The Schur complement on a bounded domain is a spectral Padé approximation about infinity of the Schur complement on the unbounded domain

Martin J. Gander, Lukáš Jakabčin, Michal Outrata

The Schur complement on a bounded domain is a spectral Padé approximation about infinity of the Schur complement on the unbounded domain

Martin J. Gander[1], Lukáš Jakabčin[2] and Michal Outrata[1]

January 23, 2021

Abstract

We show for a specific 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. In particular, using Schur complement techniques, 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 boundary 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 written with respect to the eigenbasis can be identified with a truncation of a certain continued fraction. We use the theory of continued fractions to establish an approximation result of this truncation and hence interpreting the Schur complement as the Padé approximation of the optimal boundary operator in the eigenbasis. We then look to further 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 result.

1 Introduction

In order to numerically solve a problem on an unbounded domain, we need to truncate the domain to a finite size to perform computations. This domain truncation problem was first

[1]Section de Mathématiques, Université de Genève, 2-4 rue du Lièvre, CP 64, CH-1211 Genève. Email: martin.gander@unige.ch, michal.outrata@unige.ch

[2]Laboratoire de Mécanique et d'Acoustique, Université d'Aix-Marseille, Technopôle Château-Gombert, 4, impasse Nicola Tesla, CS 40006, 13453 MARSEILLE Cedex 13, Email: jakabcin@lma.cnrs-mrs.fr

studied in [7], where the authors introduced the so called absorbing boundary conditions (ABC) for wave propagation phenomena, see also [3]. A second major technique for the truncation of infinite domains are the perfectly matched layers (PML), presented in the seminal paper [4]. The pole condition, introduced in [19], is yet another new way to construct and study domain truncations and in [11, 12] the authors relate the pole condition technique, ABC and PML, obtaining an important understanding of all of these techniques.

At the discrete level, the ABC and PML techniques can be identified with techniques approximating the Schur complement in some sense. A number of iterative solvers has been derived based on this connection, 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. 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.

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. An iteration process is used to obtain better and better approximations of the true solution on the entire domain. 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 to enhance the convergence. Based on the insight from [18] this led to a new class of optimal and optimized Schwarz methods, see [8, 9] and references therein, and also the review on Schwarz methods [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 the patch method in [17, 13]. In this sense, the overlapping domain decomposition method is shown to be equivalent to a non-overlapping domain decomposition method with a particular ABC transmission condition. Although the Schwarz method is not explicitly mentioned in what follows, it is one of the main applications for our results.

We start in Section 2 by defining the model problem and its discrete counterpart. We recall the notion of the Schur complement for this problem in Section 3. Section 4 contains the main theoretical results of this paper, obtained by spectral analysis: we show 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. We next explore numerically how the spectral approximation changes when the Dirichlet condition is replaced by a Robin condition in Section 5, and present an optimized choice for the Robin parameter, and also a generalization in volume. We give concluding remarks and discuss possible extensions in Section 7.

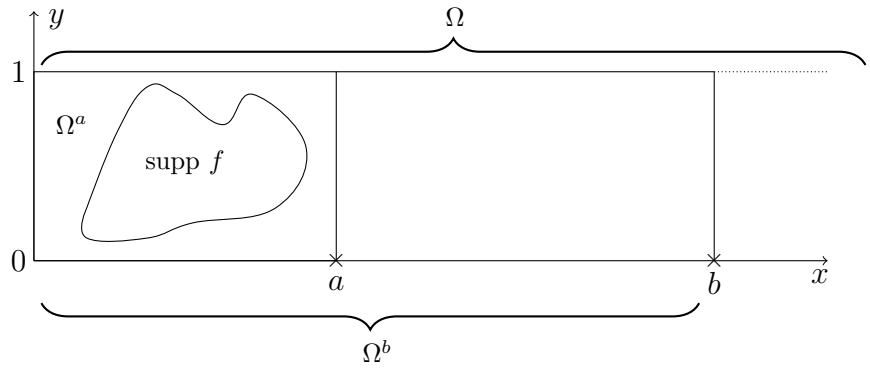

Figure 1: The unbounded strip domain in $\mathbb{R}^2$ with $\Omega = (0, +\infty) \times (0, 1)$.

2 Model Problem

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

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

where the domain $\Omega := (0, +\infty) \times (0, 1)$, see Figure 1. We assume that the support of the right-hand side function f is *localized*, namely

$$\text{supp} f \subset \Omega^a := (0, a) \times (0, 1) \subset \Omega.$$

Let $b \geq a$ and set $\Omega^b := (0, b) \times (0, 1) \subset \Omega$ as a larger region, containing Ω_a. Solving problem (1) on Ω^b and with homogeneous Dirichlet boundary condition at $x = b$ (i.e., $u(b, y) = 0$) gives a discrete approximation u^b of the true solution u on the unbounded domain Ω. Increasing b, the error of u^b as an approximation of u is decreasing.

Using a finite difference discretization of (1) on Ω^b with an equidistant mesh with mesh size $h := \frac{1}{N+1}$, the unknowns approximate the function values at the points of the grid. We can simplify the notation by gathering the unknowns with identical x coordinate into a vector of unknowns of length N. Assuming that the boundaries coincide with the mesh points, i.e.,

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

we obtain the system of linear equations

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

where $\boldsymbol{f}^b = [\boldsymbol{f}_1^T, \cdots, \boldsymbol{f}_{N^a}^T, \mathbf{0}^T, \cdots, \mathbf{0}^T]^T$, with $\boldsymbol{f}_i \in \mathbb{R}^N$ for $i = 1, 2, \ldots, N^a$ and $\mathbf{0} \in \mathbb{R}^N$. The

system matrix A^b can be written in the classical block tridiagonal form

$$A^b = \frac{1}{h^2}\begin{pmatrix} D_1 & -I_N & & & & \\ -I_N & \ddots & \ddots & & & \\ & \ddots & D_{N^a} & -I_N & & \\ & & -I_N & D_{N^a+1} & \ddots & \\ & & & \ddots & \ddots & -I_N \\ & & & & -I_N & D_{N^b} \end{pmatrix} \in \mathbb{R}^{(N\cdot N^b)\times(N\cdot N^b)}, \tag{4}$$

with I_N the identity matrix of size $N \times N$ and

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

In this notation, we can also identify the problem only on Ω^a, i.e.,

$$A^a \boldsymbol{u}^a = \boldsymbol{f}^a, \tag{6}$$

with $\boldsymbol{f}^a = [\boldsymbol{f}_1^T, \ldots, \boldsymbol{f}_{N^a}^T]^T$ and

$$A^a = \frac{1}{h^2}\begin{pmatrix} D_1 & -I_N & & \\ -I_N & \ddots & \ddots & \\ & \ddots & \ddots & -I_N \\ & & -I_N & D_{N^a} \end{pmatrix}. \tag{7}$$

Finally, we introduce also the discretization of the unbounded problem without domain truncation, i.e.,

$$A\boldsymbol{u} = \boldsymbol{f}, \tag{8}$$

where the solution as well as the right-hand side are vectors of infinite size, i.e.,

$$\begin{aligned} \boldsymbol{f} &= [\boldsymbol{f}_1^T, \cdots, \boldsymbol{f}_{N^a}^T, \boldsymbol{0}^T, \cdots, \boldsymbol{0}^T, \ldots]^T, \\ \boldsymbol{u} &= [\boldsymbol{u}_1^T, \cdots, \boldsymbol{u}_{N^a}^T, \boldsymbol{u}_{N^a+1}^T, \cdots, \boldsymbol{u}_{N^b}^T, \ldots]^T, \end{aligned}$$

and the system matrix A is also of infinite dimension, i.e.,

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

In what follows, it is enough to understand this infinite system simply as the limit of (4) as $b \to +\infty$; for more details on infinite matrices, see, e.g., the historical overview [5].

Because the data (the function f) vanishes outside of Ω^a we can formulate two new problems on Ω^a so that their solutions coincide with the solution of the problem (3) on Ω^b and with the solution of the problem (8) on Ω. This can be done both at the continuous level, using the Dirichlet-to-Neumann map (see, e.g., [10, Section 5.2]), and at the discrete level using the *Schur complement*. Here we choose the latter and introduce the Schur complement in the following section in more detail.

3 The Schur complement operator

To reduce the system (3) to a smaller one of the dimension of (6), we eliminate the variables $(\boldsymbol{u}^b_{N^a+1}, \ldots, \boldsymbol{u}^b_{N^b})$. Since those satisfy the equations in (3), i.e.,

$$\begin{aligned} -\frac{\boldsymbol{u}^b_{N^b-1}}{h^2} + \frac{D_{N^b}\boldsymbol{u}^b_{N^b}}{h^2} &= 0, \\ -\frac{\boldsymbol{u}^b_{i-1}}{h^2} + \frac{D_i\boldsymbol{u}^b_i}{h^2} - \frac{\boldsymbol{u}^b_{i+1}}{h^2} &= 0, \qquad \text{for } i = N^b-1, \ldots, N^a+1, \end{aligned} \tag{9}$$

we can recursively eliminate them. This is the well-known process of block Gaussian elimination and in the matrix formulation is known as the Schur complement approach.

Definition 3.1 (Schur complement) *The Schur complement $T^b_{N^a}$ is defined recursively by*

$$\begin{aligned} T^b_{N^b} &= \frac{D_{N^b}}{h^2} = \frac{D}{h^2}, \\ T^b_i &= \frac{D_i}{h^2} - \frac{(T^b_{i+1})^{-1}}{h^4} = \frac{D}{h^2} - \frac{(T^b_{i+1})^{-1}}{h^4}, \qquad \textit{for } i = N^b-1, \ldots, N^a. \end{aligned} \tag{10}$$

Using the Schur complement $T^b_{N^a}$ and the particular zero structure of the right-hand side in (3), we can reduce the problem (3) to a problem on Ω^a, namely

$$\tilde{A}^a \tilde{\boldsymbol{u}}^a = \boldsymbol{f}^a, \tag{11}$$

where

$$\tilde{A}^a = \begin{pmatrix} \frac{D_1}{h^2} & -\frac{\mathbb{I}_N}{h^2} & & \\ -\frac{\mathbb{I}_N}{h^2} & \ddots & \ddots & \\ & \ddots & \frac{D_{N^a-1}}{h^2} & -\frac{\mathbb{I}_N}{h^2} \\ & & -\frac{\mathbb{I}_N}{h^2} & T^b_{N^a} \end{pmatrix}. \tag{12}$$

The only difference between (6) and (11) is the last block in (12) 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 $\Omega^b \backslash \Omega^a$ that have been folded in. Since the solution $\tilde{\boldsymbol{u}}^a$ of the modified system (11) coincides with the solution $\tilde{\mathbf{u}}^b$ of (3) restricted to the points

in Ω^a, the smaller system gives a better approximation of the underlying solution $\boldsymbol{u}$ on the infinite domain compared to the solution of (6).

