Compressing variable-coefficient exterior Helmholtz problems via RKFIT

Druskin, Vladimir and Güttel, Stefan and Knizhnerman, Leonid

2016

MIMS EPrint: **2016.53**

Manchester Institute for Mathematical Sciences
School of Mathematics

The University of Manchester

Reports available from: http://eprints.maths.manchester.ac.uk/
And by contacting: The MIMS Secretary
School of Mathematics
The University of Manchester
Manchester, M13 9PL, UK

ISSN 1749-9097

COMPRESSING VARIABLE-COEFFICIENT EXTERIOR HELMHOLTZ PROBLEMS VIA RKFIT

VLADIMIR DRUSKIN*, STEFAN GÜTTEL†, AND LEONID KNIZHNERMAN‡

Abstract. The efficient discretization of Helmholtz problems on unbounded domains is a challenging task, in particular, when the wave medium is nonhomogeneous. We present a new numerical approach for compressing finite difference discretizations of such problems, thereby giving rise to efficient perfectly matched layers (PMLs) for nonhomogeneous media. This approach is based on the solution of a nonlinear rational least squares problem using the RKFIT method proposed in [M. Berljafa and S. Güttel, SIAM J. Matrix Anal. Appl., 36(2):894–916, 2015]. We show how the solution of this least squares problem can be converted into an accurate finite difference grid within a rational Krylov framework. Several numerical experiments are included. They indicate that RKFIT computes PMLs more accurate than previous analytic approaches and even works in regimes where the Dirichlet-to-Neumann functions to be approximated are highly irregular. Spectral adaptation effects allow for accurate finite difference grids with point numbers below the Nyquist limit.

Key words. finite difference grid, Helmholtz equation, Dirichlet-to-Neumann map, perfectly matched layer, rational approximation, continued fraction

AMS subject classifications. 35J05, 65N06, 65N55, 30E10

1. Introduction. Finite difference (FD) methods are widely used for the numerical solution of partial differential equations. Due to their simplicity and potential for high computational efficiency they are often preferred to more sophisticated techniques like, e.g., finite element or spectral methods. A prominent example is the finite difference time domain (FDTD) solution of Maxwell's equations [27] over a fixed time interval, which can be shown to have optimal computational efficiency in the sense that $O(n^{4/3})$ numbers are produced in $O(n^{4/3})$ operations; see [15] and [14, Section 4.6]. Here, n is the number of grid points in each coordinate direction of a three-dimensional cube. FD methods are also attractive in that it is relatively straightforward to model wave problems on unbounded domains by stretching the FD grid steps. This approach pioneered and analyzed in [15, 18, 6, 4] is nowadays known as complex coordinate stretching, leading to so-called *perfectly matched layers* (PMLs). By appropriately choosing the complex grid steps, one can even construct short-term recurrence FD grids with spectral accuracy [17, 5]. However, due to the analytic construction of these PMLs they require the medium to be invariant in the stretching direction. A more detailed review of these techniques is given in [16].

In this work we present a new approach to the compression of the FD grids with a variation in the PDE coefficients. As a prototypical problem we consider the infinite FD scheme

$$\frac{2}{h}\left(\frac{\boldsymbol{u}_1-\boldsymbol{u}_0}{h}+\boldsymbol{b}\right)=(A+c_0I)\boldsymbol{u}_0, \tag{1.1a}$$

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

*Schlumberger-Doll Research, 1 Hampshire St., Cambridge, Massachusetts, 19104-2688 (druskin1@slb.com).

†School of Mathematics, The University of Manchester, Alan Turing Building, Manchester, M13 9PL, United Kingdom (stefan.guettel@manchester.ac.uk).

‡Mathematical Modelling Department, Central Geophysical Expedition, Narodnogo Opolcheniya St., house 38, building 3, 123298, Moscow, Russia (lknizhnerman@gmail.com).

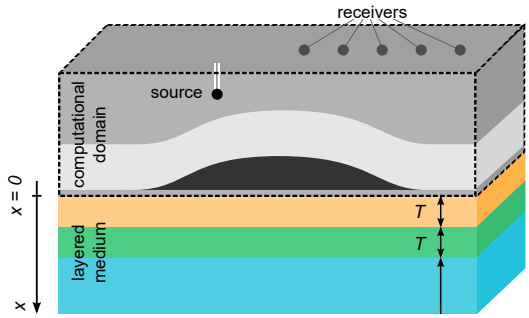

FIGURE 1.1. *Typical setup of a seismic exploration of the Earth's subsurface. We aim to compress the layered structure in $x \geq 0$ into a single PML with a small number of grid points.*

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

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

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. In this case 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. After discretization the variation of the wave number in the x-direction is modeled by varying coefficients c_j, with the "overall" wave number being $\sqrt{k_\infty^2 - c_j}$ at each grid point. Most interesting are *oscillatory* Helmholtz problems where $k_\infty^2 - c_j$ is positive on the entire domain.

Problems such as the above arise, for example, in geophysical seismic exploration; see Figure 1.1 for an illustration. 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 which then allows them 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 Figure 1.1. 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.

Referring again to the illustration in Figure 1.1, with the techniques developed in this paper it is possible to efficiently compress an FD grid for the nonhomogeneous medium in $x \geq 0$. By "compressing an FD grid" we mean the task of computing an equivalent *short-term recurrence FD grid with a small number of points* that preserves essential features of the original grid like, for example, the linear relation between Dirichlet and Neumann data at a certain grid point. The associated Dirichlet-to-Neumann (DtN) operator plays an important role in many applications, including domain decomposition methods (see, e.g., [19, 20, 13]) and the construction of PMLs. For our concrete example (1.1), the DtN operator at $x = 0$ is a matrix F such that $F\boldsymbol{u}_0 = \boldsymbol{b}$. Since (1.1) is a linear recurrence relation it is easy to verify that $F = f_h(A)$ is a matrix function in A. In the simplest case where $c_j \equiv 0$, the DtN function for (1.1) at $x = 0$ is $f_h(\lambda) = \sqrt{\lambda + (h\lambda/2)^2}$. As $h \to 0$ we obtain the DtN function $f(\lambda) = \sqrt{\lambda}$ for the continuous problem.

As will be explained in more detail below, a *compact representation* of the FD grid (1.1) can be obtained by computing a low-order rational matrix function $r_n(A) \approx f_h(A)$. In the case where A is Hermitian and $f(\lambda) = \sqrt{\lambda}$, a near-optimal rational approximant r_n to f can be constructed analytically. 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 [16] gives an explicit construction of a rational function $r_n^{(Z)}$ of type $(n, n-1)$ such that

$$\max_{\lambda \in K} |1 - r_n^{(Z)}(\lambda)/f(\lambda)| \asymp \exp\left(-n \cdot \frac{2\pi^2}{\log\left(256 \cdot \frac{a_1 b_2}{a_2 b_1}\right)}\right) \quad \text{as} \quad n \to \infty, \tag{1.2}$$

for sufficiently large interval ratios a_1/b_1 and b_2/a_2. The construction is based on combining two Zolotarev approximants (see [28] and [2, Appendix E]), one being optimal for $[a_1, b_1]$ and the other being optimal for $[a_2, b_2]$, and then balancing their degrees carefully. It can also be shown that the convergence factor in (1.2) is optimal. As a consequence, the approximation error

$$\|f(A) - r_n^{(Z)}(A)\|_2 \le C \max_{\lambda \in K} |1 - r_n^{(Z)}(\lambda)/f(\lambda)|$$

also decays exponentially at the same optimal rate. Interestingly, the continued fraction form of the rational approximants $r_n^{(Z)}$ gives rise to a geometrically meaningful three-point FD scheme, called for short the *optimal grid*. 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.

