Schur complement, Padé approximation and domain truncation

Martin J. Gander, Lukáš Jakabčin and Michal Outrata

April 9, 2022

Abstract

We show for a model problem that the truncation of an unbounded domain by an artificial Dirichlet boundary condition placed far away from the domain of interest is equivalent to a specific absorbing boundary condition at the boundary of the domain of interest. We prove that the absorbing boundary condition obtained is a spectral Padé approximation about infinity of the transparent boundary condition. We also study numerically two improvements for this boundary condition – the truncation with an artificial Robin condition placed far away from the domain of interest, and a Padé approximation about different point than infinity. Both of these give new and substantially better results compared to the artificial Dirichlet boundary condition.

Seen through the optic of linear algebra, we show that the Schur complement of our model problem can be after diagonalization identified with a truncation of a certain continued fraction. We use the theory of continued fractions to interpret the Schur complement as the Padé approximation of the *infinite Schur complement*. Next, we improve the approximation qualities by changing some of the structure of the continued fraction so that the approximation is more accurate around a point of our choice and propose two different ways of achieving this.

1 Introduction

The solution process of problems on unbounded domains usually require a domain truncation and a hence new, artificial, boundary conditions, leading to techniques such as *perfectly matched layers* (PML) or *absorbing boundary conditions* (ABC), see [4, 3]. At the discrete level, these closely relate to the problem of approximating the Schur complement in some sense, which inspired number of iterative solvers, see, e.g., [10] and the references therein. Our approach builds upon the eigendecomposition of the Schur complement, which for our model problem is very closely linked with the Fourier analysis of the Schur complement or, equivalently, the frequency domain analysis.

Domain truncation is also important in domain decomposition where a given computational domain is decomposed into many smaller subdomains, and then subdomain solutions

are computed independently in parallel. The solutions on the smaller subdomains can naturally be interpreted as solutions on truncated domains, and thus it is of interest to use ABC or PML techniques at the interfaces between the subdomains, see [8, 9, 10]. The classical Schwarz method [20] uses Dirichlet transmission conditions between subdomains and an overlap to achieve convergence [21]. In what follows the goal is to interpret the overlap as a specific ABC once the unknowns of the overlap are folded onto the interface (similarly to [17, 13]). Although the Schwarz method is not explicitly mentioned in what follows, it is one of the main applications for our results.

Notably, the question of the *optimal PML* for problems with finite difference grids has been discussed in [14, 1] for the Laplace equation and then also extended to the Helmholtz equation in [6]. Our results go in a similar direction but are qualitatively different. In [14, 1], the focus is on understanding how the underlying mesh (e.g., using a staggered finite difference mesh in 2D) affects the quality of the PML and then optimizing the mesh so that the approximation error is small. In this paper we first focus on a similar yet distinct problem, where the mesh cannot be changed but it can be extended. Thus the appropriation of the grid for a given problem is taken from a different angle. We give an *algebraic* description of the approximation quality and in this way talk about the same topic as [14, 1] but from a different viewpoint. We make substantial efforts to do that in a way that is both reasonably self-contained and also reasonably easy to follow for readers from different mathematical community. This also includes introducing continued fractions terminology and their types and properties in some detail in the Appendix but also the Schur complement. We then continue by new numerical experiments with *different types of boundary conditions at the artificial boundary* and study their effect on the approximation quality.

We start in Section 2 with some notation and definitions and continue in Section 3 by showing that there exists a limit of the Schur complement as the width of the overlap goes to infinity, and that the Schur complement of a finite width truncation with a Dirichlet condition is a spectral Padé approximation around infinity of the unbounded width limit. Next, we explore numerically how the spectral approximation changes when the Dirichlet condition is replaced by a Robin condition in Section 4, present an optimized choice for the Robin parameter and propose a new type of boundary condition in Section 5. We end with concluding remarks and possible extensions in Section 6.

2 Model Problem

We use as our model problem the partial differential equation (PDE)

$$\begin{aligned} (\eta - \Delta)u &= f \quad \text{in } \Omega := (0, +\infty) \times (0, 1),\ \eta > 0, \\ u &= 0 \quad \text{on } \partial\Omega. \end{aligned} \tag{1}$$

We assume that the support of the right-hand side function f is *localized* in $\Omega^a := (0, a) \times (0, 1)$ and having $b \geq a$ we set $\Omega^b := (0, b) \times (0, 1) \subset \Omega$ as the artificially truncated region containing Ω_a, see Figure 1.

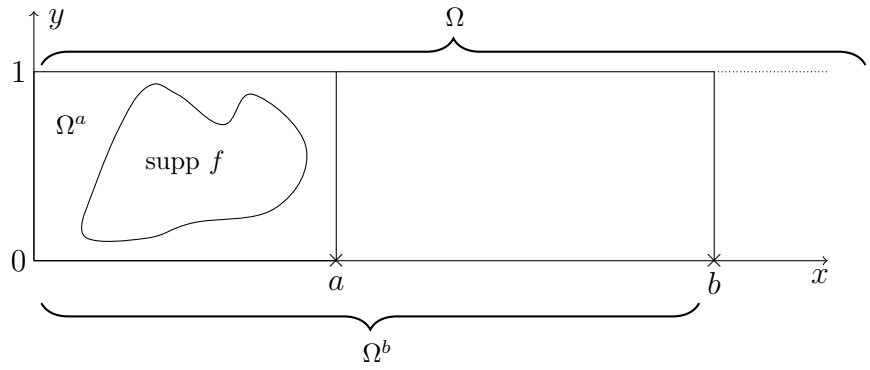

Figure 1: The unbounded strip domain in $\mathbb{R}^2$ with $\Omega = (0, +\infty) \times (0, 1)$.

Discretizing with the standard finite difference scheme, we denote by N the number of interior grid rows and obtain the mesh size $h := 1/(N+1)$. Assuming we have

$$a = (N^a + 1)h \qquad \text{and} \qquad b = (N^b + 1)h, \tag{2}$$

we obtain the discretized problems

$$A\mathbf{u} = f, \qquad A^b\mathbf{u}^b = \mathbf{f}^b, \tag{3}$$

with the right-hand side vectors $\mathbf{f}^a = [\mathbf{f}_1^T, \cdots, \mathbf{f}_{N^a}^T]^T, \mathbf{f}^b = [(\mathbf{f}^a)^T, \mathbf{0}^T, \cdots, \mathbf{0}^T]^T$ and $\mathbf{f} = [(\mathbf{f}^b)^T, \mathbf{0}^T, \cdots]^T$ and the matrices

$$A^\star = \frac{1}{h^2}\begin{pmatrix} D_1 & -I_N & & \\ -I_N & \ddots & \ddots & \\ & \ddots & D_{N^\star-1} & -I_N \\ & & -I_N & D_{N^\star} \end{pmatrix}, \quad A = \frac{1}{h^2}\begin{pmatrix} h^2A^b & & & \\ & -I_N & & \\ -I_N & D_{N^b+1} & \ddots & \\ & \ddots & \ddots & \end{pmatrix}, \tag{4}$$

where $\star$ stands for either a or b (and thus changes the number of block rows and block columns) and each block has dimension N (vectors) or $N \times N$ (matrices) and concerns only with a one particular grid column variables. The matrix I_N is the $N \times N$ identity and the diagonal blocks D_i are given by

$$D_i = D = \begin{pmatrix} \eta h^2 + 4 & -1 & \\ -1 & \ddots & -1 \\ & -1 & \eta h^2 + 4 \end{pmatrix} \in \mathbb{R}^{N\times N}. \tag{5}$$

Here, it is enough to understand the infinite-dimensional system in (3) as the limit of the finite-dimensional one as $b \to +\infty$; for more details on infinite matrices, see, e.g., the historical overview [5].

Thanks to the localization of f the solutions u, u^b (and $\mathbf{u}, \mathbf{u}^b$) restricted to Ω^a can be obtained using only the variables in Ω^a[1]. The continuous level formulation requires the

[1]This is of particular interest for the domain decomposition methods, see Section 1.

Dirichlet-to-Neumann operator (see ...) and its approximation on finite difference grids in this context has been studied in [14, 1]. We start immediately on the discrete level and in order to reduce the systems in (3) to a smaller one, we eliminate the variables $(\mathbf{u}^b_{N^a+1}, \dots, \mathbf{u}^b_{N^b})$. Recalling (3), these follow the equations

$$-\frac{\mathbf{u}^b_{i-1}}{h^2} + \frac{D_i \mathbf{u}^b_i}{h^2} - \frac{\mathbf{u}^b_{i+1}}{h^2} = 0, \qquad -\frac{\mathbf{u}^b_{N^b-1}}{h^2} + \frac{D_{N^b}\mathbf{u}^b_{N^b}}{h^2} = 0, \tag{6}$$

with $i > N^a$ where the second equation is considered only for $b < \infty$. We can recursively eliminate these, practically calculate the block Gaussian elimination and arrive at the Schur complement definition.

Definition 2.1 (Schur complement) *The Schur complement T^b_i for $i \in \{N^b, \dots, N^a\}$ is defined recursively by*

$$T^b_{N^b} := \frac{D_{N^b}}{h^2} = \frac{D}{h^2} \quad \text{and} \quad T^b_i := \frac{D_i}{h^2} - \frac{(T^b_{i+1})^{-1}}{h^4} = \frac{D}{h^2} - \frac{(T^b_{i+1})^{-1}}{h^4}. \tag{7}$$

Having a fixed $b < \infty$ we can reduce the problem in (3) to one on Ω^a only, obtaining

$$\tilde{A}^a \tilde{\boldsymbol{u}}^a = \boldsymbol{f}^a, \quad \text{with} \quad \tilde{A}^a = \frac{1}{h^2}\begin{pmatrix} D_1 & -I_N & & \\ -I_N & \ddots & \ddots & \\ & \ddots & D_{N^a-1} & -I_N \\ & & -I_N & T^b_{N^a} \end{pmatrix}, \tag{8}$$

where compared to the matrix A^a (see (4)) the only change is the last block where the Dirichlet boundary condition block has been replaced by the Schur complement $T^b_{N^a}$, representing the "far-field" domain (or the overlap) unknowns in $\Omega^b \backslash \Omega^a$. Hence $\tilde{\mathbf{u}}^a$ approaches $\mathbf{u}$ in limit as $b \to \infty$ but increasing b makes the defining recurrence in (7) longer.