To further improve the solution, we can move the boundary point b to the right to some $\tilde{b} > b$ (and thus having an $N^{\tilde{b}} > N^b$) and hence, the Schur complement matrix $T^{\tilde{b}}_{N^a}$ will be defined by a longer recurrence (see Definition 3.1). If b goes to infinity, the corresponding Schur complement matrix $T^{\infty}_{N^a}$ will be governed by the limit of (10), namely

$$T^{\infty}_{N^a} = \frac{D}{h^2} - \frac{(T^{\infty}_{N^a})^{-1}}{h^4}, \tag{13}$$

which does not depend on N^a. Thus, $T^{\infty}_{N^a} \equiv T^{\infty}$ for any N^a, where T^{∞} satisfies the same equation (13), i.e., one has

$$T^{\infty} = \frac{D}{h^2} - \frac{(T^{\infty})^{-1}}{h^4}. \tag{14}$$

Using T^{∞} instead of $T^b_{N^a}$ in (12), the corresponding solution $\tilde{\boldsymbol{u}}^a$ will coincide with $\boldsymbol{u}$ on the points in Ω^a. Equation (14) can be reformulated into a quadratic matrix equation for T^{∞} given by

$$(T^{\infty})^2 - \frac{D}{h^2}T^{\infty} + \frac{I}{h^4} = 0. \tag{15}$$

This equation has two solutions, each of them corresponding to a different underlying solution $\boldsymbol{u}$. But only one of these corresponds to a bounded solution $\boldsymbol{u}$, which is of interest. In order to solve (15), it is convenient to change the basis we work in to the eigenbasis of D, which, in this case, corresponds to the discrete sine basis[1]. We present the change of the basis and its implications for the computation in the following section.

4 Spectral analysis

The eigenpairs of the matrix D in (5) can be evaluated by writing $D = D_{yy} + 2\mathbb{I}$, where D_{yy} is the 3-point finite difference stencil discretization of $\eta - \partial_{yy}$ multiplied by h^2. The eigenpairs of D_{yy} are known in closed form, i.e., $D_{yy} = Q^T Z Q$ with

$$Z := \operatorname{diag}(z_1, \ldots, z_N), \quad z_k := \eta h^2 + 4\sin^2\left(\frac{k\pi}{2(N+1)}\right), \tag{16}$$

and Q unitary and symmetric, with the eigenvectors $\boldsymbol{q}_k$ in its columns,

$$\boldsymbol{q}_k := \left[\sqrt{\tfrac{2}{N}}\sin\left(\tfrac{k\pi}{N+1}j\right)\right]_{j=1}^{N} = \sqrt{\frac{2}{N+1}}\begin{bmatrix}\sin\left(\frac{k\pi}{N+1}\right)\\ \vdots \\ \sin\left(\frac{k\pi}{N+1}N\right)\end{bmatrix} \in \mathbb{R}^N. \tag{17}$$

Thus, we can diagonalize D with the same basis, i.e.,

$$D = Q^T \Lambda Q$$

[1]It would be possible to do the analysis that follows for a much more general PDE, but for simplicity and clarity, we use our model problem (1) throughout; more comments can be found in the concluding Section 7.

with eigenvalues given by

$$\Lambda = \mathrm{diag}(\lambda_1, \ldots, \lambda_N) \qquad \text{and} \qquad \lambda_k = 2 + z_k. \tag{18}$$

For the rest of the text we fix the index notation so that the index $k = 1, \ldots, N$ corresponds to a particular eigenmode (or frequency or Fourier mode in different terminologies).

Remark 1 *Changing the basis so that the diagonal blocks of the matrices A, A^b, A^a and also $\tilde{A}^a$ diagonalize will on the one hand make the Schur complement analysis significantly easier - we will be able to treat each eigenmode separately - but on the other hand more complicated - treating each eigenmode separately adds yet another index to the already loaded notation. Throughout the text we keep the index k reserved for the eigenmode notation.*

Although it is in our eyes not possible to simplify the notation while avoiding confusion, we recognize the difficulties associated with the notation of the following subsection.

4.1 Diagonalization of the Schur Complement

Changing the basis for the Schur complement definition in (10) gives

$$\begin{aligned} \hat{T}^b_{N^b} &= Q T^b_{N^b} Q^T = Q \frac{D}{h^2} Q^T = \frac{\Lambda}{h^2}, \\ \hat{T}^b_i &= Q T^b_i Q^T = Q \frac{D}{h^2} Q^T - Q \frac{(T^b_{i+1})^{-1}}{h^4} Q^T = \frac{\Lambda}{h^2} - \frac{(\hat{T}^b_{i+1})^{-1}}{h^4}, \end{aligned} \tag{19}$$

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 mode (frequency) also follows the analogous recurrence, namely

$$\begin{aligned} \hat{t}^b_{N^b,k} &= \frac{\lambda_k}{h^2}, \\ \hat{t}^b_{i,k} &= \frac{\lambda_k}{h^2} - \frac{1}{h^4 \hat{t}^b_{i+1,k}}, \qquad \text{for } i = N^b + 1, \ldots, N^a. \end{aligned} \tag{20}$$

Moreover, we can also write an analogue of the above recurrence for the recurrence we establish for the solution u^b in (9). In particular, having

$$\begin{aligned} -\frac{\boldsymbol{u}^b_{N^b-1}}{h^2} + \frac{Q^T \Lambda Q \boldsymbol{u}^b_{N^b}}{h^2} &= -\frac{\boldsymbol{u}^b_{N^b-1}}{h^2} + Q^T \hat{T}^b_{N^b} Q \boldsymbol{u}^b_{N^b} = 0, \\ -\frac{\boldsymbol{u}^b_{i-1}}{h^2} + \frac{Q^T \Lambda Q \boldsymbol{u}^b_i}{h^2} - \frac{\boldsymbol{u}^b_{i+1}}{h^2} &= -\frac{\boldsymbol{u}^b_{i-1}}{h^2} + \frac{Q^T \hat{T}^b_i Q \boldsymbol{u}^b_i}{h^2} = 0, \end{aligned} \tag{21}$$

with $i = N^b - 1, \ldots, N^a$, setting $\hat{\boldsymbol{u}}^b_i := Q \boldsymbol{u}^b_i$ and multiplying (21) on the left by Q, we obtain

$$\begin{aligned} -\frac{\hat{\boldsymbol{u}}^b_{N^b-1}}{h^2} + \frac{\Lambda \hat{\boldsymbol{u}}^b_{N^b}}{h^2} &= -\frac{\hat{\boldsymbol{u}}^b_{N^b-1}}{h^2} + \hat{T}^b_{N^b} \hat{\boldsymbol{u}}^b_{N^b} = 0, \\ -\frac{\hat{\boldsymbol{u}}^b_{i-1}}{h^2} + \frac{\Lambda \hat{\boldsymbol{u}}^b_i}{h^2} - \frac{\hat{\boldsymbol{u}}^b_{i+1}}{h^2} &= -\frac{\hat{\boldsymbol{u}}^b_{i-1}}{h^2} + \frac{\hat{T}^b_i \hat{\boldsymbol{u}}^b_i}{h^2} = 0, \end{aligned} \tag{22}$$

where $i = N^b + 1, \ldots, N^a$.

In the following section we use the above recurrences to analyze the the Schur complement $T^b_{N^a}$ and its limit as $b \to +\infty$. Notice that in the context of Section 1, obtaining the limit $\lim_{b\to+\infty} T^b_{N^a}$ allows us to construct $\tilde{\boldsymbol{u}}^a$ such that it coincides with $\boldsymbol{u}$ (restricted to Ω^a unknowns) and thus to obtain the best possible approximation in this framework.

4.2 Convergence of the Schur Complement

Focusing on the limit case $\lim_{b\to+\infty} T^b_{N^a}$, (19 – 20) implies that we can compute the limit for each eigenmode independently, meaning we can focus on N scalar limits

$$\lim_{b\to+\infty} \hat{t}^b_{N^a,k} =: \hat{t}^\infty_{N^a,k}, \tag{23}$$

for $k = 1, \ldots, N$. Since (20) holds for any b fixed, the limit $\hat{t}^\infty_{N^a,k}$ satisfies the same equation, meaning we have

$$(\hat{t}^\infty_{N^a,k})^2 - \frac{\lambda_k}{h^2}\hat{t}^\infty_{N^a,k} + \frac{1}{h^4} = 0. \tag{24}$$

Since the equation does not depend on N^a, neither will the solutions, i.e., the limits, and thus we can omit the N^a subscript in (23), obtaining $\hat{t}^\infty_{N^a,k} \equiv \hat{t}^\infty_k$. Returning to equation (24), there are two solutions,

$$\hat{\tau}^{\infty,1}_k = \frac{\lambda_k + \sqrt{\lambda_k^2 - 4}}{2h^2} \quad \text{and} \quad \hat{\tau}^{\infty,2}_k = \frac{\lambda_k - \sqrt{\lambda_k^2 - 4}}{2h^2}, \tag{25}$$

and using the definition of λ_k in (18), both solutions are real and positive for all $k = 1, \ldots, N$. Taking any k and using the Vieta formulas for (24), we have

$$(h^2\hat{\tau}^{\infty,1}_k)(h^2\hat{\tau}^{\infty,2}_k) = 1, \tag{26}$$

and, moreover,

$$0 < h^2\hat{\tau}^{\infty,2}_k < 1 < h^2\hat{\tau}^{\infty,1}_k. \tag{27}$$

The next step is to show that one of the solutions $\hat{\tau}^{\infty,1}_k, \hat{\tau}^{\infty,2}_k$ 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 (22) is preserved through the limit process and thus the solutions $\hat{\tau}^{\infty,1}_k, \hat{\tau}^{\infty,2}_k$ of the limit equation (24) coincide with the roots of the characteristic polynomial

$$p_k(r) = -r^2 + \lambda_k r - 1.$$

This together with the explicit formula for the solution of the recurrence relation (22) is enough to establish the following result.

