BIT Numerical Mathematics
https://doi.org/10.1007/s10543-022-00922-2

BIT

Model order reduction of layered waveguides via rational Krylov fitting

Vladimir Druskin[1] · **Stefan Güttel**[2] · **Leonid Knizhnerman**[3]

Received: 12 April 2021 / Accepted: 31 March 2022

Abstract
Rational approximation recently emerged as an efficient numerical tool for the solution of exterior wave propagation problems. Currently, this technique is limited to wave media which are invariant along the main propagation direction. We propose a new model order reduction-based approach for compressing unbounded waveguides with layered inclusions. It is based on the solution of a nonlinear rational least squares problem using the RKFIT method. We show that approximants can be converted into an accurate finite difference representation within a rational Krylov framework. Numerical experiments indicate that RKFIT computes more accurate grids than previous analytic approaches and even works in the presence of pronounced scattering resonances. Spectral adaptation effects allow for finite difference grids with dimensions near or even below the Nyquist limit.

Keywords Reduced order model · Helmholtz equation · Dirichlet-to-Neumann map · Perfectly matched layer · Rational approximation · Scattering resonance

Mathematics Subject Classification 35J05 · 65N06 · 30E10

Communicated by Ralf Hiptmair.

✉ Stefan Güttel
stefan.guettel@manchester.ac.uk

Vladimir Druskin
vdruskin1@gmail.com

Leonid Knizhnerman
lknizhnerman@gmail.com

[1] Department of Mathematical Sciences, Worcester Polytechnic Institute, 100 Institute Rd, Worcester, MA 01609, USA

[2] Department of Mathematics, The University of Manchester, Alan Turing Building, Manchester M13 9PL, UK

[3] Marchuk Institute of Numerical Mathematics, Russian Academy of Sciences, Gubkin St. 8, Moscow, Russia 119333

Published online: 08 May 2022

1 Introduction

In this work we present a new approach to the compression of Dirichlet-to-Neumann (DtN) maps of infinite waveguides with layered inclusions. This approach is inspired by rational approximation techniques from model order reduction (see, e.g., [4]), in this case the RKFIT algorithm for nonlinear rational approximation [9]. As a prototypical problem we consider the infinite finite difference (FD) scheme

$$2h^{-1}\left[h^{-1}(\boldsymbol{u}_1-\boldsymbol{u}_0)+\boldsymbol{b}\right]=(A+c_0I)\boldsymbol{u}_0 \tag{1.1a}$$

$$h^{-1}\left[h^{-1}(\boldsymbol{u}_{j+1}-\boldsymbol{u}_j)-h^{-1}(\boldsymbol{u}_j-\boldsymbol{u}_{j-1})\right]=(A+c_jI)\boldsymbol{u}_j,\quad j=1,2,\dots \tag{1.1b}$$

where either $\boldsymbol{u}_0\in\mathbb{C}^N$ or $\boldsymbol{b}\in\mathbb{C}^N$ is given, $A\in\mathbb{C}^{N\times N}$ is Hermitian, $c_j=0$ for all $j>L$, and the solution $\{\boldsymbol{u}_j\}_{j=0}^{\infty}\subset\mathbb{C}^N$ is assumed to be bounded. This problem arises from the FD discretization of the three-dimensional (indefinite) Helmholtz equation

$$\nabla^2u+(k_\infty^2-c(x))u=0$$

for $(x,y,z)\in[0,+\infty)\times[0,1]\times[0,1]$ with a compactly supported *offset function* $c(x)$ for the wave number k_∞ and appropriate boundary conditions. Here, the matrix A corresponds to the discretization of the transverse differential operator $-\partial_{yy}^2-\partial_{zz}^2-k_\infty^2$ at $x=0$ and is Hermitian indefinite. The variation of the wave number in the x-direction is modelled by varying coefficients c_j, with the "effective" wave number $\sqrt{k_\infty^2-c_j}$ at each grid point. The DtN operator F for (1.1) is defined by the relationship $F\boldsymbol{u}_0=\boldsymbol{b}$.

Since (1.1) is a linear recurrence, $F=f_h(A)$ is a matrix function in A. If $c_j\equiv0$, the DtN function for (1.1) at $x=0$ is $f_h(\lambda)=\sqrt{\lambda+(h\lambda/2)^2}$. As $h\to0$ we obtain the DtN function $f(\lambda)=\sqrt{\lambda}$ for the continuous problem. In this case, a near-optimal rational approximant to f can be constructed analytically [12, 13, 16]. More precisely, let the eigenvalues of A be contained in the union of two intervals $K=[a_1,b_1]\cup[a_2,b_2]$ with $a_1<b_1<0<a_2<b_2$. Then [12] gives an explicit construction of a compound Zolotarev rational function $r_n^{(Z)}$ of type $(n,n-1)$ such that

$$\max_{\lambda\in K}|1-r_n^{(Z)}(\lambda)/f(\lambda)|\asymp\exp\left(-2\pi^2n/\log\left(256a_1b_2/(a_2b_1)\right)\right)\ \text{as}\ n\to\infty \tag{1.2}$$

for sufficiently large interval ratios a_1/b_1 and b_2/a_2. It is also shown in [12] that the convergence factor in (1.2) is optimal. Hence, the approximation error $\|f(A)-r_n^{(Z)}(A)\|_2\le C\max_{\lambda\in K}|1-r_n^{(Z)}(\lambda)/f(\lambda)|$ decays exponentially at the same optimal rate. Interestingly, the continued fraction form of $r_n^{(Z)}$ gives rise to a geometrically meaningful three-point FD scheme. By "geometrically meaningful" we mean that the complex grid points align on a curve in the complex plane which can be interpreted as a "smooth" deformation of the original x-coordinate axis. This is

similar to the celebrated *perfectly matched layers* (PMLs) which are introduced via complex coordinate stretching [3, 5, 11, 14].

The analytic approach just outlined is essentially limited to DtN functions such as $\sqrt{\lambda}$ and $\sqrt{\lambda + (h\lambda/2)^2}$. Here we aim to overcome this limitation by numerically computing a low-order rational approximant $r_n(A) \approx f_h(A)$ and converting it into a sparse representation in form of a three-point finite difference scheme.[1] Our approach is applicable even in cases where the DtN map to be approximated is highly irregular due to the presence of scattering poles.

An illustrating example is given in Fig. 1, where the top panels show the amplitude/phase of the solution of a waveguide problem on $[0, +\infty) \times [0, 1]$, truncated and discretized by 300×150 points. The step size is $h = 1/150$ in both coordinate directions. For this problem we have chosen $k_\infty = 14$ and $c_j = -9^2$ for the grid points $j = 0, 1, \ldots, L = 150$. An absorbing boundary condition has been fitted to the right end of the domain to mimic the infinite extension $x \to \infty$. The modulus of the associated DtN function f_h is shown in the bottom of Fig. 1 (solid red curve). This function has several singularities between and close to the eigenvalues of the transverse FD matrix A (the eigenvalue positions are indicated by the black dots). In particular, one eigenvalue $\lambda_j \approx 50.5$ is extremely close to a singularity of f_h, which can be associated with the near-resonance observed in the left portion of the waveguide. These singularities make it impossible to construct a uniform approximant $r_n \approx f_h$ over the negative and positive spectral subintervals of A. Nevertheless, the RKFIT approximant r_n of order $n = 8$, also shown in the bottom of Fig. 1 (dashed blue curve), has a relative accuracy $\|f_h(A)\boldsymbol{u}_0 - r_n(A)\boldsymbol{u}_0\|_2 / \|f_h(A)\boldsymbol{u}_0\|_2 \approx 1.4 \times 10^{-6}$ for the DtN map. We see that r_n achieves this high accuracy by being close to f_h in the vicinity of the eigenvalues of A, but not necessarily in between them. This remarkable *spectral adaptation* is achieved without requiring a spectral decomposition of A explicitly; RKFIT merely requires matrix-vector products with the DtN map.

Our RKFIT approach is also applicable when A is non-Hermitian, which may result from absorbing boundary conditions in the transversal plane. We demonstrate in several experiments that the RKFIT-FD grids are exponentially accurate as an approximation to the full FD scheme, with only a small number of grid points required for practical accuracy. As a result of spectral adaptation effects, the Nyquist limit of two grid points per wavelength does not fully apply to RKFIT-FD grids. For the problem in Fig. 1, for example, we computed an RKFIT-FD grid of only $n = 8$ points which accurately (to about six digits of relative accuracy) mimics the response of the full variable-coefficient waveguide discretized by 300 grid points in the x-direction. This is a significant compression of the full grid.

The rest of this paper is structured as follows: in Sect. 2 we derive analytic expressions of DtN maps for constant- and variable-coefficient media. We relate the optimization of these DtN maps to approximation problems. Section 3 establishes a new connection between rational Krylov spaces and FD grids. In Sect. 4 we tailor the RKFIT algorithm to our specific application. Sections 5 and 6 study the convergence behaviour of the algorithm. In Sect. 7 we discuss the numerical results and

[1] Another recent approach for compressing an NtD operator for the Helmholtz equation is based on randomized matrix probing [10]. This approach has the advantage of handling a rather wide class of multidimensional variable-coefficient problems at the expense of losing the sparse representation.

