby making the integrand nonnegative in (33). Consequently this integrand is identically zero whence $g_1 = g_2$, as desired. The case where $m = 0$ and (1c) bears on z_1 is similar but easier, since we no longer need (67) and we can base the whole argument on the representing formula

$$Y_{d,g}(z) = \frac{i}{2\pi} \int_{-\infty}^{\infty} \frac{d(t)}{g(t)} \left(\frac{1}{t-z} - \frac{t-\Re(z_1)}{|t-z_1|^2} \right) dt, \qquad z \in \mathbb{C}^-, \tag{69}$$

where we note that the kernel in between parentheses behaves like $|t|^{-2}$ at infinity, locally uniformly with respect to z, so that the integral converges and defines a holomorphic function of $z \in \mathbb{C}^-$ by the boundedness of d/g. To check the validity of (69), observe on the one hand that the right side has real part the Poisson integral of $d/2g$ (recall that the Poisson kernel for $\mathbb{C}^-$ is $\pi^{-1}\Im(1/(z-t))$, compare [21, Ch. I, Eqn. (3.4)]). On the other hand, as $Y_{d,g}$ lies in $H^\infty(\mathbb{C}^-)$, it real part is a bounded harmonic function on $\mathbb{C}^-$ and therefore it is the Poisson integral of its nontangential limit [21, Ch. I, Thm. 5.3]. Therefore both sides of (69) have the same real part, thus they represent the same analytic function in $\mathbb{C}^-$, up to an additive pure imaginary constant. But since $\Im(Y_{d,g}(z_1)) = 0$, this constant must be zero because it is obvious that the right hand side of (69) is real when $z = z_1$. This confirms that (69) holds.

To prove property 3, first note that $d/g \in L^\infty(\mathbb{R})$ when $g \in \overset{\circ}{\mathbf{PE}}{}^+_{2\hat{N}}$, since the latter consists of strictly positive polynomials on $\mathbb{R}$ having exact degree $2\hat{N}$, see discussion before Proposition 2. Hence $\hat{\theta}$ is well-defined with domain an open subset of $\mathbb{R}^{2\hat{N}+1}$ and values in $\mathbb{R}\times\mathbb{R}^{2\hat{N}} = \mathbb{R}^{2N+1}$ Observe also that χ_g is strictly stable when $g \in \overset{\circ}{\mathbf{PE}}{}^+_{2\hat{N}}$, hence χ_g and χ_g^* are coprime. So, if we write

$$g(x) = g_{2\hat{N}}x^{2\hat{N}} + g_{2\hat{N}-1}x^{2\hat{N}-1} + \cdots + g_0,$$

the differentiability of χ_g with respect to the coefficients g_j follows from (15), while the differentiability of $\pi_{d,g}$ with respect to the g_j comes from the fact that it solves a nonsingular linear system of equations whose coefficients are smooth (in fact: linear) in the coefficients of χ_g. Thus, $\hat{\theta}$ is differentiable, and since it is injective by property 2 it remains to show that its differential $D\hat{\theta}(g)$ is injective (and therefore invertible) at every point g. Assume first that $m > 0$. Then, differentiating $Y_{d,g}(x_1) = d(x_1)/(2g(x_1))$, we get since evaluation at x_1 is linear that

$$D_g Y_{d,g}(x_1)[\delta g] = -\frac{\delta g(x_1) d(x_1)}{2g^2(x_1)}, \qquad \delta g \in \mathbf{P}_{\mathbb{R},2\hat{N}}, \tag{70}$$

where D_g indicates the partial differential with respect to g. Moreover, equations (36) and (37) hold for $0 \leq j \leq 2\hat{N}$, hence (38) and (39) remain valid and we obtain as in (40) that

$$\left(\frac{\partial}{\partial g_j} Y_{d,g}\right)(x) + \left(\frac{\partial}{\partial g_j} Y_{d,g}\right)^*(x) = -\frac{d(x)x^j}{g^2(x)}, \qquad 0 \leq j \leq 2\hat{N}.$$

Thus, writing the analogs of (67), (68) for $\partial Y_{d,g}/\partial g_j$ rather than $Y_{d,g}$ and combining the corresponding equations linearly, we find for all $\delta g \in \mathbf{P}_{\mathbb{R},2\hat{N}}$ and $2 \leq k \leq m$ that

$$D_g\big(Y_{d,g}(x_k)\big)[\delta g] = \frac{-d(x_k)\delta g(x_k)}{2g^2(x_k)} - \frac{i(x_k - x_1)}{2\pi} \lim_{\varepsilon \to 0} \int_{\substack{\varepsilon < |t-x_1| \\ \varepsilon < |t-x_k|}} \frac{d(t)\delta g(t)}{g(t)^2} \frac{dt}{(t-x_1)(t-x_k)}, \tag{71}$$

while for $1 \leq \ell \leq l$ it holds that

$$D_g\big(Y_{d,g}(z_\ell)\big)[\delta g] = -\frac{i(z_\ell - x_1)}{2\pi} \int_{\varepsilon < |t-x_1|} \frac{d(t)\delta g(t)}{g(t)^2} \frac{dt}{(t-x_1)(t-z_\ell)}. \tag{72}$$

Assume now that δg lies in the kernel of $D\hat{\theta}(g)$:

$$D_g Y_{d,g}(x_k)[\delta g] = 0, \quad 1 \leq k \leq m, \tag{73}$$

$$D_g Y_{d,g}(z_\ell)[\delta g] = 0, \quad 1 \leq \ell \leq l. \tag{74}$$

□

In view of (70), and since d vanishes at no x_k by assumption, we deduce firstly from (73) that $\delta g(x_1) = 0$, and secondly taking real parts in (73) and (71) that $\delta g(x_k) = 0$ for $2 \leq k \leq m$. In particular $\hat{\delta g}(x) = \delta g(x)/\Pi_k(x - x_k)$ is a real polynomial, and the principal part of the integral can be omitted in (71) and (72). Next, combining linearly (with arbitrary complex coefficients) the equations (73) for $2 \leq k \leq m$ together with the equations (74) augmented with their conjugates, we get upon substituting therein (71) and (72) while making use of $\delta g(x_k) = 0$ that (45) holds, where this time, $P_{a,b,c}$ ranges over $\mathbf{P}_{2l+m-2}$. Since $\hat{\delta g}$ has degree at most $2\hat{N} - m = 2l + m - 2$, we can pick $P_{a,b,c} = \hat{\delta g}$ thereby making the integrand in (45)non-negative. Therefore the integrand is identically zero, implying that $\delta g = 0$, as desired. The case where $m = 0$ is similar but easier, applying the analog of (69) to $\partial Y_{d,g}/\partial g_j$ rather than $Y_{d,g}$.

As to property 4, remark that $D\hat{\theta}(g)$ is surjective at every $g \in \overset{\circ}{\mathbf{PE}}{}^+_{2\hat{N}}$ by property 3. Therefore, the preimage $\hat{\theta}^{-1}(\{1\} \times \mathbb{C}^{\hat{N}})$ of the affine submanifold $\{1\} \times \mathbb{C}^{\hat{N}}$ of $\subset \mathbb{R} \times \mathbb{C}^{\hat{N}}$ is a smooth embedded submanifold of $\overset{\circ}{\mathbf{PE}}{}^+_{2\hat{N}}$ with the same codimension, namely 1 (see [24, Ch. 1, p.28]). This shows that $\mathscr{M}_{2\hat{N}}(d)$ is a smooth embedded submanifold of $\overset{\circ}{\mathbf{PE}}{}^+_{2\hat{N}}$ of real dimension $2\hat{N}$. Moreover, the tangent space $\mathscr{T}_g\mathscr{M}_{2\hat{N}}(d)$ to $\mathscr{M}_{2\hat{N}}(d)$ at g is the preimage under $D\hat{\theta}(g)$ of the tangent space to $\{1\} \times \mathbb{C}^{\hat{N}}$ at $\hat{\theta}(g)$ which is but $\{0\} \times \mathbb{C}^{\hat{N}}$ (see [24, Ch. 1, p.32, ex. 5]). This implies that $G \circ D\hat{\theta}(g)(v) \neq 0$ whenever $0 \neq v \in \mathscr{T}_g\mathscr{M}_{2\hat{N}}(d)$, otherwise $D\hat{\theta}(g)(v)$ would be zero (since the first component is already known to vanish), thereby contradicting the injectivity of $D\hat{\theta}(g)$. Hence the restriction of $D\hat{\theta}(g)$ to $\mathscr{T}_g\mathscr{M}_{2\hat{N}}(d)$ is injective, therefore an isomorphism onto $\{0\} \times \mathbb{C}^{\hat{N}}$. In view of the local inversion theorem, this proves that $\hat{\theta}_{red}$, which is already known to be a homeomorphism $\mathscr{M}_{2\hat{N}}(d) \to \{1\} \times \mathbb{C}^{\hat{N}}$ (being a restriction of $\hat{\theta}$), is in fact a diffeomorphism.

