Geosystems Mathematics

Domenico Lahaye
Jok Tang
Kees Vuik
Editors

Modern Solvers for Helmholtz Problems

Birkhäuser

Geosystems Mathematics

Birkhäuser

This series provides an ideal frame and forum for the publication of mathematical key technologies and their applications to geoscientific and geo-related problems. Current understanding of the highly complex system Earth with its interwoven subsystems and interacting physical, chemical, and biological processes is not only driven by scientific interest but also by the growing public concern about the future of our planet, its climate, its environment and its resources. In this situation mathematics provides concepts, tools, methodology and structures to characterize, model and analyze this complexity at various scales. Modern high speed computers are increasingly entering all geo-disciplines. Terrestrial, airborne as well as spaceborne data of higher and higher quality become available. This fact has not only influenced the research in geosciences and geophysics, but also increased relevant mathematical approaches decisively as the quality of available data was improved.

Geosystems Mathematics showcases important contributions and helps to promote the collaboration between mathematics and geo-disciplines. The closely connected series *Lecture Notes in Geosystems Mathemactics and Computing* offers the opportunity to publish small books featuring concise summaries of cutting-edge research, new developments, emerging topics and practical applications. Also PhD theses may be evaluated, provided that they represent a significant and original scientific advance.

For more information about this series at http://www.springer.com/series/13389

Domenico Lahaye • Jok Tang • Kees Vuik
Editors

Modern Solvers for Helmholtz Problems

Birkhäuser

Editors

Domenico Lahaye
Delft Institute of Applied Mathematics
Delft University of Technology
Delft, The Netherlands

Jok Tang
Delft Institute of Applied Mathematics
Delft University of Technology
VORtech B.V.
Delft, The Netherlands

Kees Vuik
Delft Institute of Applied Mathematics
Delft University of Technology
Delft, The Netherlands

ISSN 2510-1544 ISSN 2510-1552 (electronic)
Geosystems Mathematics
ISBN 978-3-319-28831-4 ISBN 978-3-319-28832-1 (eBook)
DOI 10.1007/978-3-319-28832-1

Library of Congress Control Number: 2016962706

Mathematics Subject Classification (2010): 35-XX, 65FXX, 35J05

Printed on acid-free paper

This book is published under the trade name Birkhäuser, www.birkhauser-science.com
The registered company is Springer International Publishing AG
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

Foreword

The Helmholtz equation represents the time-independent part of the wave equation for electromagnetic, seismic, and acoustic waves. It was named after the scientist Hermann von Helmholtz (1821–1894). Today, it is one of the most used partial differential equations in numerical simulation.

In the multibillion seismic imaging industry, being my field of expertise, longitudinal wave fields are generated in complex geological media. These wave fields show a broad range of space-variant wavenumbers. This broad range is caused by seismic sources with a bandwidth up to five octaves (starting at a few Hz) and by rock velocities that range from a few hundred m/s to many thousand m/s. In such a challenging natural environment, accurate solutions need to be computed that are used for the design of effective data collection geometries, for understanding the very complex seismic responses, and for making imaging algorithms that utilize solvers in reverse time.

Notorious problems are the accuracy of high-wavenumber solutions, where avoiding numerical dispersion requires very fine spatial sampling. For the very sizeable geological models, this makes these traditional solvers economically not feasible.

I complement the authors for giving an elegant overview of Helmholtz solvers, with emphasis on the latest developments. The book is particularly valuable by showing the reader how to derive and use solvers that are independent of the wavenumber. This could find wide application in all wave field simulations where sizeable models and high wavenumbers are of large interest.

I hope that the new insights in this book will be widely used in academics and industry to better solve the multiple forward and inverse problems that play a critical role in the increasing amount of wave field applications worldwide.

Professor of Geosciences Emeritus, TU Delft
Director of the Centre for Global
Socio-Economic Change

Dr. A.J. (Guus) Berkhout

Contents

Part I Algorithms: New Developments and Analysis

Recent Results on Domain Decomposition Preconditioning for the High-Frequency Helmholtz Equation Using Absorption 3
Ivan G. Graham, Euan A. Spence, and Eero Vainikko

High Order Transparent Boundary Conditions for the Helmholtz Equation 27
Lothar Nannen

On the Optimality of Shifted Laplacian in a Class of Polynomial Preconditioners for the Helmholtz Equation 53
Siegfried Cools and Wim Vanroose

Part II Algorithms: Practical Methods and Implementations

How to Choose the Shift in the Shifted Laplace Preconditioner for the Helmholtz Equation Combined with Deflation 85
D. Lahaye and C. Vuik