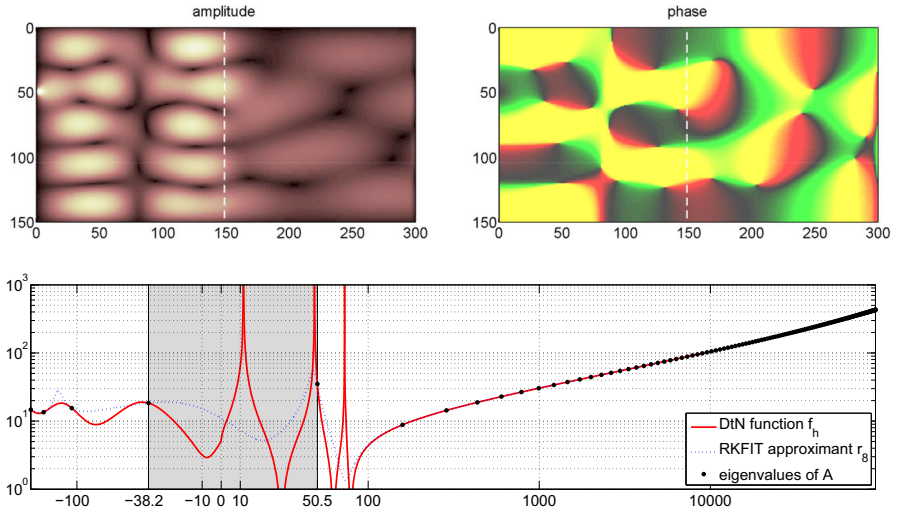

Fig. 1 A waveguide with varying wave number in the x-direction (piecewise constant over the first 150 grid points and the remaining grid points until infinity). The top row shows the amplitude and phase of the solution, with the position of the coefficient jump highlighted by vertical dashed line. The bottom shows a plot of the exact DtN function f_h (solid red line) over the spectral interval of the indefinite matrix A. The plot is logarithmic on both axes, with the x-axis showing a negative and positive part of the real axis, glued together by the gray linear part in between. The RKFIT approximant of degree $n = 8$ (dotted blue curve) exhibits spectral adaptation to some of A's eigenvalues (black dots)

compare them to the Nyquist limit and other (spectral) discretization schemes. In the "Appendix" we give a rational approximation interpretation of the Nyquist limit and explain why this limit is not necessarily strict for RKFIT-FD grids.

2 From DtN maps to continued fractions and FD grids

There is an intimate connection between FD grids and rational functions. To see this, let us first consider the scalar ODE $u''(x) = \lambda u(x)$ on $x \geq 0$ and its FD discretization

$$h^{-1}\left[h^{-1}(u_{j+1} - u_j) - h^{-1}(u_j - u_{j-1})\right] = \lambda u_j, \quad j = 1, 2, \ldots, \tag{2.1}$$

where λ and u_0 are given constants and we demand that u_n remains bounded as $n \to \infty$. This linear recurrence is a scalar version of (1.1b) with $c \equiv 0$. It can easily be solved by computing the roots of the characteristic polynomial $p(t) = (t^2 - (2 + h^2\lambda)t + 1)/h^2$ and choosing the solution $u_j = (1 + h^2\lambda/2 - h\sqrt{\lambda + h^2\lambda^2/4})^j \cdot u_0$. Indeed this is the only solution that decays for $\lambda > 0$. Moreover, this solution is bounded under the condition[2] $\lambda \geq -4/h^2$ and unbounded for $\lambda < -4/h^2$.

[2] This is an interesting condition in the indefinite Helmholtz case, where the role of λ is played by the eigenvalues of the shifted Laplacian $-\nabla^2 - k^2$ and k is the wave number. Because we require $\lambda \geq -4/h^2$, we have a condition $k^2 \leq 4/h^2$ on the wave number, which is equivalent to $kh \leq 2$. The solution of the Helmholtz equation in a homogeneous medium has wave length $\ell = 2\pi/k$. Hence the number of FD grid

We can use the explicit solution $\{u_j\}$ to extract interesting information about the problem. For example, from the FD relation $2h^{-1}\left[h^{-1}(u_1-u_0)+b\right]=\lambda u_0$, the scalar version of (1.1a), we obtain an approximation b to the Neumann boundary data $-u'(x=0)$ for the continuous analogue of the FD scheme. Eliminating u_1 using the above formula, we can directly relate u_0 and b via $b=\sqrt{\lambda+h^2\lambda^2/4}\,u_0=:f_h(\lambda)u_0$. We refer to f_h as the *DtN function* or *discrete impedance function*. By letting $h\to\infty$ we recover the DtN relation $b=\sqrt{\lambda}u_0=:f(\lambda)u_0$ and indeed $b=-u'(0)$ for the continuous solution $u(x)=\exp(-x\sqrt{\lambda})u_0$.

Now let us turn to the variable-coefficient problem (1.1) in scalar form:

$$2h^{-1}\left[h^{-1}(u_1-u_0)+b\right]=(\lambda+c_0)u_0 \tag{2.2a}$$

$$h^{-1}\left[h^{-1}(u_{j+1}-u_j)-h^{-1}(u_j-u_{j-1})\right]=(\lambda+c_j)u_j,\quad j=1,2,\ldots. \tag{2.2b}$$

By eliminating the grid points with indices $j>L$ (where $c_j=0$) we find the DtN relation $b/u_0=f_h(\lambda)$ in continued fraction form

$$f_h(\lambda)=\frac{h(\lambda+c_0)}{2}+\cfrac{1}{h+\cfrac{1}{h(\lambda+c_1)+\cfrac{1}{h+\cdots+\cfrac{1}{h(\lambda+c_L)+\cfrac{1}{h+\cfrac{1}{\frac{h\lambda}{2}+\sqrt{\lambda+\frac{h^2\lambda^2}{4}}}}}}}}. \tag{2.3}$$

In view of the original vector-valued problem (1.1), the role of λ is played by the eigenvalues of the matrix A. When employing a rational approximant $r_n\approx f_h$ it hence seems reasonable to be accurate on the *spectral region of* A. For example, if A is diagonalizable as $A=X\mathrm{diag}(\lambda_1,\lambda_2,\ldots,\lambda_N)X^{-1}$, we have $\|f_h(A)-r_n(A)\|_2\le\|X\|_2\|X^{-1}\|_2\max_{1\le j\le N}|f_h(\lambda_j)-r_n(\lambda_j)|$. Hence if the condition number $\kappa(X)=\|X\|_2\|X^{-1}\|_2$ is moderate, we can bound the accuracy of $r_n(A)$ using a scalar approximation problem on the eigenvalues λ_j. The rational approximant r_n can be viewed as a reduced order model of (2.2) where the spectral parameter λ of the transversal operator is an equivalent of the temporal (Laplace) frequency in linear time invariant dynamical systems (see, e.g., [4]).

3 From FD grids to rational Krylov spaces

The crucial observation for optimizing the rational approximant $r_n\approx f_h$ of a DtN function, or equivalently its associated FD grid, is that the grid steps do not need to

Footnote 2 continued
points per wavelength, $n=\ell/h$, must satisfy $n=\ell/h=2\pi/(kh)\ge\pi$ in order to approximate a bounded oscillatory solution.

be equispaced, and not even real-valued. Consider the FD scheme

$$\widehat{h}_0^{-1}\left[(u_1-u_0)+b\right]=\lambda u_0 \tag{3.1a}$$

$$\widehat{h}_j^{-1}\left[h_{j+1}^{-1}(u_{j+1}-u_j)-h_j^{-1}(u_j-u_{j-1})\right]=\lambda u_j,\quad j=1,\dots,n-1 \tag{3.1b}$$

with arbitrary complex-valued primal and dual grid steps h_j and $\widehat{h}_{j-1}$ ($j=1,2,\dots,n$), respectively. The continued fraction form of the associated DtN maps, derived in exactly the same manner as for the case of constant h in Sect. 2, is

$$r_n(\lambda)=\widehat{h}_0\lambda+\cfrac{1}{h_1+\cfrac{1}{\widehat{h}_1\lambda+\cfrac{1}{h_2+\cdots+\cfrac{1}{\widehat{h}_{n-1}\lambda+\cfrac{1}{h_n}}}}}. \tag{3.2}$$

This is a rational function of type $(n,n-1)$, i.e., a quotient p_n/q_{n-1} of polynomials of degree n and $n-1$, respectively. By choosing the free grid steps we can optimize it for our purposes. In particular, we can tune (3.1) so that it implements a rational approximation to any DtN map, even if the associated analytic DtN function f_h is complicated. To this end, we need a robust method for computing such rational approximants and a numerical conversion into continued fraction form.

The vector form of (3.1) is

$$\widehat{h}_0^{-1}\left[h_1^{-1}(\boldsymbol{u}_1-\boldsymbol{u}_0)+\boldsymbol{b}\right]=A\boldsymbol{u}_0 \tag{3.3a}$$

$$\widehat{h}_j^{-1}\left[h_{j+1}^{-1}(\boldsymbol{u}_{j+1}-\boldsymbol{u}_j)-h_j^{-1}(\boldsymbol{u}_j-\boldsymbol{u}_{j-1})\right]=A\boldsymbol{u}_j,\quad j=1,\dots,n-1. \tag{3.3b}$$

Again, $\boldsymbol{b}=r_n(A)\boldsymbol{u}_0$ with a rational function $r_n=p_n/q_{n-1}$ whose continued fraction form (3.2) involves the grid steps h_j and $\widehat{h}_{j-1}$. The vectors $\boldsymbol{u}_j$ and $\boldsymbol{b}=r_n(A)\boldsymbol{u}_0$ satisfy a *rational Krylov decomposition*

$$AU_{n+1}\underline{\widetilde{K}_n}=U_{n+1}\underline{\widetilde{H}_n}, \tag{3.4}$$

where $U_{n+1}=[\,r_n(A)\boldsymbol{u}_0\,|\,\boldsymbol{u}_0\,|\,\boldsymbol{u}_1\,|\,\cdots\,|\,\boldsymbol{u}_{n-1}\,]\in\mathbb{C}^{N\times(n+1)}$ and $\underline{\widetilde{K}_n},\underline{\widetilde{H}_n}\in\mathbb{C}^{(n+1)\times n}$ are

$$\underline{\widetilde{K}_n}=\begin{bmatrix}0&&&\\ \widehat{h}_0&&&\\ &\widehat{h}_1&&\\ &&\ddots&\\ &&&\widehat{h}_{n-1}\end{bmatrix},\quad \underline{\widetilde{H}_n}=\begin{bmatrix}1&&&&\\ -h_1^{-1}&h_1^{-1}&&&\\ h_1^{-1}&-h_1^{-1}-h_2^{-1}&\ddots&&\\ &\ddots&\ddots&h_{n-1}^{-1}&\\ &&h_{n-1}^{-1}&-h_{n-1}^{-1}-h_n^{-1}\end{bmatrix}. \tag{3.5}$$

The entries in $(\underline{\widetilde{H}_n}, \underline{\widetilde{K}_n})$ encode the recursion coefficients in (3.1) and the columns of U_{n+1} all correspond to rational functions in A multiplied by the vector $\boldsymbol{u}_0$. More precisely,

$$\operatorname{colspan}(U_{n+1}) = q_{n-1}(A)^{-1}\operatorname{span}\{\boldsymbol{u}_0, A\boldsymbol{u}_0, \ldots, A^n\boldsymbol{u}_0\}$$

for some denominator polynomial q_{n-1} of degree at most $n-1$ and with no roots at any of A's eigenvalues. Such a space is also known as a *rational Krylov space* [18]. In the next section we will show how to generate decompositions of the form (3.4) numerically and how to interpret them as FD grids.