Theorem 4.1 *The Schur complement $T^b_{N^a}$ defined in* (10) *converges to $T^{\infty,1}$, i.e., the solution of the formal limit equation* (15)*, as $b \to +\infty$. To be more specific, the eigenvectors of those matrices are equal and the eigenvalues of $\hat{T}^b_{N^a}$ converge to the ones of $\hat{T}^{\infty,1}$, i.e.,*

$$\hat{t}^{\infty}_k \equiv \lim_{N^b\to\infty} \hat{t}^b_{N^a,k} = \hat{\tau}^{\infty,1}_k = \frac{\lambda_k + \sqrt{\lambda_k^2 - 4}}{2h^2}, \tag{28}$$

where $\lambda_k = \eta h^2 + 2 + 4\sin^2\left(\frac{hk\pi}{2}\right)$, see (18)*.*

Proof Fixing b, we take the solution subvector $\hat{\boldsymbol{u}}^b_i$ of length N for any $i = N^a, \ldots, N^b$ and denote its scalar entries by $\hat{u}^b_{i,k}$ for $k = 1, \ldots, N$. These entries follow the system of difference equations in (22) and as such they can be written as a linear combination of powers of the roots of the characteristic polynomial of the difference equation. In other words, there exist two constants ν^b_k and μ^b_k such that the solution $\hat{u}^b_{i,k}$ of (22) is given by

$$\hat{u}^b_{i,k} = \mu^b_k \left(h^2\hat{\tau}^{\infty,1}_k\right)^{i-N^a} + \nu^b_k \left(h^2\hat{\tau}^{\infty,2}_k\right)^{i-N^a}, \tag{29}$$

where the constants μ^b_k, ν^b_k depend on N^b and N^a. Recalling the inequality of the roots $\hat{\tau}^{\infty,1}_k, \hat{\tau}^{\infty,2}_k$ in (27) it follows that

$$\left(h^2\hat{\tau}^{\infty,1}_k\right)^{N^b-N^a} \to +\infty.$$

Since we look for the limit of solutions satisfying a homogeneous Dirichlet boundary condition at $x = b$, the limit has to decay at $+\infty$. Thus the bounded solution of the form (29) has to satisfy

$$\mu^b_k \to 0 \quad \text{as} \quad b \to +\infty.$$

Hence, setting ν^{∞}_k as the limit of ν^b_k as $b \to +\infty$ we obtain

$$\hat{u}^{\infty}_{i,k} = \nu^{\infty}_k \left(h^2\hat{\tau}^{\infty,2}_k\right)^{i-N^a}. \tag{30}$$

Returning to the system of difference equations (22), the scalar equations become

$$-\hat{u}^b_{N^a,k} + \lambda_k\hat{u}^b_{N^a+1,k} - \hat{u}^b_{N^a+2,k} = -\hat{u}^b_{N^a,k} + h^2\hat{t}^b_{N^a+1,k}\hat{u}^b_{N^a+1,k} = 0,$$

and solving for $h^2\hat{t}^b_{N^a+1,k}$, we get

$$h^2\hat{t}^b_{N^a+1,k} = \frac{\hat{u}^b_{N^a,k}}{\hat{u}^b_{N^a+1,k}}.$$

Inserting the values of $\hat{u}^b_{i,k}$ from (30), we find

$$\lim_{N^b\to\infty} h^2\hat{t}^b_{N^a+1,k} = \frac{\lim_{N^b\to\infty}\hat{u}^b_{N^a,k}}{\lim_{N^b\to\infty}\hat{u}^b_{N^a+1,k}} = \frac{\hat{u}^{\infty}_{N^a,k}}{\hat{u}^{\infty}_{N^a+1,k}} = \frac{\nu^{\infty}_k\left(h^2\hat{\tau}^{\infty,2}_k\right)^{i-N^a}}{\nu^{\infty}_k\left(h^2\hat{\tau}^{\infty,2}_k\right)^{i-N^a+1}} = \frac{1}{h^2\hat{\tau}^{\infty,2}_k} = h^2\hat{\tau}^{\infty,1}_k,$$

where the last step follows from (26), finishing the proof. □

Remark 2 *All of the computations starting at the beginning of this subsection with* (23) *and finishing by Theorem 4.1 were focusing on computations for the full Schur complement, i.e., the full recurrence in* (20)*. However, starting in* (23) *with* N^a+1, N^a+2 *or any fixed* N^a+n*, all of the computations remain the same. This confirms that indeed the limit result is independent of* N^a *as the same result holds also if we stop the recurrence elimination process one, two or any other fixed number of steps earlier (that is before reaching the domain of interest* Ω^a*).*

The limit $\hat{t}_k^\infty$ can be written in a more convenient form, using the notation $\lambda_k = z_k + 2$ from (18). A direct computation gives

$$\begin{aligned}\hat{t}_k^\infty &= \frac{\lambda_k + \sqrt{\lambda_k^2 - 4}}{2h^2} = \frac{2 + z_k + \sqrt{(2+z_k)^2 - 4}}{2h^2} = \frac{1}{h^2} + \frac{z_k}{2h^2} + \frac{\sqrt{z_k^2 + 4z_k}}{2h^2} \\ &= \frac{1}{h^2} + \frac{z_k}{2h^2} + \frac{\sqrt{z_k^2(1+\frac{4}{z_k})}}{2h^2} = \frac{1}{h^2} + \frac{z_k}{2h^2} + \frac{z_k}{2h^2}\sqrt{1+\frac{4}{z_k}},\end{aligned} \tag{31}$$

and thus, for any $k = 1, .., N$, the limit $\hat{t}_k^\infty$ can be written in the form

$$\hat{t}_k^\infty = \frac{1}{h^2}\left(1 + \frac{z_k}{2}\left(1 + \sqrt{1 + \frac{4}{z_k}}\right)\right), \tag{32}$$

with $z_k := \eta h^2 + 4\sin^2(hk\pi/2)$. Each limit mode $\hat{t}_k^\infty$ can thus be seen as a function of an argument z,

$$\hat{t}^\infty(z) = \frac{1}{h^2}\left(1 + \frac{z}{2}\left(1 + \sqrt{1 + \frac{4}{z}}\right)\right), \tag{33}$$

evaluated at $z = z_k \in [\eta h^2 + 4\sin^2(h\pi/2), \eta h^2 + 4\sin^2(hN\pi/2)]$. In the same way, the recurrence relation (19) for $\hat{t}_{i,k}^b$ can be interpreted as a function evaluation at the points $z = z_k$,

$$\hat{t}_{i,k}^b = \hat{t}_i^b(z_k) \quad \text{for } i = N^a, \ldots, N^b.$$

The first three functions are

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

and by the recursive definition in (20), the general term is

$$\hat{t}_i^b(z) = \frac{2+z}{h^2} - \cfrac{\frac{1}{h^2}}{2+z-\cfrac{1}{2+z-\ddots-\frac{1}{2+z}}}. \tag{34}$$

Objects 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 4.4 later on. In the light of this observation we adjust our notation, clarifying everything in Remark 3 below.

Remark 3 *Notice that in the continued fraction representation of $\hat{t}_i^b(z)$ in* (34)*, 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*

$$\begin{aligned}
\hat{t}_0^b(z) &= \frac{2+z}{h^2}, \\
\hat{t}_1^b(z) &= \frac{2+z}{h^2} - \frac{1}{h^4 \frac{2+z}{h^2}} = \frac{1}{h^2}\left(2+z-\frac{1}{2+z}\right), \\
\hat{t}_2^b(z) &= \frac{2+z}{h^2} - \frac{1}{h^4 \hat{t}_{N^b-1}^b(z)} = \frac{2+z}{h^2} - \frac{1}{h^4 \frac{1}{h^2}\left(2+z-\frac{1}{2+z}\right)} \\
&= \frac{1}{h^2}\left(2+z-\frac{1}{2+z-\frac{1}{2+z}}\right),
\end{aligned}$$

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 $\Omega^b \setminus \Omega^a$. Whenever i increases, we understand it as* extension of the domain Ω^b*, rather than the mesh size changing.*

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

Remark 4 *By a direct computation we obtain*

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

and by re-insertion we also get

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

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

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

and the functions $\hat{t}_i^b(z)$ in (34) *are approximations in the sense of a truncation of this infinite continued fraction after i levels.*

Since the continued fractions are not the primary focus, which is to say that 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 with Section 4.3 where we recall some standard terminology and results of the field and give some auxiliary lemmas in the following subsection.

4.3 Padé Approximation and Continued Fractions

We start by recalling the Padé approximation theory. We follow the notation from [2], i.e., the $[M/L]$-Padé approximant of $f(z)$ is denoted by $[M/L]_f \equiv [M/L]_f(z)$.

Theorem 4.2 ([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 $m \geq 1$ be fixed and assume $f(z) = \sum_{j=0}^{+\infty} c_j z^j$ to be a formal power series (we do not consider the convergence question here). 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. *Assume $f(0) \neq 0$ and set $g(z) = 1/f(z)$. Then $[M/L]_g(z) = 1/[L/M]_f(z)$.*

We continue by introducing the basic terminology of continued fractions following [2, Chapter 4].