The analytic approach is essentially limited to the scalar approximation of DtN functions such as $\sqrt{\lambda}$ and $\sqrt{\lambda + (h\lambda/2)^2}$. If the coefficients c_j in (1.1) are nonconstant, f_h is more involved and an explicit construction of a rational approximant $r_n \approx f_h$ is generally impossible. In the case of non-oscillatory boundary value problems (i.e., when $k_\infty^2 - c(x)$ is nonpositive for all $x \ge 0$), variable-coefficient media have been considered in the context of inverse problems; see, e.g., [11, 12]. In this case f_h is analytic and of Stieltjes–Markov type on the spectral interval of A, and rational approximants can be obtained efficiently via (multi-point) Padé techniques. The coefficients of the continued fraction representation of these approximants can again be interpreted as geometrically meaningful FD grids.

The approximation problems become much more difficult in the oscillatory case. An illustrating example is given in Figure 1.2, 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 Figure 1.2 (solid red curve). Apparently 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.

Further complications arise when A is non-Hermitian, in which case the problem $r_n(A) \approx F = f_h(A)$ may require rational approximation in the complex plane.

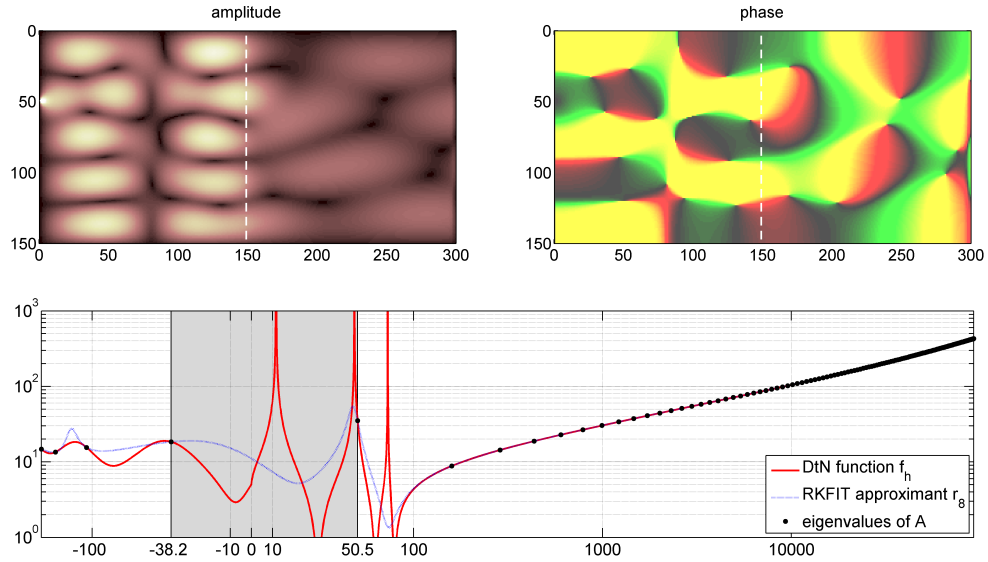

FIGURE 1.2. *A waveguide with varying coefficient (wave number) in the x-direction (piecewise constant over the first 150 grid points and the remaining grid points until infinity). The top row shows the amplitude and phase of the solution, with the position of the coefficient jump highlighted by vertical 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 doubly logarithmic on both axis, 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 the eigenvalues of A (black dots).*

In order to overcome these problems, we propose a new numerical approach using RKFIT, which is an iterative rational Krylov-based algorithm for computing a rational function r_n such that, for a given nonzero *training vector* $\boldsymbol{v} \in \mathbb{C}^N$, $r_n(A)\boldsymbol{v} \approx F\boldsymbol{v}$ in the Euclidean norm [7, 9]. The RKFIT approximant naturally incorporates the spectral weights present in $\boldsymbol{v}$ and it exploits the discreteness of the spectrum of A. In particular the latter property is crucial for solving the aforementioned approximation problems where f_h has singularities in or nearby the spectral region of A. This is exemplified by the RKFIT approximant r_n shown on the bottom of Figure 1.2 (dashed blue curve). This approximant is of degree $n = 8$ and has a relative accuracy of $\|F\boldsymbol{u}_0 - r_n(A)\boldsymbol{u}_0\|_2/\|F\boldsymbol{u}_0\|_2 \approx 1.4 \cdot 10^{-6}$. As the plot illustrates, 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. We emphasize that this remarkable *spectral adaptation* is achieved *without requiring a spectral decomposition of A explicitly*; RKFIT merely requires the repeated computation of matrix-vector products with F.

The rational Krylov framework is very natural not only for the efficient computation of r_n, but we also present a rational Krylov-based algorithm for its direct conversion into an implementable three-point FD scheme. We refer to the resulting grid as an *RKFIT-FD grid*. Unlike the above mentioned optimal grids for constant-coefficient oscillatory and variable-coefficient non-oscillatory problems, RKFIT-FD grids may not have a nice geometric interpretation, but they can nevertheless be used as efficient PMLs for nonhomogeneous media.

We typically observe 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. In fact, we will demonstrate that the Nyquist limit of two grid points per wavelength does not fully apply to RKFIT-FD grids due to spectral adaptation effects. For the problem in Figure 1.2, 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 outline of this paper is as follows: in section 2 we derive analytic expressions of DtN maps for constant- and variable-coefficient media. We also show how the optimization of DtN approximants relates to rational approximation problems. In section 3 we establish a new connection between rational Krylov spaces and FD grids. In section 4 we briefly review the RKFIT algorithm and tailor it to our specific application of FD grid optimization. A pseudocode of our algorithm, together with computational considerations, is given in section 5. Sections 6 and 7 are dedicated to convergence comparisons of our FD approximants, with relations made to convergence results from the literature whenever possible. In section 8 we discuss the numerical results and 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.

All our numerical experiments have been performed in MATLAB using the Rational Krylov Toolbox (RKToolbox) [8], which has been extended for this paper; for details see section 5. Example files for reproducing our results are available online at `http://guettel.com/rktoolbox/examples/html/example_ehcompress.html`.

2. Analytic forms of DtN maps. There is a beautiful connection between FD grids and rational functions. In order to illustrate this connection we will first consider a scalar constant-coefficient FD grid and show how to convert it into an equivalent continued fraction. We will then discuss the variable-coefficient case and finally introduce the problem of grid optimization.

2.1. The constant-coefficient case. Consider the ODE $u''(x) = \lambda u(x)$ on $x \geq 0$ and its associated FD scheme

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

where λ and u_0 are given constants, and we demand that u_n remains bounded as $n \to \infty$. This linear recurrence relation is a scalar version of (1.1b) with $c \equiv 0$. It can easily be solved by computing the roots of the associated characteristic polynomial $p(t) = (t^2 - (2 + h^2\lambda)t + 1)/h^2$ and choosing the solution

$$u_j = \left(1 + \frac{h^2\lambda}{2} - h\sqrt{\lambda + \frac{h^2\lambda^2}{4}}\right)^j u_0.$$

Indeed this is the only solution that decays for $\lambda > 0$. Moreover, this solution is bounded under the condition[1] $\lambda \geq -4/h^2$ and unbounded for $\lambda < -4/h^2$.