4 The RKFIT approach

Assume that $F, A \in \mathbb{C}^{N\times N}$ are given matrices and $\boldsymbol{v} \in \mathbb{C}^N$ with $\|\boldsymbol{v}\|_2 = 1$. Our aim is to find a rational approximant $r_n(A)\boldsymbol{v}$ such that

$$\|F\boldsymbol{v} - r_n(A)\boldsymbol{v}\|_2 \to \min. \tag{4.1}$$

For the purpose of this paper, F is the linear DtN map and the sought rational function $r_n = p_n/q_{n-1}$ is of type $(n, n-1)$. As (4.1) is a nonconvex optimization problem it may have many solutions, exactly one solution, or no solution at all. However, this difficulty has not prevented the development of algorithms for the (approximate) solution of (4.1); see [9] for a discussion of various algorithms. The RKFIT algorithm [7, 9] is particularly suited for this task and in this section we shall briefly review it and adapt it to our application.

4.1 Search and target spaces

Given a set of *poles* $\xi_1, \xi_2, \ldots, \xi_{n-1} \in \mathbb{C}$ and an associated nodal polynomial $q_{n-1}(\lambda) = \prod_{j=1}^{n-1}(\lambda - \xi_j)$, RKFIT makes use of two spaces, namely an n-dimensional *search space* $\mathscr{V}_n$ defined as $\mathscr{V}_n := q_{n-1}(A)^{-1}\mathscr{K}_n(A, \boldsymbol{v})$, and an $(n+1)$-dimensional *target space* $\mathscr{W}_{n+1}$ defined as $\mathscr{W}_{n+1} := q_{n-1}(A)^{-1}\mathscr{K}_{n+1}(A, \boldsymbol{v})$. Here, $\mathscr{K}_j(A, \boldsymbol{v}) = \operatorname{span}\{\boldsymbol{v}, A\boldsymbol{v}, \ldots, A^{j-1}\boldsymbol{v}\}$ is the standard (polynomial) Krylov space of dimension j for the matrix A and starting vector $\boldsymbol{v}$. Let $V_n \in \mathbb{C}^{N\times n}$ and $W_{n+1} \in \mathbb{C}^{N\times(n+1)}$ be orthonormal bases for $\mathscr{V}_n$ and $\mathscr{W}_{n+1}$, respectively.

The space $\mathscr{V}_n$ is a rational Krylov space with starting vector $\boldsymbol{v}$ and the poles $\xi_1, \ldots, \xi_{n-1}$, i.e., a linear space of type $(n-1, n-1)$ rational functions $(p_j/q_{n-1})(A)\boldsymbol{v}$, all sharing the same denominator q_{n-1}. As a consequence, we can arrange the columns of V_n such that $V_n\boldsymbol{e}_1 = \boldsymbol{v}$ and a rational Krylov decomposition

$$AV_n\underline{K_{n-1}} = V_n\underline{H_{n-1}} \tag{4.2}$$

is satisfied. The existence of such a decomposition under the assumption that $\mathscr{V}_n$ is a rational Krylov space is shown in [7, Thm. 2.5]. For a given sequence of poles

$\xi_1, \ldots, \xi_{n-1}$, decompositions of this form are computed by Ruhe's rational Krylov sequence (RKS) algorithm [18, Section 2] and its variant described in [8, Algorithm 2.1]. Here, $(\underline{H_{n-1}}, \underline{K_{n-1}})$ is an unreduced upper Hessenberg pair of size $n \times (n-1)$, i.e., both $\underline{H_{n-1}}$ and $\underline{K_{n-1}}$ are upper Hessenberg matrices which do not share a common zero element on the subdiagonal. The following result, established in [7, Thm. 2.5], relates the generalized eigenvalues of the lower $(n-1) \times (n-1)$ submatrices of $(\underline{H_{n-1}}, \underline{K_{n-1}})$, the poles of the rational Krylov space, and its starting vector.

Theorem 4.1 *The generalized eigenvalues of the lower* $(n-1) \times (n-1)$ *submatrices of* $(\underline{H_{n-1}}, \underline{K_{n-1}})$ *of* (4.2) *are the poles* $\xi_1, \ldots, \xi_{n-1}$ *of the rational Krylov space* $\mathscr{V}_n$ *with starting vector* $\boldsymbol{v}$.

Conversely, let a decomposition $A\widehat{V}_n\underline{\widehat{K}_{n-1}} = \widehat{V}_n\underline{\widehat{H}_{n-1}}$ *with* $\widehat{V}_n \in \mathbb{C}^{N\times n}$ *of full column rank and an unreduced upper Hessenberg pair* $(\underline{\widehat{H}_{n-1}}, \underline{\widehat{K}_{n-1}})$ *be given. Assume further that none of the generalized eigenvalues* $\widehat{\xi}_j$ *of the lower* $(n-1) \times (n-1)$ *submatrices of* $(\underline{\widehat{H}_{n-1}}, \underline{\widehat{K}_{n-1}})$ *coincides with an eigenvalue of* A. *Then the columns of* $\widehat{V}_n$ *form a basis for a rational Krylov space with starting vector* $\widehat{V}_n\boldsymbol{e}_1$ *and poles* $\widehat{\xi}_j$.

4.2 Pole relocation and projection step

The main component of RKFIT is a pole relocation step based on Theorem 4.1. Assume that a guess for the denominator polynomial q_{n-1} is available and orthonormal bases V_n and W_{n+1} for the spaces $\mathscr{V}_n$ and $\mathscr{W}_{n+1}$ have been computed. Then we can identify a vector $\widehat{\boldsymbol{v}} \in \mathscr{V}_n$, $\|\widehat{\boldsymbol{v}}\|_2 = 1$, such that $F\widehat{\boldsymbol{v}}$ is best approximated by some vector in $\mathscr{W}_{n+1}$. More precisely, we can find a coefficient vector $\boldsymbol{c}_n \in \mathbb{C}^n$, $\|\boldsymbol{c}_n\|_2 = 1$, such that $\|(I_N - W_{n+1}W_{n+1}^*)FV_n\boldsymbol{c}_n\|_2 \to \min$. The vector $\boldsymbol{c}_n$ is given as a right singular vector of $(I_N - W_{n+1}W_{n+1}^*)FV_n$ corresponding to a smallest singular value.

Assume that a "sufficiently good" denominator q_{n-1} of $r_n = p_n/q_{n-1}$ has been found. Then the problem of finding the numerator p_n such that $\|F\boldsymbol{v} - r_n(A)\boldsymbol{v}\|_2$ is minimal becomes a linear one. Indeed, the vector $r_n(A)\boldsymbol{v} := W_{n+1}W_{n+1}^*F\boldsymbol{v}$ corresponds to the orthogonal projection of $F\boldsymbol{v}$ onto $\mathscr{W}_{n+1}$ and its representation in the rational Krylov basis W_{n+1} is

$$r_n(A)\boldsymbol{v} = W_{n+1}\boldsymbol{c}_{n+1}, \quad \text{where} \quad \boldsymbol{c}_{n+1} := W_{n+1}^*F\boldsymbol{v}. \tag{4.3}$$

The pseudocode for a single RKFIT iteration is given in Algorithm 4.1. A MATLAB implementation is contained in the Rational Krylov Toolbox [6] which is available online at http://rktoolbox.org.

4.3 Conversion to continued fraction form

Similarly to what we did in (4.2), we can arrange the columns of W_{n+1} so that $W_{n+1}\boldsymbol{e}_1 = \boldsymbol{v}$ and a rational Krylov decomposition

$$AW_{n+1}\underline{K_n} = W_{n+1}\underline{H_n} \tag{4.4}$$

Algorithm 4.1 One RKFIT iteration for superdiagonal approximants.

Require: Matrices $A, F \in \mathbb{C}^{N\times N}$, nonzero $\boldsymbol{v} \in \mathbb{C}^N$, and initial poles $\xi_1, \xi_2, \ldots, \xi_{n-1} \in \mathbb{C} \setminus \Lambda(A)$ (in the first iteration it is recommended to initialize all poles at ∞).

Ensure: Improved poles $\widehat{\xi}_1, \widehat{\xi}_2, \ldots, \widehat{\xi}_{n-1}$.

1. Compute a rational Krylov decomposition $AW_{n+1}\underline{K_n} = W_{n+1}\underline{H_n}$ with $W_{n+1}\boldsymbol{e}_1 = \boldsymbol{v}/\|\boldsymbol{v}\|_2$ and poles $\xi_1, \xi_2, \ldots, \xi_{n-1}, \infty$.
2. Define $V_n = W_{n+1}[\, I_n \,|\, \boldsymbol{0}\,]^T$.
3. Compute a right singular vector $\boldsymbol{c}_n \in \mathbb{C}^n$ of $(I - W_{n+1}W_{n+1}^*)FV_n$ corresponding to a smallest singular value.
4. Form $A\widehat{V}_n\underline{\widehat{H}_{n-1}} = \widehat{V}_n\underline{\widehat{K}_{n-1}}$ spanning $\mathscr{R}(V_n)$ with $\widehat{V}_n\boldsymbol{e}_1 = V_n\boldsymbol{c}_n$.
5. Compute $\widehat{\xi}_1, \widehat{\xi}_2, \ldots, \widehat{\xi}_{n-1}$ as the generalized eigenvalues of the lower $(n-1)\times(n-1)$ part of $(\underline{\widehat{H}_{n-1}}, \underline{\widehat{K}_{n-1}})$.

is satisfied, where $(\underline{H_n}, \underline{K_n})$ is an unreduced upper Hessenberg pair of size $(n+1)\times n$. Indeed, we have $\mathscr{V}_n \subset \mathscr{W}_{n+1}$ and $\mathscr{W}_{n+1}$ is a rational Krylov space with starting vector $\boldsymbol{v}$, finite poles $\xi_1, \ldots, \xi_{n-1}$, and a formal additional "pole" at ∞.