Definition 4.3 *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*

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

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*

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

with $k_j \neq 0$ for all j then it is called J-fraction *(short for Jacobi continued fraction).*

Next, we give some remarks on the definition above. First, let us emphasize that we have ignored the questions of convergence of infinite continued fractions and we refer the reader to [16] and [22]. Also, notice that one function can be represented by two seemingly different continued fractions (different possibly in type and/or in the coefficient values) and one way to recognize the equality of two continued fractions is via the *three-term recurrence relation* of the numerators and denominators of the continued fraction truncations (convergents), see [2, Theorem 4.1.1, pp.106]. We have that

$$\begin{aligned} A_{-1} &= 1, \quad A_0 = b_0, \quad A_n = b_n A_{n-1} + a_n A_{n-2}, \\ B_{-1} &= 0, \quad B_0 = 1, \quad B_n = b_n B_{n-1} + a_n B_{n-2}. \end{aligned} \tag{35}$$

and assuming the n-th truncation (convergent) of two continued fractions are equal for any n, the infinite continued fractions are equal as well. 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 to [16] and [22] and references therein. Last but not least, we want to 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 4.4 ([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 4.3. 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]. We state the result in Theorem 4.5, referencing 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 derivation we refer the reader to [22, Chapter XVIII] or [16, Chapter VI].

Theorem 4.5 ([2, Section 4.6, Theorem 4.4.3 and formula (6.4) on pp. 139]) *For any $\alpha \in (-1, +\infty)$[3] we have*

$$\sqrt{1+\alpha} = 1 + \cfrac{\frac{\alpha}{2}}{1 + \cfrac{\frac{\alpha}{2}}{2 + \cfrac{\frac{\alpha}{2}}{1 + \cfrac{\frac{\alpha}{2}}{\ddots \; b_{n-2} + \cfrac{a_n}{b_{n-1} + \cfrac{a_{n+1}}{b_n + \cfrac{}{\cdots}}}}}}}} = 1 + \cfrac{\frac{\alpha}{2}}{1} + \cfrac{\frac{\alpha}{2}}{2} + \cfrac{\frac{\alpha}{2}}{1} + \cdots + \cfrac{a_n}{b_n} + \cdots, \tag{36}$$

where the denominator sequence is given by $b_0 = 1$, $b_j = \frac{3+(-1)^j}{2}$ and the numerator sequence is given by $a_j = \frac{\alpha}{2}$, $j \geq 1$.

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 (36) 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 (36).

Remark 5 *By a direct computation we see that*

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

and thus the above representation in (36) can be equivalently written as a cyclic S-fraction[4] with $a_j = 1/2$ for all j.

We finish this subsection by proving some preparatory results, the first of which will be useful in linking a truncation of the S-fraction introduced in Theorem 4.5 and a truncation of the J-fraction from Remark 4. However, notice that the continued fraction considered there is not identical with the one in (36) – they obey the same recurrence but they differ at the beginning.

Lemma 4.6 *Let α be real and consider the continued fraction*

$$\tau(\alpha) := \cfrac{\frac{\alpha}{2}}{2 + \cfrac{\frac{\alpha}{2}}{1 + \cfrac{\frac{\alpha}{2}}{2 + \cfrac{\frac{\alpha}{2}}{1 + \cdots}}}}$$

[2]Although we do not use them, there are many ways to create a continued fraction representation of a function based on its formal power series. We do not consider these here in more detail but rather refer the interested reader to [2, Section 4.2, 4.4, 4.5], [16, Chapter V] or [22, Part II].

[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 such that the sequences $\{a_j\}, \{b_j\}$ are periodic are called cyclic continued fractions.

and denote its n-th truncation by $A_n(\alpha)/B_n(\alpha)$. Considering the J-fraction

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

with n-th truncation $C_n(\alpha)/D_n(\alpha)$ we have

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

for any $n = 1, 2, \dots$.

Proof We start by transforming the continued fraction τ by the rules of Lemma 4.4 and without further relabeling we obtain

$$\tau(\alpha) := \cfrac{1}{\frac{4}{\alpha}+\cfrac{1}{1+\cfrac{1}{\frac{4}{\alpha}+\cfrac{1}{1+\cfrac{1}{\frac{4}{\alpha}+\cfrac{1}{1+\dots}}}}}}. \tag{37}$$

First, by a direct computation, we confirm the equality for $n = 1$, obtaining

$$\frac{1}{\frac{4}{\alpha}+\frac{1}{1}} = \frac{1}{\frac{4}{\alpha}+1},$$

and next, we notice that the continued fraction (37) can be written in a cyclic form with *the core* R given by

$$R = \frac{4}{\alpha} + \frac{1}{1+\frac{1}{R}}. \tag{38}$$

That is, the continued fraction can be obtained by a successive re-insertion of the core equality (38) into itself, e.g.,

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

Notice that in this way every re-insertion adds two elements of the numerator and denominator sequences as highlighted by the indices of the truncations. Using the algebraic identity

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

we reformulate the core equality (38) to obtain

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

or, more conveniently,

$$1 + R = 2 + \frac{4}{\alpha} - \frac{1}{1+R}. \tag{39}$$

Notice that using the algebraic identity above we *contracted* the two-level core that is to be re-inserted to only a one-level core, while not changing the resulting value of the fraction. Moreover, the core equality in (39) is the one that generates the J-fraction $\sigma(\alpha)$.

Hence we have shown that for any $n \geq 2$ the $2n$ re-insertions of the core R in the core equality (38) is equal to only n re-insertions of the core $1 + R$ in the equality (39). It follows that the $2n$-th convergent $A_{2n}(z)/B_{2n}(z)$ is equal to the n-th convergent $C_n(z)/D_n(z)$, yielding the result. □

We build upon Lemma 4.6 with the following result that contracts the S-fraction in (36) into a J-fraction.

Proposition 4.7 *Let α be real and set the continued fractions $\tau(\alpha)$ and $\sigma(\alpha)$ as in Lemma 4.6. Moreover, we define the continued fractions*

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

with n-th truncations $\tilde{A}_n(\alpha)/\tilde{B}_n(\alpha)$ and $E_n(\alpha)/F_n(\alpha)$ respectively. Then

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

for any $n = 0, 1, 2, \dots$ with $E_0 = F_0 = 1$.

Proof The equality for $n = 0$ holds by inspection. Taking $n \geq 1$, we use Lemma 4.6 for the continued fraction $\tilde{\tau}(\alpha)$ and we 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 it remains to show that

$$\frac{1}{1 + C_n(\alpha)/D_n(\alpha)} = 1 - E_n(\alpha)/F_n(\alpha), \tag{40}$$

where $C_n(\alpha), D_n(\alpha)$ are the truncations of the J-fraction $\sigma(\alpha)$ from Lemma 4.6. We first notice that the cyclic parts of both J-fractions $\sigma(\alpha)$ and $\phi(\alpha)$ coincide and we denote them by $\tilde{\sigma}(\alpha)$,

$$\tilde{\sigma}(\alpha) := \frac{1}{2 + \frac{4}{\alpha} - \frac{1}{2+\frac{4}{\alpha}-\dots}} \tag{41}$$

We then notice that

$$\sigma(\alpha) = \frac{1}{1 + \frac{1}{1+\frac{4}{\alpha} - \tilde{\sigma}(\alpha)}},$$

and

$$\phi(\alpha) = 1 - \frac{1}{2 + \frac{4}{\alpha} - \tilde{\sigma}(\alpha)}.$$

Therefore, in order to show (40) it is enough to show that

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

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

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

and

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

finishing the proof. □

4.4 Approximation Properties of the Schur Complement

We now show that the function $\hat{t}_i^b(z)$ representing the Schur complements T_i^b is a Padé approximation about $z = +\infty$ of the function $\hat{t}^\infty(z)$ representing the infinite Schur complement T^∞. To obtain this result in Theorem 4.8 we use a similar technique as in [11] where the authors compute a Padé approximation of the Dirichlet to Neumann operator. This similarity is not a coincidence: the Schur complement and the Dirichlet-to-Neumann map have a very deep connection, see, e.g., [10, Section 5.2].

Also, in the proof and in the rest of the document we change the variables back to the ones from our application, i.e.,

$$\alpha = \frac{4}{z}. \tag{42}$$

Notice that the term change of variables here is somewhat misleading. Usually, change of variables in approximation theory then requires a re-computation of the approximation – because of the way derivation of a composite function works. However, our result considers expansion about $+\infty$, which is defined by considering the expansion of the same function about zero but of the reciprocal argument, e.g., not of z but of $1/z$. Recalling this convention, our change of variables in fact consists only of multiplying by 4. Also, notice that such change of variables does not require a re-computation of the derivatives.