Finally, characterization (61) follows directly from formula (63), while characterization (62) is obtained upon evaluating (69) at z_1 and equating the result to 1.

Proof. (of Theorem 10) If $m>0$, then pp^*+rr^* cannot vanish at x_1 since $r(x_1)\neq 0$ by assumption. Thus, the continuity of $\hat\psi$ follows from (59) and the continuity of φ_{x_1} (resp. φ_{z_1} if $m=0$) in Proposition 2. Injectivity is proved like in Proposition 5 upon choosing ξ so that G_j defined by (47) lies in $H^\infty(\mathbb{C}^-)$ for $j=1,2$, and appealing to property 2 of Theorem 11 (rather than of Theorem 4) with $d=2rr^*$, $g_j=(q_j+\xi p_j)(q_j+\xi p_j)^*$, $\chi_{g_j}=(q_j+\xi p_j)$ and $\pi_{d,g_j}=(q_j-\xi p_j)$. To secure the choice of ξ, as p_j, q_j are no longer monic, we trade the requirement made in Proposition 5 that $\xi\neq -1$ for the requirement that $\deg(q_j+\xi p_j)=\deg q_j$, which is obviously possible since $\deg q_j\geq \deg p_j$ by (59) and (6). Properness of $\hat\psi$ is established as in Proposition 6, noting that now $\deg h_n\leq l-1$ by construction. Then, reasoning as in Proposition 7 shows that $\hat\psi$ is a homeomorphism. Finally, let $\mathbf{P}_{\hat N}(r)\subset\mathbf{P}_{\hat N}$ be the subset of polynomials having no common real root with r including at infinity, which is easily seen to be open. As in corollary 3 one checks that $\hat\psi$ is differentiable on $\mathbf{P}_{\hat N}(r)$. To prove that $\hat\psi$ restricted to $\mathbf{P}_{\hat N}(r)$ is a local diffeomorphism, we write it locally as a composition of local diffeomorphisms, like we did to obtain (57). The arguments, however, are a little different and we detail them below. We consistently denote with q_p (or simply with q if p is understood) the polynomial defined by (59).

As in Theorem 11, let G be the canonical projection from $\mathbb{R}\times\mathbb{C}^{\hat N}$ onto $\mathbb{C}^{\hat N}$. We define

$$\mathbf{P}_{\hat N,x_1}=\{p\in\mathbf{P}_{\hat N},\ p(x_1)=0\},\qquad \big(\text{resp. }\ \mathbf{P}_{\hat N,z_1}=\{p\in\mathbf{P}_{\hat N},\ p(z_1)=0\}\text{ if }m=0\big),$$

and

$$\mathbf{P}_{\hat N,x_1}(r)=\{p\in\mathbf{P}_{\hat N}(r),\ p(x_1)=0\},\qquad \big(\text{resp. }\ \mathbf{P}_{\hat N,z_1}(r)=\{p\in\mathbf{P}_{\hat N}(r),\ p(z_1)=0\}\text{ if }m=0\big).$$

Note that $\mathbf{P}_{\hat N,x_1}$ (resp. $\mathbf{P}_{\hat N,z_1}$) is isomorphic to $\mathbb{R}^{2\hat N}$ and that $\mathbf{P}_{\hat N,x_1}(r)$ (resp. $\mathbf{P}_{\hat N,z_1}(r)$) is an open subset thereof. In a first step, we prove that the map $\hat\psi_{red}\overset{def}{=}G\circ\hat\psi$ defines a homeomorphism from $\mathbf{P}_{\hat N,x_1}$ onto $\{0\}\times\mathbb{D}^{m-1}\times\mathbb{P}^+$ (resp. $\{0\}\times\mathbb{P}^+_{z_2,\cdots,z_{\hat N+1}}$ if $m=0$) and a diffeomorphism onto its image when restricted to $\mathbf{P}_{\hat N,x_1}(r)$ (resp. $\mathbf{P}_{\hat N,z_1}(r)$). In fact, as $q(x_1)>0$ (resp. $q(z_1)>0$ if $m=0$), the relation $p(x_1)=0$ (resp. $p(z_1)=0$) amounts to $(p/q)(x_1)=0$ (resp. $(p/q)(z_1)=0$), and since we know $\hat\psi$ is a homeomorphism $\mathbf{P}_{\hat N}\to\mathbb{D}^m\times\mathbb{P}^+$ it follows from its very definition that it induces by restriction a homeomorphism from $\mathbf{P}_{\hat N,x_1}$ onto $\{0\}\times\mathbb{D}^{m-1}\times\mathbb{P}^+$ (resp. $\{0\}\times\mathbb{P}^+_{z_2,\cdots,z_{\hat N+1}}$). Obviously then, $\hat\psi_{red}$ is a homeomorphism from $\mathbf{P}_{\hat N,x_1}$ onto $\mathbb{D}^{m-1}\times\mathbb{P}^+$ (resp. $\mathbb{P}^+_{z_2,\cdots,z_{\hat N+1}}$).

Next, the differentiability of $\hat\psi_{red}$ on $\mathbf{P}_{\hat N,x_1}(r)$ (resp. $\mathbf{P}_{\hat N,z_1}(r)$) follows from the differentiability of $\hat\psi$ on $\mathbf{P}_{\hat N}(r)$, and it remains to show that $\hat\psi_{red}$ has non-singular differential there.

If $p\in\mathbf{P}_{\hat N,x_1}(r)$ (resp. $\mathbf{P}_{\hat N,z_1}(r)$), then $pp^*+rr^*\in\overset{\circ}{\mathbf{PE}}{}^+_{2\hat N}$ (for p in $\mathbf{P}_{\hat N}(r)$ has no common root with r including at infinity), hence q belongs to $\mathbf{SE}_{\hat N}$ and is a smooth function of p by Proposition 2. In a neighborhood V of $p_0\in\mathbf{P}_{\hat N,x_1}(r)$ (resp. $\mathbf{P}_{\hat N,z_1}(r)$), define $\hat\eta(p)=\xi p+q$ where $\xi\in\mathbb{C}$ is such that $|\xi|=1$ and $\xi p_0+q_{p_0}\in\mathbf{SE}_{\hat N}$. That ξ exists can be shown as in the beginning of the proof of Proposition 8, replacing the condition $\xi\neq -1$ by $\xi p_0[\hat N]\neq -q_{p_0}[\hat N]$ which is clearly an open condition by the continuity of $p\mapsto q_p$ (here, the symbol $[\hat N]$ means that we select the coefficient of degree $\hat N$). Shrinking V is necessary, we may assume that $\xi p+q_p\in\mathbf{SE}_{\hat N}$ for all $p\in V$. If $dp\in\mathbf{P}_{\hat N,x_1}$ (resp. $\mathbf{P}_{\hat N,z_1}$) lies in the kernel of the derivative $D\hat\eta(p)$, then (55) holds. As $(\xi p+q)^*$ is coprime to $\xi p+q$ by stability of the latter, and since $\deg dp\leq\hat N=\deg(\xi p+q)$, it follows that $dp=\lambda(\xi p+q)$ for some $\lambda\in\mathbb{C}$. Evaluating at x_1 (resp. z_1) yields $\lambda=0$ for $q_p(x_1)>0$ (resp. $q_p(z_1)>0$). The derivative $D\hat\eta(p):\mathbf{P}_{\hat N,x_1}\to\mathbf{P}_{\hat N}$ is therefore injective at p_0 and so, for sufficiently small V, the map $\hat\eta$ is a diffeormorphism from V onto a smooth embedded submanifold $\hat\eta(V)\subset\mathbf{SE}_{\hat N}$ of dimension $2\hat N$ [38, Ch. 1, Cor. (f)]. Note that, by construction, $\hat\eta(V)\subset\{P\in\mathbf{SE}_{\hat N},\ P(x_1)>0\}$ (resp. $\{P\in\mathbf{SE}_{\hat N},\ P(z_1)>0\}$).