If b goes to infinity, the corresponding Schur complement matrix $T^\infty_{N^a}$ will still be governed by of (7), namely

$$T^\infty_{N^a} = \frac{D}{h^2} - \frac{(T^\infty_{N^a})^{-1}}{h^4}, \quad \text{i.e.,} \quad (T^\infty)^2 - \frac{D}{h^2}T^\infty + \frac{I}{h^4} = 0. \tag{9}$$

Notably, this equation is independent on N^a and hence so will be the solution $T^\infty_{N^a} \equiv T^\infty$. To solve (9), we start the following section by changing the basis we work in to the eigenbasis of D, effectively applying the discrete Fourier transform in the y variable.

3 Spectral analysis

Writing D from (5) as $D = D_{yy} + 2I$, where D_{yy} is the 3-point finite difference stencil discretization of $\eta - \partial_{yy}$ multiplied by h^2, we recall that $D_{yy} = Q^T \mathrm{diag}(z_1, \dots, z_N) Q$ with

$$z_k := \eta h^2 + 4\sin^2\left(\frac{k\pi}{2(N+1)}\right) \quad \text{and} \quad \boldsymbol{q}_k := \left[\sqrt{\tfrac{2}{N+1}}\sin\left(\tfrac{k\pi}{N+1}j\right)\right]_{j=1}^N \in \mathbb{R}^N, \tag{10}$$

where Q is unitary and symmetric, with the eigenvectors $\boldsymbol{q}_k$ in its columns. Thereby we can write $D = Q^T \Lambda Q$ with $\Lambda = \mathrm{diag}(\lambda_1, \ldots, \lambda_N)$ and $\lambda_k = 2 + z_k$ as its eigendecomposition.

Remark 1 *Calculating in the eigenbasis of D is a necessity for our Schur complement analysis but in treating each eigenmode separately we would add yet another index to the already loaded notation. That is why, instead of referring to the particular eigenvalues $\lambda_k = 2 + z_k$ we will introduce a new variable $\lambda = 2 + z$ and treat all quantities dependent on the eigenvalues as functions of λ. This way we avoid the index k, mostly λ_k and z_k, whenever we can. Nonetheless, in some places the reference to a particular eigenvalue or eigenmode is unavoidable and we keep the index k reserved for the eigenmode notation throughout the text.*

3.1 Diagonalization and convergence of the Schur Complement

Changing the basis for the Schur complement definition in (7) gives

$$\hat{T}^b_{N^b} = Q\frac{D}{h^2}Q^T = \frac{\Lambda}{h^2} \quad \text{and} \quad \hat{T}^b_i = Q\frac{D}{h^2}Q^T - Q\frac{(T^b_{i+1})^{-1}}{h^4}Q^T = \frac{\Lambda}{h^2} - \frac{(\hat{T}^b_{i+1})^{-1}}{h^4},$$

where $i = N^b + 1, \ldots, N^a$ and all of the matrices $\hat{T}^b_i$ are diagonal. Working with the diagonal entries only, each of them becomes a function of its corresponding eigenvalues (frequency) λ_k and also follows the recurrence. Recalling Remark 1, we can write

$$\hat{t}^b_{N^b}(\lambda) = \frac{\lambda}{h^2} \quad \text{and} \quad \hat{t}^b_i(\lambda) = \frac{\lambda}{h^2} - \frac{1}{h^4 \hat{t}^b_{i+1}(\lambda)} \quad \text{for } i = N^b - 1, \ldots, N^a. \tag{11}$$

We obtain an analogous recurrence for the solution $\mathbf{u}^b$ in (6). Setting $\hat{\boldsymbol{u}}^b_i := Q\boldsymbol{u}^b_i$ we get

$$-\frac{\hat{\boldsymbol{u}}^b_{N^b-1}}{h^2} + \frac{\Lambda\hat{\boldsymbol{u}}^b_{N^b}}{h^2} = -\frac{\hat{\boldsymbol{u}}^b_{N^b-1}}{h^2} + \hat{T}^b_{N^b}\hat{\boldsymbol{u}}^b_{N^b} = 0 \quad \text{and} \quad -\frac{\hat{\boldsymbol{u}}^b_{i-1}}{h^2} + \frac{\Lambda\hat{\boldsymbol{u}}^b_i}{h^2} - \frac{\hat{\boldsymbol{u}}^b_{i+1}}{h^2} = -\frac{\hat{\boldsymbol{u}}^b_{i-1}}{h^2} + \frac{\hat{T}^b_i\hat{\boldsymbol{u}}^b_i}{h^2} = 0, \tag{12}$$

with $i = N^b + 1, \ldots, N^a$. Focusing on the limit case $b \to +\infty$, we can now treat each mode separately, obtaining a scalar problem instead of (9). Setting

$$\lim_{b\to+\infty} \hat{t}^b_{N^a}(\lambda) =: \hat{t}^\infty_{N^a}(\lambda), \tag{13}$$

we observe that

$$(\hat{t}^\infty_{N^a})^2(\lambda) - \frac{\lambda}{h^2}\hat{t}^\infty_{N^a}(\lambda) + \frac{1}{h^4} = 0, \tag{14}$$

and we find the two solutions

$$\hat{\tau}^{\infty,1}(\lambda) = \frac{\lambda + \sqrt{\lambda^2 - 4}}{2h^2} \quad \text{and} \quad \hat{\tau}^{\infty,2}(\lambda) = \frac{\lambda - \sqrt{\lambda^2 - 4}}{2h^2},$$

and, moreover, find that

$$\left(h^2\hat{\tau}^{\infty,1}(\lambda)\right)\left(h^2\hat{\tau}^{\infty,2}(\lambda)\right) = 1 \quad \text{and} \quad 0 < h^2\hat{\tau}^{\infty,2}(\lambda) < 1 < h^2\hat{\tau}^{\infty,1}(\lambda). \tag{15}$$

Next, we show that one of the solutions $\hat{\tau}^{\infty,1}(\lambda), \hat{\tau}^{\infty,2}(\lambda)$ indeed acts as the limit Schur complement for our solution vector $\tilde{\boldsymbol{u}}^a$. The key observation is that the characteristic polynomial of the recurrence relation in (12) is preserved through the limit process and thus the solutions $\hat{\tau}^{\infty,1}(\lambda), \hat{\tau}^{\infty,2}(\lambda)$ of the limit equation (14) coincide with the roots of the characteristic polynomial of the recurrence relation in (12) given by

$$p_\lambda(r) = -r^2 + \lambda r - 1.$$

This together with the explicit formula for the solution of the recurrence relation (12) is enough to solve the matrix equation defining T^∞ in 9 (in order to do so, we will evaluate the functions of λ at the particular points of interest λ_k, i.e., at the eigenvalues of the matrix D).

Theorem 3.1 *The Schur complement $T^b_{N^a}$ defined in (7) converges to $T^{\infty,1}$, i.e., the solution of the formal limit equation (9), as $b \to +\infty$. That is to say, the eigenvectors of those matrices are equal and for any $k = 1, \ldots, N$ we have*

$$\hat{t}^\infty(\lambda_k) \equiv \lim_{N^b \to \infty} \hat{t}^b_{N^a}(\lambda_k) = \hat{\tau}^{\infty,1}(\lambda_k) = \frac{\lambda_k + \sqrt{\lambda_k^2 - 4}}{2h^2}. \tag{16}$$

Proof Fixing b large, we take a particular $i = N^a, \ldots, N^b$ and the solution subvector $\hat{\mathbf{u}}^b_i = [\hat{u}^b_{i,1}, \ldots, \hat{u}^b_{i,N}]^T \in \mathbb{R}^N$ and recall it follows the recurrence in (12) that has a closed formula solution, i.e., there exist pairs of constants $(\nu^b_1, \mu^b_1), \ldots, (\nu^b_N, \mu^b_N)$ independent on i such that

$$\hat{\boldsymbol{u}}^b_i = \begin{bmatrix} \mu^b_1 \left(h^2 \hat{\tau}^{\infty,1}(\lambda_1)\right)^{i-N^a} + \nu^b_1 \left(h^2 \hat{\tau}^{\infty,2}(\lambda_1)\right)^{i-N^a} \\ \vdots \\ \mu^b_N \left(h^2 \hat{\tau}^{\infty,1}(\lambda_N)\right)^{i-N^a} + \nu^b_N \left(h^2 \hat{\tau}^{\infty,2}(\lambda_N)\right)^{i-N^a} \end{bmatrix}.$$

Furthermore, recalling (15) it follows that

$$\left(h^2 \hat{\tau}^{\infty,1}(\lambda_k)\right)^{N^b - N^a} \to +\infty \quad \text{as} \quad b \to +\infty,$$

for all k and as there is homogeneous Dirichlet boundary condition at $x = b$ we obtain

$$\lim_{b\to\infty} \mu^b_1 = \ldots = \lim_{b\to\infty} \mu^b_N = 0 \quad \text{and} \quad \hat{\boldsymbol{u}}^\infty_i \equiv \lim_{b\to\infty} \hat{\boldsymbol{u}}^b_i = \begin{bmatrix} \nu^\infty_1 \left(h^2 \hat{\tau}^{\infty,2}(\lambda_1)\right)^{i-N^a} \\ \vdots \\ \nu^\infty_N \left(h^2 \hat{\tau}^{\infty,2}(\lambda_N)\right)^{i-N^a} \end{bmatrix}. \tag{17}$$

Taking $i = N^a + 1$ we can solve the k-th entry of the recurrence in (12) for $h^2 \hat{t}^b_{N^a+1}(\lambda)$ and using the finite difference stencil, we obtain

$$h^2 \hat{t}^b_{N^a+1}(\lambda_k) = \frac{\lambda \hat{u}^b_{N^a+1,k} - \hat{u}^b_{N^a+2,k}}{\hat{u}^b_{N^a+1,k}} = \frac{\hat{u}^b_{N^a,k}}{\hat{u}^b_{N^a+1,k}},$$

and hence letting $b \to +\infty$ and recalling (17) we have

$$h^2 \hat{t}^{\infty}_{N^a+1}(\lambda_k) = \frac{\hat{u}^{\infty}_{N^a,k}}{\hat{u}^{\infty}_{N^a+1,k}} = \frac{1}{h^2 \hat{\tau}^{\infty,2}_k(\lambda)} = h^2 \hat{\tau}^{\infty,1}_k.$$

□

Recalling $\lambda = 2 + z$ we can also write $\hat{t}^{\infty}(\lambda)$ as

$$\hat{t}^{\infty}(z) = \frac{2+z+\sqrt{(2+z)^2-4}}{2h^2} = \frac{1}{h^2} + \frac{z}{2h^2} + \frac{\sqrt{z^2+4z}}{2h^2} = \frac{1}{h^2}\left(1 + \frac{z}{2}\left(1 + \sqrt{1+\frac{4}{z}}\right)\right), \tag{18}$$

where the evaluation points of interest are $z_k := \eta h^2 + 4\sin^2(hk\pi/2)$, see Remark 1.