Our aim is to transform the decomposition (4.4) so that it can be identified with (3.4) when $\boldsymbol{u}_0 = \boldsymbol{v}$. This transformation should not alter the space $\mathscr{W}_{n+1}$ but merely transform the basis W_{n+1} into the continued fraction basis U_{n+1} and the pair $(\underline{H_n}, \underline{K_n})$ into the tridiagonal-and-diagonal form of (3.5).

First we transform (4.4) so that $r_n(A)\boldsymbol{v}$ defined in (4.3) becomes the first vector in the rational Krylov basis, and $\boldsymbol{v}$ the second. To this end, we define the transformation matrix $X = [\, \boldsymbol{c}_{n+1} \,|\, \boldsymbol{e}_1 \,|\, \boldsymbol{x}_3 \,|\, \cdots\, \boldsymbol{x}_{n+1}] \in \mathbb{C}^{(n+1)\times(n+1)}$ with the columns $\boldsymbol{x}_3, \ldots, \boldsymbol{x}_{n+1}$ chosen freely but so that X is invertible, and rewrite (4.4) by inserting XX^{-1}:

$$A W_{n+1}^{(0)} \underline{K_n^{(0)}} = W_{n+1}^{(0)} \underline{H_n^{(0)}}, \tag{4.5}$$

where $W_{n+1}^{(0)} = W_{n+1}X$, $\underline{K_n^{(0)}} = X^{-1}\underline{K_n}$ and $\underline{H_n^{(0)}} = X^{-1}\underline{H_n}$. By construction, the transformed rational Krylov basis $W_{n+1}^{(0)}$ is of the form $W_{n+1}^{(0)} = \left[\, r_n(A)\boldsymbol{v} \,|\, \boldsymbol{v} \,|\, * \,|\, \cdots \,|\, * \,\right] \in \mathbb{C}^{N\times(n+1)}$. The transformation to (4.5) has potentially destroyed the upper Hessenberg structure of the decomposition and $(\underline{H_n^{(0)}}, \underline{K_n^{(0)}})$ generally is a dense $(n+1)\times n$ matrix pair. Here is a pictorial view of decomposition (4.5) for the case $n = 4$:

$$A W_{n+1}^{(0)} \begin{bmatrix} * & * & * & * \\ * & * & * & * \\ * & * & * & * \\ * & * & * & * \\ * & * & * & * \end{bmatrix} = W_{n+1}^{(0)} \begin{bmatrix} * & * & * & * \\ * & * & * & * \\ * & * & * & * \\ * & * & * & * \\ * & * & * & * \end{bmatrix}. \tag{4.6}$$

We now transform $(\underline{H_n^{(0)}}, \underline{K_n^{(0)}})$ into tridiagonal-and-diagonal form by successive right and left multiplication, giving rise to pairs $(\underline{H_n^{(j)}}, \underline{K_n^{(j)}})$ $(j = 1, 2, \ldots, 5)$ all corresponding to the same rational Krylov space $\mathscr{W}_{n+1}$ and all without the two leading vectors in $W_{n+1}^{(0)}$ being altered. More precisely, the allowed transformations are:

– right-multiplication of the pair by any invertible matrix $R \in \mathbb{C}^{n\times n}$,
– left-multiplication of the pair by an invertible matrix $L \in \mathbb{C}^{(n+1)\times(n+1)}$, the first two columns of which are $[\, \boldsymbol{e}_1 \,|\, \boldsymbol{e}_2 \,]$. This ensures that inserting $L^{-1}L$ into the decomposition will not alter the leading two vectors $[\, r_n(A)\boldsymbol{v} \,|\, \boldsymbol{v} \,]$ in the rational Krylov basis.

Here are the transformations we perform:

1. We right-multiply the pair $(\underline{H_n^{(0)}}, \underline{K_n^{(0)}})$ by the inverse of the lower $n \times n$ part of $\underline{K_n^{(0)}}$, giving rise to $(\underline{H_n^{(1)}}, \underline{K_n^{(1)}})$ (we now only show a pictorial view of the transformed pairs):

$$A W_{n+1}^{(1)} \begin{bmatrix} 0 & * & * & * \\ 1 & & & \\ & 1 & & \\ & & 1 & \\ & & & 1 \end{bmatrix} = W_{n+1}^{(1)} \begin{bmatrix} * & * & * & * \\ * & * & * & * \\ * & * & * & * \\ * & * & * & * \\ * & * & * & * \end{bmatrix}.$$

   The Krylov basis matrix $W_{n+1}^{(1)} = W_{n+1}^{(0)} = [\, r_n(A)\boldsymbol{v} \,|\, \boldsymbol{v} \,|\, * \,|\, \cdots \,|\, * \,]$ has not changed. The $(1, 1)$ element of the transformed matrix $\underline{K_n^{(1)}} = [k_{ij}^{(1)}]$ is automatically zero because the decomposition states that the linear combination $k_{11}^{(1)} A r_n(A)\boldsymbol{v} + k_{21}^{(1)}\boldsymbol{v}$ is in the column span of $W_{n+1}^{(1)}$, a space of type $(n, n-1)$ rational functions. This linear combination is a type $(n+1, n-1)$ rational function unless $k_{11} = 0$.
2. We left-multiply the pairs to zero the first row of $\underline{K_n^{(1)}}$ completely. This can be done by adding multiples of the 3rd, 4th, …, $(n+1)$th row to the first. As a result we obtain

$$A W_{n+1}^{(2)} \begin{bmatrix} 0 & & & & \\ 1 & & & & \\ & 1 & & & \\ & & 1 & & \\ & & & 1 & \end{bmatrix} = W_{n+1}^{(2)} \begin{bmatrix} * & * & * & * \\ * & * & * & * \\ * & * & * & * \\ * & * & * & * \\ * & * & * & * \end{bmatrix}. \tag{4.7}$$

   This left-multiplication does not affect the leading two columns of the Krylov basis, hence $W_{n+1}^{(2)}$ is still of the form $W_{n+1}^{(2)} = [\, r_n(A)\boldsymbol{v} \,|\, \boldsymbol{v} \,|\, * \,|\, \cdots \,|\, * \,]$.
3. We right-multiply the pair to zero all elements in the first row of $\underline{H_n^{(2)}}$ except the $(1, 1)$ entry, which we can assume to be nonzero (see Remark 4.1). This can be done by adding multiples of the first column to the 2nd, 3rd, …, nth column. As a result we have

$$A W_{n+1}^{(3)} \begin{bmatrix} 0 & & & \\ 1 & * & * & * \\ & 1 & & \\ & & 1 & \\ & & & 1 \end{bmatrix} = W_{n+1}^{(3)} \begin{bmatrix} * & & & \\ * & * & * & * \\ * & * & * & * \\ * & * & * & * \\ * & * & * & * \end{bmatrix}.$$

   Again, this right-multiplication has not affected $W_{n+1}^{(3)} = W_{n+1}^{(2)}$.

4. With a further left-multiplication, adding multiples of the 3rd, 4th,…,$(n+1)$st row to the second row, we can zero all the entries in the second row of $\underline{K_n^{(3)}}$, except the entry in the $(2,1)$ position:

$$A W_{n+1}^{(4)} \begin{bmatrix} 0 & & \\ 1 & & \\ & 1 & \\ & & 1 \\ & & \end{bmatrix} = W_{n+1}^{(4)} \begin{bmatrix} * & & & \\ * & * & * & * \\ * & * & * & * \\ * & * & * & * \\ * & * & * & * \end{bmatrix}.$$

   Note that $\underline{H_n^{(4)}}$ still has zero entries in its first row. Also, $W_{n+1}^{(4)}$ is still of the form $W_{n+1}^{(4)} = [\, r_n(A)\boldsymbol{v} \,|\, \boldsymbol{v} \,|\, * \,|\, \cdots \,|\, * \,]$.
5. We apply the two-sided Lanczos algorithm with the lower $n \times n$ part of $\underline{H_n^{(4)}}$, using $\boldsymbol{e}_1$ as the left and right starting vector. This produces biorthogonal matrices $Z_L, Z_R \in \mathbb{C}^{n\times n}$, $Z_L^H Z_R = I_n$. Left-multiplying the decomposition with $\mathrm{blkdiag}(1, Z_L^H)$ and right-multiplication with Z_R results in the demanded structure:

$$A W_{n+1}^{(5)} \begin{bmatrix} 0 & & \\ 1 & & \\ & 1 & \\ & & 1 \\ & & \end{bmatrix} = W_{n+1}^{(5)} \begin{bmatrix} * & & \\ * & * & \\ * & * & * \\ & * & * \\ & & * \end{bmatrix}. \tag{4.8}$$

6. Finally, let the nonzero entries of $\underline{H_n^{(5)}}$ be denoted by $\eta_{i,j}$ $(1 \le j \le n,\ j \le i \le j+2)$, then we aim to scale these entries so that they are matched with those of the matrix $\underline{\widetilde{H}_n}$ in (3.5). This can be achieved by left multiplication of the pair with $L = \mathrm{diag}(1, 1, \ell_3, \ldots, \ell_{n+1}) \in \mathbb{C}^{(n+1)\times(n+1)}$ and right multiplication with $R = \mathrm{diag}(\rho_1, \rho_2, \ldots, \rho_n) \in \mathbb{C}^{n\times n}$. The diagonal entries of L and R are found by equating $\underline{\widetilde{H}_n}$ in (3.5) and $L\underline{H_n^{(5)}}R$, starting from the $(1,1)$ entry and going down columnwise. We obtain $r_1 = 1/\eta_{1,1}$, $h_1 = -1/(\eta_{2,1}\rho_1)$, $\ell_3 = 1/(\eta_{3,1}h_1\rho_1)$, and for $j = 2, 3, \ldots$ $r_j = 1/(\ell_j \eta_{j,j} h_{j-1})$, $h_j = -1/(1/h_{j-1} + \ell_{j+1}\eta_{j+1,j}\rho_j)$, $\ell_{j+2} = 1/(\eta_{j+2,j} h_j \rho_j)$. The diagonal entries of $\underline{\widetilde{K}_n}$ in (3.5) satisfy $\widehat{h}_{j-1} = \ell_{j+1}\rho_j$, $j = 1, \ldots, n$, and thus the pair has been transformed exactly into the form (3.5).