Consider now the map $\hat{\mathfrak{m}}:\hat\eta(V)\to\overset{\circ}{\mathbf{PE}}{}^+_{2\hat N}$ given by $\hat{\mathfrak{m}}(v)=vv^*$. Clearly $\hat{\mathfrak{m}}$ is the restriction to $\hat\eta(V)$ of $\varphi_{x_1}^{-1}$ (resp. $\varphi_{z_1}^{-1}$), and we get from Proposition 2 that it is a diffeomorphism onto an embedded submanifold $W\subset\overset{\circ}{\mathbf{PE}}{}^+_{2\hat N}$ of dimension $2\hat N$. By construction, the elements of W can uniquely be written as $(\xi p+q_p)(\overline{\xi}p^*+q_p^*)$ with $p\in V\subset\mathbf{P}_{\hat N,x_1}(r)$. Thus, the elementary computation

$$(q-\xi p)(\xi p+q)^*+(q-\xi p)^*(\xi p+q)=2rr^*$$

together with the fact that $(\xi p+q)(x_1)=q(x_1)>0$ (resp. $(\xi p+q)(z_1)=q(z_1)>0$) and the relation

$$\frac{q-\xi p}{q+\xi p}(x_1)=1\qquad\left(\text{resp. }\frac{q-\xi p}{q+\xi p}(z_1)=1\right)$$

allow us for an application of Theorem 11 with $d = 2rr^*$ and $g = (\xi p + q_p)(\overline{\xi} p^* + q_p^*) = \hat{\mathfrak{m}} \circ \hat{\eta}(p)$. In the notation of that theorem we have that $\pi_{d,g} = q - \xi p$ and $\chi_g = q + \xi p$, hence $Y_{d,g} = (q - \xi p)/(q + \xi p)$. In particular, we deduce from point 4 that $\hat{\mathfrak{m}} \circ \hat{\eta}(V) \subset \mathscr{M}_{2\hat{N}}(d)$. Moreover, letting $\hat{\tau} : (\mathbb{C}^+)^{\hat{N}} \to \mathbb{D}^{\hat{N}}$ act componentwise as $z \mapsto (1-z)/(1+z)$, we easily check that

$$\hat{\psi}_{red} = \frac{1}{\xi} \hat{\tau} \circ \hat{\theta}_{red} \circ \hat{\mathfrak{m}} \circ \hat{\eta}$$

where $\hat{\theta}_{red}$ is the diffeomorphism from $\mathscr{M}_{2\hat{N}}(d)$ into $\mathbb{C}^{\hat{N}}$ introduced in Theorem 11 point 4. This indicates that $\hat{\psi}_{red}$ admits a local representation as a composition of diffeomorphisms on the neighborhood V of p_0. Since the latter was arbitrary in $\mathbf{P}_{\hat{N},x_1}$ (resp. $\mathbf{P}_{\hat{N},z_1}$) and $\hat{\psi}_{red}$ is already known to be a homeomorphism, it follows that it is a diffeomorphism from $\mathbf{P}_{\hat{N},x_1}(r)$ (resp. $\mathbf{P}_{\hat{N},z_1}(r)$) onto its image in in $\mathbb{C}^{\hat{N}}$. This completes the first step.

In a second step, we pass from $\hat{\psi}_{red}$ to $\hat{\psi}$. For this, observe that for any $\alpha \in \mathbb{D}$

$$\begin{aligned}(q - \overline{\alpha} p)(q - \overline{\alpha} p)^* - (\alpha q - p)(\alpha q - p)^* &= (1 - |\alpha|^2)(qq^* - pp^*) \\ &= (1 - |\alpha|^2) rr^*,\end{aligned} \tag{75}$$

and that

$$\frac{\alpha q - p}{q - \overline{\alpha} p} = \frac{\alpha - \frac{p}{q}}{1 - \overline{\alpha} \frac{p}{q}} = M_\alpha(p/q) \tag{76}$$

where $M_\alpha(z) \overset{def}{=} (\alpha - z)/1 - \bar{\alpha} z)$ is the familiar automorphism of the unit disk swaping 0 and α. Together, (75) and (76) imply that if (p,q) is the pair of polynomials solving for problem $\hat{\mathscr{P}}$ with interpolation values $(0, \gamma_2, \dots \gamma_m, \beta_1 \dots \beta_l)$ (resp. $(0, \beta_2 \dots \beta_{\hat{N}+1})$ if $m = 0$), then $\frac{1}{\sqrt{1-|\alpha|^2}}(\alpha q - p, q - \overline{\alpha} p)$ is the pair of polynomials solving for $\hat{\mathscr{P}}$ with interpolation values

$$\left(\alpha, \frac{\alpha - \gamma_2}{1 - \overline{\alpha}\gamma_2} \dots \frac{\alpha - \gamma_m}{1 - \overline{\alpha}\gamma_m}, \frac{\alpha - \beta_1}{1 - \overline{\alpha}\beta_1}, \dots, \frac{\alpha - \beta_l}{1 - \overline{\alpha}\beta_l} \right) \qquad \left(resp. \left(\alpha, \frac{\alpha - \beta_2}{1 - \overline{\alpha}\beta_2}, \dots, \frac{\alpha - \beta_{\hat{N}+1}}{1 - \overline{\alpha}\beta_{\hat{N}+1}} \right) \right), \tag{77}$$

where it should be observed that if $(\beta_1, \cdots, \beta_l) \in \mathbb{P}^+$ then $(M_\alpha(\beta_1), \cdots, M_\alpha(\beta_l)) \in \mathbb{P}^+$ also, because if f is a Schur function on $\mathbb{C}^-$ which is strictly less than 1 in modulus on a subset of $\mathbb{R}$ of positive measure, then so is $M_\alpha(f)$. Thus, if we let $\hat{\mathfrak{f}}_\alpha : \mathbb{D}^{\hat{N}} \to \mathbb{D}^{\hat{N}}$ act componentwise as M_α, we find since the latter is involutive that

$$\hat{\psi}(p) = \begin{pmatrix} \frac{p}{q}(x_1) \\ \hat{\mathfrak{f}}_{\frac{p}{q}(x_1)} \circ \hat{\psi}_{red} \left(\frac{1}{\sqrt{1 - |\frac{p}{q}(x_1)|^2}} (\frac{p}{q}(x_1) q - p) \right) \end{pmatrix} \qquad \forall p \in \mathbf{P}_{\hat{N}} \quad \text{if } m > 0, \tag{78}$$

$$\hat{\psi}(p) = \begin{pmatrix} \frac{p}{q}(z_1) \\ \hat{\mathfrak{f}}_{\frac{p}{q}(z_1)} \circ \hat{\psi}_{red} \left(\frac{1}{\sqrt{1 - |\frac{p}{q}(z_1)|^2}} (\frac{p}{q}(z_1) q - p) \right) \end{pmatrix} \qquad \forall p \in \mathbf{P}_{\hat{N}} \quad \text{if } m = 0. \tag{79}$$

This completes the second step.