[1] 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 points per wavelength, $n = \ell/h$, must satisfy $n = \ell/h = 2\pi/(kh) \geq \pi$ in order to approximate a bounded oscillatory solution.

We can use the explicit solution $\{u_j\}$ to extract interesting information about the problem. For example, from the FD relation

$$\frac{2}{h}\left(\frac{u_1-u_0}{h}+b\right)=\lambda u_0 \tag{2.2}$$

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

$$b=\sqrt{\lambda+\frac{h^2\lambda^2}{4}}u_0 =: f_h(\lambda)u_0.$$

This is the *Dirichlet-to-Neumann (DtN) relation*, with f_h being referred to as the *DtN 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 $u(x)=\exp(-x\sqrt{\lambda})u_0$.

While closed formulas of DtN maps are certainly useful for the theoretical analysis of solutions, they are not suitable for practical implementation in an FD framework. Hence our aim is to approximate DtN maps by FD grids with a small (and finite) number n of grid points. One approach to obtain a finite FD scheme is to simply truncate (2.1) after its first $n-1$ terms, setting $u_n=0$. Together with (2.2) and the auxiliary variables $\widehat{u}_{j-1}=(u_j-u_{j-1})/h$ we can then form a linear system

$$\begin{bmatrix}
\frac{h\lambda}{2} & -1 & & & & & \\
1 & h & -1 & & & & \\
 & 1 & h\lambda & -1 & & & \\
 & & 1 & h & \ddots & & \\
 & & & \ddots & \ddots & -1 & \\
 & & & & 1 & h\lambda & -1 \\
 & & & & & 1 & h
\end{bmatrix}
\begin{bmatrix}
u_0 \\ \widehat{u}_0 \\ u_1 \\ \widehat{u}_1 \\ \vdots \\ u_{n-1} \\ \widehat{u}_{n-1}
\end{bmatrix}
=
\begin{bmatrix}
b \\ 0 \\ 0 \\ 0 \\ \vdots \\ 0 \\ 0
\end{bmatrix}.$$

By row-wise Gaussian elimination of the -1's on the superdiagonal, starting from the bottom-right and going up to the left, we find that

$$b/u_0=\frac{h\lambda}{2}+\cfrac{1}{h+\cfrac{1}{h\lambda+\cfrac{1}{h+\cdots+\cfrac{1}{h\lambda+\cfrac{1}{h}}}}} =: r_n(\lambda).$$

The rational function r_n is of type $(n,n-1)$, that is, numerator and denominator degree n and $n-1$, respectively, and we expect that $r_n\approx f_h$ in some sense. Indeed, one can verify that r_n is the type $(n,n-1)$ Padé approximant to f_h with expansion point $\lambda=\infty$. Hence r_n can be expected to be a good approximation to f_h for large values of λ, i.e., for rapidly decaying solutions of (2.1).

2.2. Variable-coefficient case. The scalar form of the variable-coefficient FD scheme (1.1) is

$$\begin{aligned}\frac{2}{h}\left(\frac{u_1-u_0}{h}+b\right) &= (\lambda+c_0)u_0,\\ \frac{1}{h}\left(\frac{u_{j+1}-u_j}{h}-\frac{u_j-u_{j-1}}{h}\right) &= (\lambda+c_j)u_j, \quad j=1,2,\dots.\end{aligned}$$

By eliminating the grid points with indices $j > L$ (at which we assumed $c_j = 0$) in the same manner as above, we find the DtN relation $b/u_0 = f_h(\lambda)$ with

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

2.3. Optimizing FD grids. 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 $A = X\,\mathrm{diag}(\lambda_1,\lambda_2,\dots,\lambda_N)X^{-1}$, 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 directly relate the accuracy of $r_n(A)$ as an approximation to the DtN map $f_h(A)$ to a scalar approximation problem on the eigenvalues λ_j.

The crucial observation for optimizing the rational approximant r_n, or equivalently an FD grid, is that the grid steps do not need to be equispaced, and not even real-valued. Consider the FD scheme

$$\frac{1}{\widehat{h}_0}\left(\frac{u_1-u_0}{h_1}+b\right) = \lambda u_0, \tag{2.4a}$$

$$\frac{1}{\widehat{h}_j}\left(\frac{u_{j+1}-u_j}{h_{j+1}}-\frac{u_j-u_{j-1}}{h_j}\right) = \lambda u_j, \quad j=1,\dots,n-1, \tag{2.4b}$$

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 above, is

$$r_n(\lambda) = \widehat{h}_0\lambda + \cfrac{1}{h_1+\cfrac{1}{\widehat{h}_1\lambda+\cfrac{1}{h_2+\cdots+\cfrac{1}{\widehat{h}_{n-1}\lambda+\cfrac{1}{h_n}}}}}. \tag{2.5}$$

As before, r_n is a rational function of type $(n, n-1)$, and by choosing the free grid steps we can optimize it for our purposes. In particular, we can tune (2.4) 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 way to convert them into continued fraction form numerically. This will be subject of the following two sections.

3. From FD grids to rational Krylov spaces. Let us turn to the vector form of (2.4), which is

$$\frac{1}{\widehat{h}_0}\left(\frac{\boldsymbol{u}_1-\boldsymbol{u}_0}{h_1}+\boldsymbol{b}\right)=A\boldsymbol{u}_0, \tag{3.1a}$$

$$\frac{1}{\widehat{h}_j}\left(\frac{\boldsymbol{u}_{j+1}-\boldsymbol{u}_j}{h_{j+1}}-\frac{\boldsymbol{u}_j-\boldsymbol{u}_{j-1}}{h_j}\right)=A\boldsymbol{u}_j,\quad j=1,\dots,n-1. \tag{3.1b}$$

In the previous section we have derived that $\boldsymbol{b}=r_n(A)\boldsymbol{u}_0$ with a rational function $r_n=p_n/q_{n-1}$ whose continued fraction form (2.5) 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*

$$AU_{n+1}\underline{\widetilde{K}_n}=U_{n+1}\underline{\widetilde{H}_n}, \tag{3.2}$$

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 given as

$$\underline{\widetilde{K}_n}=\begin{bmatrix}0&&&\\ \widehat{h}_0&&&\\ &\widehat{h}_1&&\\ &&\ddots&\\ &&&\widehat{h}_{n-1}\end{bmatrix},\quad \underline{\widetilde{H}_n}=\begin{bmatrix}1&&&&\\ -h_1^{-1}&h_1^{-1}&&&\\ h_1^{-1}&-h_1^{-1}-h_2^{-1}&\ddots&&\\ &\ddots&\ddots&h_{n-1}^{-1}&\\ &&h_{n-1}^{-1}&-h_{n-1}^{-1}-h_n^{-1}\end{bmatrix}. \tag{3.3}$$

The entries in $(\underline{\widetilde{H}_n},\underline{\widetilde{K}_n})$ encode the recursion coefficients in (2.4), and the columns of U_{n+1} all correspond to rational functions in A multiplied by the vector $\boldsymbol{u}_0$. The span of such vectors is called a *rational Krylov space* [25]. In the next section we will show how to generate decompositions of the form (3.2) 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 operator for a BVP with possibly varying coefficients, and the sought rational function r_n is of type $(n,n-1)$, i.e., $r_n=p_n/q_{n-1}$ with $p_n\in\mathcal{P}_n$ and $q_{n-1}\in\mathcal{P}_{n-1}$. Note that (4.1) is a nonconvex optimization problem which 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 developed in [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* $\mathcal{V}_n$ defined as

$$\mathcal{V}_n := q_{n-1}(A)^{-1}\mathcal{K}_n(A, \boldsymbol{v}),$$

and an $(n+1)$-dimensional *target space* $\mathcal{W}_{n+1}$ defined as

$$\mathcal{W}_{n+1} := q_{n-1}(A)^{-1}\mathcal{K}_{n+1}(A, \boldsymbol{v}).$$

Here, $\mathcal{K}_j(A, \boldsymbol{v}) = \operatorname{span}\{\boldsymbol{v}, A\boldsymbol{v}, \ldots, A^{j-1}\boldsymbol{v}\}$ is the standard (polynomial) Krylov space for $(A, \boldsymbol{v})$ of dimension j. Let $V_n \in \mathbb{C}^{N\times n}$ and $W_{n+1} \in \mathbb{C}^{N\times(n+1)}$ be orthonormal bases for $\mathcal{V}_n$ and $\mathcal{W}_{n+1}$, respectively.