Returning to the case $b < \infty$, the recurrence relations in (11) as functions of z reads

$$\hat{t}^{b}_{N^b}(z) = \frac{2+z}{h^2}, \quad \hat{t}^{b}_{N^b-1}(z) = \frac{2+z}{h^2} - \frac{1}{h^4 \frac{2+z}{h^2}} = \frac{1}{h^2}\left(2 + z - \frac{1}{2+z}\right),$$
$$\hat{t}^{b}_{N^b-2}(z) = \frac{2+z}{h^2} - \frac{1}{h^4 \hat{t}^{b}_{N^b-1}(z)} = \frac{1}{h^2}\left(2 + z - \frac{1}{2+z-\frac{1}{2+z}}\right),$$

and by the recursive definition in (11), the i-th one is given by

$$\hat{t}^{b}_{i}(z) = \frac{2+z}{h^2} - \cfrac{\frac{1}{h^2}}{2+z-\cfrac{1}{2+z-\cfrac{\ddots}{2+z-\cfrac{1}{2+z}}}}. \tag{19}$$

Objects (functions) of this form are called *continued fractions*. Their theory links various areas of mathematics, e.g., Padé approximations, orthogonal polynomials, Vorobyev's moment matching problem, Gauss quadrature and the method of conjugate gradients (see [16] and also [15, Section 3.3.2 - 3.3.6] for further references) and we will use some of these links to establish our main result in Section 3.3 later on. In the light of this observation we adjust our notation, clarifying everything in Remark 2 below.

Remark 2 *Notice that in the continued fraction representation of $\hat{t}^{b}_{i}(z)$ in* (19)*, the continued fraction has exactly $N^b - i$ levels. In order to simplify the notation, we will from now on change the subscript i to correspond to the "number of levels" or "depth" of the continued fraction. Hence, for the rest of the text we will write*

$$\hat{t}^{b}_{0}(z) = \frac{2+z}{h^2}, \quad \hat{t}^{b}_{1}(z) = \frac{2+z}{h^2} - \frac{1}{h^4 \frac{2+z}{h^2}} = \frac{1}{h^2}\left(2 + z - \frac{1}{2+z}\right),$$
$$\hat{t}^{b}_{2}(z) = \frac{2+z}{h^2} - \frac{1}{h^4 \hat{t}^{b}_{N^b-1}(z)} = \frac{1}{h^2}\left(2 + z - \frac{1}{2+z-\frac{1}{2+z}}\right),$$

and so on. This means that the index i changes the meaning from *the number of grid columns in the domain Ω^b* to *the number of grid columns in the domain $\Omega^b \setminus \Omega^a$.*

We continue by a simple observation regarding the functions $\hat{t}^{\infty}$ and $\hat{t}_i^b$.

Remark 3 *By a direct computation and subsequent by re-insertion we obtain*

$$\hat{t}^{\infty}(z) = 2 + z - \frac{1}{\hat{t}^{\infty}(z)}, \quad \hat{t}^{\infty}(z) = 2 + z - \frac{1}{2 + z - \frac{1}{\hat{t}^{\infty}(z)}}, \dots$$

and so on. This suggests that the function $\hat{t}^{\infty}(z)$ is equal to the infinite continued fraction

$$\hat{t}^{\infty}(z) = 2 + z - \frac{1}{2 + z - \frac{1}{2+z-\dots}},$$

and $\hat{t}_i^b(z)$ in (19) *are approximations in the sense of a truncation after i levels.*

As we will use the continued fractions only as a tool to arrive at our main result, we choose [2] as the main reference, which is a book written as an overview for Padé approximations and the continued fractions are approached mainly from that perspective. We refer the interested readers to [16] and [22] for more detailed expositions of the theory of continued fractions.

We continue in Section 3.2 with a concise summary of the continued fractions results and the connected simple algebraic calculations. We formulate these in terms of an auxiliary variable α, given by

$$\alpha = \frac{4}{z}. \tag{20}$$

This change of variables is unavoidable as we will need to expand about ∞ and the standard way of defining this is to expand the same function but of reciprocal argument about 0 – hence (20). This is why we do not consider (20) as a proper change of variables, which would otherwise necessitates tedious calculations of the derivatives of the function composition. In fact, the *true* change of variables consists only of multiplying by 4 and hence does not require a re-computation of the derivatives.

3.2 Padé Approximation and Continued Fractions

We follow the notation from [2], i.e., the $[M/L]$-Padé approximant of $f(z)$ is denoted by $[M/L]_f \equiv [M/L]_f(z)$. We start with Padé theory and proceed with continued fractions ([2, Chapter 4]).

Theorem 3.2 ([2, Theorem 1.5.3, 1.5.4, 1.5.1]) *Let $f(z)$ be a real function of a real variable. Then the following holds provided the Padé approximants exist :*

1. *Let $\alpha, \beta \in \mathbb{R}$. Then $\alpha + \beta[M/L]_f = [M/L]_{\alpha+\beta f}$.*
2. *Let $m \geq 1$ and $f(z) = \sum_{j=0}^{+\infty} c_j z^j$ be a formal power series. Setting $g(z) = \frac{1}{z^m}\left(f(z) - \sum_{j=0}^{m-1} c_j z^j\right)$ and assuming $M - m \geq L - 1$ we have*

$$[M - m/L]_g(z) = \frac{1}{z^m}\left([M/L]_f(z) - \sum_{j=0}^{m-1} c_j z^j\right).$$

3. Let $f(0) \neq 0$ and set $g(z) = 1/f(z)$. Then $[M/L]_g(z) = 1/[L/M]_f(z)$.

Definition 3.3 *A* continued fraction *is given by sequences of real numbers $\{a_j\}_j, \{b_j\}_j$ – the numerator and the denominator sequence of the continued fraction – and has the general form*

$$b_0 + \cfrac{a_1}{b_1 + \cfrac{a_2}{b_2 + \cfrac{a_3}{\ddots}}} =: b_0 + \sum_{j=1}^{+\infty} \frac{a_j}{b_j+} \equiv b_0 + \frac{a_1}{b_1} + \frac{a_2}{b_2} + \dots,$$

where the sum is to be understood only formally. The continued fraction is called infinite *as long as $a_j, b_j \neq 0$ for all j. The n-th truncation (or convergent) of a continued fraction is given by*

$$\frac{A_n}{B_n} = b_0 + \sum_{j=1}^{n} \frac{a_j}{b_j+} = b_0 + \cfrac{a_1}{b_1 + \cfrac{a_2}{\ddots \cfrac{}{b_{n-2} + \cfrac{a_n}{b_{n-1} + \cfrac{a_n}{b_n}}}}},$$

where A_n and B_n are the n-th truncation (or convergent) numerator and denominator.

Replacing the scalars a_j and/or b_j by linear (or affine) functions of a real variable z, A_n and B_n become polynomials in z and the n-th truncation of the continued fraction becomes a rational function in z. Different settings of this framework lead to different types of continued fractions. Most notably, a continued fraction is called regular C-fraction *(short for regular classical continued fraction), provided it has the form*

$$b_0 + \cfrac{a_1 z}{1 + \cfrac{a_2 z}{1 + \cfrac{a_3 z}{\ddots}}} \equiv b_0 + \frac{a_1 z}{1} + \frac{a_2 z}{1} + \dots,$$

with $a_j \neq 0$ for all j. If, moreover, $a_j > 0$ for all j, then it is called S-fraction *(short for Stieltjes continued fraction). If the continued fraction takes the form*

$$b_0 + \cfrac{r_1}{z + s_1 - \cfrac{r_2}{z + s_2 - \cfrac{r_3}{\ddots}}} \equiv b_0 + \frac{r_1}{z + s_1} - \frac{r_2}{z + s_2} - \dots,$$

with $r_j \neq 0$ for all j then it is called J-fraction *(short for Jacobi continued fraction). For more details on the introduced types of continued fractions as well as other types of continued fractions (e.g., non-regular C-fraction, T-fraction, P-fraction,...) we refer also to [16] and [22] and references therein.*

First, we note that we have ignored the questions of convergence of infinite continued fractions and we refer the reader to [16] and [22]. Next, notice that one function can be represented by two seemingly different continued fractions (different in type and/or in the coefficient values) and one way to recognize their equality is via the *three-term recurrence relation* (see [2, Theorem 4.1.1, pp.106]). We have that

$$\begin{aligned} &A_{-1} = 1, \quad A_0 = b_0, \quad A_n = b_n A_{n-1} + a_n A_{n-2}, \\ &B_{-1} = 0, \quad B_0 = 1, \quad B_n = b_n B_{n-1} + a_n B_{n-2}. \end{aligned} \tag{21}$$

and assuming the n-th truncation (convergent) of two continued fractions are equal for any n, the infinite continued fractions are equal as well. Last but not least, we note that some authors will call a continued fraction an S-fraction even though the fraction itself does not meet the definition above but can be *transformed* into a continued fraction that does. We next recall a basic transformation rule of continued fractions.

Lemma 3.4 ([2, Section 4.1, pp. 105-106]) *Let $\{a_j\}_j, \{b_j\}_j$ be two real sequences of the numerators and denominators of a continued fraction as in Definition 3.3. Let $\{e_j\}_j$ be a sequence of real numbers different from zero. Then we have*

$$b_0 + \frac{a_1}{b_1} + \frac{a_2}{b_2} + \frac{a_3}{b_3} + \cdots = b_0 + \frac{e_1 a_1}{e_1 b_1} + \frac{e_1 e_2 a_2}{e_2 b_2} + \frac{e_2 e_3 a_3}{e_3 b_3} + \cdots,$$

For purposes of this text we present immediately the continued fraction result for the square root function, which is of interest to us[2].