Finally, we make use of the previous two steps to compute $\hat{\psi}^{-1}$ and show that it is differentiable at $\psi(p)$ when $p \in \mathbf{P}_{\hat{N}}(r)$. This will achieve the proof. We give the argument when $m > 0$ only, as the case $m = 0$ is entirely similar, replacing formally φ_{x_1} by φ_{z_1} and $\mathbf{P}_{\hat{N},x_1}$ by $\mathbf{P}_{\hat{N},z_1}$.

Define $\kappa : \mathbb{D}^m \times \mathbb{P}^+ \to \mathbf{P}_{\hat{N},x_1}$ by

$$\kappa(y) = \hat{\psi}_{red}^{-1} \circ \hat{\mathfrak{f}}_{y_1}(G(y)), \qquad y = (y_1, \dots, y_{\hat{N}+1})^t \in \mathbb{D}^m \times \mathbb{P}^+. \tag{80}$$

Note that $(\kappa(y), q_{\kappa(y)})$ is the solution to $\hat{\mathscr{P}}$ with interpolation values $(0, \hat{\mathfrak{f}}_{y_1}(y_2 \dots y_{\hat{N}+1})^T)$. Therefore, it is readily checked from (77) that the inverse of $\hat{\psi} : \mathbf{P}_{\hat{N}} \to \mathbb{D}^m \times \mathbb{P}^+$ is given by

$$\hat{\psi}^{-1}(y) = \frac{1}{\sqrt{1 - |y_1|^2}} (y_1 \varphi_{x_1}(\kappa(y)\kappa(y)^* + rr^*) - \kappa(y)), \qquad y = (y_1, \dots, y_{\hat{N}+1})^t \in \mathbb{D}^m \times \mathbb{P}^+. \tag{81}$$

To prove that $\hat{\psi}^{-1}$ is differentiable at every $y \in \hat{\psi}(\mathbf{P}_{\hat{N}}(r))$, observe from (75) that p has no real common root with r if, and only if $\alpha q - p$ does. Applying this with $\alpha = (p/q)(x_1)$, we deduce from (78) that $y \to \hat{\mathrm{f}}_{y_1} \circ G(y)$ maps $\hat{\psi}(\mathbf{P}_{\hat{N}}(r))$ into $\hat{\psi}_{red}(\mathbf{P}_{\hat{N},x_1}(r))$. The k^{th} component of this map is:

$$\frac{y_1 - y_{k+1}}{1 - \overline{y_1} y_{k+1}}, \tag{82}$$

which is differentiable with respect to the components of y as the denominator of (82) is locally bounded away from zero. The differentiability of $\hat{\psi}_{red}^{-1}$ on $\hat{\psi}_{red}(\mathbf{P}_{\hat{N},x_1}(r))$ is then to the effect that κ is differentiable on $\psi(\mathbf{P}_{\hat{N}}(r))$. The differentiability of φ_{x_1} and formula (81) then yield that $\hat{\psi}^{-1}$ is differentiable at each $y \in \hat{\psi}(\mathbf{P}_{\hat{N}}(r))$, as desired. □

To conclude this section, let us point out an interesting relation between problems $\mathscr{P}$ and $\hat{\mathscr{P}}$. In the statement below, we write ψ_r and $\hat{\psi}_r$ to emphasize the dependency of ψ and $\hat{\psi}$ with respect to the polynomial r.

PROPOSITION 12. *Suppose that $y \in \mathbb{D}^m \times \mathbb{P}^+$ and that (α_k) is a sequence of real numbers tending to $+\infty$. Let $p_k = \psi_{\alpha_k r}^{-1}(y)$ and put $e^{i\beta_k}$ for the leading term of $\varphi_{x_1}(\alpha_k^2 rr^* + p_k p_k^*)$ (resp. $\varphi_{z_1}(\alpha_k^2 rr^* + p_k p_k^*)$ if $m = 0$), noting that $\beta_k \in \mathbb{R}$ (because $\deg r < m + l$ while $p_k \in \mathbf{PM}_{m+l}$). Then, it holds that*

$$\lim_{k\to\infty} e^{i\beta_k} \frac{p_k}{\alpha_k} = \hat{\psi}_r^{-1}(y).$$

Proof. Set $q_k = \varphi_N(\alpha_k^2 rr^* + p_k p_k^*)$ and note that $e^{i\beta_k} q_k = \varphi_{x_1}(\alpha_k^2 rr^* + p_k p_k^*)$ (resp. $varphi_{x_1}(\alpha_k^2 rr^* + p_k p_k^*)$ if $m = 0$). The polynomials (p_k/α_k) verify the Feldtkeller equation:

$$\frac{p_k p_k^*}{\alpha_k^2} + rr^* = \frac{q_k q_k^*}{\alpha_k^2},$$

and argueing as in Proposition 6 we see that (p_k/α_k) is bounded independently of k, for otherwise y would lie on the boundary of $\mathbb{D}^m \times \mathbb{P}^+$. Thus, we can extract from any subsequence $(e^{i\beta_{k_n}} \frac{p_{k_n}}{\alpha_{k_n}})$ and $(e^{i\beta_{k_n}} \frac{q_{k_n}}{\alpha_{k_n}})$ a subsequence that converges to some polynomials p and q. Of necessity, p and q have degree strictly less than N because p_{k_n} and q_{k_n} are monic. By continuity we get that p/q verifies (11) with $y = (\gamma_1, \ldots, \gamma_m, \beta_1, \ldots, \beta_l)^T$, and that $q = \varphi_{x_1}(rr^* + pp^*)$ (resp. $\varphi_{z_1}(rr^* + pp^*)$) which indicates that (p,q) is the solution of $\hat{\mathscr{P}}$. □

5. Numerical experiments. In order to invert the maps ψ and $\hat{\psi}$, a continuation method has been implemented as follows. Suppose we want to compute ψ^{-1} (resp.$\hat{\psi}^{-1}$) at $\nu_1 \in \mathbb{D}^m \times \mathbb{P}_Z^+$. We pick an aritrary $p_0 \in \mathbf{PM}_N$ devoid of common real zero with r, and we compute $\nu_0 = \psi(p_0)$ (resp. $\nu_0 = \hat{\psi}(p_0)$). Then, we select γ to be a smooth path in $\mathbb{D}^m \times \mathbb{P}_Z^+$ joining ν_0 to ν_1. Now, using a classical predictor-corrector method, we lift γ to the path $\lambda = \psi^{-1}(\gamma)$ by numerically integrating the differential equation:

$$\frac{d\lambda}{dt} = D\psi^{-1}(\gamma(t)) \left[\frac{d\gamma}{dt}\right] \tag{83}$$

with initial condition $\lambda(0) = p_0$. Note that $D\psi$, thus also $D\psi^{-1}$ is easily computed from Proposition 2, and that Proposition 9 ensures the integration process will run smoothly along γ at the cost of jiggling the latter slightly if near-singular places are met.

Below we consider the case of an antenna functioning around 2.4 Ghz. The red curve on Figure 2 represents the reflexion coefficient $L_{1,1}$ of the antenna. The latter was designed to match well a load of 50Ω at the frquency $2.454Ghz$, whith a value of $-23.54dB$. *Our objective here is to improve this match on the whole frequency pass-band $I = [2.2, 2.5]\,Ghz$, while requiring strong rejection outside of this band.* For this we solve Problem $\hat{\mathscr{P}}$ in degree 5, choosing r to have two transmission zeros at $2.17Ghz$ and $2.53Ghz$ respectively. The interpolation points are initially placed as Tchebychev nodes on the frequency interval I (*i.e.* the roots of the Tchebychev polynomial of the first kind with degree 5 on I). Then, we iteratively adjust the interpolation points by feeding the whole process to a blackbox optimizer from *Matlab* so as to minimize the maximum of the reflexion level (see equation 2):

$$|G_{1,1}(w)| = \left| \frac{p/q - \overline{L}_{1,1}}{1 - p/q\overline{L}_{1,1}} \right| \tag{84}$$

over the segment I. The obtained reflexion level $G_{1,1}$ is presented on Figure 2, showing a clear improvement with respect to the initial reflexion level of $L_{1,1}$, while exhibiting strong selectivity at both ends of the pass-band. The whole procedure takes less than 3 *sec*. on a pc equipped with a *PentiumI*7 cpu, which makes

it perfectly suited to design matching circuit responses. If needed, these can be further adjusted using dedicated local optimization procedures.