The space $\mathcal{V}_n$ 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. Here, $(\underline{H_{n-1}}, \underline{K_{n-1}})$ is an unreduced upper Hessenberg pencil 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. Decompositions of the form (4.2) can be computed by Ruhe's rational Krylov sequence (RKS) algorithm [25]. The following result, established in [7, Thm. 2.5], relates the generalized eigenvalues of the lower $(n-1)\times(n-1)$ subpencil 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 $(n-1)\times(n-1)$ subpencil 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 $\mathcal{V}_n$ 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 pencil $(\underline{\widehat{H}_{n-1}}, \underline{\widehat{K}_{n-1}})$ be given. Assume further that none of generalized eigenvalues $\widehat{\xi}_j$ of the lower $(n-1)\times(n-1)$ subpencil 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 $\mathcal{V}_n$ and $\mathcal{W}_{n+1}$ have been computed. Then we can identify a vector $\widehat{\boldsymbol{v}} \in \mathcal{V}_n$, $\|\widehat{\boldsymbol{v}}\|_2 = 1$, such that $F\widehat{\boldsymbol{v}}$ is best approximated by some vector in $\mathcal{W}_{n+1}$. 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 $\mathcal{W}_{n+1}$ and its representation in the rational Krylov basis W_{n+1} is

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

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

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

is satisfied, where $(\underline{H_n}, \underline{K_n})$ is an unreduced upper Hessenberg pencil of size $(n+1)\times n$. Indeed, we have $\mathcal{V}_n \subset \mathcal{W}_{n+1}$ and $\mathcal{W}_{n+1}$ 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 (3.2) when $\boldsymbol{u}_0 = \boldsymbol{v}$. This transformation should not alter the space $\mathcal{W}_{n+1}$ but merely transform the basis W_{n+1} into the continued fraction basis U_{n+1} and the pencil $(\underline{H_n}, \underline{K_n})$ into the tridiagonal-and-diagonal form of (3.3).

As a first step 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, 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}:

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

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 pencil. Here is a pictorial view of decomposition (4.5) for the case $n = 4$:

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

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 pencils $(\underline{H_n^{(j)}}, \underline{K_n^{(j)}})$ $(j = 1, 2, \ldots, 5)$ all corresponding to the same rational Krylov space $\mathcal{W}_{n+1}$ and all without the two leading vectors in $W_{n+1}^{(0)}$ being altered. More precisely, the transformations we are allowed to perform are:

- right-multiplication of the pencil by any invertible matrix $R \in \mathbb{C}^{n\times n}$,
- left-multiplication of the pencil 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 pencil $(\underline{H_n^{(0)}}, \underline{K_n^{(0)}})$ 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 pencils):

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

   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 rational Krylov 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 pencil to zero the first row of $\underline{K_n^{(1)}}$ completely. This can be done by adding multiples of the 3rd, 4th, $\ldots$, $(n+1)$th row to the first. As a result we obtain

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

   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 pencil 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, $\ldots$, nth column. As a result we have

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

   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,$\ldots$, $(n+1)$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:

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

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 blkdiag$(1, Z_L^H)$ and right-multiplication with Z_R results in the demanded structure:

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

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 (3.3). This can be achieved by left multiplication of the pencil 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 (3.3) and $L\underline{H_n^{(5)}}R$, starting from the $(1,1)$ entry and going down columnwise. We obtain

$$r_1 = \frac{1}{\eta_{1,1}}, \qquad h_1 = \frac{-1}{\eta_{2,1}\rho_1}, \qquad \ell_3 = \frac{1}{\eta_{3,1}h_1\rho_1},$$

and for $j = 2, 3, \ldots$

$$r_j = \frac{1}{\ell_j \eta_{j,j} h_{j-1}}, \quad h_j = \frac{-1}{1/h_{j-1} + \ell_{j+1}\eta_{j+1,j}\rho_j}, \quad \ell_{j+2} = \frac{1}{\eta_{j+2,j}h_j\rho_j}.$$

The diagonal entries of $\underline{\widetilde{K}_n}$ in (3.3) satisfy

$$\widehat{h}_{j-1} = \ell_{j+1}\rho_j, \quad j = 1, \ldots, n,$$

and thus the pencil has been transformed exactly into the form (3.3).

The above six-step procedure allows us to convert the rational function r_n computed via the RKFIT iteration into continued fraction form, and hence reinterpret it 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 is only interested in the continued fraction parameters.

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* $\mathcal{V}_n \subset \mathcal{W}_{n+1}$, *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 (2.5) *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* [26] *are reviewed in* [23]*; see Theorem 1.39 therein.) Even if our conjecture were 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 numerical experiments.*

5. Computational aspects.

5.1. Pseudocode and implementation. The pseudocode for a single RKFIT iteration is given in Algorithm 5.1. A MATLAB implementation is contained in the RKToolbox which is available online at `http://rktoolbox.org`. The provided `rkfit` method is very easy to use. For example, the following three lines of MATLAB code will compute an RKFIT approximant $r_n(A)\boldsymbol{v} \approx \sqrt{A}\boldsymbol{v}$ for $A = \text{tridiag}(-1, 2, -1)$ and a random vector $\boldsymbol{v}$ of size 100:

```
A = gallery('tridiag', 100); F = @(V) sqrtm(full(A))*V;
v = randn(100, 1); xi = inf(1, 9); param.k = +1;
[misfit, ratfun] = rkfit(F, A, v, xi, param);
```

Note that `F` is a function handle for computing the action of $\sqrt{A}$ onto a block of vectors, and typically this can be done more efficiently than using dense matrix algorithms like `sqrtm`. In particular, rational Krylov techniques themselves can be used for this purpose (see, e.g., [22, 21]). The degree $(10, 9)$ of r_n is specified by 9 initial poles at infinity (the variable `xi`) and the numerator degree offset parameter `k = +1` given to RKFIT. In all our numerical experiments we choose all the initial poles to be at ∞. When used with its default options, `rkfit` performs 10 iterations and then returns the solution r_n (represented by the output `ratfun`) with the smallest relative misfit

$$\texttt{misfit} = \frac{\|F\boldsymbol{v} - r_n(A)\boldsymbol{v}\|_2}{\|F\boldsymbol{v}\|_2}.$$

For the numerical experiments in the following sections we report the number of RKFIT iterations required to achieve a relative misfit below 1.01 times the overall minimum achieved in (at most) 30 iterations. This is to avoid artificially high iteration numbers being reported in the case that RKFIT stagnates on its final accuracy level (where tiny `misfit` fluctuations may occur due to floating point arithmetic).