Theorem 3.5 ([2, Section 4.6, Theorem 4.4.3 and formula (6.4) on pp. 139]) *For any $\alpha \in (-1, +\infty)$[3] we have*

$$\sqrt{1+\alpha} = 1 + \cfrac{\frac{\alpha}{2}}{1 + \cfrac{\frac{\alpha}{2}}{2 + \cfrac{\frac{\alpha}{2}}{1 + \cfrac{\frac{\alpha}{2}}{\ddots \\ b_{n-2} + \cfrac{a_n}{b_{n-1} + \cfrac{a_n}{b_n + \cfrac{a_{n+1}}{\cdots}}}}}}}} = 1 + \frac{\frac{\alpha}{2}}{1} + \frac{\frac{\alpha}{2}}{2} + \frac{\frac{\alpha}{2}}{1} + \cdots + \frac{a_n}{b_n} + \cdots, \quad (22)$$

with $b_0 = 1$, $b_j = \frac{3+(-1)^j}{2}$ and $a_j = \frac{\alpha}{2}$, $j \geq 1$. Moreover, for any n the $[n,n]$-Padé approximation of $\sqrt{1+\alpha}$ expanded about $\alpha = 0$ is given by the $(2n)$-th truncation of the continued fraction in (22) and the $[n+1,n]$-Padé approximation of $\sqrt{1+\alpha}$ expanded about $\alpha = 0$ is given by the $(2n+1)$-st truncation of the continued fraction in (22).

Remark 4 *By a direct computation we see that*

$$\sqrt{1+\alpha} = 1 + \frac{\alpha}{2} + \frac{\alpha}{2} + \frac{\alpha}{2} + \cdots,$$

i.e., the representation in (22) can be written as a cyclic S-fraction[4] with $a_j = 1/2$ for all j.

[2]We refer to the book of Baker but the original result is due to Gauss, who showed a much more general result for the hypergeometric function ${}_2F_1$; for more details see [22, Chapter XVIII] or [16, Chapter VI].

[3]There is a misprint in [2, equation (6.4), page 139]. The authors state the convergence "for all z except $-\infty < z \leq 1$)" but the the result also holds for $z \in (-1, 1]$.

[4]Infinite continued fractions with periodic sequences $\{a_j\}, \{b_j\}$ are called cyclic continued fractions.

The rest of this section is devoted to auxiliary results, the first of which links a truncation of the S-fraction from Theorem 3.5 and a truncation of the J-fraction from Remark 3. Notice that the continued fractions are not identical but rather differ in the absolute term.

Lemma 3.6 *Let α be real and consider the two continued fractions*

$$\tau(\alpha) := \cfrac{\frac{\alpha}{2}}{2+\cfrac{\frac{\alpha}{2}}{1+\cfrac{\frac{\alpha}{2}}{2+\cfrac{\frac{\alpha}{2}}{1+\ldots}}}} \quad \text{and} \quad \sigma(\alpha) := \cfrac{1}{1+\frac{4}{\alpha}-\cfrac{1}{2+\frac{4}{\alpha}-\cfrac{1}{2+\frac{4}{\alpha}-\cfrac{1}{2+\frac{4}{\alpha}-\ldots}}}},$$

and denote their n-th truncations by $A_n(\alpha)/B_n(\alpha)$ and $C_n(\alpha)/D_n(\alpha)$. For any $n = 1, 2, \ldots$ we have

$$A_{2n}(\alpha)/B_{2n}(\alpha) = C_n(\alpha)/D_n(\alpha).$$

Proof Using Lemma 3.4 we transform $\tau(\alpha)$ and without further relabeling we obtain

$$\tau(\alpha) := \cfrac{1}{\frac{4}{\alpha}+\cfrac{1}{1+\cfrac{1}{\frac{4}{\alpha}+\cfrac{1}{1+\cfrac{1}{\frac{4}{\alpha}+\frac{1}{1+\ldots}}}}}}, \tag{23}$$

and by a direct computation confirm that the equality holds for $n = 1$. Next, we notice that the continued fraction (23) can be written in a cyclic form with *the core* R given by

$$R = \frac{4}{\alpha} + \frac{1}{1+\frac{1}{R}}, \tag{24}$$

i.e., the continued fraction can be obtained by a successive re-insertion of the core equality (24) into itself, e.g.,

$$\underbrace{\cfrac{1}{\frac{4}{\alpha}+\frac{1}{1}}}_{=\frac{A_2(\alpha)}{B_2(\alpha)}}, \quad \underbrace{\cfrac{1}{\frac{4}{\alpha}+\cfrac{1}{1+\cfrac{1}{\frac{4}{\alpha}+\frac{1}{1}}}}}_{=\frac{A_4(\alpha)}{B_4(\alpha)}}, \quad \underbrace{\cfrac{1}{\frac{4}{\alpha}+\cfrac{1}{1+\cfrac{1}{\frac{4}{\alpha}+\cfrac{1}{1+\cfrac{1}{\frac{4}{\alpha}+\frac{1}{1}}}}}}}_{=\frac{A_6(\alpha)}{B_6(\alpha)}}, \quad \ldots .$$

In this way every re-insertion adds two elements of the numerator and denominator sequences and using the algebraic identity

$$\frac{1}{1+\frac{1}{R}} = 1 - \frac{1}{1+R},$$

we reformulate the core equality (24) to obtain

$$1 + R = 2 + \frac{4}{\alpha} - \frac{1}{1+R}. \tag{25}$$

and notice that the core equality in (25) is the one that generates the J-fraction $\sigma(\alpha)$.

Hence we have shown that for $n \geq 2$ the $2n$ re-insertions of the core R in the equality (24) is equal to n re-insertions of the core $1 + R$ in the equality (25), finishing the proof. □

We build upon Lemma 3.6 by contracting the S-fraction in (22) into a J-fraction.

Proposition 3.7 *Let α be real and set the continued fractions $\tau(\alpha)$ and $\sigma(\alpha)$ as in Lemma 3.6. Moreover, we define the continued fractions*

$$\tilde{\tau}(\alpha) := \frac{1}{1+\tau(\alpha)} \quad \text{and} \quad \phi(\alpha) := 1 - \cfrac{1}{2+\frac{4}{\alpha} - \cfrac{1}{2+\frac{4}{\alpha} - \cfrac{1}{2+\frac{4}{\alpha} - \dots}}}$$

with n-th truncations $\tilde{A}_n(\alpha)/\tilde{B}_n(\alpha)$ and $E_n(\alpha)/F_n(\alpha)$ with $E_0 = F_0 = 1$. Then for $n \geq 0$

$$A_{2n+1}(\alpha)/B_{2n+1}(\alpha) = E_n(\alpha)/F_n(\alpha).$$

Proof The equality for $n = 0$ holds by inspection. Taking $n \geq 1$, we use Lemma 3.6 for the continued fraction $\tilde{\tau}(\alpha)$, obtain

$$\tilde{A}_{2n+1}(\alpha)/\tilde{B}_{2n+1}(\alpha) = \frac{1}{1+A_{2n}(\alpha)/B_{2n}(\alpha)} = \frac{1}{1+C_n(\alpha)/D_n(\alpha)}$$

and having the truncations $C_n(\alpha), D_n(\alpha)$ of $\sigma(\alpha)$ from Lemma 3.6 it remains to show that

$$\frac{1}{1+C_n(\alpha)/D_n(\alpha)} = 1 - E_n(\alpha)/F_n(\alpha). \tag{26}$$

The cyclic parts of both $\sigma(\alpha)$ and $\phi(\alpha)$ coincide and we denote them by $\tilde{\sigma}(\alpha)$,

$$\tilde{\sigma}(\alpha) := \cfrac{1}{2+\frac{4}{\alpha} - \cfrac{1}{2+\frac{4}{\alpha} - \dots}}. \tag{27}$$

In turn, this gives

$$\sigma(\alpha) = \cfrac{1}{1 + \cfrac{1}{1 + \frac{4}{\alpha} - \tilde{\sigma}(\alpha)}} \quad \text{and} \quad \phi(\alpha) = 1 - \frac{1}{2 + \frac{4}{\alpha} - \tilde{\sigma}(\alpha)},$$

and thus to show (26) it is enough to prove

$$\cfrac{1}{1 + \cfrac{1}{1 + \frac{4}{\alpha} - \tilde{\sigma}}} = 1 - \frac{1}{2 + \frac{4}{\alpha} - \tilde{\sigma}},$$

as $\tilde{\sigma}$ contains the common part. By a direct computation we obtain

$$\cfrac{1}{1 + \cfrac{1}{1 + \frac{4}{\alpha} - \tilde{\sigma}}} = \frac{1 + \frac{4}{\alpha} - \tilde{\sigma}(\alpha)}{2 + \frac{4}{\alpha} - \tilde{\sigma}(\alpha)} \quad \text{and} \quad 1 - \frac{1}{2 + \frac{4}{\alpha} - \tilde{\sigma}} = \frac{1 + \frac{4}{\alpha} - \tilde{\sigma}(\alpha)}{2 + \frac{4}{\alpha} - \tilde{\sigma}(\alpha)},$$

finishing the proof. □

3.3 Approximation Properties of the Schur Complement

Let us first recall the function $\hat{t}_i^b(z)$ and $\hat{t}^\infty(z)$ in (19) and (18) representing the Schur complements T_i^b and T^∞

$$\hat{t}_i^b(z) = \frac{1}{h^2}\left(2 + z - \cfrac{1}{2 + z - \cfrac{1}{2 + z - \cfrac{\ddots}{2 + z - \cfrac{1}{2 + z}}}}\right), \quad \hat{t}^\infty(z) = \frac{1}{h^2}\left(1 + \frac{z}{2} + \frac{z}{2}\sqrt{1 + \frac{4}{z}}\right).$$

In Theorem 3.8 we show an important approximation property of these functions. We use a similar technique as in [11] where the authors compute a Padé approximation of the Dirichlet to Neumann operator. This is not a coincidence: the Schur complement and the Dirichlet-to-Neumann map have a deep connection, see, e.g., [10, Section 5.2].