The above six-step procedure converts the RKFIT approximant r_n into continued fraction form and hence allows its interpretation as an FD scheme. This scheme is referred to as an RKFIT-FD grid. Note that all transformations only act on small matrices of size $(n+1) \times n$ and the computation of the tall skinny matrices $W_{n+1}^{(j)}$ is not required if one only needs the continued fraction parameters. We have extended the Rational Krylov Toolbox by the `contfrac` method, which implements the conversion of an RKFUN, the fundamental data type to represent and work with rational functions r_n, into continued fraction form following the above transformations. Numerically, these transformations may be ill conditioned and the use of multiple precision arithmetic is recommended. The toolbox supports MATLAB's Variable Precision Arithmetic and the Advanpix Multiprecision Toolbox [1].

Remark 4.1 In Step 3 we have assumed that the (1, 1) element of $\underline{H_n^{(2)}}$ is nonzero. This assumption is always satisfied: assuming to the contrary that the (1, 1) element of $\underline{H_n^{(2)}}$ vanishes, the first column of (4.7) reads $A\boldsymbol{v} = W_{n+2}^{(2)}[0, *, \ldots, *]^T$. This is a contradiction as the left-hand side of this equation is a superdiagonal rational function in A times $\boldsymbol{v}$, whereas the trailing n columns of $W_{n+1}^{(2)}$ can be taken to be a basis for $\mathscr{V}_n \subset \mathscr{W}_{n+1}$, which only contains diagonal (and subdiagonal) rational functions in A times $\boldsymbol{v}$ (provided that all poles $\xi_1, \ldots, \xi_{n-1}$ are finite).

Remark 4.2 In Step 5 we have assumed that the lower $n \times n$ part of $\underline{H_n^{(4)}}$ can be tridiagonalized by the two-sided Lanczos algorithm. While this conversion can potentially fail, we conjecture that if r_n admits a continued fraction form (3.2) then such an unlucky breakdown cannot occur. (The conditions for the rational function $(r_n(\lambda) - \widehat{h}_0\lambda)$ to posses this so-called Stieltjes continued fraction form [19] are reviewed in [15]; see Theorem 1.39 therein.) Even if our conjecture was false, the starting vector $\boldsymbol{v}$ will typically be chosen at random in our application. So if an unlucky breakdown occurs, trying again with another vector $\boldsymbol{v}$ would easily solve the problem. We have not encountered any unlucky breakdowns in our experiments.

5 Numerical tests: constant-coefficient case

The nonlinear rational least squares problem (4.1) is nonconvex and there is no guarantee that a minimizing solution exists, nor that such a solution would be unique. As a consequence of these theoretical difficulties and due to the nonlinear nature of RKFIT's pole relocation procedure, a comprehensive convergence analysis seems currently intractable. (An exception is [9, Corollary 3.2], which states that in exact arithmetic RKFIT converges within a single iteration if F itself is a rational matrix function of appropriate type.) However, for some special cases we can compare the RKFIT approximants to analytically constructed near-best approximants. Here we provide such comparisons to the compound Zolotarev approach in [12] and the approximants studied by Newman and Vjacheslavov [17, Section 4].

Throughout this section we assume that A is Hermitian with eigenvalues $\lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_N$. In our discussion of available convergence bounds we will usually focus on the function $f(\lambda) = \sqrt{\lambda}$, however, as has been argued in [12, Section 5.1], it is possible to obtain similar bounds for the discrete impedance function $f_h(\lambda) = \sqrt{\lambda + (h\lambda/2)^2}$. Some of our numerical experiments will be for the latter function, illustrating that the convergence behavior is indeed similar to that for the former.

5.1 Two-interval approximation with coarse spectrum

Our first test concerns the approximation of $F = f_h(A)$, $f_h(\lambda) = \sqrt{\lambda + (h\lambda/2)^2}$, where A is a nonsingular indefinite Hermitian matrix with relatively large gaps between neighboring eigenvalues. We recall the convergence result (1.2) from the introduction, which states that the geometric convergence factor is governed by the ratios of the spectral subintervals $[a_1, b_1]$ and $[a_2, b_2]$, $a_1 < b_1 < 0 < a_2 < b_2$.

Example 5.1 In Fig. 2 (top left) we show the relative errors $\|F\boldsymbol{u}_0 - r_n(A)\boldsymbol{u}_0\|_2 / \|F\boldsymbol{u}_0\|_2$ of the type $(n, n-1)$ rational functions obtained by RKFIT (dashed red curve) and the two-interval Zolotarev approach (dotted blue) for varying degrees $n = 1, 2, \ldots, 25$. Here the matrix A is defined as $A = L/h^2 - k_\infty^2 I \in \mathbb{R}^{N\times N}$, where $N = 150, h = 1/N$, $k_\infty = 15$, and

$$L = \begin{bmatrix} 1 & -1 & & & \\ -1 & 2 & -1 & & \\ & \ddots & \ddots & \ddots & \\ & & -1 & 2 & -1 \\ & & & -1 & 1 \end{bmatrix}. \tag{5.1}$$

The matrix L corresponds to a scaled FD discretization of the 1D Laplace operator with homogeneous Neumann boundary conditions. The spectral subintervals of A are $[a_1, b_1] \approx [-225, -67.2]$ and $[a_2, b_2] \approx [21.5, 8.98 \times 10^4]$. The vector $\boldsymbol{u}_0 \in \mathbb{R}^N$ is chosen at random with normally distributed entries. To compute the RKFIT approximant r_n we have used another random training vector $\boldsymbol{v}$ with normally distributed entries. The corresponding errors $\|F\boldsymbol{v} - r_n(A)\boldsymbol{v}\|_2 / \|F\boldsymbol{v}\|_2$ together with the number of required RKFIT iterations are also shown in the plot (solid red curve). For all degrees n at most 5 RKFIT iterations where required until stagnation occurred. Note that the two RKFIT convergence curves (for the vectors $\boldsymbol{u}_0$ and $\boldsymbol{v}$) are very close together, indicating that the random choice for the training vector does not affect much the computed RKFIT approximant. Note further that the RKFIT convergence follows the geometric rate predicted by (1.2) (dotted black curve) very closely initially (up to a degree $n \approx 10$), but then the convergence becomes superlinear. This convergence acceleration is due to the spectral adaptation of the RKFIT approximant.

The spectral adaptation is illustrated in the graph on the top right of Fig. 2, which plots the error curve $|f_h(\lambda) - r_{10}(\lambda)|$ of the RKFIT approximant r_{10} (solid red curve) over the spectral interval of A, together with the attained values at the eigenvalues of A (red crosses). In particular, close to $\lambda = 0$, there are two eigenvalues at which the error curve attains a relatively small value in comparison to the other eigenvalues farther away (meaning that r_n interpolates f_h nearby). These eigenvalues have started to become "deflated" by RKFIT, effectively shrinking the spectral subintervals $[a_1, b_1]$ and $[a_2, b_2]$, and thereby leading to the observed superlinear convergence.

In the bottom of Fig. 2 we show the poles and residues of the RKFIT approximant r_{10} (left) and the associated continued fraction parameters (right), giving rise to the RKFIT-FD grid. All the involved quantities have been computed using the new `contfrac` method in the Rational Krylov Toolbox.

5.2 Two-interval approximation with dense spectrum

The superlinear convergence effects observed in the previous example should disappear when the spectrum of A is dense enough so that, for the order n under consideration, no eigenvalues of A are deflated by interpolation nodes of r_n. The next example demonstrates this.

Example 5.2 In Fig. 3 we show the relative errors $\|F\boldsymbol{u}_0 - r_n(A)\boldsymbol{u}_0\|_2 / \|F\boldsymbol{u}_0\|_2$ of the type $(n, n-1)$ rational functions obtained by RKFIT and the Zolotarev approach for

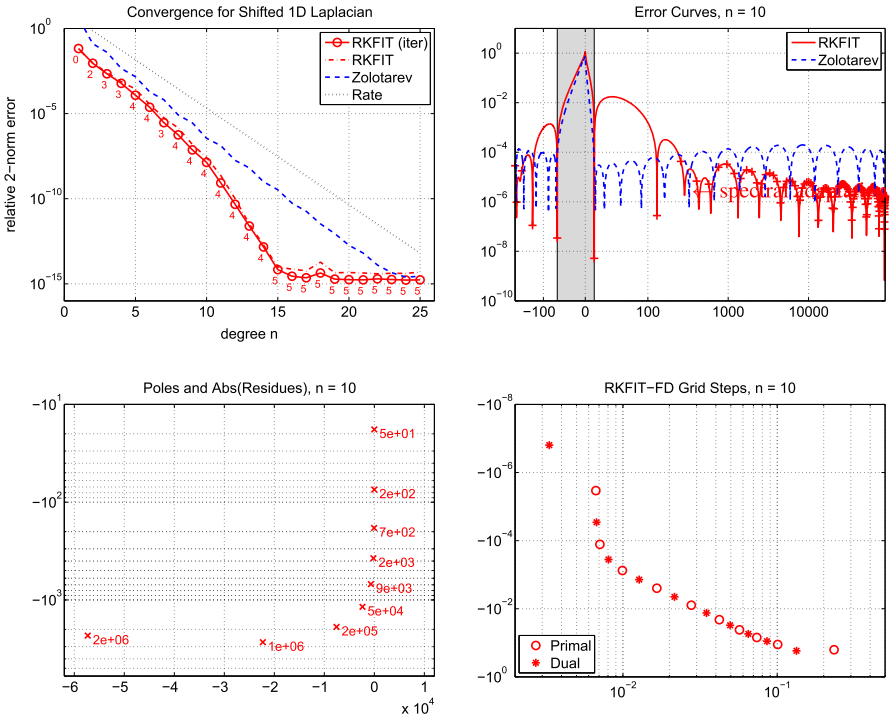

Fig. 2 Top: Accuracy comparison of RKFIT and Zolotarev approximants for a shifted 1D Laplacian which has a rather coarse spectrum, hence resulting in superlinear RKFIT convergence. The DtN function is $f_h(\lambda) = \sqrt{\lambda + (h\lambda/2)^2}$. The small numbers on the solid red convergence curve on the left indicate the number of required RKFIT iterations. Bottom: The poles and residues of the RKFIT approximant r_{10} (left) and the associated continued fraction parameters (right)