For the purpose of this paper we have extended the RKToolbox by the `contfrac` method, which allows for the conversion of an RKFUN, the fundamental data type to represent and work with rational functions r_n, into continued fraction form. The implementation follows exactly the procedure given in section 4.3. Numerically, these transformations may be ill conditioned and the use of multiple precision arithmetic is recommended. The RKToolbox supports both MATLAB's Variable Precision Arithmetic (`vpa`) and, preferably, the Advanpix Multiprecision Toolbox (`mp`) [1]. To compute the continued fraction coefficients h_j and $\widehat{h}_{j-1}$ in (2.5) for the above `ratfun` object one simply types `[h, hath] = contfrac(mp(ratfun))`.

5.2. Training and testing vectors. In section 2 we established that many DtN maps can be written in the form $\boldsymbol{b} = f_h(A)\boldsymbol{u}_0$, where $\boldsymbol{u}_0$ is the Dirichlet data at the interface $x = 0$ for (1.1b), and $\boldsymbol{b}$ is the corresponding Neumann data. In section 4 we have replaced $\boldsymbol{u}_0$ by a general vector $\boldsymbol{v}$ because in many applications the Dirichlet data $\boldsymbol{u}_0$ may actually be unknown. For example, if the DtN approximation

Algorithm 5.1 One RKFIT iteration for superdiagonal approximants.

Input: 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 typically all chosen at ∞).

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

1. Compute 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 $\mathcal{R}(V_n)$ 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 incorporated into an existing FD scheme to mimic a perfectly matched layer, then $\boldsymbol{u}_0$ appears only implicitly as an unknown variable in the FD grid. One may not know a priori which spectral components will be present in $\boldsymbol{u}_0$, or in terms of the indefinite Helmholtz problem mentioned in the introduction, it may not be clear a priori which wave modes will arrive at the $x = 0$ interface.

In order to still apply RKFIT in such situations and to compute a DtN approximation $r_n(A) \approx f_h(A)$ independent of $\boldsymbol{u}_0$, we will compute the RKFIT approximant for a *training vector* $\boldsymbol{v}$ and then assume that it is accurate for all *testing vectors* $\boldsymbol{u}_0$. In all our experiments we choose both vectors at random, so that almost surely all spectral components of A enter as weights into the computation of r_n. We observed in the numerical experiments that if $r_n(A)\boldsymbol{v} \approx f_h(A)\boldsymbol{v}$ is a good approximation, then typically also $r_n(A)\boldsymbol{u}_0 \approx f_h(A)\boldsymbol{u}_0$ is a good approximation.

5.3. Surrogate approximation. In some cases $A \in \mathbb{C}^{N\times N}$ may be too large to compute $F = f(A)$ or FV_n directly. In this case we perform the RKFIT computation with a surrogate matrix $\widetilde{A} \in \mathbb{C}^{\widetilde{N}\times\widetilde{N}}$ and a surrogate training vector $\widetilde{\boldsymbol{v}} \in \mathbb{C}^{\widetilde{N}}$, $\widetilde{N} \ll N$, instead of $(A, \boldsymbol{v})$. Similar approaches have been applied successfully in [10, 9]. In the case where A is Hermitian, for example, the surrogate may be chosen as a diagonal matrix with sufficiently dense eigenvalues in the spectral interval of A (which requires eigenvalue estimates) and the vector $\widetilde{\boldsymbol{v}} = [1, 1, \ldots, 1]^T$. The main operations in the RKFIT algorithm (like matrix-vector products and linear system solves) involving a diagonal matrix become trivial $O(\widetilde{N})$ operations. Moreover, the application of $f_h(\widetilde{A})$ reduces to $\widetilde{N}$ scalar evaluations of the DtN function. In some of our numerical experiments we will use this surrogate approach to test the performance of RKFIT when being applied with a matrix A that has "essentially dense" spectrum, thereby mimicking uniform approximation over its spectral interval.

6. Convergence comparisons for the 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. Concrete examples of nonlinear rational least squares problems with no or highly sensitive solutions are given in the introduction of [9]. 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. The aim of this section is to provide such a comparison to the two-interval Zolotarev approach in [16], and the one-interval approximants studied by Newman and Vjacheslavov [24, 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 [16, 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.

6.1. Two-interval approximation with coarse spectrum. Our first example focuses on 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 6.1. *In Figure 6.1 (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*

$$L = \begin{bmatrix} 1 & -1 & & & \\ -1 & 2 & -1 & & \\ & \ddots & \ddots & \ddots & \\ & & -1 & 2 & -1 \\ & & & -1 & 1 \end{bmatrix}. \tag{6.1}$$

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*

$$[a_1, b_1] \approx [-225, -67.2] \quad \textit{and} \quad [a_2, b_2] \approx [21.5, 8.98 \cdot 10^4].$$

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 to satisfy the stagnation criterion described in section 5.1. 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.*

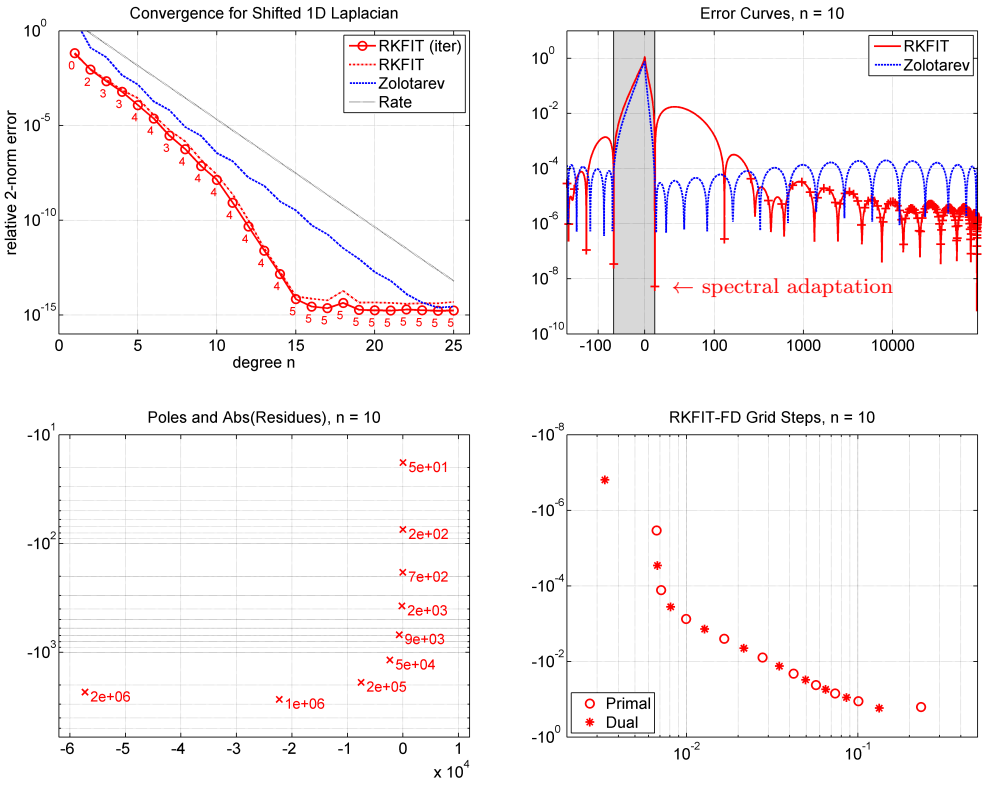

FIGURE 6.1. *Top: Accuracy comparison of RKFIT and Zolotarev approximants for a shifted 1D Laplacian which has a rather coarse spectrum, hence resulting in superlinear RKFIT convergence. The DtN function is* $f_h(\lambda) = \sqrt{\lambda + (h\lambda/2)^2}$. *The small numbers on the solid red convergence 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).*

The spectral adaptation is illustrated in the graph on the top right of Figure 6.1, 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 Figure 6.1 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 steps. All the involved quantities have been calculated using the RKFUN calculus of the RKToolbox as described in section 5.1.*

6.2. 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.