Theorem 3.8 *The function $\hat{t}_i^b(z)$ is the $[i, i]$-Padé approximation about the expansion point $z = +\infty$ of $\hat{t}^\infty(z)$.*

Proof First, we drop the $1/h^2$ factor for both of the functions and transpose the expansion point $z = +\infty$ to $\alpha = 0$ as in (20) and without further relabeling we obtain

$$\hat{t}_i^b(\alpha) = 2 + \frac{4}{\alpha} - \cfrac{1}{2 + \frac{4}{\alpha} - \cfrac{1}{2 + \frac{4}{\alpha} - \cfrac{\ddots}{2 + \frac{4}{\alpha} - \cfrac{1}{2 + \frac{4}{\alpha}}}}}, \quad \hat{t}^\infty(\alpha) = 1 + \frac{2}{\alpha} + \frac{2}{\alpha}\sqrt{1 + \alpha}.$$

Recalling (15), we have

$$\{\hat{t}^{\infty}\}^{-1}(\alpha) := \frac{1}{\hat{t}^{\infty}(\alpha)} = 1 + \frac{2}{\alpha} - \frac{2}{\alpha}\sqrt{1+\alpha} \tag{28}$$

and using Theorem 3.2(3.) we get

$$[i/i]_{\hat{t}^{\infty}}(\alpha) = \frac{1}{[i+1/i]_{\{\hat{t}^{\infty}\}^{-1}}(\alpha)}$$

for any $i \geq 1$. By a direct computation we obtain

$$\{\hat{t}^{\infty}\}^{-1}(\alpha) = 1 + \frac{2}{\alpha} - \frac{2}{\alpha}\sqrt{1+\alpha} = 1 - 2\frac{1}{\alpha}\left(\sqrt{1+\alpha} - 1\right),$$

and hence by the Padé approximant calculus (see Theorem 3.2(1. and 2.)) we obtain

$$[i/i]_{\{\hat{t}^{\infty}\}^{-1}}(\alpha) = 1 - 2\frac{1}{\alpha}\left([i+1/i]_{\sqrt{1+\alpha}}(\alpha) - 1\right).$$

Using the continued fraction representation from Theorem 3.5, we obtain

$$[i/i]_{\{\hat{t}^{\infty}\}^{-1}}(\alpha) = 1 - \frac{2}{\alpha}\left(1 + 1 + \frac{A_{2i+1}(\alpha)}{B_{2i+1}(\alpha)} - 1\right) = 1 - \frac{2}{\alpha}\left(\cfrac{\frac{\alpha}{2}}{1+\cfrac{\frac{\alpha}{2}}{2+\cfrac{\frac{\alpha}{2}}{1+\cfrac{\frac{\alpha}{2}}{\ddots\, b_{2i-1}+\cfrac{}{b_{2i}+\cfrac{a_{2i+1}}{b_{2i+1}}}}}}}\right),$$

where the sequences $\{a_j\}_j, \{b_j\}_j$ are given as in Theorem 3.5 and $A_{2i+1}(\alpha)/B_{2i+1}(\alpha)$ is the $(2i+1)$-st truncation of the continued fraction $\tau(\alpha)$ from Lemma 3.6. Hence we have

$$[i/i]_{\{\hat{t}^{\infty}\}^{-1}}(\alpha) = 1 - \cfrac{1}{1+\cfrac{\frac{\alpha}{2}}{2+\cfrac{\frac{\alpha}{2}}{1+\cfrac{\frac{\alpha}{2}}{\ddots\, b_{2i-1}+\cfrac{}{b_{2i}+\frac{a_{2i+1}}{b_{2i+1}}}}}}}, \tag{29}$$

and by a straight-forward manipulation (see Proposition 3.7) we observe that the continued fraction on the left-hand side in (29) is the $(2i+1)$-st truncation of the continued fraction

$$\tilde{\tau}(\alpha) := \frac{1}{1+\tau(\alpha)}.$$

Finally, Proposition 3.7 gives a *J-fraction representation* of the continued fraction $\tilde{\tau}$ and its $(2i+1)$-st truncation denoted by $\tilde{C}_i(\alpha)/\tilde{D}_i(\alpha)$, obtaining

$$[i/i]_{\{\hat{t}^\infty\}^{-1}}(\alpha) = 1 - \left(1 - \frac{\tilde{C}_n(\alpha)}{\tilde{D}_n(\alpha)}\right) = \underbrace{\cfrac{1}{2+\frac{4}{\alpha} - \cfrac{1}{2+\frac{4}{\alpha} - \cfrac{1}{2+\frac{4}{\alpha} - \ddots - \cfrac{1}{2+\frac{4}{\alpha} - \cfrac{1}{2+\frac{4}{\alpha}}}}}}}_{i-1 \text{ "levels"}}.$$

As a result, we get that for any $i \geq 1$

$$[i/i]_{\hat{t}^\infty}(\alpha) = \cfrac{1}{\cfrac{1}{2+\frac{4}{\alpha} - \cfrac{1}{2+\frac{4}{\alpha} - \cfrac{1}{2+\frac{4}{\alpha} - \ddots - \cfrac{1}{2+\frac{4}{\alpha} - \cfrac{1}{2+\frac{4}{\alpha}}}}}}} = 2 + \frac{4}{\alpha} - \underbrace{\cfrac{1}{2+\frac{4}{\alpha} - \cfrac{1}{2+\frac{4}{\alpha} - \cfrac{1}{2+\frac{4}{\alpha} - \ddots - \cfrac{1}{2+\frac{4}{\alpha} - \cfrac{1}{2+\frac{4}{\alpha}}}}}}}_{i-1 \text{ "levels"}}$$

finishing the proof. □

Setting the approximation the error[5]

$$err_D(z,i) := \hat{t}^\infty(z) - \hat{t}^b_i(z),$$

where i denotes the number of grid columns that were folded into the Schur complement (see Remark 2), we plot it for small i in Figure 2.

As expected, we see that $err_D(z,i)$ quickly decreases as z grows and this become more pronounced for larger i, i.e., for for higher order Padé approximation, i.e., when b increases. We see that the error is still large for small z, i.e., we struggle with the low frequency modes approximation. We try improving this in the next section by considering a Robin boundary condition at $x = b$.

4 Robin boundary condition for truncation

Recalling the interval bounding the spectrum of D_{yy} (and thus the natural domain of the variable z) is given by

$$\left[\eta h^2 + 4\sin^2\left(\frac{\pi}{2}\frac{1}{N+1}\right), \eta h^2 + 4\sin^2\left(\frac{\pi}{2}\frac{N}{N+1}\right)\right] \approx \left[\eta h^2, \eta h^2 + 4\right], \tag{30}$$

[5]The subscript D stands for the "Dirichlet" boundary condition at the end point $x = b$.

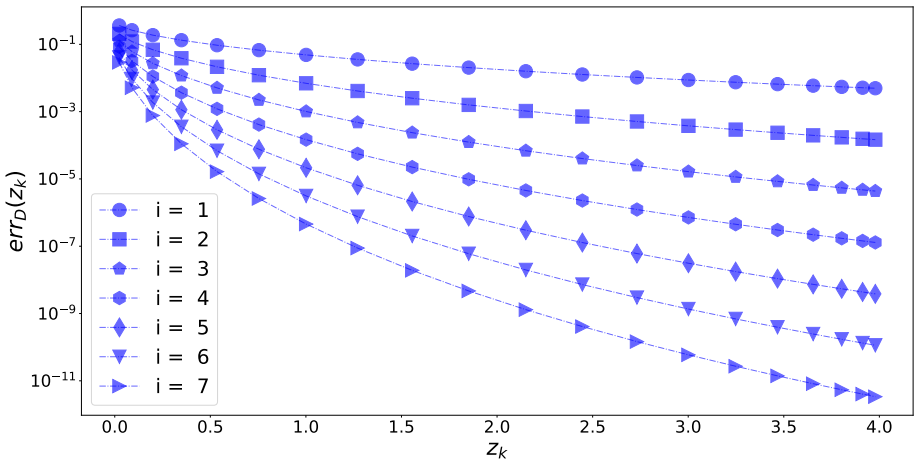

Figure 2: Plots of the function $err_D(z, i)$ at the points z_k with the parameters set to $N = 20$ and $\eta = 2$. The value i corresponds to the number of grid columns in $\Omega^b \setminus \Omega^a$, see Remark 2.

we see that the expansion point $z = +\infty$ of the Padé approximation is far from optimal. By replacing the Dirichlet boundary condition at $x = b$ by a homogeneous Robin boundary condition[6] with the Robin parameter $p \geq 0$ at b, i.e., with

$$\frac{\partial u}{\partial n} + pu = 0 \quad \text{at} \quad x = b,$$

we hope to improve this. Using a centered finite difference approximation as before, the Robin condition can be discretized with the so-called *ghost point* trick. Discretizing the Robin condition we get

$$\frac{\boldsymbol{u}_{N^b+1} - \boldsymbol{u}_{N^b-1}}{2h} + p\boldsymbol{u}_{N^b} = 0, \quad \text{i.e.,} \quad \boldsymbol{u}_{N^b+1} = \boldsymbol{u}_{N^b-1} - 2phu_{N^b},$$

we use the discretized equation at b, $-\boldsymbol{u}_{N^b+1} + D_{N^b}\boldsymbol{u}_{N^b} - \boldsymbol{u}_{N^b-1} = 0$, and eliminate the unknowns $\boldsymbol{u}_{N^b+1}$ – the ghost points – to get $(D_{N^b} + 2hpI)\boldsymbol{u}_{N^b} - 2\boldsymbol{u}_{N^b-1} = 0$. Hence the new system matrix denoted by $\bar{A}^b$ becomes

$$\bar{A}^b = \frac{1}{h^2}\begin{pmatrix} D_1 & -I_N & & & & \\ -I_N & \ddots & \ddots & & & \\ & \ddots & D_{N^a} & -I_N & & \\ & & -I_N & D_{N^a+1} & \ddots & \\ & & & \ddots & \ddots & -I_N \\ & & & & -I_N & \bar{D}_{N^b} \end{pmatrix} \tag{31}$$

[6] A Robin boundary condition is a simple approximation to the transparent boundary condition and works in general substantially better; for subdomain truncation in domain decomposition see [8].

with $\bar{D}_{N^b} = \frac{1}{2}\left(D_{N^b} + (2ph)I_N\right)$. This also modifies the Schur complement, yielding

$$\bar{T}^b_{N^b} = \frac{\bar{D}_{N^b}}{2h^2} \quad \text{and} \quad \bar{T}_j = \frac{D}{h^2} - \frac{\bar{T}^{-1}_{j+1}}{h^4}, \quad \text{for } j = N^b - 1, \ldots, N^a, \tag{32}$$

After the diagonalization (see (11)) and without relabeling, the first three functions representing the diagonal entries are

$$\bar{t}^b_0(z) = \frac{1 + ph + \frac{z}{2}}{h^2}, \quad \bar{t}^b_1(z) = \frac{2+z}{h^2} - \frac{1}{h^4 \frac{1+ph+\frac{z}{2}}{h^2}}, = \frac{1}{h^2}\left(2 + z - \frac{1}{1 + ph + \frac{z}{2}}\right),$$

$$\bar{t}^b_2(z) = \frac{2+z}{h^2} - \frac{1}{h^4 \tilde{t}^b_{N^b-1}(z)} = \frac{1}{h^2}\left(2 + z - \frac{1}{2 + z - \frac{1}{1+ph+\frac{z}{2}}}\right),$$

and by the recursive definition in (32) we obtain

$$\tilde{t}^b_i(z) = \frac{2+z}{h^2} - \cfrac{\frac{1}{h^2}}{2 + z - \cfrac{1}{2+z-\cfrac{\ddots}{2+z-\cfrac{1}{1+ph+\frac{z}{2}}}}}. \tag{33}$$

Notice that if $p \to +\infty$ we recover the original Dirichlet boundary condition with one less level of the continued fraction, i.e., corresponding to the physical domain $(b, b+h) \times (0, 1)$. With (33), we can numerically explore the effect of the Robin parameter p on the behavior of the error[7]

$$err_R(z, i) := \hat{t}^\infty(z) - \bar{t}^b_i(z),$$

and we illustrate this in Figure 3.