Theorem 4.8 *The function $\hat{t}_i^b(z)$ defined in* (34) *as*

$$\hat{t}_i^b(z) = \frac{1}{h^2} \left(2 + z - \cfrac{1}{2 + z - \cfrac{1}{2 + z - \ddots - \cfrac{1}{2 + z}}} \right)$$

is the $[i,i]$-Padé approximation about the expansion point $z=+\infty$ of the Schur complement function $\hat{t}^{\infty}(z)$ defined in (33) *as*

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

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 (42) and without a further relabeling of the functions we obtain

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

$$\hat{t}^{\infty}(\alpha)=1+\frac{2}{\alpha}+\frac{2}{\alpha}\sqrt{1+\alpha}.$$

Recalling (26), we have

$$\{\hat{t}^{\infty}\}^{-1}(\alpha):=\frac{1}{\hat{t}^{\infty}(\alpha)}=1+\frac{2}{\alpha}-\frac{2}{\alpha}\sqrt{1+\alpha} \tag{43}$$

and thus by Theorem 4.2 part 3., we have

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

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 by application of Theorem 4.2 parts 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 of the Padé approximant from Theorem 4.5, we obtain

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

where $A_{2i+1}(\alpha), B_{2i+1}(\alpha)$ are the $(2i+1)$-st truncation numerator and denominator of the continued fraction $\tau(\alpha)$ in (36) and the sequences $\{a_j\}_j, \{b_j\}_j$ are taken as in (36). Hence we have

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

and recalling Proposition 4.7 we notice that the continued fraction is the $(2i+1)$-st truncation of the continued fraction $\tilde{\tau}(\alpha)$ defined there. Therefore, Proposition 4.7 gives that

$$[i/i]_{\{\hat{t}^\infty\}^{-1}}(\alpha) = 1 - \left(1 - \frac{\tilde{C}_n(\alpha)}{\tilde{D}_n(\alpha)}\right) = \underbrace{\cfrac{1}{2 + \frac{4}{\alpha} - \cfrac{1}{2+\frac{4}{\alpha} - \cfrac{1}{\ddots - \cfrac{1}{2+\frac{4}{\alpha}}}}}}_{i-1 \text{ ``levels''}},$$

where $\tilde{C}_i(\alpha), \tilde{D}_i(\alpha)$ are the $(2i+1)$-st truncation numerator and denominator of the J-fraction $\tilde{\sigma}(\alpha)$ in (41) in Proposition 4.7. As a result we have that

$$[i/i]_{\hat{t}^\infty}(\alpha) = \cfrac{1}{2+\frac{4}{\alpha} - \cfrac{1}{2+\frac{4}{\alpha} - \cfrac{1}{\ddots - \cfrac{1}{2+\frac{4}{\alpha}}}}} = 2 + \frac{4}{\alpha} - \underbrace{\cfrac{1}{2 + \frac{4}{\alpha} - \cfrac{1}{2+\frac{4}{\alpha} - \cfrac{1}{\ddots - \cfrac{1}{2+\frac{4}{\alpha}}}}}}_{i-1 \text{ ``levels''}}$$

for any $i \geq 1$, finishing the proof. □

Theorem 4.8 allows us to quantify the error[5]

$$err_D(z, i) := \hat{t}^\infty(z) - \hat{t}^b_{N^a+i}(z),$$

where i denotes the number of grid columns that were folded into the Schur complement, i.e., the number of layers between a and b, and thus also the degree of the Padé approximation. We show plots of the function $err_D(z, i)$ for small i in Figure 2.

As expected, we see that the error $err_D(z, i)$ decreases when more grid columns are added, i.e., when i and thus also b increases. Also, for any fixed i, the error is decreasing when z is increasing. This behavior was established for $z \to +\infty$ by Theorem 4.8.

On the other hand, the error is still quite large for small z. Recalling the role of z, this can be identified with poor quality of the approximation for low frequency modes following from our spectral analysis. Reducing the maximum of the error would yield a better approximation of the true solution $\mathbf{u}$ by the approximation $\tilde{\mathbf{u}}^a$, see (11). We explore this direction in the next section.

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

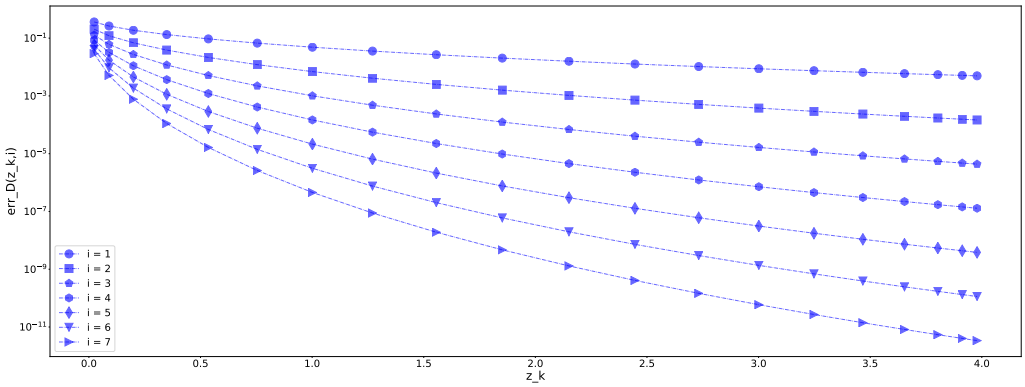

Figure 2: Plots of the function $err_D(z, i)$ at the points z_k with the parameters set to $N = 20$ and $\eta = 2$. The value i corresponds to the number of grid columns in $\Omega^b \setminus \Omega^a$, see Remark 3.

5 Robin boundary condition for truncation

Theorem 4.8 shows that the choice of the Dirichlet boundary condition at b corresponds to truncating the continued fraction interpretation, and leads to a the Padé approximation of $\hat{t}^\infty(z)$ about the point $z = +\infty$. Writing the interval containing the spectrum of D_{yy} as

$$\left[\eta h^2 + 4\sin^2\left(\frac{\pi}{2}\frac{1}{N+1}\right), \eta h^2 + 4\sin^2\left(\frac{\pi}{2}\frac{N}{N+1}\right)\right] \approx \left[\eta h^2, \eta h^2 + 4\right], \tag{44}$$

the expansion point $z = +\infty$ of the Padé approximation is far away from the interval where we want to approximate $\hat{t}^\infty$, i.e., this expansion point is not well-chosen and far from optimal. An idea is to replace the Dirichlet boundary condition used to truncate the recurrence relation by a homogeneous Robin boundary condition[6] with Robin parameter $p \geq 0$ at b, i.e.,

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

The Dirichlet condition is the limit of the Robin condition as $p \to +\infty$. Taking any finite positive p will change the approximation error function in Figure 2 and can give a better approximation of $\hat{t}^\infty(z)$ in the interval (44), as we show next by introducing a discretization.

Using a centered finite difference approximation as before, the Robin condition can be discretized with the so-called *ghost point* trick: first, we discretize the Robin condition,

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

Then in the discretized equation at b, $-\boldsymbol{u}_{N^b+1} + D_{N^b}\boldsymbol{u}_{N^b} - \boldsymbol{u}_{N^b-1} = 0$, we can eliminate the unknowns $\boldsymbol{u}_{N^b+1}$ – the ghost points – to get $(D_{N^b} + 2hp\mathbb{I})\boldsymbol{u}_{N^b} - 2\boldsymbol{u}_{N^b-1} = 0$, and thus the

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

modified matrix A^b becomes

$$\tilde{A}^b = \frac{1}{h^2}\begin{pmatrix} D_1 & -\mathbb{I}_N & & & & \\ -\mathbb{I}_N & \ddots & \ddots & & & \\ & \ddots & D_{N^a} & -\mathbb{I}_N & & \\ & & -\mathbb{I}_N & D_{N^a+1} & \ddots & \\ & & & \ddots & \ddots & -\mathbb{I}_N \\ & & & & -\mathbb{I}_N & \tilde{D}_{N^b} \end{pmatrix} \tag{45}$$

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

$$\begin{aligned} \tilde{T}^b_{N^b} &= \frac{\tilde{D}_{N^b}}{2h^2}, \quad \text{with the diagonal entries } \tilde{t}^b_{N^b,k} = \frac{\lambda_k}{2h^2} + \frac{p}{h}, \\ \tilde{T}_i &= \frac{D_i}{h^2} - \frac{\tilde{T}^{-1}_{i+1}}{h^4} = \frac{D}{h^2} - \frac{\tilde{T}^{-1}_{i+1}}{h^4}, \quad \text{for } i = N^b - 1, \ldots, N^a. \end{aligned} \tag{46}$$

The first three functions representing the diagonal entries are

$$\begin{aligned} \tilde{t}^b_0(z) &= \frac{1 + ph + \frac{z}{2}}{h^2}, \\ \tilde{t}^b_1(z) &= \frac{2+z}{h^2} - \frac{1}{h^4 \frac{1+ph+\frac{z}{2}}{h^2}}, = \frac{1}{h^2}\left(2 + z - \frac{1}{1 + ph + \frac{z}{2}}\right), \\ \tilde{t}^b_2(z) &= \frac{2+z}{h^2} - \frac{1}{h^4 \tilde{t}^b_{N^b-1}(z)} = \frac{2+z}{h^2} - \frac{1}{h^4 \frac{1}{h^2}\left(2 + z - \frac{1}{1+ph+\frac{z}{2}}\right)} \\ &= \frac{1}{h^2}\left(2 + z - \frac{1}{2 + z - \frac{1}{1+ph+\frac{z}{2}}}\right), \end{aligned}$$

where we keep the subscript notation (see Remark 3) and by the recursive definition in (20), the general form is

$$\tilde{t}^b_i(z) = \frac{2+z}{h^2} - \cfrac{\frac{1}{h^2}}{2 + z - \cfrac{1}{2+z-\cfrac{\ddots}{2+z-\cfrac{1}{1+ph+\frac{z}{2}}}}}. \tag{47}$$

Notice that by letting $p \to +\infty$ we can recover the original Dirichlet boundary condition with one less level of the continued fraction. In other words, having h fixed and the Robin parameter p large enough, we can interpret this as a Dirichlet boundary condition in a geometry where Ω^b was extended by one additional grid column after b, i.e., by the strip $(b, b+h) \times (0, 1)$.

We see that the Robin condition just changes the last denominator in the continued fraction representation. With (47), we can numerically explore the effect of the Robin

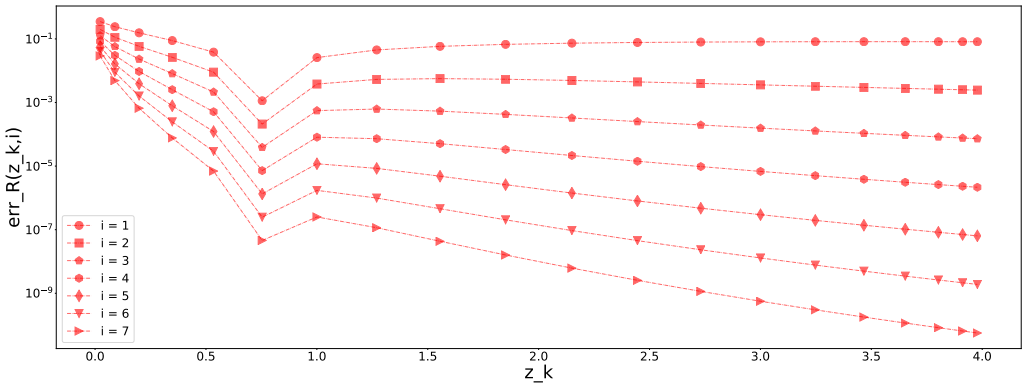

Figure 3: Plots of the function $err_R(z, i)$ at the points z_k evaluated for different values of i, for $p = 20$, $N = 20$ and $\eta = 2$. The value i corresponds to the number of grid columns in $\Omega^b \setminus \Omega^a$, see Remark 3.