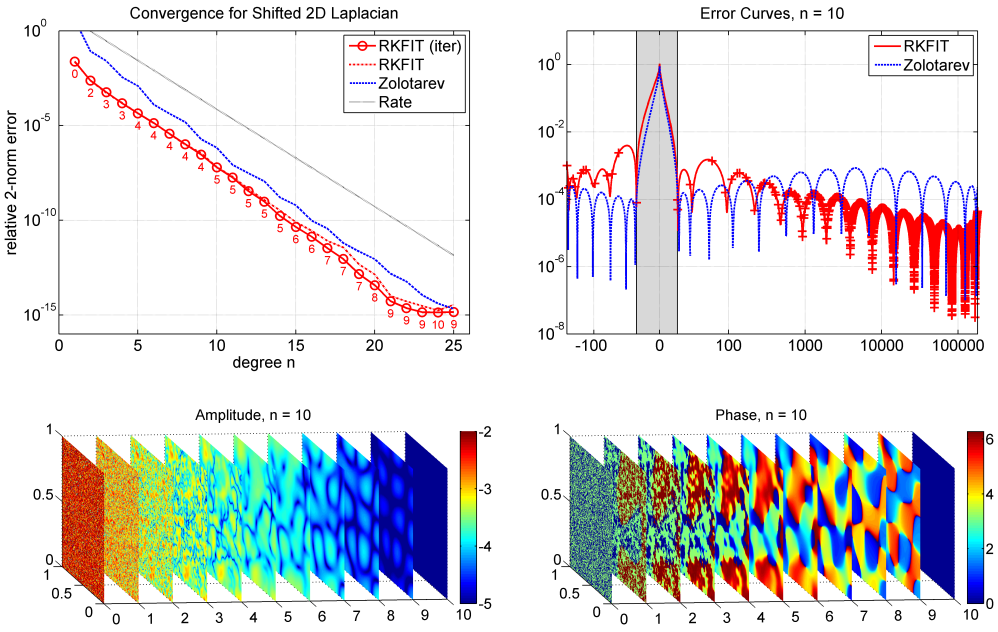

FIGURE 6.2. *Top: Comparison of RKFIT and Zolotarev approximants for a shifted 2D Laplacian. Bottom: The* $\log_{10}$ *of the amplitude and phase of the grid vectors* $\boldsymbol{u}_j$ $(j = 0, 1, \ldots, n = 10)$. *Qualitatively, the poles and residues and the complex grid steps for the associated RKFIT approximant* r_{10} *look similar to those in Figure 6.1 and are therefore omitted.*

EXAMPLE 6.2. *In Figure 6.2 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 varying degrees* $n = 1, 2, \ldots, 25$. *Now the matrix* A *corresponds to a shifted 2D Laplacian* $A = (L \otimes L)/h^2 - k_\infty^2 I \in \mathbb{R}^{N\times N}$, *where* $N = 150^2$, $h = 1/150$, $k_\infty = 15$, *and with* L *defined in* (6.1). *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*

$$[a_1, b_1] \approx [-225, -27.7] \quad \textit{and} \quad [a_2, b_2] \approx [21.5, 1.80 \cdot 10^5].$$

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 Figure 6.2 we show the grid vectors $\boldsymbol{u}_j$ *satisfying the FD relation* (3.1) *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 random signal travels further to the right in the grid, illustrating the good absorption property of this grid for outgoing waves.*

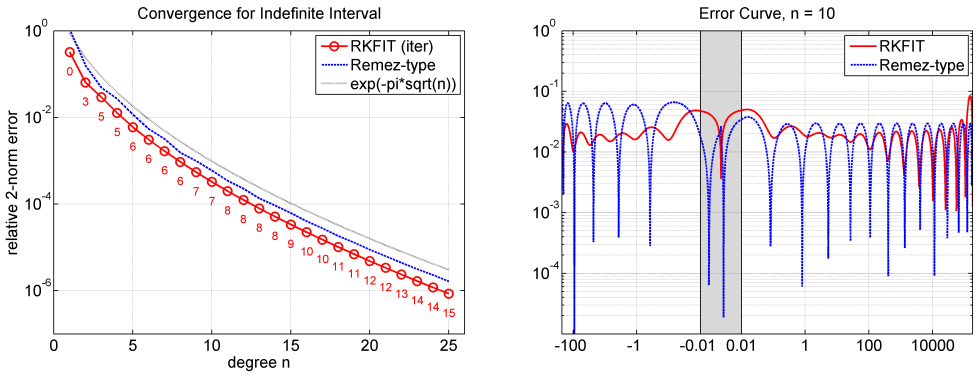

FIGURE 6.3. *RKFIT approximation of* $f(\lambda) = \sqrt{\lambda}$ *on an indefinite interval* $[a_1, b_2]$, $a_1 < 0 < b_2$, *compared to a two-interval Remez-type approximant. Qualitatively, the poles/residues and the complex grid steps associated with* r_{10} *look similar to those in Figure 6.1 and are therefore omitted.*

6.3. Approximation on an indefinite interval. In order to remove superlinear convergence effects and the spectral gap $[b_2, a_1]$ from which the previous two examples benefited, we now consider the approximation on an indefinite interval. The following test uses a diagonal matrix with sufficiently dense eigenvalues and hence is an example for the surrogate strategy described in section 5.3.

EXAMPLE 6.3. *We approximate* $f(\lambda) = \sqrt{\lambda}$ *on the indefinite interval*

$$[a_1, b_2] = [-225, 1.80 \cdot 10^5].$$

Note that $[a_1, b_2]$ *is the same as in the previous Example 6.2, 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 Figure 6.3 (left).*

We also compare RKFIT to a two-interval Remez-type approximant which is obtained by using the interpolation nodes of numerically computed best approximants to $\sqrt{\lambda}$ *on* $[0, 1]$*, 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 6.1. *The uniform rational approximation of* $\sqrt{\lambda}$ *on a semi-definite interval* $[0, b_2]$ *has been studied by Newman and Vjacheslavov. In particular, it is known that the error of the best rational approximant reduces like* $\exp(-\pi\sqrt{2n})$ *with the degree* n*; see [24, Section 4]. Based on the observations in Figure 6.3 we conjecture that the error of the best rational approximant to* $\sqrt{\lambda}$ *on an indefinite interval* $[a_1, b_2]$ *reduces like* $\exp(-\pi\sqrt{n})$*.*

7. Variable-coefficient case. We now consider the case of a variable coefficient function c. Here analytic results are not available and numerical approximation ap-