We see that the behavior around the right endpoint of the interval is analogous to the one in Figure 2 but the Robin condition introduced a new point z_p around which the approximation is accurate, e.g., in Figure 3 we see that $z_p \approx 0.82$. Assuming z_p is a solution of

$$err_R(z, i) = 0, \tag{34}$$

and that $err_R(z, i)$ is smooth except at a finite number of points, equation (34) defines z_p as an implicit function of p and the other parameters of the problem. For $i = 1$ we get

$$err_R(z, i) = \frac{1}{h^2}\left(1 + \frac{z}{2}\left(1 + \sqrt{1 + \frac{4}{z}}\right)\right) - \frac{1 + ph + \frac{z}{2}}{h^2} = \frac{1}{h^2}\left(\frac{z}{2}\sqrt{1 + \frac{4}{z}} - ph\right),$$

which gives z_p as the positive root of the quadratic equation

$$z_p^2 + 4z_p - 4p^2h^2 = 0 \quad \Longrightarrow \quad z_p = -2 + 2\sqrt{1 + p^2h^2}. \tag{35}$$

Numerically, this formula worked for all different settings we have tried and, e.g., the numerical independence of z_p on i is already visible on the example in Figure 3.

Next, we try numerically optimizing p so that the infinity norm of $err_R(z)$ is minimized, i.e., we search for p that equioscilates the maximum of $err_R(z)$ on the left and on the right of z_p, and show the results in Figure 4. The relative improvement in the infinity norm of replacing the Dirichlet condition with the Robin one for that setting is well over a 5 fold.

[7]The subscript R stands for the "Robin" boundary condition at the end point $x = b$.

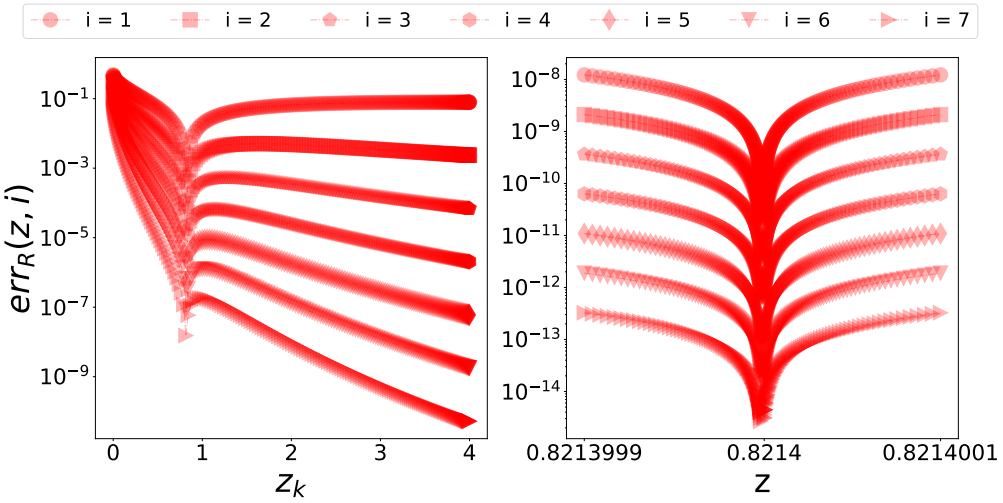

Figure 3: Left: plots of $err_R(z, i)$ at the points z_k (see (10)) evaluated for different number of grid columns i in $\Omega^b \setminus \Omega^a$ (see Remark 2), with $p = 200$, $N = 200$ and $\eta = 2$. Right: plots of the same functions under the same settings but zoomed in on the cusp (and thus plotted over artificial variables z rather then the eigenvalues z_k).

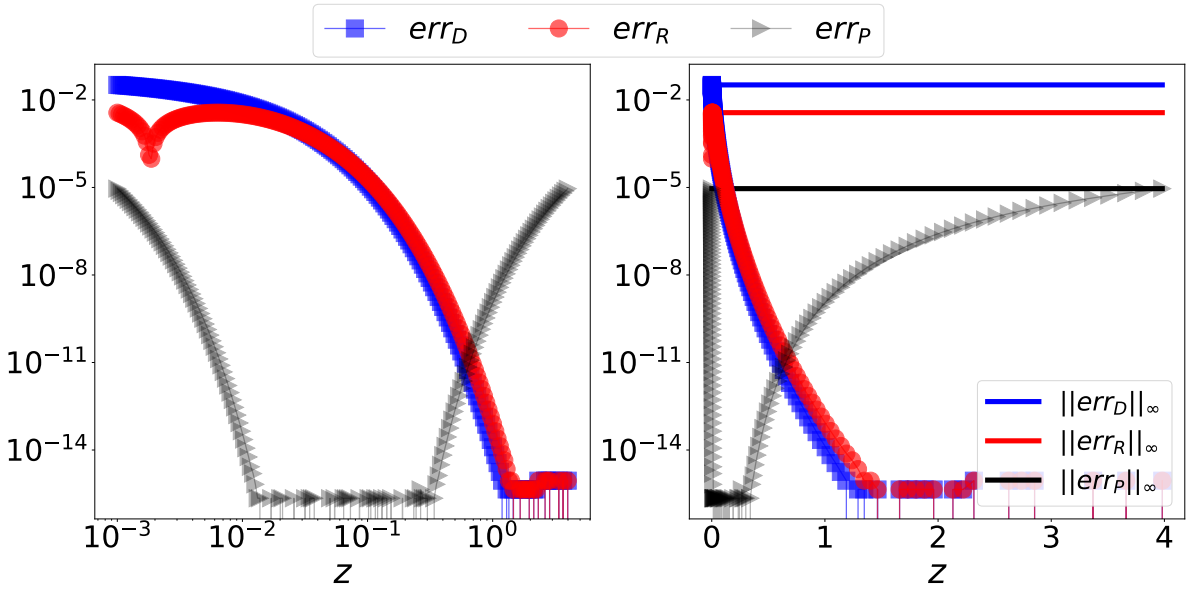

Figure 4: Left: minimization over p of the infinity norm of the Robin condition error, clearly showing the equioscillation. Right: optimized error compared with the corresponding Dirichlet condition error. We set $N = 200, i = 5$ and $\eta = 2$ and note that instead of z_k from (10) we take logarithmically equidistant z from the interval (30).

i	$p^*(i)$	$\frac{\|err_D\|_\infty}{\|err_R\|_\infty}$
1	27.4013	2.569
2	13.7783	3.924
4	8.2295	5.167
8	5.6016	6.598
16	4.3271	8.940

Table 1: Evolution of the optimized Robin parameter $p^*(i)$ depending on the number of layers i and the improvement ratio from the Dirichlet condition error to the Robin condition error in the infinity norm.

i	optimal z_0	$\frac{\|err_D\|_\infty}{\|err_P\|_\infty}$	$\frac{\|err_R\|_\infty}{\|err_P\|_\infty}$
1	0.4356	3.691	1.441
2	0.2101	10.091	2.572
4	0.1409	18.446	3.569
8	0.0932	86.163	13.058
16	0.0680	3595.822	402.186

Table 2: Evolution of the optimized expansion point z_0 depending on the number of layers i and the improvement ratio from the Dirichlet and Robin boundary condition error to the error of the approximation $\breve{t}^i_{z_0}$.

Running the optimization while varying i, i.e., the number of grid columns from a to b, we obtain Table 1, again for $N = 200$ and $\eta = 2$. We see that the improvement over the Dirichlet truncation increases with increasing number of layers. The corresponding results over a larger range of i are shown graphically in Figure 5.

In Figure 5 we varied i as powers of 2 from $2^1 = 2$ to $2^8 = 256$ on the left and then up to 2^{15} on the right and observe a linear dependence in the log-log scale on the left, i.e., for values $i \leq 256$, and fitting the line gives the law

$$p^*(i) \sim C \cdot i^q, \quad \text{with} \quad C \approx 11,\ q \approx -1. \tag{36}$$

The range $i \leq 256$ (and hence also the approximation (36)) in our eyes well covers the practically interesting values of i but it is clear that in general $p^*(i)$ does not follow the proposed relation (36).

Although the change and optimization of the Robin condition at $x = b$ offers a considerable improvement over the Dirichlet condition, we still observe for both of these the qualitatively identical behavior for large z. This is not particularly useful for our application as we simply want to minimize the error *only* over the interval (10). A possible way to do this is to *take the continued fraction and instead changing only the bottom-most denominator, change the entire sequence of numerators and denominator.* In this way we could *shift the expansion point of the Padé approximation* so that the error function no longer decreases around the point $z = +\infty$. At the same time, this will not only change the boundary condition at $x = b$ but in fact change the problem itself on the entire eliminated domain $x \in (a, b]$. We explore this direction further in the following section.

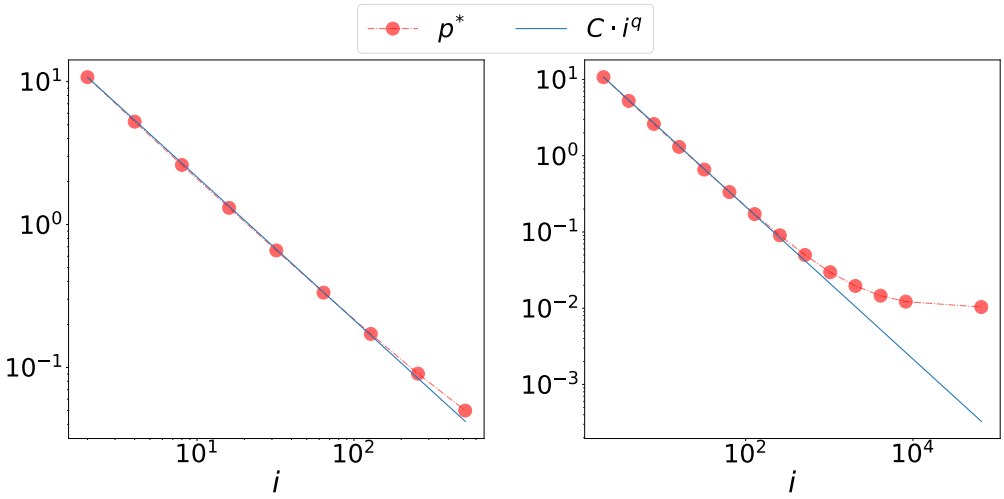

Figure 5: Dependence of the optimized Robin parameter $p^*(i)$ on the number of layers i added after a compared with the predicted behavior. The value i corresponds to the number of grid columns in $\Omega^b \setminus \Omega^a$, see Remark 2.