REFERENCES

[1] M. Abrahamse, *The Pick interpolation theorem for finitely connected domains*, Michigan Math. J., 26 (1979), pp. 195–203.
[2] J. Agler and J. E. McCarthy, *Nevanlinna-Pick interpolation on the bidisk*, Journal für die reine und angewandte Mathematik, 506 (1999), pp. 191–124.
[3] J. Agler and N. J. Young, *Boundary Nevanlinna-Pick interpolation via reduction and augmentation*, Mathematische Zeitschrift, 268 (2011), pp. 791–817.
[4] A. Arias and G. Popescu, *Noncommutative interpolation and poisson transforms*, Israel J. Math., 115 (2000), pp. 205–234.
[5] J. Ball, I. Gohberg, and L. Rodman, *Interpolation of Rational Matrix Functions*, vol. 45 of Operator Theory: Advances and Applications, Birkhäuser, 1990.
[6] J. A. Ball and V. Bolotnikov, *Nevanlinna-Pick interpolation for Schur-Agler class functions on domains with matrix polynomial defining function*, New York J. Math., 11 (2005).
[7] J. A. Ball and V. Bolotnikov, *Interpolation in the noncommutative Schur-Agler class*, J. Operator Theory, 58 (2007), pp. 83–126.
[8] J. A. Ball and S. ter Horst, *Robust Control, Multidimensional Systems and Multivariable Nevanlinna-Pick Interpolation*, vol. 203 of Operator Theory: Advances and Applications, Birkhaüser, 2010, ch. 2, pp. 13–88.
[9] L. Baratchart, S. Chevillard, and T. Qian, *Minimax principle and lower bounds in H^2-rational approximation*, Journal of Approximation Theory, (2015). DOI: 10.1016/j.jat.2015.03.004.
[10] L. Baratchart and M. Olivi, *Critical points and error rank in best H^2 matrix rational approximation*, Constructive Approximation, 14 (1998), pp. 273–300.
[11] L. Baratchart, M. Olivi, and F. Seyfert, *Generalized Nevanlinna-Pick interpolation on the boundary. Application to impedance matching*, in Proceedings of the MTNS (Groninguen, Netherlands), 2014.
[12] V. Belevitch, *Elementary applications of scattering formalism to network design*, IRE Transactions on Circuit Theory, 3 (1956), pp. 97–104.
[13] V. Belevitch, *Classical Network Theory*, Holden-Day, 1968.
[14] C. Byrnes, T. Georgiou, and A. Lindquist, *A generalized entropy criterion for Nevanlinna-Pick interpolation with degree constraint*, IEEE Transactions on Automatic Control, 46 (2001), pp. 822–839.
[15] C. Byrnes, T. T. Georgiou, A. Lindquist, and A. Megretski, *Generalized interpolation in H^∞ with a complexity constraint*, Trans. Amer. Math. Soc., 358 (2006), pp. 965–988.
[16] H. Carlin and P. Civalleri, *Wideband Circuit Design*, CRC Press, 1997.
[17] K. R. Davidson and D. R. Pitts, *Nevanlinna-Pick interpolation for noncommutative analytic Toeplitz algebras*, Integral Equations and Operator Theory, 31 (1998), pp. 321–337.

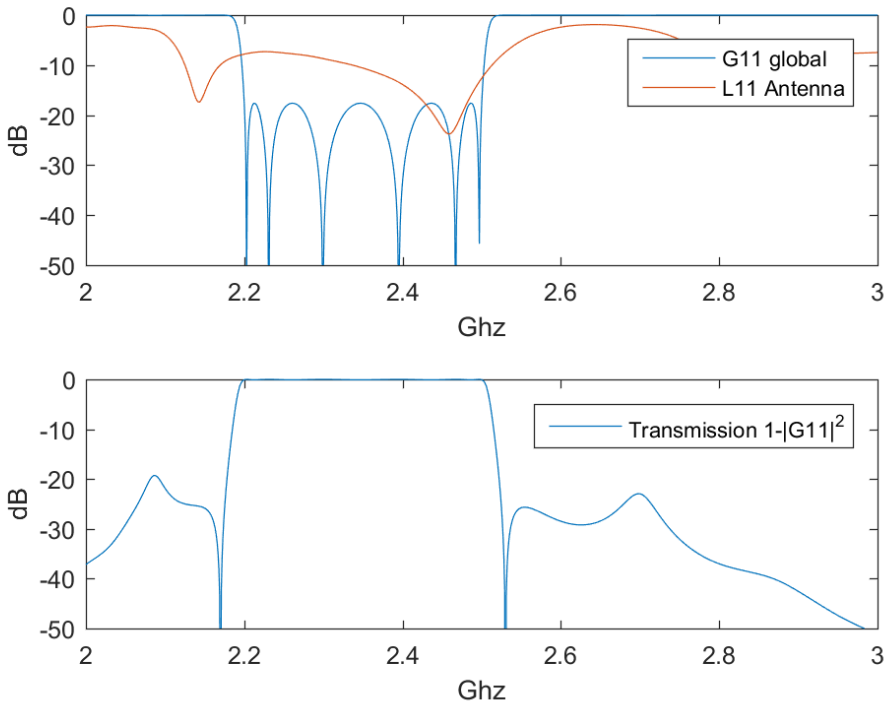


FIG. 2. *Antenna reflexion $L_{1,1}$ and global reflexion $G_{1,1}$*

[18] H. DYM, *J-contractive matrix functions, reproducing kernel spaces and interpolation*, vol. 71 of CBMS lecture notes, American Mathematical Society, Rhode Island, 1989.
[19] B. A. FRANCIS, J. HELTON, AND G. ZAMES, *H^∞-optimal feedback controllers for linear multivariable systems*, IEEE Transactions on Automatic Control, 29 (1984), pp. 888–900.
[20] P. FUHRMANN, *Linear Systems and Operators in Hilbert Spaces*, Mc Graw Hill, 1981.
[21] J. GARNETT, *Bounded Analytic Functions*, Academic Press, 1981.
[22] T. GEORGIOU, *A topological approach to Nevanlinna-Pick interpolation*, SIAM J. Math. Anal., 18(5) (1987), pp. 1248–1260.
[23] T. GEORGIOU, *The interpolation problem with a degree constraint*, IEEE Transactions on Automatic Control, (1999).
[24] V. GUILLEMIN AND A. POLLACK, *Differential Topology*, Prentice-Hall, 1974.
[25] J. HELTON, *Broadbanding:gain equalization directly from data*, IEEE Transactions on Circuits and Systems, 28 (1981), pp. 1125–1137.
[26] T. T. G. J. KARLSSON AND A. G. LINDQUIST, *The inverse problem of analytic interpolation with degree constraint and weight selection for control synthesis*, IEEE Transactions on Automatic Control, 55 (2010), pp. 405–418.
[27] R. KALMAN, P. FALB, AND A. ARBIB, *Topics in Mathematical System Theory*, McGraw-Hill, New-York, 1969.
[28] P. KHARGONEKAR AND A. TANNENBAUM, *Noneuclidean metrics and the robust stabilization of systems with parameter uncertainty*, IEEE Transactions on Automatic Control, 30 (1985), pp. 1005–1013.
[29] H. KIMURA, *Directional interpolation approach to H^∞-optimization and robust stabilization*, IEEE Transactions on Automatic Control, 32 (1987), pp. 1095–1093.
[30] H. KIMURA, *Conjugation, interpolation and model-matching in H^∞*, Int. J. Control, 49 (1989), pp. 269–307.
[31] D. J. N. LIMEBEER AND B. D. O. ANDERSON, *An interpolation theory approach to H^∞ controller degree bounds*, Linear Algebra and its Applications, 98 (1988), pp. 347–386.
[32] J. MUNKRES, *Elements of algebraic topology*, Addison-Wesley, 1984.
[33] Y. G. P.H. DELSARTE AND Y. KAMP, *On the role of the Nevanlinna-Pick problem in circuit and system theory*, Circuit Theory and Appl., 9 (1981), pp. 177–187.
[34] F. RIESZ AND B. SZŐKEFALVI-NAGY, *Functional Analysis*, Dover, 1990.
[35] D. SARASON, *Nevanlinna–Pick interpolation with boundary data*, Integral Equations Operator Theory, 30 (1998), pp. 231–250.
[36] M. S. TAKYAR AND T. T. GEORGIOU, *Analytic interpolation with a degree constraint for matrix-valued functions*, IEEE Transactions on Automatic Control, 55 (2010), pp. 1075–1088, http://dx.doi.org/10.1109/TAC.2010.2042004.
[37] A. TANNENBAUM, *Feedback stabilization of linear dynamical plants with uncertainty in the gain factor*, Int. Journal Control, 32 (1980), pp. 1–16.
[38] F. W. WARNER, *Foundations of Differential Manifolds and Lie Groups*, vol. 94 of Graduate Texts in Mathematics, Springer, 1983.
[39] H. YONG-JIAN, K. D. A. BOUBAKAR, AND C. GONG-NING, *On boundary Nevanlinna–Pick interpolation for Carathéodory matrix functions*, Linear Algebra and its Applications, 423 (2007), pp. 209–229.
[40] D. C. YOULA AND M. SAITO, *Interpolation with positive real functions*, J. Franklin Institute, (1967).
[41] O. ZAMES AND B. A. FRANCIS, *Feedback, minimax sensitivity, and optimal robustness*, IEEE Transactions on Automatic Control, 28 (1983), pp. 585–600.