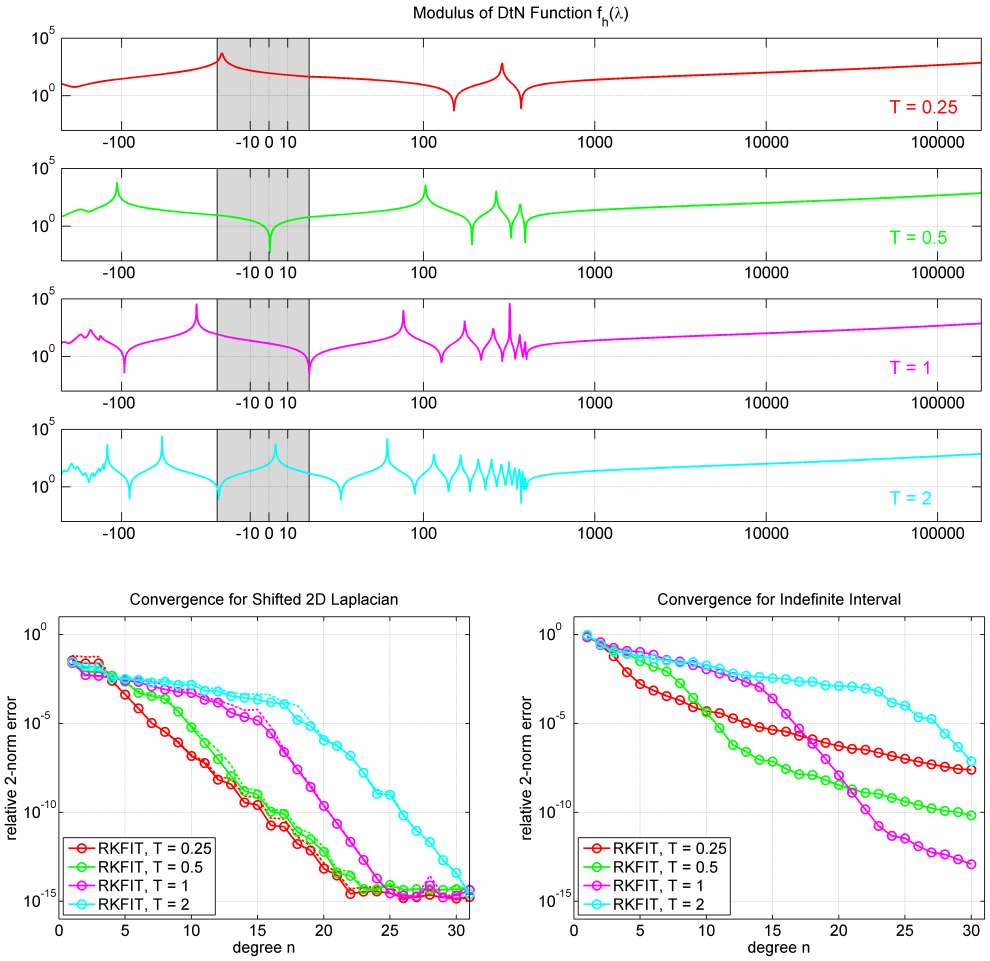

FIGURE 7.1. *Top: The four panels show the modulus of the discrete variable-coefficient DtN function f_h for varying thickness T of the two finite layers. Bottom: The two plots show the RKFIT convergence for approximating $f_h(A)\boldsymbol{v}$ when A is a shifted 2D Laplacian (left) and a diagonal matrix with dense eigenvalues in the same spectral interval (right), respectively.*

pears to be the only option. In this section we mainly present the numerical results, which are then discussed in more detail in the following section 8.

EXAMPLE 7.1. *We consider the seismic exploration setup in Figure 1.1. At the $x = 0$ interface of the computational domain we assume to have a 2D Laplacian $A = (L \otimes L)/h^2 - k_\infty^2 I$ with L defined in (6.1), 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 given by*

$$c(x) = \begin{cases} -400 & \text{if } 0 \le x < T, \\ +125 & \text{if } T \le x < 2T, \\ 0 & \text{if } 2T \le x < \infty, \end{cases}$$

where the thickness of the two finite layers T is varied over $\{0.25, 0.5, 1, 2\}$. For each thickness T, the four panels in the top of Figure 7.1 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 in the plot.*

The convergence of the RKFIT approximants for varying degree parameter n *is shown in Figure 7.1 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*. While we cannot report on all these tests here, we highlight again that the codes for producing our examples are available online and can easily be modified to other coefficient functions.*

EXAMPLE 7.2. *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 Figure 7.1. Again the RKFIT behavior is very robust even for high approximation degrees* n*, but compared to the previous example the convergence is delayed, indicating that spectral adaptation effects have led to the faster convergence for the shifted 2D Laplacian.*

8. Discussion and conclusions. An obvious alternative to our grid compression approach in the two examples of section 7 would be to use an efficient discretization method on c's support, and then to append it to the constant-coefficient PML of [16]. In principle such an approach requires at least the integer part of

$$\mathrm{N} = \int_0^H \frac{\sqrt{k_\infty^2 - c(x)}}{\pi}\,\mathrm{d}x$$

discretization points according to the Nyquist sampling rate, where H is the total thickness of c's support. In fact, the classical spectral element method (SEM) with polynomial local basis requires at least $\frac{\pi}{2}\mathrm{N}$ grid points [3]. (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 section 7, 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:

	$T = 0.25$	$T = 0.5$	$T = 1$	$T = 2$
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 7.1)	8	10	16	19
RKFIT-FD (Example 7.2)	14	11	17	28

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 being slowed down in convergence by singularities of the DtN function well separated from the eigenvalues of A. In the appendix we give some insight into this phenomenon.

REFERENCES

[1] Advanpix LLC., *Multiprecision Computing Toolbox for MATLAB*, ver 3.8.3.8882, Tokyo, Japan, 2015. `http://www.advanpix.com/`.

[2] N. I. Akhiezer, *Theory of Approximation*, Dover, 1992.

[3] M. Ainsworth and H. A. Wajid, *Dispersive and dissipative behavior of the spectral element method*, SIAM J. Numer. Anal., 47 (2009), pp. 3910–3937.

[4] D. Appelö, T. Hagstrom, and G. Kreiss, *Perfectly matched layers for hyperbolic systems: General formulation, well-posedness, and stability*, SIAM J. Appl. Math., 67 (2006), pp. 1–23.

[5] S. Asvadurov, V. Druskin, M. Guddati, and L. Knizhnerman, *On optimal finite-difference approximation of PML*, SIAM J. Numer. Anal., 41 (2003), pp. 287–305.

[6] J. P. Berenger, *A perfectly matched layer for the absorption of electromagnetic waves*, J. Comp. Phys., 114 (1994), pp. 185–200.

[7] M. Berljafa and S. Güttel, *Generalized rational Krylov decompositions with an application to rational approximation*, SIAM J. Matrix Anal. Appl., 36 (2015), pp. 894–916.

[8] M. Berljafa and S. Güttel, *A Rational Krylov Toolbox for MATLAB*, The University of Manchester, MIMS Eprint 2014.56, 2014.

[9] M. Berljafa and S. Güttel, *The RKFIT algorithm for nonlinear rational approximation*, The University of Manchester, MIMS Eprint 2015.38, 2015.

[10] R.-U. Börner, O. G. Ernst, and S. Güttel, *Three-dimensional transient electromagnetic modelling using rational Krylov methods*, Geophys. J. Int., 202 (2015), pp. 2025–2043.

[11] L. Borcea, V. Druskin, F. G. Vasquez, and A. Mamonov, *Resistor network approaches to electrical impedance tomography*, in Inverse Problems and Applications: Inside Out II. Vol. 60. Cambridge University Press, 2013, pp. 55–119.