5 Shifting the Padé expansion point

Taking some $\alpha_0 > 0$ and denoting the new variable

$$\tilde{\alpha} := \frac{\alpha - \alpha_0}{1 + \alpha_0} \quad \text{and hence} \quad \alpha(\tilde{\alpha}) = \tilde{\alpha} \cdot (1 + \alpha_0) + \alpha_0, \tag{37}$$

a direct computation gives

$$\sqrt{1+\alpha} = \sqrt{1+\alpha_0}\sqrt{1+\tilde{\alpha}},$$

and expanding the right-hand side about 0 then relates to expanding the left-hand side about α_0. Using Theorem 3.5 we get

$$\sqrt{1+\alpha} = \sqrt{1+\alpha_0}\left(1 + \cfrac{\frac{\tilde{\alpha}}{2}}{1 + \cfrac{\frac{\tilde{\alpha}}{2}}{2 + \cfrac{\frac{\tilde{\alpha}}{2}}{1 + \cfrac{\frac{\tilde{\alpha}}{2}}{2+\dots}}}}}\right) = \sqrt{1+\alpha_0}\left(1 + \frac{\tilde{\alpha}}{2}\left(1 - \cfrac{1}{2 + \frac{4}{\tilde{\alpha}} - \cfrac{1}{2+\frac{4}{\tilde{\alpha}}-\dots}}\right)\right).$$

Notice that the equality is valid only for the *formal, infinite* continued fraction and once we truncate, the correspondence follows from Proposition 3.7. Setting $\breve{t}^{\infty}_{\alpha_0}(z)(\tilde{\alpha}) := \left(\hat{t}^{\infty} \circ \alpha\right)(\tilde{\alpha})$ we get

$$\breve{t}^{\infty}_{\alpha_0}(\tilde{\alpha}) = 1 + \frac{2}{\tilde{\alpha}(1+\alpha_0)+\alpha_0}\left(1 + \left(1 + \frac{\tilde{\alpha}}{2}\right)\sqrt{1+\alpha_0}\right) - \frac{2}{\tilde{\alpha}(1+\alpha_0)+\alpha_0} \cdot \frac{\tilde{\alpha}}{2} \cdot \cfrac{1}{2 + \frac{4}{\tilde{\alpha}} - \cfrac{1}{2+\frac{4}{\tilde{\alpha}}-\dots}}.$$

and based on Theorem 3.8 the truncation after i levels of $\breve{t}^{\infty}_{\alpha_0}$ results in the $[i+1, i+1]$-Padé approximant of $\hat{t}^{\infty}$ about α_0. We define $\breve{t}^{i}_{\alpha_0}(\tilde{\alpha})$ as

$$\breve{t}^{i}_{\alpha_0}(\tilde{\alpha}) := 1 + \frac{2}{\tilde{\alpha}(1+\alpha_0)+\alpha_0}\left(1+\left(1+\frac{\tilde{\alpha}}{2}\right)\sqrt{1+\alpha_0}\right) - \frac{2}{\tilde{\alpha}(1+\alpha_0)+\alpha_0}\cdot\underbrace{\frac{\tilde{\alpha}}{2}\cdot\frac{1}{2+\frac{4}{\tilde{\alpha}}-\frac{\ddots}{2+\frac{4}{\tilde{\alpha}}}}}_{i \text{ "levels"}},$$

and continue by focusing on the formulation of $\breve{t}^{i}_{\alpha_0}$ as a function of z rather than $\tilde{\alpha}$. Recalling the definition of $\tilde{\alpha}$ in (37) we have

$$z = \frac{4}{\alpha} = \frac{4}{\tilde{\alpha}(1+\frac{4}{z_0})+\frac{4}{z_0}},$$

obtaining

$$\tilde{\alpha} = \frac{4\frac{z_0}{z}-4}{4+z_0} \quad \text{and hence} \quad \frac{4}{\tilde{\alpha}} = \frac{4+z_0}{\frac{z_0}{z}-1}.$$

Without relabeling the function[8] we can write

$$\breve{t}^{i}_{z_0}(z) = 1 + \frac{z}{2}\left(1+\left(1+2\frac{\frac{z_0}{z}-1}{4+z_0}\right)\sqrt{1+\frac{4}{z_0}}\right) - \underbrace{\frac{1}{2+\frac{4+z_0}{\frac{z_0}{z}-1}-\cfrac{1}{\ddots\;\cfrac{}{2+\frac{4+z_0}{\frac{z_0}{z}-1}-\cfrac{1}{2+\frac{4+z_0}{\frac{z_0}{z}-1}}}}}}_{i \text{ "levels"}}. \tag{38}$$

and thereby define the error function $err_P(z,i)$ (P for Padé) by

$$err_P(z,i) := |\hat{t}^{\infty}(z) - \breve{t}^{i}_{z_0}(z)|.$$

The expectation is that the error function $err_P(z,i)$ should have one root at $z_0 = 4/\alpha_0$, which should get numerically more pronounced as i increases, in contrast to $err_R(z,i)$ and indeed, this is fully supported by the numerical results which we illustrate in Figure 6.

Again, we turn our attention to finding the optimal z_0, i.e., such that equioscilates the error on the left and on the right of z_0 and we present the results first in Figure 4, then we present the improvement when increasing i in Table 2 and finally in Figure 7 where we plot the evolution of the optimal z_0 as a function of i.

Figure 7. We can see that for $i \leq 64$ there seems to be a trajectory for the optimal choice of z_0, possibly convergent. But for i around 80 the error function becomes numerically equal to zero on the entire interval (10) and thus the optimization routine converges very close to or exactly at the initial guess, which was taken as 1.

We conclude this section by linking the above proposed approximation back to the physical problem and its solution methods by introducing a new PML technique that stems from the above approximation in the following section.

[8]However we do signal the variable by the expansion point in subscript from α_0 to z_0.

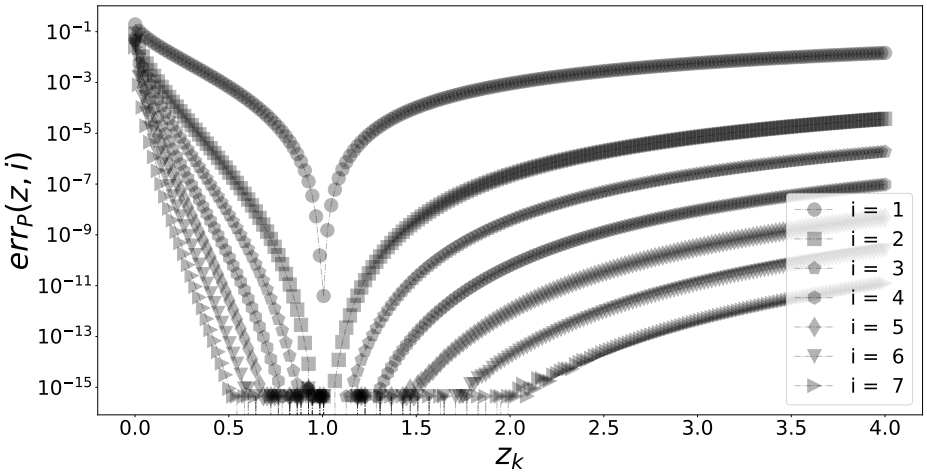

Figure 6: Plots of the function $err_P(z, i)$ at points equally spaced in the interval $[0, 4]$ evaluated for different values of i, for $\alpha_0 = 4$ (and thus $z_0 = 1$), $N = 200$ and $\eta = 2$. The value i corresponds to the number of grid columns in $\Omega^b \setminus \Omega^a$, see Remark 2.

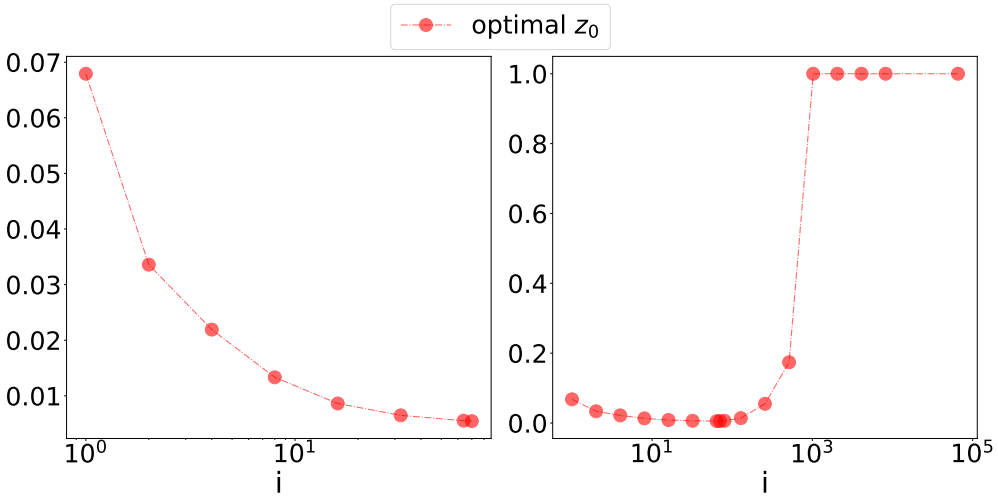

Figure 7: Dependence of the optimal choice of z_0 (and consequently $\alpha_0 = 4/z_0$) on the number of layers i added after a. We used again $N = 200$ and $\eta = 2$.