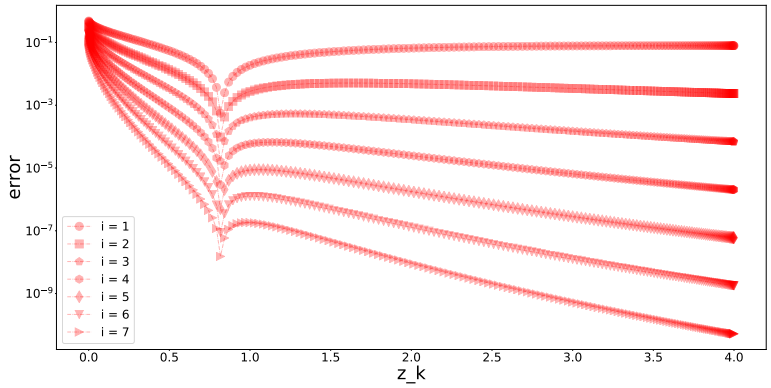

Figure 4: Plots of the function $err_R(z, i)$ at the points z_k evaluated for different values of i, for $p = 200$, $N = 200$ and $\eta = 2$. The value i corresponds to the number of grid columns in $\Omega^b \setminus \Omega^a$, see Remark 3.

parameter p on the behavior of the error[7]

$$err_R(z, i) := \hat{t}^{\infty}(z) - \tilde{t}_i^b(z),$$

where i denotes again the number of grid columns that were folded into the Schur complement, i.e., the number of layers added after the point a. Similarly to Figure 2 we show the evolution of the error function $err_R(z, i)$ in Figure 3.

We see that the behavior around the right endpoint of the interval is still present but the Robin condition introduced another point z_p around which the approximation is accurate. In Figure 3 we see that for $z_p \approx 0.75$. To confirm the hypothesis, we use a finer mesh in z in Figure 4.

This confirms that the Robin parameter p affects the approximation error by minimizing it around a particular point z_p. Shifting z_p towards 0, i.e., towards the area where both

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

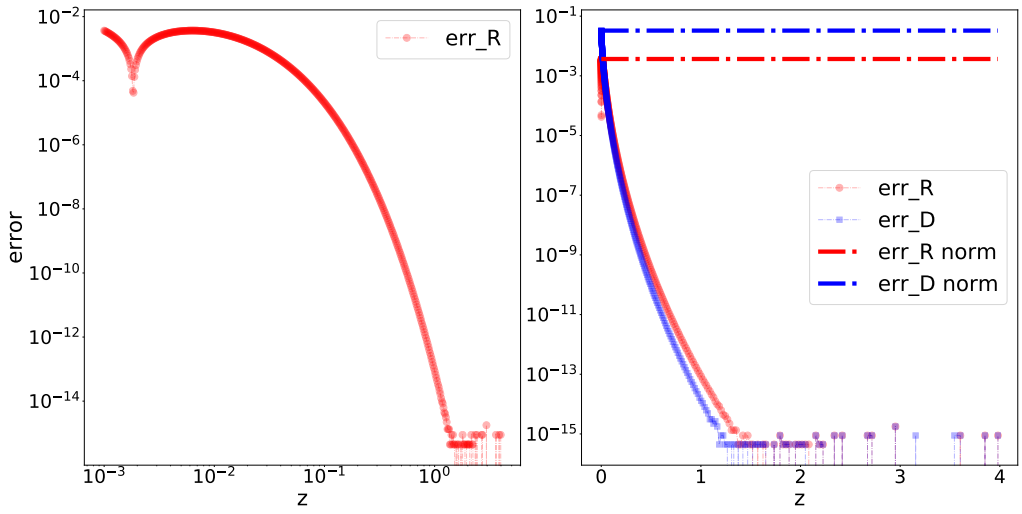

Figure 5: Left: minimization over p of the infinity norm of the Robin condition error, clearly showing the equioscillation. Right: optimized error compared with the corresponding Dirichlet condition error. We set $N = 200, i = 5$ and $\eta = 2$. The value i corresponds to the number of grid columns in $\Omega^b \setminus \Omega^a$, see Remark 3.

$err_D(z, i)$ and $err_R(z, i)$ attain their maximum, could thus reduce the overall maximum of the error. By definition, z_p is a solution of the equation

$$err_R(z, i) = 0 \tag{48}$$

for z with a given p. Assuming that $err_R(z, i)$ is smooth except at a finite number of points, which is the case based on the numerical experiments above, equation (48) defines z_p as an implicit function of p and also the other parameters of the problem. For example, 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

$$z_p^2 + 4z_p - 4p^2h^2 = 0 \quad \Longrightarrow \quad z_p = -2 + 2\sqrt{1 + p^2h^2}. \tag{49}$$

Based on the observation in Figure 4, varying the number of layers i does not substantially change the value of z_p.

In order to optimize the parameter p in the Robin condition, one can minimize the infinity norm of the error in Figures 3 and/or 4. Using an optimization routine to minimize the infinity norm of the error over p, the local maxima equioscillate, see Figure 5.

Here we did not use the natural values of z_k but rather chose to span the entire interval with logarithmically equidistant points. We used $N = 200$, $i = 5$, $\eta = 2$, and the optimal

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. The value i corresponds to the number of grid columns in $\Omega^b \setminus \Omega^a$, see Remark 3.

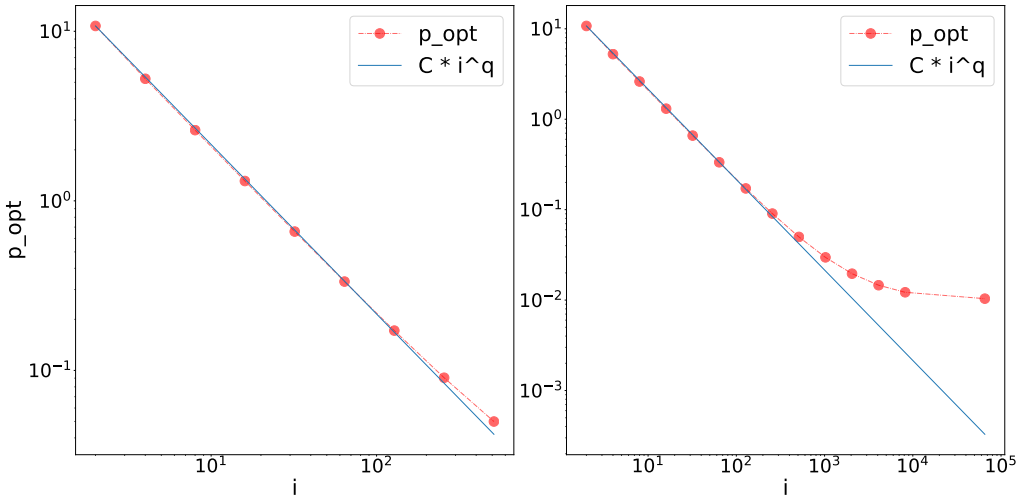

Figure 6: Dependence of the optimized Robin parameter $p^*(i)$ on the number of layers i added after a compared with the predicted behavior. The value i corresponds to the number of grid columns in $\Omega^b \setminus \Omega^a$, see Remark 3.

value of p was found to be $p = 7.167$. The Dirichlet error in the infinity norm is equal to 0.1371 while the Robin error in the infinity norm is equal to 0.0245, an almost 5.6 fold improvement.

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 $\eta = 2$.

This shows 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 6.

Again we minimized the maximum norm $\|err_R(z, i)\|_\infty$ over the interval in (44). We used again $N = 200$, $\eta = 2$ and 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. One can see 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

$$p^*(i) \sim C \cdot i^q \tag{50}$$

with

$$C \approx 11 \qquad \text{and} \qquad q \approx -1$$

for $i \leq 256$. For practical purposes having $i > 256$ is not desirable and thus the above law basically covers a sufficient domain of i. However, Figure 6 shows that in general $p^*(i)$ does not follow the proposed relation (50).

Although we have achieved considerable improvement over the Dirichlet boundary condition, the Robin boundary condition still could be improved. Observing Figure 4 the Robin boundary condition error function still decreases for large z and it seems that it inherited most of the approximation qualities around $z = +\infty$. This is, however, not useful for our application as we simply want to minimize the error *only* over the interval spanned by the eigenvalues, i.e., over $[\eta h^2, \eta h^2 + 4]$. A possible way to do this is to *shift the expansion point of the Padé approximation* - an approach discussed in the following section.

6 Shifting the Padé expansion point

We start with shifting the expansion point of $\sqrt{1+\alpha}$. Taking some $\alpha_0 > 0$, a direct computation gives

$$\sqrt{1+\alpha} = \sqrt{1+\alpha_0}\sqrt{1+\frac{\alpha-\alpha_0}{1+\alpha_0}},$$

and denoting

$$\tilde{\alpha} := \frac{\alpha-\alpha_0}{1+\alpha_0},$$

we can write

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

where $\tilde{\alpha}$ is small for values of α around[8] α_0. Thus, using Theorem 4.5 we obtain the continued fraction representation

$$\sqrt{1+\alpha} = \sqrt{1+\alpha_0}\left(1+\cfrac{\frac{\tilde{\alpha}}{2}}{1+\cfrac{\frac{\tilde{\alpha}}{2}}{2+\cfrac{\frac{\tilde{\alpha}}{2}}{1+\cfrac{\frac{\tilde{\alpha}}{2}}{2+\ddots}}}}\right),$$

and analogously to the proof of Theorem 4.8, we combine Lemma 4.6 and Proposition 4.7 to obtain the J-fraction representation

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

[8]Meaning we do not consider α, α_0 such that $|\alpha - \alpha_0| > 1$ in order to ensure the argument of the square root being positive.