varying degrees $n = 1, 2, \ldots, 25$. Now the matrix A corresponds to a shifted 2D Laplacian $A = (L \otimes L)/h^2 - k_\infty^2 I \in \mathbb{R}^{N\times N}$ with $N = 150^2$, $h = 1/150$, $k_\infty = 15$, and with L defined in (5.1). The special structure of L (and A) allows for the use of the 2D discrete cosine transform for computing $F = f_h(A)$. The spectral subintervals of A are $[a_1, b_1] \approx [-225, -27.7]$ and $[a_2, b_2] \approx [21.5, 1.80 \times 10^5]$. The vector $\boldsymbol{u}_0 \in \mathbb{R}^N$ is chosen at random with normally distributed entries. We also show the relative error of the RKFIT approximant $r_n(A)\boldsymbol{v}$ with another randomly chosen training vector $\boldsymbol{v}$, and the number of required RKFIT iterations. As in the previous example there is no big difference in accuracy when evaluating the RKFIT approximant for $\boldsymbol{u}_0$ or $\boldsymbol{v}$, however, the number of required RKFIT iterations is slightly higher in this example. As the eigenvalues of the matrix A are relatively dense in its spectral interval, we now observe that no spectral adaptation takes place and both the RKFIT and the Zolotarev approximants converge at the rate predicted by (1.2).

In the bottom of Fig. 3 we show the grid vectors $\boldsymbol{u}_j$ satisfying the FD relation (3.3) for $n = 10$, with the RKFIT-FD grid parameters h_j and $\widehat{h}_{j-1}$ ($j = 0, 1, \ldots, 10$) extracted from r_{10}. The entries of $\boldsymbol{u}_j$ are complex-valued, hence we show the $\log_{10}$ of the amplitude and phase separately. Note how the amplitude decays very quickly as the

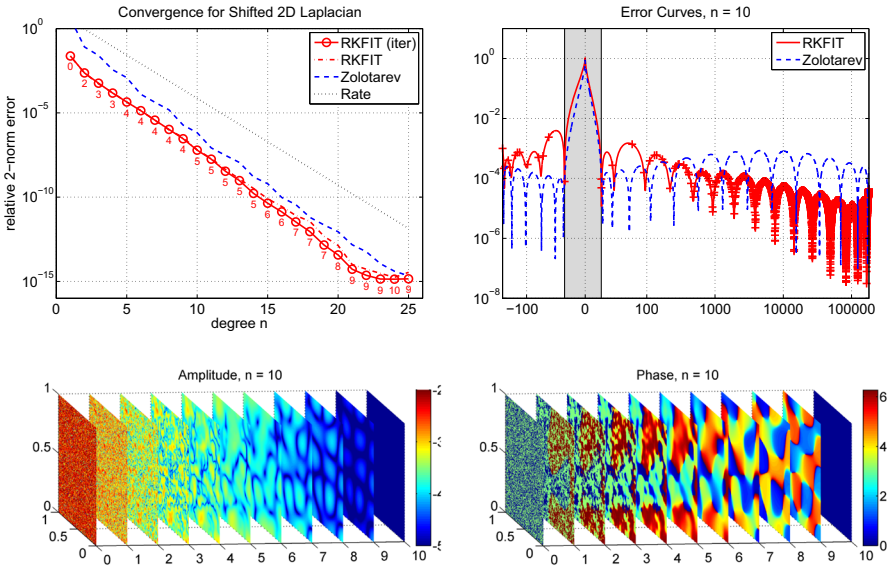

Fig. 3 Top: Comparison of RKFIT and Zolotarev approximants for a shifted 2D Laplacian. Bottom: The $\log_{10}$ of the amplitude and phase of the grid vectors $\boldsymbol{u}_j$ $(j = 0, 1, \ldots, n = 10)$. Qualitatively, the poles and residues and the complex grid steps for the associated RKFIT approximant r_{10} look similar to those in Fig. 2 and are therefore omitted

random signal travels further to the right in the grid, illustrating the good absorption property of this grid.

5.3 Approximation on an indefinite interval

In order to remove the spectral gap $[b_1, a_2]$ from which the previous two examples benefited, we now consider the approximation on an indefinite interval.

Example 5.3 We approximate $f(\lambda) = \sqrt{\lambda}$ on the indefinite interval $[a_1, b_2] = [-225, 1.80 \times 10^5]$. Note that $[a_1, b_2]$ is the same as in the previous Example 5.2, but without the spectral gap about zero. This problem is of interest as, in the variable-coefficient case, one cannot easily exploit a spectral gap between the eigenvalues of A which are closest to zero. This is because a varying coefficient $c(x)$ can be thought of as a variable shift of the eigenvalues of A; hence an eigenvalue-free interval $[b_1, a_2]$ may not always exist.

To mimic continuous approximation on an interval, we use for A a surrogate diagonal matrix of size $N = 200$ having 100 logspaced eigenvalues in $[a_1, -10^{-16}]$ and $[10^{-16}, b_2]$, respectively. The training vector $\boldsymbol{v}$ is chosen as $[1, 1, \ldots, 1]^T$. We run RKFIT for degrees $n = 1, 2, \ldots, 25$. The relative error of the RKFIT approximants $\|F\boldsymbol{v} - r_n(A)\boldsymbol{v}\|_2 / \|F\boldsymbol{v}\|_2$ seems to reduce like $\exp(-\pi\sqrt{n})$; see Fig. 4 (left).

We also compare RKFIT to a two-interval Remez-type approximant obtained by using the interpolation nodes of numerically computed best approximants to $\sqrt{\lambda}$ on

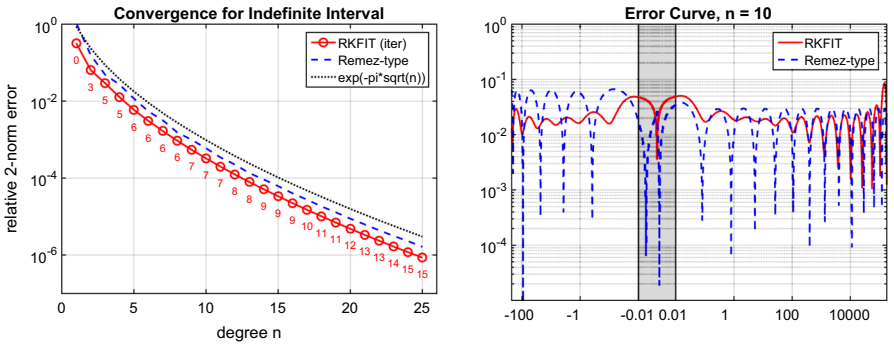

Fig. 4 RKFIT approximation of $f(\lambda) = \sqrt{\lambda}$ on an indefinite interval $[a_1, b_2]$, $a_1 < 0 < b_2$, compared to a two-interval Remez-type approximant. Qualitatively, the poles/residues and the complex grid steps associated with r_{10} look similar to those in Fig. 2 and are therefore omitted

$[0, 1]$, scaling them appropriately, and unifying them for the intervals $[a_1, 0]$ and $[0, b_2]$. The number of interpolation nodes on both intervals is balanced so that the resulting error curve is closest to being equioscillatory on the whole of $[a_1, b_2]$. Again the error of the so-obtained Remez-type approximant seems to reduce like $\exp(-\pi\sqrt{n})$.

Remark 5.1 The uniform rational approximation of $\sqrt{\lambda}$ on a semi-definite interval $[0, b_2]$ has been studied by Newman and Vjacheslavov. It is known that the error reduces like $\exp(-\pi\sqrt{2n})$ with the degree n; see [17, Section 4]. Based on the observations in Fig. 4 we conjecture that the error of the best rational approximant to $\sqrt{\lambda}$ on an indefinite interval $[a_1, b_2]$ reduces like $\exp(-\pi\sqrt{n})$.

6 Numerical tests: variable-coefficient case

We now consider a variable-coefficient function c motivated by a geophysical seismic exploration setup as shown in Fig. 5. Here a pressure wave signal of a single frequency is emitted by an acoustic transmitter in the Earth's subsurface, travels through the underground, and is then logged by receivers on the surface. From these measurements geophysicists try to infer variations in the wave speed to draw conclusions about the subsurface composition. The computational domain of interest is a three-dimensional portion of the Earth and we might have knowledge about the sediment layers below this domain, i.e., for $x \geq 0$ in Fig. 5. While the acoustic waves in $x \geq 0$ may not be of interest on their own, the layers might cause wave reflections back into the computational domain and hence need be part of the model.