5.1 A new PML technique

Recalling the progress from (7) and (8) to (11), we need to move now in the opposite direction, starting with (38) and working up to the new block matrix we will call $\breve{A}^b$ and as $\breve{t}^i_{z_0}(z)$ is structurally identical J-fraction to $\vec{t}^b_i(z)$, we get

$$\breve{A}^b = \frac{1}{h^2}\begin{pmatrix} D_1 & -I_N & & & & \\ -I_N & \ddots & \ddots & & & \\ & \ddots & \breve{D}_{N^a} & -I_N & & \\ & & -I_N & \breve{D}_{N^a+1} & \ddots & \\ & & & \ddots & \ddots & -I_N \\ & & & & -I_N & \breve{D}_{N^b} \end{pmatrix}. \tag{39}$$

Recalling (10), we denote the eigenvalues of D_{yy} by μ_k, obtaining

$$z_k := \eta h^2 + \mu_k.$$

The point z_0 now translates to some μ_0 and the denominator of the cyclic part of the continued fraction $\breve{t}^i_{z_0}(z)$ in (38) becomes

$$2 + \frac{4+z_0}{\frac{z_0}{z_k}-1} = 2 + z\frac{4+z_0}{z_0-z_k} = 2 + \left(\eta h^2+\mu_k\right)\frac{4+\eta h^2+\mu_k}{\mu_0-\mu_k} \tag{40}$$

and hence we obtain $\breve{t}^i_{z_0}(z)$ as

$$\breve{t}^i_{z_0}(z) = 1 + \frac{\eta h^2+\mu_k}{2} + \frac{\eta h^2+\mu_k}{2}\sqrt{1+\frac{4}{\eta h^2+\mu_0}} + \frac{\mu_0-\mu_k}{\eta h^2+\mu_0+4}\sqrt{1+\frac{4}{\eta h^2+\mu_0}}. \tag{41}$$

Thereby we have $\breve{D}_{N^b} = \breve{D}_{N^b-1} = \ldots = \breve{D}_{N^a+1} \neq \breve{D}_{N^a}$ and for any $i = N^b, \cdots, N^a+1$ we must have

$$\begin{aligned}\breve{D}_i &= Q^T \begin{pmatrix} 2+z_1\frac{4+z_0}{\mu_0-\mu_1} & & \\ & \ddots & \\ & & 2+z_N\frac{4+z_0}{\mu_0-\mu_N} \end{pmatrix} Q \\ &= 2I + (4+\eta h^2+\mu_0)(D-2I)(\mu_0 I - D_{yy})^{-1},\end{aligned} \tag{42}$$

with $Z := \mathrm{diag}(z_1,\ldots,z_N)$ and Q given in (10)) and D being the diagonal block of the original problem, see (5). For $i = N^a$ we observe

$$\begin{aligned}\breve{D}_{N^a} &= Q^T \begin{pmatrix} 1+\frac{z_1}{2}+\frac{z_1}{2}\sqrt{1+\frac{4}{z_0}}+\frac{\mu_0-\mu_1}{z_0+4}\sqrt{1+\frac{4}{z_0}} & & \\ & \ddots & \\ & & 1+\frac{z_N}{2}+\frac{z_N}{2}\sqrt{1+\frac{4}{z_0}}+\frac{\mu_0-\mu_N}{z_0+4}\sqrt{1+\frac{4}{z_0}} \end{pmatrix} Q \\ &= D + \frac{\sqrt{1+\frac{4}{z_0}}}{2}Z + \frac{\sqrt{1+\frac{4}{z_0}}}{z_0+4}(\mu_0 I - D_{yy}).\end{aligned}$$

We finish this section with the following remark.

Remark 5 *The formula* (42) *contains an explicit inverse, which is clearly unpractical but can be easily avoided by multiplying the block-rows* $N^a+1,\dots,N^b-1,N^b$ *in* (39) *by the matrix* $M:=\mu_0 I-D_{yy}$*, obtaining*

$$\frac{1}{h^2}\begin{pmatrix} D_1 & -I_N & & & & & \\ -I_N & \ddots & \ddots & & & & \\ & \ddots & \bar{D}_{N^a} & -I_N & & & \\ & & -M & \bar{D}_{N^a+1}M & \ddots & & \\ & & & \ddots & \ddots & -M & \\ & & & & -M & \bar{D}_{N^b}M \end{pmatrix},$$

where no inverse of a matrix appears. An overall deeper understanding of $\breve{A}^b$ *and its continuous counterpart are clearly of interest and will be discussed in future work.*

6 Conclusion and future work

We proved for a model problem that truncation of the unbounded computational domain by a Dirichlet boundary conditions at a certain distance from the domain of interest is a spectral Padé approximation about infinity of the transparent boundary condition at the boundary of the domain of interest, and that the degree of the Padé approximation increases with the distance. We then replaced the Dirichlet truncation condition by a Robin truncation condition and showed that this greatly improves the behavior around a different point in the spectrum. We showed how to optimize the Robin parameter leading to an equioscillation property, but this is not a Padé approximation of the transparent boundary condition any more.

Aiming to obtain the Padé approximation about a different point we have proposed a different approximant in the eigenspace (leading to a new PML method for this problem), which poses a significant improvement over the Robin truncation. However, the theoretical proof of the approximation property is an open problem, which needs to be addressed properly on its own. We showed numerical results on the optimal choice of the parameter z_0, i.e., the shifted expansion point.

Recognizing we worked with a very particular problem, there are some straightforward generalizations. First, none of the computations required the particular choice of D in (5). As long as D is symmetric and positive-definite, all of the computations still work and the only change is in the interval of interest for the minimization of the Robin parameter p and the shifted expansion point z_0 in Section 4 and Section 5. This even holds if D is only symmetric, non-singular and with eigenvalues outside the interval $(-\infty,-1]$. If the spectrum intersected the interval $(-\infty,-1]$, the square root becomes a complex number and the computations move to the complex domain – in fact this is true for any diagonalizable

non-singular normal matrix D. The Helmholtz problem is the cannonical example and in fact a very similar technique was used to establish a similar result to Theorem 3.8 in [12]. If D is not normal, then the eigenvectors cannot be chosen to form an orthonormal basis of $\mathbb{R}^N$ (or $\mathbb{C}^N$) but the formulas would follow (based on the spectrum) one of the above mentioned cases in the same way, but one could not use the results directly. For example, the improvement factor would not be of immediate interest as the condition number of the eigenbasis would play an important role in computing the optimal Robin parameter p. If the matrix is diagonalizable and singular, then the modes corresponding to the zero eigenvalues do not admit the formulation of the function $\hat{t}_i^b(z)$ as in (18) but the analysis would work for the rest of the modes, based on the normality and spectrum of the matrix. In the case that the matrix is not diagonalizable, it is not immediately clear how to generalize any of the results based on the available Jordan form.

There are many further roads of exploration opened up by our approach: first, obtaining the asymptotic formulas for $p^*(i)$, z_p and z_0 as $h \to 0$ and, e.g., compare these with the known results in the optimized Schwarz methods. Nest, there is the open question about the nature of the approximation produced by the Robin truncation, which from the numerical results gives an error function with two roots. Last but not least, putting the above in the context of the work on the Zolotarev approximation in [14, 1] seems also beneficial. We intend to address these in a future work.

Finally, as we mentioned in Section 3, the three term recurrence (and thus the continued fraction formulation) has a deep, non-trivial connection with many other areas of mathematics, such as orthogonal polynomials, Gauss quadrature and the conjugate gradient method. Investigating this further would certainly be a worthwhile effort.

7 Acknowledgement

We would like to thank prof. Zdeněk Strakoš and the anonymous referees for their very useful comments and references to the literature.

References

[1] S. Asvadurov, V. Druskin, M. N. Guddati, and L. Knizhnerman. On optimal finite-difference approximation of PML. *SIAM Journal on Num. Anal.*, 41(1):287–305, 2003.

[2] G.A. Baker. Padé Approximants Part I: Basic theory. Addison-Wesley, 1981.

[3] A. Bayliss and E. Turkel. Radiation boundary conditions for wave-like equations. *Comm. Pure and Appl. Math.*, 33(6):707–725, 1980.

[4] J. P. Bérenger. A perfectly matched layer for the absorption of electromagnetic waves. *J. Comput. Phys.*, 114(2):185–200, 1994.

[5] M. Bernkopf. A history of infinite matrices. *Archive for History of Exact Sciences*, 4 (4):308–358, 1968.

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

[7] B. Engquist and A. Majda. Absorbing boundary conditions for the numerical simulation of waves. *Math. Comp.*, 31(139):629–651, 1977.

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

[9] M. J. Gander. Schwarz methods over the course of time. *Electron. Trans. Numer. Anal*, 31(5):228–255, 2008.

[10] M. J. Gander and H. Zhang. A class of iterative solvers for the Helmholtz equation: factorizations, sweeping preconditioners, source transfer, single layer potentials, polarized traces, and optimized schwarz methods. *SIAM Review*, 61(1):3–76, 2019.

[11] M.J. Gander and A. Schädle. The Pole condition: A Padé approximation of the Dirichlet to Neumann operator. In *Domain Decomposition Methods in Science and Engineering XIX, Lecture Notes in Computational Science and Engineering*. Springer-Verlag, 2010.

[12] M.J. Gander and A. Schädle. On the relationship between the pole condition, absorbing boundary conditions and perfectly matched layers. *In preparation*, 2016.

[13] M.J. Gander, L. Halpern, and F. Magoules. Analysis of patch substructuring methods. *Int. J. Appl. Math. Comput. Sci.*, 17(3):395–402, 2007.

[14] D. Ingerman, V. Druskin, and L. Knizhnerman. Optimal finite difference grids and rational approximations of the square root : I. Elliptic problems. *Communications on Pure and Applied Mathematics*, 53(8):1039–1066, 2000.

[15] J. Liesen and Z. Strakoš. *Krylov Subspace Methods: Principles and Analysis*. Oxford University Press, 2013.

[16] L. Lorentzen and H. Waadeland. *Continued Fractions with Applications*. North Holland, 1992.

[17] F. Magoulès, F.-X. Roux, and L. Series. Algebraic approximation of Dirichlet-to-Neumann maps for the equations of linear elasticity. *Comp. Meth. in Appl. Mech. and Eng.*, 195(29–32):3742–3759, 2006.

[18] F. Nataf, F. Rogier, and E. de Sturler. Optimal interface conditions for domain decomposition methods. *CMAP (Ecole Polytechnique)*, 301:1–18, 1994.

[19] F. Schmidt, T. Hohage, R. Klose, A. Schädle, and L. Zschiedrich. Pole condition: A numerical method for Helmholtz-type scattering problems with inhomogeneous exterior domain. *J. Comput. Appl. Math.*, 218(1):61–69, 2008.

[20] H. A. Schwarz. Über einen Grenzübergang durch alternierendes Verfahren. *Vierteljahrsschrift der Naturforschenden Gesellschaft in Zürich*, 15:272–286, 1870.

[21] A. Toselli and O. Widlund. *Domain Decomposition Methods - Algorithms and Theory*. Springer, 2004.

[22] H. S. Wall. *Analytic Theory of Continued Fractions*. Courier Dover Publ., 2018.