Let us emphasize here that the equality is valid only for the *infinite* continued fraction and once we truncate, the correspondence is again as established in Proposition 4.7.

Similarly, we have that

$$\alpha = \tilde{\alpha} \cdot (1 + \alpha_0) + \alpha_0, \tag{51}$$

so that we can rewrite the function $\hat{t}^{\infty}(z)$ as

$$\begin{aligned}
\hat{t}^{\infty}(\alpha) &= 1 + \frac{2}{\alpha} + \frac{2}{\alpha}\sqrt{1+\alpha} \\
&= 1 + \frac{2}{\tilde{\alpha}(1+\alpha_0)+\alpha_0} + \frac{2}{\tilde{\alpha}(1+\alpha_0)+\alpha_0}\sqrt{1+\alpha_0}\cdot\sqrt{1+\tilde{\alpha}} \\
&= 1 + \frac{2}{\tilde{\alpha}(1+\alpha_0)+\alpha_0}\left(1 + \left(1+\frac{\tilde{\alpha}}{2}\right)\sqrt{1+\alpha_0}\right) \\
&\quad - \frac{2}{\tilde{\alpha}(1+\alpha_0)+\alpha_0}\cdot\frac{\tilde{\alpha}}{2}\cdot\frac{1}{2+\frac{4}{\tilde{\alpha}}-\frac{1}{2+\frac{4}{\tilde{\alpha}}-\ldots}}
\end{aligned}$$

and set

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

If we want to approximate $\hat{t}^{\infty}$ around α_0 it is intuitive to take some truncation of the continued fraction in $\bar{t}^{\infty}_{\alpha_0}$ and based on Theorem 4.8 the expectation is that taking i levels will result in an $[i+1, i+1]$-Padé approximant of $\hat{t}^{\infty}$ around α_0.

Following the development in Section 5, we define the approximation

$$\bar{t}^{i}_{\alpha_0}(\tilde{\alpha}) := 1 + \frac{2}{\tilde{\alpha}(1+\alpha_0)+\alpha_0}\left(1 + \left(1+\frac{\tilde{\alpha}}{2}\right)\sqrt{1+\alpha_0}\right) - \frac{2}{\tilde{\alpha}(1+\alpha_0)+\alpha_0}\cdot\underbrace{\frac{\tilde{\alpha}}{2}\cdot\frac{1}{2+\frac{4}{\tilde{\alpha}}-\frac{\ddots}{2+\frac{4}{\tilde{\alpha}}}}}_{i \text{ ``levels''}},$$

and continue by focusing on the formulation of $\bar{t}^{i}_{\alpha_0}$ as a function of z rather than $\tilde{\alpha}$. Recalling the definition of $\tilde{\alpha}$ in (51) we have

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

obtaining

$$\tilde{\alpha} = \frac{4\frac{z_0}{z} - 4}{4 + z_0},$$

and hence

$$\frac{4}{\tilde{\alpha}} = \frac{4 + z_0}{\frac{z_0}{z} - 1}.$$

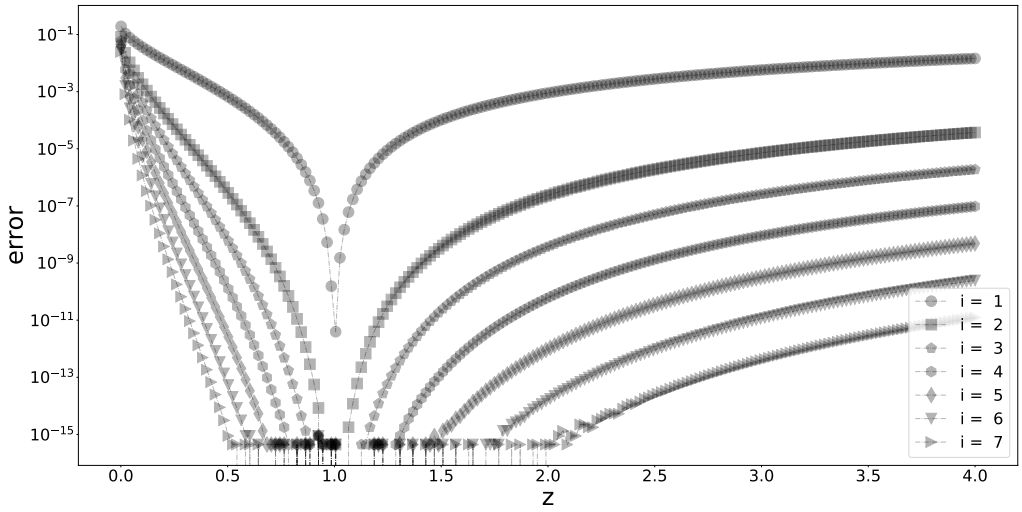

Figure 7: Plots of the function $err_P(z,i)$ at points equally spaced in the interval $[0,4]$ evaluated for different values of i, for $\alpha_0 = 4$ (and thus $z_0 = 1$), $N = 200$ and $\eta = 2$. The value i corresponds to the number of grid columns in $\Omega^b \setminus \Omega^a$, see Remark 3.

Without relabeling the function[9] we can write

$$\bar{t}^{\,i}_{z_0}(z) = 1 + \frac{z}{2}\left(1 + \left(1 + 2\frac{\frac{z_0}{z} - 1}{4 + z_0}\right)\sqrt{1 + \frac{4}{z_0}}\right) - \underbrace{\cfrac{1}{2 + \cfrac{4+z_0}{\frac{z_0}{z}-1} - \cfrac{1}{\ddots \; - \cfrac{1}{2 + \cfrac{4+z_0}{\frac{z_0}{z}-1} - \cfrac{1}{2 + \cfrac{4+z_0}{\frac{z_0}{z}-1}}}}}}_{i \text{ “levels”}}. \tag{52}$$

Hence we can write the error function $err_P(z,i)$ (P for Padé) as

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

The expectation is that the error function $err_P(z,i)$ should have one root at $z_0 = 4/\alpha_0$, which should get more pronounced as i increases. Indeed, the numerical results shown in Figure 7 support this fully.
However, as Remark below emphasizes, this doesn't constitute the result yet.

Remark 6 *The results above do not prove that the function $\bar{t}^{\,i}_{z_0}$ is a Padé approximant of either $\bar{t}^{\infty}_{z_0}$ or $\hat{t}^{\infty}$ about the point z_0 (and then analogously in the α variable) because the construction took into account only the approximant of $\sqrt{1+\tilde{\alpha}}$ in the α domain but neglected the rest of the function $\bar{t}^{\infty}_{\alpha_0}$. However, the numerical results suggest that in spite of that the function $\bar{t}^{\,i}_{z_0}$ has good approximation qualities.*

[9]However we do change the expansion point α_0 to z_0 in the subscript.

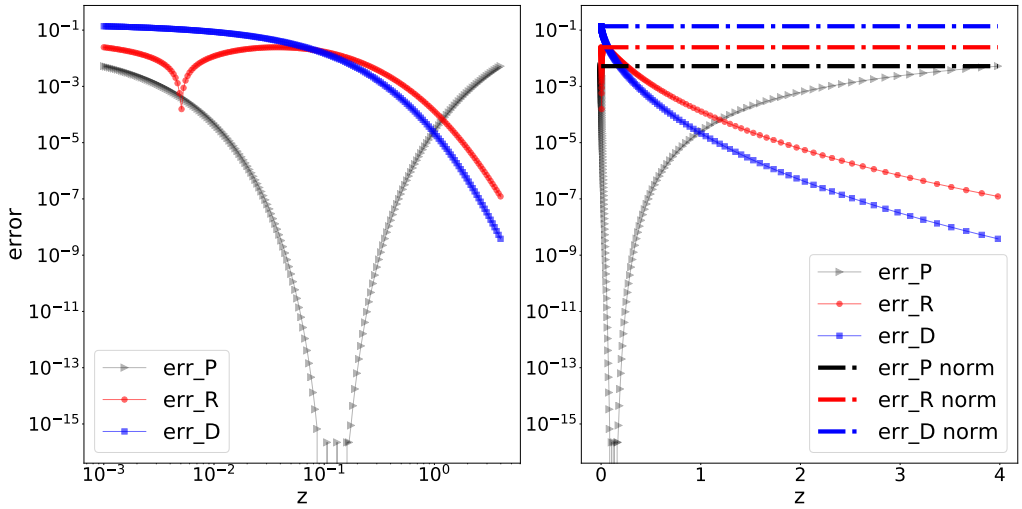

Figure 8: Left: minimization over z_0 (or equivalently over α_0) of the infinity norm of the error function err_P, clearly showing the equi-oscillation. Right: optimized error compared with the error functions err_R for the optimal p^* and the error function err_D. We set $N = 200, i = 5$ and $\eta = 2$. The value i corresponds to the number of grid columns in $\Omega^b \setminus \Omega^a$, see Remark 3.

We end this section by comparing the error function $err_P(z, i)$ to the other ones, i.e., to $err_D(z, i)$ and $err_R(z, i)$. First, we present the equi-oscillation result in Figure 8, analogously to Figure 5.

We again did not use the natural values of z_k but rather chose to span the entire interval with logarithmically equidistant points on the left and equidistant points on the right. We used $N = 200$, $i = 5$, $\eta = 2$, obtaining the optimal value of z_0 to be $z_0 \approx 0.12$ (the optimal value of α_0 amounts to $\alpha_0 \approx 32.57$). The Dirichlet error in the infinity norm is equal to 0.1371 the Robin error in the infinity norm is equal to 0.0245 while the error of the "Shifted Padé" (see Remark 6 is equal to 0.0052, an almost 4.7 fold improvement over the optimal Robin parameter p^* approximation and an overall 26.2 fold improvement over the original Dirichlet approximation.

Running the optimization while varying i, i.e., the number of grid columns from after a we obtain Table 2 (where we keep $N = 200$ and $\eta = 2$).