Notations. The main notations used in this paper are listed below.

$\mathbb{C}$	field of complex number
$\mathbb{C}^+$	open upper half-plane
$\mathbb{C}^-$	open lower half-plane
$\mathbb{T}$	unit circle
$\mathbb{D}$	open unit disk
$P(Z,\beta)$	Pick matrix associated with the sequence of interpolation data (z_k,β_k)
$\mathbb{P}_Z^+$	the set of interpolation values $\beta \in \mathbb{C}^l$ such that $P(Z,\beta) > 0$
$\mathbf{P}_N$	complex polynomials of degree at most N
$\mathbf{PE}_N$	complex polynomials of exact degree N
$\mathbf{PM}_N$	monic complex polynomials of degree N
$\mathbf{P}_{2N}^+$	non negative real polynomials of degree at most $2N$
$\mathbf{PE}_{2N}^+$	non negative real polynomials of exact degree $2N$
$\mathbf{PM}_{2N}^+$	non negative real monic polynomials of degree $2N$
$\mathbf{S}_N$	stable (no roots in $\overline{\mathbb{C}^-}$) complex polynomials of degree at most N
$\mathbf{SE}_N$	stable complex polynomials of exact degree N
$\mathbf{SM}_N$	stable monic complex polynomials of degree N
$\mathbf{SB}_N$	polynomials of degree at most N stable in the broad sense
$\mathbf{SBM}_N$	monic polynomials of degree N stable in the broad sense
$\mathbf{PM}_N(r)$	the subset of $\mathbf{PM}_N$ of polynomials having no common root with r
$\mathbf{P}_{\hat{N},x_1}$	the set of polynomials $p \in \mathbf{P}_{\hat{N}}$ vanishing at x_1
$\mathbf{P}_{\hat{N},x_1}(r)$	the set of polynomials $p \in \mathbf{P}_{\hat{N}}$ vanishing at x_1, having no common root with r
$H^\infty(\mathbb{C}^-)$	the space of bounded holomorphic functions in the lower half-plane
$H^2(\mathbb{C}^-)$	the Hardy space of exponent 2 of the lower half-plane
$L^2(\mathbb{R})$	the space of square integrable functions on the real line
$F^*(s) = F(\bar{s})^*$	the para-Hermitian conjugate of a rational matrix function $F(s)$
$\mathring{V}$	denotes the interior of a set V in a topological space

Chapter 6

Main Bibliography

[1] W. Rudin. *Real and complex analysis*. McGraw–Hill, 1987.

[2] J. Garnett. *Bounded Analytic Functions*. Academic Press, 1981.

[3] K. Hoffman. *Banach spaces of analytic functions*. Dover, 1962.

[4] P.L. Duren. *Theory of H^p spaces*. Academic Press, 1970.

[5] L. Baratchart, J. Leblond, and J.R. Partington. "Hardy approximation to L^∞ functions on subsets of the circle". In: *Constructive Approximation* 12 (1996), pp. 423–436.

[6] L. Baratchart, J. Leblond, and J.R. Partington. "Problems of Adamjan–Arov–Krein type on subsets of the circle and minimal norm extensions". In: *Constructive Approximation* 16 (2000), pp. 333–357.

[7] L. Baratchart, J. Leblond, and J.R. Partington. *Hardy approximation to L^p functions on subsets of the circle*. INRIA research report no. 2377. 1994.

[8] M.G. Krein and P.Y. Nudel'man. "Approximation of $L^2(\omega_1, \omega_2)$ functions by minimum– energy transfer functions of linear systems". In: *Problemy Peredachi Informatsii* 11.2 (1975). English translation, pp. 37–60.

[9] D. Alpay, L. Baratchart, and J. Leblond. "Some extremal problems linked with identification from partial frequency data". In: *10th conference on analysis and optimization of systems, Sophia–Antipolis 1992*. Ed. by J.L. Lions R.F. Curtain A. Bensoussan. Vol. 185. Lect. Notes in Control and Information Sc. Springer-Verlag, 1993, pp. 563–573.

[10] L. Baratchart and J. Leblond. "Hardy approximation to L^p functions on subsets of the circle with $1 \leq p < \infty$". In: *Constructive Approximation* 14 (1998), pp. 41–56.

[11] L. Baratchart, J. Grimm, J. Leblond, M. Olivi, F. Seyfert, and F. Wielonsky. *Identification d'un filtre hyperfréquences par approximation dans le domaine complexe*. INRIA technical report no. 0219. 1998.

[12] F. Seyfert. "Problèmes extrémaux dans les espaces de Hardy. Application à l'identification de filtres hyperfréquences à cavités couplées". These de Doctorat, Ecole des Mines de Paris. 1998.

[13] Arne Schneck. "Constrained Hardy space approximation". In: *Journal of Approximation Theory* 162.8 (2010), pp. 1466–1483. ISSN: 0021-9045. DOI: https://doi.org/10.1016/j.jat.2010.03.006. URL: http://www.sciencedirect.com/science/article/pii/S0021904510000584.

[14] Laurent Baratchart, Juliette Leblond, and Fabien Seyfert. "Constrained L^2-approximation by polynomials on subsets of the circle". In: *New Trends in Approximation Theory. In Memory of André Boivin*. Ed. by Javad Mashreghi, Myrto Manolaki, and Paul M. Gauthier. Vol. 81. Fields Institute Communications. Springer, 2017, pp. 1–14. URL: https://hal.archives-ouvertes.fr/hal-01671183.

[15] L. Baratchart, J. Leblond, and Fabien Seyfert. " Constrained extremal problems in H^2 and Carleman's formulae". In: *Mat. Sbornik* 209 (7 2018), pp. 922–957.

[16] Fabien Seyfert, Martine Olivi, and J.P Marmorat. *Software Presto-HF*. https://project.inria.fr/presto-hf/.