Example 6.1 At the $x = 0$ interface of the computational domain, shown in Fig. 5, we assume to have a 2D Laplacian $A = (L \otimes L)/h^2 - k_\infty^2 I$ with L defined in (5.1), and $N = 150^2$, $h = 150$, and $k_\infty = 15$. Now the function f_h of interest is (2.3), with the coefficients c_j obtained by discretizing the piecewise-constant coefficient function c which equals -400 on $[0, T)$, $+125$ on $[T, 2T)$ and 0 on $[2T, \infty)$. The thickness of the

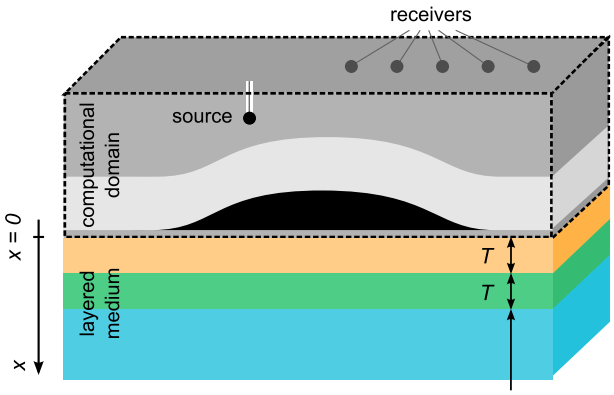

Fig. 5 Typical setup of a seismic exploration of the Earth's subsurface. It is of practical interest to compress the layered medium in $x \geq 0$ into a single PML with a small number of grid points

two finite layers T is varied in $\{0.25, 0.5, 1, 2\}$. For each thickness T, the four panels in the top of Fig. 6 show the modulus of f_h over the spectral subintervals $[a_1, b_1]$ and $[a_2, b_2]$ of A, glued together with the gray linear region $[b_1, a_2]$. It becomes apparent that with increasing T the function f_h exhibits more poles on or nearby the spectral interval of A, indicated by the upward spikes.

The convergence of the RKFIT approximants for increasing degree n is shown in Fig. 6 on the bottom left. For each thickness T there are two curves very nearby: a solid curve showing the relative 2-norm approximation error for $F\boldsymbol{v}$ (where $\boldsymbol{v}$ is a random training vector) and a dashed curve for $F\boldsymbol{u}_0$ (where $\boldsymbol{u}_0$ is another random testing vector). We observe that RKFIT converges very robustly for this piecewise constant-coefficient problem. Similar behavior has been observed in many numerical tests with other offset functions c. We refer to the example collection of the Rational Krylov Toolbox which contains further examples. The codes for producing our examples are available online and can easily be modified to other coefficient functions.

Example 6.2 Here we consider a diagonal matrix A with the same indefinite spectral interval as the matrix in the previous example but with dense eigenvalues, namely 100 logspaced eigenvalues in $[a_1, -10^{-16}]$ and $[10^{-16}, b_2]$, respectively. The convergence is shown on the bottom right of Fig. 6. Again the RKFIT behavior is very robust even for high approximation degrees n, but compared to the above Example 6.1 the convergence is delayed, indicating that spectral adaptation has been prevented here.

7 Discussion and conclusions

An obvious alternative to our grid compression approach in the two examples of Sect. 6 would be to use an efficient discretization method on c's support, and then to append it to the constant-coefficient PML of [12]. In principle such an approach requires at least the integer part of $\mathrm{N} = \pi^{-1} \int_0^H \sqrt{k_\infty^2 - c(x)}\, \mathrm{d}x$ discretization points according to the Nyquist sampling rate, where H is the total thickness of c's support. In fact, the

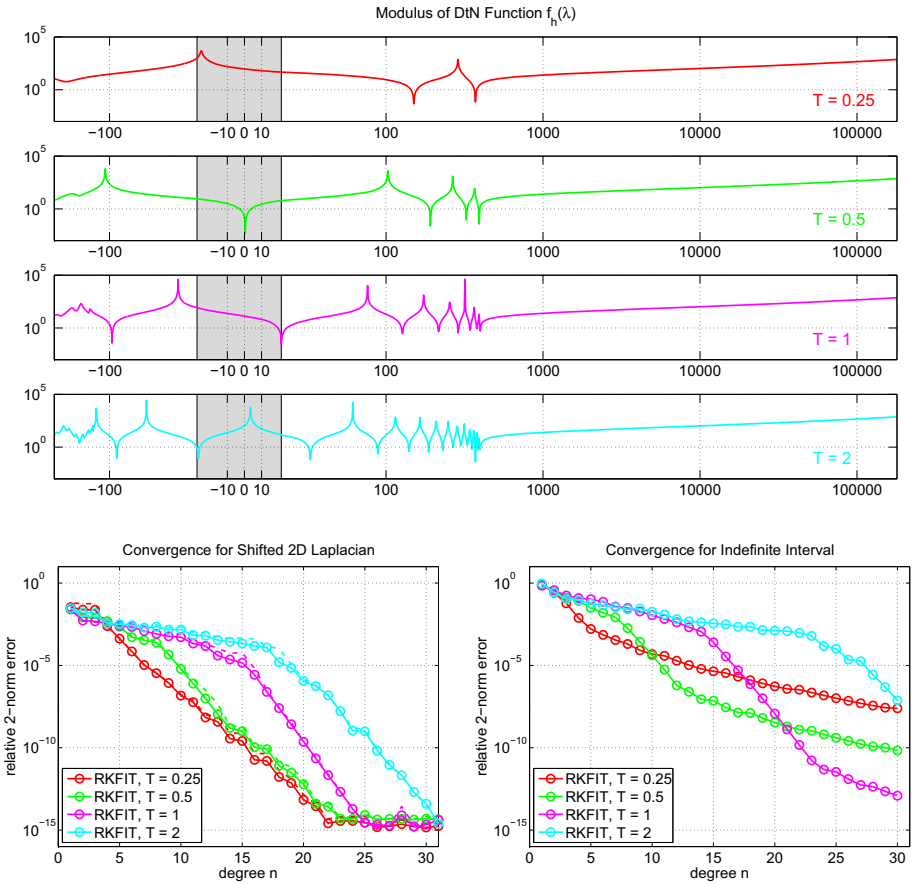

Fig. 6 Top: The four panels show the modulus of the discrete variable-coefficient DtN function f_h for varying thickness T of the two finite layers. Bottom: The two plots show the RKFIT convergence for approximating $f_h(A)\boldsymbol{v}$ when A is a shifted 2D Laplacian (left) and a diagonal matrix with dense eigenvalues in the same spectral interval (right), respectively

classical spectral element method (SEM) with polynomial local basis requires at least $\frac{\pi}{2}$N grid points [2]. (The downside of SEM compared to our FD approach is its high linear solver cost per unknown caused by the dense structure of the resulting linear systems.) The following table shows the minimal number of grid points required for discretizing the two finite layers in the examples of Sect. 6, depending on the layer thickness T, as well as the number of RKFIT-FD grid points to achieve a relative accuracy of 10^{-5} for the same problem:

Although we also observe with RKFIT-FD a tendency that the DtN functions become more difficult to approximate when the layer thickness increases (an increase of the coefficient jumps between the layers will have a similar effect), the number of required grid points can be significantly smaller than the Nyquist limit N. A possible explanation for this phenomenon is RKFIT's ability to adapt to the spectrum of A, not

	$T = 0.25$	$T = 0.5$	$T = 1$	$T = 2$
Nyquist minimum N	8.75	17.5	35	70
SEM minium $\frac{\pi}{2}$N	13.7	27.5	55.0	110.0
RKFIT-FD (Example 6.1)	8	10	16	19
RKFIT-FD (Example 6.2)	14	11	17	28

being slowed down in convergence by singularities of the DtN function well separated from the eigenvalues of A. In the "Appendix" we analyze this phenomenon.

Acknowledgements We thank Ralf Hiptmair and the anonymous referees for constructive comments that have significantly improved the presentation. Druskin was partially supported by the Air Force Office of Scientific Research (AFOSR) Grant FA9550-20-1-0079. Güttel was partially supported by The Alan Turing Institute, Engineering and Physical Sciences Research Council (EPSRC) Grant EP/W001381/1. Knizhnerman was supported by the Moscow Center of Fundamental and Applied Mathematics (Agreement 075-15-2019-1624 with the Ministry of Education and Science of the Russian Federation).

A Nyquist limit-type criterion for rational approximation

The top-four panels in Fig. 6 suggest that the DtN function f_h, specified in (2.3), develops more and more poles on the real axis as the thickness of the finite layers increases. These poles are also known as scattering resonances. In order to analyze this behavior, we consider a two-layer waveguide problem with piecewise constant wave numbers similar to the one in Fig. 1, but now in the continuous setting without any FD approximation. This problem is governed by the equations

$$u''(x) = (\lambda + c)u(x) \text{ for } x \in [0, T), \qquad u''(x) = \lambda u(x) \text{ for } x \in [T, \infty),$$

with given $u(0) = u_0$ and the decay condition $u(x) \to 0$ as $x \to \infty$. Here, T is the thickness of the first layer with an offset coefficient c. In terms of the Helmholtz equation, a value $c = -k_0^2 < 0$ means that the wave number on the first layer is larger than on the second, whereas $c > 0$ means that the wave number on the first layer is smaller than on the second. If $c = 0$ we have a homogeneous infinite waveguide.

Our aim is to solve for u explicitly and to determine a formula for the DtN function f satisfying $f(\lambda)u_0 = -u'(0)$. For $x \in [0, T]$ we have

$$u(x) = \alpha e^{x\sqrt{\lambda+c}} + (u_0 - \alpha)e^{-x\sqrt{\lambda+c}} 2\alpha \sinh\left(x\sqrt{\lambda + c}\right) + e^{-x\sqrt{\lambda+c}} u_0,$$

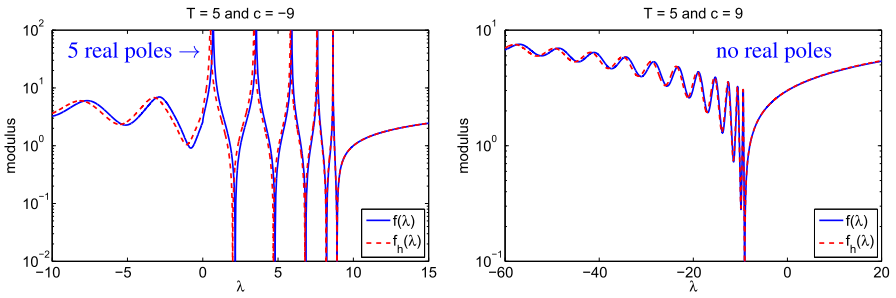

Fig. 7 The DtN function f defined in (A.1), as well as its discrete counterpart (2.3), for two different choices of the parameters (T, c)

where the square roots are understood as the analytical continuation through the upper half plane from the axis $\lambda > -c$. For $x \in [T, \infty)$ we require a decaying solution, hence $u(x) = \beta e^{-x\sqrt{\lambda}}$ there. By continuity of $u(x)$ at $x = T$ we have

$$\beta = \left(2\alpha \sinh\left(T\sqrt{\lambda + c}\right) + e^{-T\sqrt{\lambda+c}}u_0\right) \cdot e^{T\sqrt{\lambda}}.$$