This shows that the improvement over both of the previous options increases with increasing number of layers and does so quite rapidly.

The corresponding results over a larger range of i are shown graphically in Figure 9.

Again we minimized the maximum norm $\|err_P(z, i)\|_\infty$ over the interval in (44). We used again $N = 200$, $\eta = 2$ and varied i as powers of 2 from $2^0 = 1$ to $2^8 = 256$ on the left and then up to 2^{15} on the right. We can see that for $i \leq 64$ there seems to be a trajectory for the optimal choice of z_0. The evolution afterwards is caused by the finite precision. For $i \geq 80$ the optimization routine (we used the `Nelder-Mead` algorithm) reaches ϵ_{mach} on the entire interval $[\eta h^2, 4 + \eta h^2]$ seemingly for any z_0 beneath a certain threshold. As we increase the value of i this threshold increases as well up to a point where the initial guess $z_0 = 1$ already

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 $\bar{t}^i_{z_0}$. The value i corresponds to the number of grid columns in $\Omega^b \setminus \Omega^a$, see Remark 3.

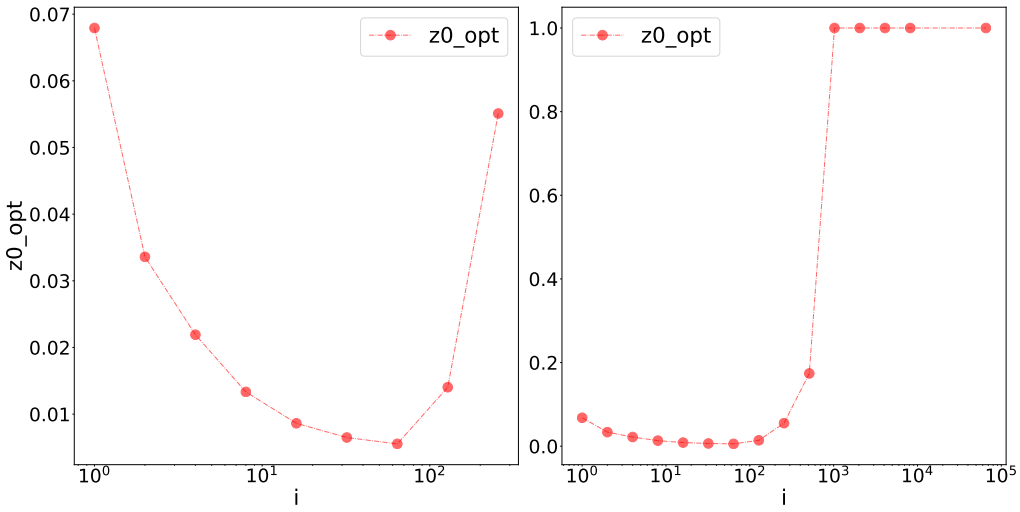

Figure 9: Dependence of the optimal choice of z_0 (and consequently $\alpha_0 = 4/z_0$) on the number of layers i added after a.

suffices.

If the approximant $\bar{t}^i_{z_0}$ would be indeed the $[i+1, i+1]$-Padé approximant around $z_0 = 4/\alpha_0$ (see Remark 6), then this behavior was to be expected from some value of i onward. The fact that the optimal "expansion point" z_0 tends towards the left endpoint was also expected as we have seen that the error is largest around that point.

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.

6.1 A new PML technique

Although the function $\bar{t}^i_{z_0}(z)$ in (52) may not be the Shifted Padé approximant the numerical results suggest that it still leads to a very good approximation and by the construction we will be able to propose good and very accurate PML compared to the previously proposed ones in this paper.

Recalling the definition of A^b in (4), the definition of the Schur complement in Definition 3.1 and the process of its transformation into a continued fraction, we need to reverse this process (starting with the eigenvalue reccurences and working up to the block matrix A^b). This requires returning to z_k from the artificial z and linking the formulas containing the variables z_k to the blocks in A^b and thus obtain the PML method.

First, recalling

$$z_k := \eta h^2 + 4\sin^2\left(\frac{k\pi}{2(N+1)}\right),$$

as in (16) we denote now the eigenvalues of the matrix D_{yy} by μ_k, obtaining

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

having an artificial z_0 as we have above can now be translated as having an artificial μ_0 and the denominator of the cyclic part of the continued fraction in $\bar{t}^i_{z_0}(z)$ in (52) can be now rewritten back in terms of the variables η, h, μ_k that are of interest. Returning to z_k instead of z, a direct computation gives

$$2 + \frac{4+z_0}{\frac{z_0}{z_k} - 1} = 2 + z\frac{4+z_0}{z_0 - z_k} = 2 + \left(\eta h^2 + \mu_k\right)\frac{4 + \eta h^2 + \mu_k}{\mu_0 - \mu_k} \tag{53}$$

for the reoccurring term of the continued fraction in (52) and

$$\begin{aligned} &1 + \frac{z_k}{2}\left(1 + \left(1 + 2\frac{\frac{z_0}{z_k} - 1}{4+z_0}\right)\sqrt{1 + \frac{4}{z_0}}\right) \\ &= 1 + \frac{z_k}{2} + \frac{z_k}{2}\sqrt{1 + \frac{4}{z_0}} + \frac{z_0 - z_k}{z_0 + 4}\sqrt{1 + \frac{4}{z_0}} \\ &= 1 + \frac{\eta h^2 + \mu_k}{2} + \frac{\eta h^2 + \mu_k}{2}\sqrt{1 + \frac{4}{\eta h^2 + \mu_0}} + \frac{\mu_0 - \mu_k}{\eta h^2 + \mu_0 + 4}\sqrt{1 + \frac{4}{\eta h^2 + \mu_0}} \end{aligned} \tag{54}$$

for the so-called absolute term in (52). Already here we can notice that the structure of the absolute term seems quite similar to the exact formula in (33), especially compared with the formulas (34) and (47) corresponding to the approximations given by the Dirichlet and Robin boundary conditions at b.

Recalling the elimination process (19) - (22), we recover the last $N^b - N^a - 1$ block rows in A^b (i.e., the block rows governing the unknowns in the interior of Ω^b) from (53) and the $(N^a - 1)$-st block row (i.e., the block row governing the interface of Ω^a and Ω^b) from (54). In particular, denoting the modified matrix by $\bar{A}^b$ we have

$$\bar{A}^b = \frac{1}{h^2} \begin{pmatrix} D_1 & -\mathbb{I}_N & & & & \\ -\mathbb{I}_N & \ddots & \ddots & & & \\ & \ddots & \bar{D}_{N^a} & -\mathbb{I}_N & & \\ & & -\mathbb{I}_N & \bar{D}_{N^a+1} & \ddots & \\ & & & \ddots & \ddots & -\mathbb{I}_N \\ & & & & -\mathbb{I}_N & \bar{D}_{N^b} \end{pmatrix} \tag{55}$$

where $\bar{D}_{N^b} = \bar{D}_{N^b-1} = \ldots = \bar{D}_{N^a+1} \neq \bar{D}_{N^a}$. Reversing the process of diagonalization of the Schur complement from Section 4.1 and realizing that μ_0, η and h are only real constants, we get for any $i = N^b - 1, \ldots, N^a + 1$

$$\begin{aligned} \bar{D}_i &= Q^T \begin{pmatrix} 2 + z_1 \frac{4+z_0}{\mu_0 - \mu_1} & & \\ & \ddots & \\ & & 2 + z_N \frac{4+z_0}{\mu_0 - \mu_N} \end{pmatrix} Q \\ &= Q^T \left(2 + (4 + z_0) Z \begin{pmatrix} \mu_0 - \mu_1 & & \\ & \ddots & \\ & & \mu_0 - \mu_N \end{pmatrix}^{-1} \right) Q \\ &= 2I + (4 + \eta h^2 + \mu_0)(D - 2I)(\mu_0 I - D_{yy})^{-1}, \end{aligned} \tag{56}$$

with Z defined as the diagonal matrix $Z = \mathrm{diag}(z_1, \ldots, z_N)$ (see (16)), Q being the discrete Fourier sine basis (see (17)), D being the diagonal block of the original problem (see (5)) and D_{yy} being the three-point finite difference stencil discretization of the second derivative in y (see (16) and above).

Focusing on the block $\bar{D}_{N^a}$ we obtain

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

with the notation as in (56). We finish this section with the following remark.

Remark 7 *Notice that the formula* (56) *contains an explicit inverse. This raises the question whether it is more reasonable to perform the inverse operation (or, in practice, the solve operation) in the original Schur complement (and thereby obtain the exact result) rather than on this approximation.*

This and other practical challenges as well as an overall deeper understanding of $\bar{A}^b$ *and its continuous counterpart are clearly of interest and will be discussed in future work.*

7 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 away 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.

There are many further roads of exploration opened up by our approach: first, one could try to obtain an asymptotic formula for $p^*(i)$ as $h \to 0$, which would require to obtain the first expansion terms in our closed form formula for the error function $err_R(z, i, p)$ for

the Robin condition. Similarly one could try to obtain an asymptotic formula for the best parameter in volume. Both would need expansions of the finite continued fraction (47), which could be quite technical. Analogously, the same direction is worth exploring also for the approximant with shifted expansion point z_0. Moreover, the question about the nature of the approximation produced by the Robin truncation is still open as well as the question whether the second method matches the Padé approximation about the point z_0. 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.

Recognizing that our results were developed for a very particular problem, namely for the $\eta - \Delta$ equation on an unbounded strip in $\mathbb{R}^2$, there are some straightforward generalizations as none of our 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 5 and Section 6.

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 we would have to move to the complex domain with the continued fractions. The same holds in fact for any diagonalizable non-singular normal matrix D. If D is not normal, then the eigenvectors cannot be chosen to form an orthonormal basis of $\mathbb{R}^N$ (or $\mathbb{C}^N$). In that case, 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 $\hat{t}_k^i(z)$ as in (31) 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.

Finally, as we mentioned in Section 4, 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.

8 Acknowledgement

We would like to thank prof. Zdeněk Strakoš for his very useful comments and references to the continued fraction 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.