[17] V.M. Adamjan, D.Z. Arov, and M.G. Krein. "Analytic properties of Schmidt pairs for a Hankel operator and the generalized Schur–Takagi problem". In: *Math. USSR Sbornik* 15 (1971), pp. 31–73.

[18] N.J. Young. *An introduction to Hilbert space*. Cambridge University Press, 1988.

[19] Jonathan Partington. *Linear operators and linear systems*. Student texts 60. London Math. Soc., 2004.

[20] K. Glover. "All optimal Hankel norm approximations of linear multivariable systems and their L^∞-error bounds". In: *Int. J. Contr.* 39 (1984), pp. 1115–1193.

[21] J. P. Marmorat and M. Olivi. *RARL2: a Matlab based software for H^2 rational approximation*. http://www-sop.inria.fr/apics/RARL2/rarl2.html. 2004.

[22] P. Fulcheri and M. Olivi. "Identification and matrix rational H^2–approximation: a gradient algorithm based on Schur analysis." Rapport de recherche INRIA, N° 2520, Avril 1995.

[23] Martine Olivi, Fabien Seyfert, and Jean-Paul Marmorat. "Identification of microwave filters by analytic and rational H^2 approximation". In: *Automatica* (2012). DOI: 10.1016/j.automatica.2012.10.005. URL: http://hal.inria.fr/hal-00753824.

[24] F Seyfert, M. Oldoni, G. Macchiarella, and D. Pacaud. "De-embedding response of filters from diplexer measurements". In: *International Journal of RF and Microwave Computer-Aided Engineering* (2012).

[25] Fabien Seyfert, Matteo Oldoni, Giuseppe Macchiarella, and Damien Pacaud. "De-embedding response of filters from diplexer measurements". In: *International Journal of RF and Microwave Computer-Aided Engineering* (Aug. 2012), D. (2012). DOI: 10.1002/mmce.20664. URL: https://hal.archives-ouvertes.fr/hal-00763665.

[26] Sanda Lefteriu, Martine Olivi, Fabien Seyfert, and Matteo Oldoni. "System identification of microwave filters from multiplexers by rational interpolation". In: *Automatica* (Aug. 2016). URL: https://hal.inria.fr/hal-01357934.

[27] Matteo Oldoni, Giuseppe Macchiarella, and Fabien Seyfert. *Synthesis and Modelling Techniques for Microwave Filters and Diplexers: Advances in Analytical Methods with Applications to Design and Tuning*. Scholars' Press, Feb. 2014. URL: https://hal.inria.fr/hal-01096252.

[28] Adam Cooman, Fabien Seyfert, Martine Olivi, Sylvain Chevillard, and Laurent Baratchart. "Model-Free Closed-Loop Stability Analysis: A Linear Functional Approach". In: *IEEE Transactions on Microwave Theory and Techniques* (Sept. 2017). DOI: 10.1109/TMTT.2017.2749222. URL: https://hal.inria.fr/hal-01381731.

[29] V. Belevitch. *Classical Network Theory*. Holden-Day, 1968.

[30] H.J. Carlin and P.P. Civalleri. *Wideband Circuit Design*. CRC Press, 1997.

[31] J. Grimm. *Rational approximation of transfer functions in the Hyperion software*. Tech. rep. 4002. INRIA, 2000.

[32] T. Kailath. *Linear Systems*. Prentice-Hall, 1980.

[33] P. Fuhrmann. *Linear Systems and Operators in Hilbert Spaces*. Mc Graw Hill, 1981.

[34] R.J. Cameron, R. Mansour, and C.M. Kudsia. *Microwave Filters for Communication Systems: Fundamentals, Design and Applications*. Wiley, 2007. ISBN: 9780471450221. URL: https://books.google.fr/books?id=GyVTAAAAMAAJ.

[35] V. Lunot, F. Seyfert, S. Bila, and A. Nasser. "Certified computation of optimal multiband filtering functions". In: *IEEE Transactions on Microwave Theory and Techniques* 56.1 (2008), pp. 105–112.

[36] Vincent Lunot. "Techniques d'approximation rationnelle en synthèse fréquentielle : problème de Zolotarov et algorithme de Schur". PhD Thesis. Université de Provence, 2008.

[37] A.N. Kolmogorov and A.P. Yushkevich. *Mathematics of the 19th Century: Function Theory According to Chebyshev Ordinary Differential Equations Calculus of Variations Theory of Finite Differences*. Mathematics of the 19th Century. Birkhäuser Basel, 1998. ISBN: 9783764358457. URL: https://books.google.fr/books?id=Mw6JMdZQO-wC.

[38] K.L. Su. *Handbook of Tables for Elliptic-Function Filters*. Springer US, 1990. ISBN: 9780792391098.

[39] D. Braess. *Nonlinear approximation theory*. Vol. 7. Series in computational mathematics. Springer–Verlag, 1986.

[40] M. J. D. Powell. *Approximation Theory and Methods*. Cambridge University Press, 1981. DOI: 10.1017/CBO9781139171502.

[41] V. Lunot, S. Bila, and F. Seyfert. "Optimal synthesis for multi-band microwave filters". In: *2007 IEEE MTT-S International Microwave Symposium Digest*. 2007, pp. 115–118.

[42] S. Bila, R.J. Cameron, P. Lenoir, V. Lunot, and F. Seyfert. "Chebyshev synthesis for multi-band microwave filters". In: *2006 IEEE MTT-S International Microwave Symposium Digest*. 2006, pp. 1221–1224.

[43] Abdallah Nasser, V. Lunot, Stéphane Bila, Dominique Baillargeat, Serge Verdeyme, and Fabien Seyfert. "Design of multiband filters in waveguide technology for space applications." In: *Workshop on Design and Implementation Techniques for Multiband Filters IEEE MTT-S International Microwave Symposium Digest, IMS 2008*. Atlanta, United States, June 2008. URL: https://hal.archives-ouvertes.fr/hal-00358460.

[44] J. C. Slater. "Microwave Electronics". In: *Rev. Mod. Phys.* 18 (4 Oct. 1946), pp. 441–512. DOI: 10.1103/RevModPhys.18.441. URL: https://link.aps.org/doi/10.1103/RevModPhys.18.441.

[45] K. Kurokawa. *An introduction to the theory of microwave circuits*. Academic press, 1969.

[46] R.F. Harrington. *Time-Harmonic Electromagnetic Fields*. IEEE Press Series on Electromagnetic Wave Theory. Wiley, 2001. ISBN: 9780471208068. URL: https://books.google.fr/books?id=4-6kNAEACAAJ.

[47] G.L. Matthaei. *Microwave filters, impedance-matching networks, and coupling structures*. vol. 1. McGraw-Hill, 1964. URL: https://books.google.fr/books?id=GVqNkauCv1UC.

[48] I. Hunter and Institution of Electrical Engineers. *Theory and Design of Microwave Filters*. Electromagnetics and Radar Series. Institution of Engineering and Technology, 2001. ISBN: 9780852967775. URL: https://books.google.fr/books?id=4dOYvsZ6uDQC.

[49] P. Faurre, M. Clerget, and F. Germain. *Opérateurs Rationnels positifs*. Méthodes Mathématiques de l'Informatique vol. 8. Dunod, 1979.

[50] R.J. Cameron. "General coupling matrix synthesis methods for chebyshev filtering functions". In: *IEEE Transaction on Microwave Theory and Techniques* 47.4 (1999), pp. 433–442.

[51] R. J. Cameron. "Advanced Filter Synthesis". In: *IEEE Microwave Magazine* 12.6 (Oct. 2011), pp. 42–61. ISSN: 1527-3342. DOI: 10.1109/MMM.2011.942007.

[52] R. J. Cameron, A. R. Harish, and C. J. Radcliffe. "Synthesis of advanced microwave filters without diagonal cross-couplings". In: *IEEE Transactions on Microwave Theory and Techniques* 50.12 (Dec. 2002), pp. 2862–2872. ISSN: 0018-9480. DOI: 10.1109/TMTT.2002.805141.