By continuity of $u'(x)$ at $x = T$ we further require

$$\sqrt{\lambda + c} \cdot \left(2\alpha \cosh(T\sqrt{\lambda + c}) - e^{-T\sqrt{\lambda+c}}u_0\right) = -\beta\sqrt{\lambda} \cdot e^{-T\sqrt{\lambda}},$$

hence

$$\begin{aligned}\sqrt{\lambda + c} \cdot \left(2\alpha \cosh(T\sqrt{\lambda + c}) - e^{-T\sqrt{\lambda+c}}u_0\right) \; &= -\left(2\alpha \sinh\left(T\sqrt{\lambda + c}\right)\right. \\ &\quad \left. + e^{-T\sqrt{\lambda+c}}u_0\right) \cdot \sqrt{\lambda},\end{aligned}$$

from which α can be determined as

$$\alpha = \frac{u_0}{2} \cdot \frac{\left(\sqrt{\lambda + c} - \sqrt{\lambda}\right)e^{-T\sqrt{\lambda+c}}}{\sqrt{\lambda + c}\cosh(T\sqrt{\lambda + c}) + \sqrt{\lambda}\sinh\left(T\sqrt{\lambda + c}\right)}.$$

Note that $\alpha = \alpha_\lambda$ is a function of λ. Using the fact that $u'(0) = (2\alpha_\lambda - u_0)\sqrt{\lambda + c}$, the DtN function f satisfying $f(\lambda)u_0 = -u'(0)$ is given as

$$f(\lambda) = \frac{\sqrt{\lambda + c} \cdot \sinh(T\sqrt{\lambda + c}) + \sqrt{\lambda} \cdot \cosh\left(T\sqrt{\lambda + c}\right)}{\sqrt{\lambda + c} \cdot \cosh(T\sqrt{\lambda + c}) + \sqrt{\lambda} \cdot \sinh\left(T\sqrt{\lambda + c}\right)} \cdot \sqrt{\lambda + c}. \tag{A.1}$$

A plot of this function for two different parameter choices $T = 5$ and $c = \pm 9$ is shown in Fig. 7. We observe that this function is smooth over the whole real axis when $c \geq 0$, while it develops singularities when $c < 0$. The following lemma shows that the number of real poles is proportional to c and T.

Lemma A.1 *The function f defined in (A.1) can be analytically continued from $\lambda > \max\{0, -c\}$ through the upper half plane to the whole real axis except for two ramification points $\lambda = 0$ and $\lambda = -c$ and possibly a finite number of poles. For $c > 0$, the function f has no poles on the real axis. For $c < 0$, the function f has $\left\lfloor \frac{T\sqrt{-c}}{\pi} \right\rfloor + q$ real poles, where $q \in \{0, 1\}$, all located in the interval $(0, -c)$.*

Proof We investigate the roots of the denominator $g(\lambda) = \sqrt{\lambda + c} \cdot \cosh(T\sqrt{\lambda + c}) + \sqrt{\lambda} \cdot \sinh(T\sqrt{\lambda + c})$. We first consider the case $c < 0$ and argue that there are no real roots of g outside $[0, -c]$. For $\lambda < 0$, the factors $\sqrt{\lambda + c}$ and $\sqrt{\lambda}$ are purely imaginary and nonzero, while $\cosh(T\sqrt{\lambda + c}) = \cos(Tz)$ is purely real and $\sinh(T\sqrt{\lambda + c}) = i \sin(Tz)$ purely imaginary (here and throughout the proof $z = \operatorname{imag}(\sqrt{\lambda + c})$). Hence, λ can only be a root of g if $\cos(Tz) = \sin(Tz) = 0$, but this cannot happen as $\cos(\cdot)$ and $\sin(\cdot)$ do not have any roots in common. A similar argument shows that there are no roots of g for $\lambda > -c$.

For $\lambda \in (0, -c)$, $z = \operatorname{imag}(\sqrt{\lambda + c})$ varies in $(0, \sqrt{-c})$ and we want to count the number of roots of the purely imaginary function $h(z) = g(\lambda) = iz \cos(Tz) + \sqrt{z^2 + c} \cdot \sin(Tz)$ on that interval. Consider the subintervals $I_k = ((k-1)\pi/T, k\pi/T]$ for $k = 1, 2, \ldots, K = \lfloor T\sqrt{-c}/\pi \rfloor$. Then on the first half of each I_k the function $\operatorname{imag}(h)$ is strictly positive (or negative), while on the second half it is strictly monotonically decreasing (increasing) with a sign change. Therefore each I_k contributes exactly one root of h. The final interval $(K\pi/T, \sqrt{-c})$ may or may not contain a further root of h. By the same argument one shows that the roots of the numerator of f are located on the first half's of I_k, and hence the roots of the denominator do not cancel out.

For $c \geq 0$ one argues similarly to the first part of the proof that the denominator function g has no roots for all real values of λ. □

To interpret this result in terms of the indefinite Helmholtz equation $(\partial_{yy} + \partial_{zz})u + (k_\infty^2 - c(x))u = 0$ for $c < 0$, first note that the DtN function (A.1) does not depend on k_∞, but merely on the offset c. We may therefore set $k_\infty = 0$, in which case the wave number on the first finite layer is simply $k = \sqrt{-c}$. Furthermore, $\ell = 2\pi/k = 2\pi/\sqrt{-c}$ is the corresponding wavelength. Using this notation, Lemma A.1 states that f has $\approx 2T/\ell$ poles on the real axis, that is, *two real poles per wavelength!*

Although Lemma A.1 is stated for the continuous waveguide problem, discrete DtN functions f_h seem to have poles very close to those of their continuous counterparts f. An example is shown in Fig. 7 (dashed red curve), which corresponds to (2.3) with "piecewise" constant coefficients c_j and $h = 0.05$.

Returning to the RKFIT convergence, we observed in the experiments in Sect. 6 that the minimal number n of RKFIT-FD grid points required to achieve convergence does not seem to be directly linked to the Nyquist criterion. Although f_h may have a large number N of singularities on the spectral interval of A, RKFIT's spectral adaptation capabilities mean that r_n does not need to resolve them all, and therefore the degree n can be significantly smaller than N. Although Lemma A.1 effectively states a Nyquist-type criterion for the layered waveguide, from a rational approximation point of view RKFIT-FD grids can outperform it in case of a favorable spectral distribution of the matrix A.

References

1. Advanpix LLC., Tokyo, Japan. Multiprecision Computing Toolbox for MATLAB (2015)
2. Ainsworth, M., Wajid, H.A.: Dispersive and dissipative behavior of the spectral element method. SIAM J. Numer. Anal. **47**, 3910–3937 (2009)
3. Appelö, D., Hagstrom, T., Kreiss, G.: Perfectly matched layers for hyperbolic systems: general formulation, well-posedness, and stability. SIAM J. Appl. Math. **67**, 1–23 (2006)
4. Beattie, C.A., Gugercin, S.: Model reduction by rational interpolation. Model reduction and algorithms: theory and applications. Comput. Sci. Eng. **15**, 297–334 (2017)
5. Berenger, J.P.: A perfectly matched layer for the absorption of electromagnetic waves. J. Comput. Phys. **114**, 185–200 (1994)
6. M. Berljafa, S. Elsworth, and S. Güttel. A Rational Krylov Toolbox for MATLAB. Technical Report MIMS Eprint 2014.56, The University of Manchester, 2020
7. Berljafa, M., Güttel, S.: Generalized rational Krylov decompositions with an application to rational approximation. SIAM J. Matrix Anal. **36**, 894–916 (2015)
8. Berljafa, M., Güttel, S.: Parallelization of the rational Arnoldi algorithm. SIAM J. Sci. Comput. **39**(5), S197–S221 (2017)
9. Berljafa, M., Güttel, S.: The RKFIT algorithm for nonlinear rational approximation. SIAM J. Sci. Comput. **39**(5), A2049–A2071 (2017)
10. Bélanger-Rioux, R., Demanet, L.: Compressed absorbing boundary conditions via matrix probing. SIAM J. Numer. Anal. **53**(5), 2441–2471 (2015)
11. Chew, W., Weedon, B.: A 3D perfectly matched medium from modified Maxwell's equations with stretched coordinates. Microw. Opt. Technol. Lett. **7**, 599–604 (1994)
12. Druskin, V., Güttel, S., Knizhnerman, L.: Near-optimal perfectly matched layers for indefinite Helmholtz problems. SIAM Rev. **58**(1), 90–116 (2016)
13. Druskin, V., Knizhnerman, L.: Gaussian spectral rules for the three-point second differences: I. A two-point positive definite problem in a semi-infinite domain. SIAM J. Numer. Anal. **37**(2), 403–422 (1999)
14. Engquist, B., Majda, A.: Radiation boundary conditions for acoustic and elastic wave calculations. Commun. Pure Appl. Math. **32**, 313–357 (1979)
15. Holtz, O., Tyaglov, M.: Structured matrices, continued fractions, and root localization of polynomials. SIAM Rev. **54**(3), 421–509 (2012)
16. Ingerman, D., Druskin, V., Knizhnerman, L.: Optimal finite difference grids and rational approximations of the square root. I. Elliptic problems. Commun. Pure Appl. Math. **53**(8), 1039–1066 (2000)
17. Petrushev, P.P., Popov, V.A.: Rational Approximation of Real Functions. Cambridge University Press, Cambridge (1987)
18. Ruhe, A.: Rational Krylov: a practical algorithm for large sparse nonsymmetric matrix pencils. SIAM J. Sci. Comput. **19**(5), 1535–1551 (1998)
19. Stieltjes, T.J.: Recherches sur les fractions continues. Ann. Fac. Sci. Toulouse **8**, 1–122 (1894)