[12] L. Borcea, F. G. Vasquez, and A. Mamonov, *A discrete Liouville identity for numerical reconstruction of Schrödinger potentials*, arXiv preprint, arXiv:1601.07603, 2016.

[13] Y. Boubendir, X. Antoine, and C. Geuzaine, *A quasi-optimal non-overlapping domain decomposition algorithm for the Helmholtz equation*, J. Comp. Phys., 231 (2012), pp. 262–280.

[14] W. C. Chew, *Waves and Fields in Inhomogeneous Media,* IEEE Press New York, 1995.

[15] W. Chew and B. Weedon, *A 3D perfectly matched medium from modified Maxwell's equations with stretched coordinates*, Microwave Opt. Technol. Lett., 7 (1994), pp. 599–604.

[16] V. Druskin, S. Güttel, and L. Knizhnerman, *Near-optimal perfectly matched layers for indefinite Helmholtz problems*, SIAM Rev., 58:1 (2016), pp. 90–116.

[17] V. Druskin and L. Knizhnerman, *Gaussian spectral rules for the three-point second differences: I. A two-point positive definite problem in a semiinfinite domain*, SIAM J. Numer. Anal., 37 (1999), pp. 403–422.

[18] B. Engquist and A. Majda, *Radiation boundary conditions for acoustic and elastic wave calculations*, Comm. Pure Appl. Math., 32 (1979), pp. 313–357.

[19] B. Engquist and L. Ying, *Sweeping preconditioner for the Helmholtz equation: hierarchical matrix representation*, Comm. Pure Appl. Math., 64:5 (2011), pp. 697–735.

[20] M J. Gander, *Optimized Schwarz methods*, SIAM J. Numer. Anal., 44:2 (2006), pp. 699–731.

[21] S. Güttel, *Rational Krylov approximation of matrix functions: Numerical methods and optimal pole selection*, GAMM-Mitt., 36:1 (2013), pp. 8–31.

[22] S. Güttel and L. Knizhnerman, *A black-box rational Arnoldi variant for Cauchy–Stieltjes matrix functions*, BIT Numer. Math., 53:3 (2013), pp. 595–616.

[23] O. Holtz and M. Tyaglov, *Structured matrices, continued fractions, and root localization of polynomials*, SIAM Rev., 54:3 (2012), pp. 421–509.

[24] P. P. Petrushev and V. A. Popov, *Rational Approximation of Real Functions*, Cambridge Univ. Press, Cambridge, 1987.

[25] A. Ruhe, *Rational Krylov: A practical algorithm for large sparse nonsymmetric matrix pencils*, SIAM J. Sci. Comput., 19:5 (1998), pp. 1535–1551.

[26] T. J. Stieltjes, *Recherches sur les fractions continues*, Annales de la Faculté des Sciences de Toulouse, 8 (1984), pp. 1–122, 9 (1894), pp. 1–47.

[27] K. S. Yee, *Numerical solution of initial boundary value problems involving Maxwell's equations in isotropic media.* IEEE Trans. Antennas Propag., 14 (1966), pp. 302–307.

[28] Y. I. Zolotarev, *Collection of Works*, Saint Petersburg Academy of Sciences, 30:5, 1877.

Appendix A. A Nyquist limit-type criterion for rational approximation.

The plots in Figure 7.1 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. In order to obtain a better understanding of this behavior, we consider a two-layer waveguide problem with piecewise constant wave numbers similar to the one in Figure 1.2, but now in the continuous setting without any FD approximation. This problem is governed by the equations

$$\begin{aligned} u''(x) &= (\lambda + c)u(x), \quad && x \in [0, T), \\ u''(x) &= \lambda u(x), && x \in [T, \infty), \end{aligned}$$

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

$$\begin{aligned} u(x) &= \alpha e^{x\sqrt{\lambda+c}} + (u_0 - \alpha)e^{-x\sqrt{\lambda+c}} \\ &= 2\alpha \sinh\left(x\sqrt{\lambda + c}\right) + e^{-x\sqrt{\lambda+c}}u_0. \end{aligned}$$

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

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

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 Figure A.1. It appears that for $c \geq 0$, this function is smooth over the whole real axis, while it develops singularities for $c < 0$. The following lemma shows that the number of real singularities is proportional to c and T.

LEMMA A.1. *For $c \geq 0$, the function f defined in* (A.1) *has no singularities on the real axis. For $c < 0$, f has $\left\lfloor \frac{T\sqrt{-c}}{\pi} \right\rfloor + q$ real poles, where $q \in \{0, 1\}$, all located in the interval $[0, -c]$.*

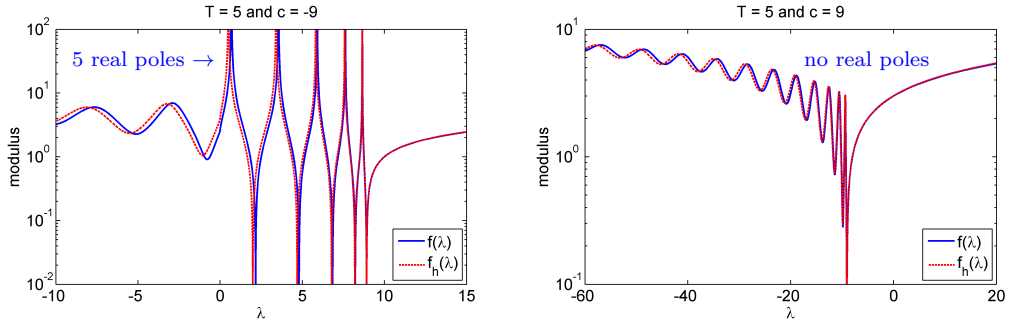

FIGURE A.1. *The DtN function f defined in* (A.1)*, as well as its discrete counterpart* (2.3)*, for two different choices of the parameters (T, c).*

Proof. We investigate the roots of the denominator function $g(\lambda) = \sqrt{\lambda + c} \cdot \cosh(T\sqrt{\lambda + c}) + \sqrt{\lambda} \cdot \sinh(T\sqrt{\lambda + c})$. 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 $\cosh(T\sqrt{\lambda + c}) = \cos(T\mathrm{imag}(\sqrt{\lambda + c}))$ is purely real and $\sinh(T\sqrt{\lambda + c}) = i \sin(T\mathrm{imag}(\sqrt{\lambda + c}))$ purely imaginary. Hence, λ can only be a root of g if $\cos(z) = \sin(z) = 0$, $z = T\mathrm{imag}(\sqrt{\lambda + c})$, 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 $\lambda \in [0, -c]$, $z = T\mathrm{imag}(\sqrt{\lambda + c})$ varies in $[0, T\sqrt{-c}]$. As both of the factors $\sqrt{\lambda + c}$ and $\sqrt{\lambda}$ do not change sign for $\lambda \in [0, -c]$, we simply need to count the intersections of $\cos(z)$ and $\sin(z)$ on $[0, T\sqrt{-c}]$ in order to determine the roots of g. Note that $\cos(z)$ and $\sin(z)$ have exactly one intersection on $[(k-1)\pi, k\pi]$ for $k = 1, 2, \ldots, K$, hence there are $K = \lfloor T\sqrt{-c}/\pi \rfloor$ guaranteed intersections on $[0, K\pi]$. The final interval $[K\pi, T\sqrt{-c}]$ may or may not contain a further root.

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 Figure A.1 (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 section 7 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.