[53] S. Tamiazzo and G. Macchiarella. "An analytical technique for the synthesis of cascaded N-tuplets cross-coupled resonators microwave filters using matrix rotations". In: *IEEE Transactions on Microwave Theory and Techniques* 53.5 (May 2005), pp. 1693–1698. ISSN: 0018-9480. DOI: 10.1109/TMTT.2005.847065.

[54] S. Amari. "On the maximum number of finite transmission zeros of coupled resonator filters with a given topology". In: *IEEE Microwave and Guided Wave Letters* 9.9 (Sept. 1999), pp. 354–356. ISSN: 1051-8207. DOI: 10.1109/75.790472.

[55] S. Amari. "Synthesis of cross-coupled resonator filters using an analytical gradient-based optimization technique". In: *IEEE Transactions on Microwave Theory and Techniques* 48.9 (Sept. 2000), pp. 1559–1564. ISSN: 0018-9480. DOI: 10.1109/22.869008.

[56] G. Macchiarella. "A powerful tool for the synthesis of prototype filters with arbitrary topology". In: *IEEE MTT-S International Microwave Symposium Digest, 2003*. Vol. 3. June 2003, 1467–1470 vol.3. DOI: 10.1109/MWSYM.2003.1210382.

[57] Williams A.E. and A.E Atia. "New type of waveguide bandpass filters for satellite transponders". In: *COMSAT Tech. Rev.* 1 (1971).

[58] W. A. Atia, K. A. Zaki, and A. E. Atia. "Synthesis of general topology multiple coupled resonator filters by optimization". In: *1998 IEEE MTT-S International Microwave Symposium Digest (Cat. No.98CH36192)*. Vol. 2. June 1998, 821–824 vol.2. DOI: 10.1109/MWSYM.1998.705116.

[59] K. Kendig. *Elementary Algebraic Geometry*. 44. Springer-Verlag New York, 1977.

[60] A. Pollack V. Guillemin. *Differential Topology*. Prentice–Hall, 1974.

[61] David A. Cox, John Little, and Donal O'Shea. *Ideals, Varieties, and Algorithms: An Introduction to Computational Algebraic Geometry and Commutative Algebra, 3/e (Undergraduate Texts in Mathematics)*. Berlin, Heidelberg: Springer-Verlag, 2007. ISBN: 0387356509.

[62] E. Arbarello and D. Mumford. *The Red Book of Varieties and Schemes: Includes the Michigan Lectures (1974) on Curves and their Jacobians*. Lecture Notes in Mathematics. Springer Berlin Heidelberg, 1999. ISBN: 9783540632931. URL: https://books.google.fr/books?id=K1hFauNsmR4C.

[63] Igor R Shafarevich. *Basic algebraic geometry; 3rd ed.* Berlin: Springer, 2013.

[64] J.R. Sendra, F. Winkler, and S. Pérez-Diaz. *Rational Algebraic Curves: A Computer Algebra Approach*. Algorithms and Computation in Mathematics. Springer Berlin Heidelberg, 2007. ISBN: 9783540737254. URL: https://books.google.fr/books?id=-IqDwQ1co4EC.

[65] Jean-Charles Faugère. "FGb: A Library for Computing Gröbner Bases". In: *Mathematical Software - ICMS 2010*. Ed. by Komei Fukuda, Joris Hoeven, Michael Joswig, and Nobuki Takayama. Vol. 6327. Lecture Notes in Computer Science. Kobe, Japan: Springer Berlin / Heidelberg, Sept. 2010, pp. 84–87. DOI: 10.1007/978-3-642-15582-6_17.

[66] Fabien Seyfert. *Software Dedale-HF*. https://www-sop.inria.fr/apics/Dedale/WebPages/.

[67] Richard J. Cameron, Jean-Charles Faugère, and Fabien Seyfert. "Coupling matrix synthesis for a new class of microwave filter configuration". In: *2005 IEEE MTT-S International Microwave Symposium*. Vol. 1. Long Beach, United States, June 2005, pp. 119–124. DOI: 10.1109/MWSYM.2005.1516536. URL: https://hal.inria.fr/hal-00663550.

[68] Richard J. Cameron, Jean-Charles Faugère, Fabrice Rouillier, and Fabien Seyfert. "Exhaustive approach to the coupling matrix synthesis problem and application to the design of high degree asymmetric filters". In: *International Journal of RF and Microwave Computer-Aided Engineering* 17.1 (Jan. 2007), pp. 4–12. DOI: 10.1002/mmce.20190. URL: https://hal.inria.fr/hal-00663777.

[69] Fabien Seyfert and Stéphane Bila. "General synthesis techniques for coupled resonator networks". In: *IEEE Microwave Magazine* 8.5 (2007), pp. 98–104. DOI: 10.1109/MMW.2007.4383440. URL: https://hal.inria.fr/hal-00663533.

[70] Philippe Lenoir, Stéphane Bila, Fabien Seyfert, Dominique Baillargeat, and Serge Verdeyme. "Synthesis and design of asymmetrical dual-band bandpass filters based on equivalent network simplification". In: *IEEE Transactions on Microwave Theory and Techniques* 54.7 (2006), pp. 3090–3097. DOI: 10.1109/TMTT.2006.877037. URL: https://hal.inria.fr/hal-00663496.

[71] Smain Amari, Fabien Seyfert, and M. Bekheit. "Theory of Coupled Resonator Microwave Bandpass Filters of Arbitrary Bandwidth". In: *IEEE Transactions on Microwave Theory and Techniques* 58.8 (2010), pp. 2188–2203. DOI: 10.1109/TMTT.2010.2052874. URL: https://hal.inria.fr/hal-00663513.

[72] Laurent Baratchart, Martine Olivi, and Fabien Seyfert. "Boundary Nevanlinna-Pick interpolation with prescribed peak points. Application to impedance matching". In: *SIAM Journal on Mathematical Analysis* (2017). DOI: 10.1137/16M1085577. URL: https://hal.inria.fr/hal-01377782.

[73] D. Youla. "A New Theory of Broad-band Matching". In: *IEEE Transactions on Circuit Theory* 11.1 (Mar. 1964), pp. 30–50. ISSN: 0018-9324. DOI: 10.1109/TCT.1964.1082267.

[74] R.M. Fano. "Theoretical limitations on the broadband matching of arbitrary impedances". In: *Journal of the Franklin Institute* 249.1 (1950), pp. 57–83. ISSN: 0016-0032. DOI: https://doi.org/10.1016/0016-0032(50)90006-8. URL: http://www.sciencedirect.com/science/article/pii/0016003250900068.

[75] J. William Helton. "Non-Euclidean functional analysis and electronics". In: *Bull. Amer. Math. Soc. (N.S.)* 7.1 (July 1982), pp. 1–64. URL: https://projecteuclid.org:443/euclid.bams/1183549048.

[76] J.W. Helton. "Broadbanding:Gain equalization directly from data". In: *IEEE Transactions on Circuits and Systems* 28.12 (1981), pp. 1125–1137. ISSN: 0098-4094.

[77] Laurent Baratchart, Martine Olivi, and Fabien Seyfert. "Generalized Nevanlinna-Pick interpolation on the boundary. Application to impedance matching". working paper or preprint. Dec. 2015. URL: https://hal.inria.fr/hal-01249330.

[78] L.V. Ahlfors. *Conformal invariants*. AMS Chelsea, 1973.

[79] J. Partington. *Interpolation, Identification and Sampling*. Oxford University Press, 1997.

[80] T. Georgiou. “A topological approach to Nevanlinna-Pick interpolation”. In: *SIAM J. Math. Anal.* 18(5) (1987), pp. 1248–1260.

[81] T. Georgiou. “The interpolation problem with a degree constraint”. In: *IEEE Transactions on Automatic Control* (1999).

[82] A. Linquist, C. Byrnes, and T. Georgiou. “A generalized entropy criterion for Nevanlinna-Pick interpolation with degree constraint”. In: *IEEE Transactions on Automatic Control* AC-46 (2001), pp. 822–839.