The Multilevel Krylov-Multigrid Method for the Helmholtz Equation Preconditioned by the Shifted Laplacian 113
Yogi A. Erlangga, Luis García Ramos, and Reinhard Nabben

A Geometric Multigrid Preconditioner for the Solution of the Helmholtz Equation in Three-Dimensional Heterogeneous Media on Massively Parallel Computers 141
H. Calandra, S. Gratton, and X. Vasseur

Part III Implementations and Industrial Applications

Some Computational Aspects of the Time and Frequency Domain Formulations of Seismic Waveform Inversion 159
René-Édouard Plessix

Optimized Schwarz Domain Decomposition Methods for Scalar and Vector Helmholtz Equations 189
X. Antoine and C. Geuzaine

Computationally Efficient Boundary Element Methods for High-Frequency Helmholtz Problems in Unbounded Domains 215
Timo Betcke, Elwin van ’t Wout, and Pierre Gélat

Contributors

X. Antoine Institut Elie Cartan de Lorraine, Université de Lorraine, Inria Nancy-Grand Est EPI SPHINX, Vandoeuvre-lès-Nancy, France

Timo Betcke Department of Mathematics, University College London, London, UK

H. Calandra TOTAL E&P Research and Technology USA, Houston, TX, USA

Siegfried Cools Department of Mathematics and Computer Science, University of Antwerp, Antwerp, Belgium

Yogi A. Erlangga Mathematics Department, Nazarbayev University, Astana, Kazakhstan

Luis García Ramos Institut für Mathematik, TU Berlin, Berlin, Germany

Pierre Gélat Department of Mechanical Engineering, University College London, London, UK

C. Geuzaine Institut Montefiore B28, Université de Liège, Liège, Belgium

Ivan G. Graham Department of Mathematical Sciences, University of Bath, Bath, UK

S. Gratton INPT-IRIT, University of Toulouse and ENSEEIHT, Toulouse, France

D. Lahaye DIAM, TU Delft, Delft, The Netherlands

Reinhard Nabben Institut für Mathematik, TU Berlin, Berlin, Germany

Lothar Nannen Technische Universität Wien, Wien, Austria

René-Édouard Plessix Shell Global Solutions International, Rijswijk, The Netherlands

Euan A. Spence Department of Mathematical Sciences, University of Bath, Bath, UK

Eero Vainikko Institute of Computer Science, University of Tartu, Tartu, Estonia

Wim Vanroose Department of Mathematics and Computer Science, University of Antwerp, Antwerp, Belgium

X. Vasseur ISAE-SUPAERO, Toulouse, France

C. Vuik DIAM, TU Delft, Delft, The Netherlands

Elwin van 't Wout School of Engineering and Faculty of Mathematics, Pontificia Universidad Católica de Chile, Santiago, Chile

Introduction

We are very pleased to introduce this book on modern solvers for Helmholtz problems. Ten years ago, in 2006, Erlangga, Oosterlee, and Vuik published their seminal paper [1] that introduced the complex shifted Laplace preconditioner to solve the discretized Helmholtz equation. To celebrate this tenth birthday, we organized the TU Delft workshop entitled "Recent Developments in Fast Helmholtz Solvers" in the spring of 2015.[1] This seminar has motivated us to edit this book. A number of speakers at the workshop contributed to this volume. The authors of other chapters in this book accepted our invitation to share their unique insights into the recent developments in fast and robust solvers for Helmholtz problems. By collecting their contributions into a single volume, this unique and timely monograph that overviews the field was created.

This book is subdivided into the following three parts, each consisting of three chapters:

- **Part I Algorithms: New Developments and Analysis**
  - In the first chapter of the book, Graham, Spence, and Vainikko discuss the choice of the change in the shifted Laplace preconditioner that is approximately inverted by a domain decomposition approach.
  - In the second chapter, Nannen reviews both perfectly matched layers and Hardy space infinite elements for the treatment of unbounded domains.
  - In the third chapter, Cools and Vanroose present their experiences with a polynomial extension of the shifted Laplace preconditioner.
- **Part II Algorithms: Practical Methods**
  - In the fourth chapter, Lahaye and Vuik elaborate on how deflation allows to choose the shift in the shifted Laplace preconditioner to significantly accelerate the iterative convergence.

[1] http://ta.twi.tudelft.nl/nw/users/domenico/ten_years_shifted_laplacians/index.html.

  - In the fifth chapter, Erlangga, Garcia, and Nabben provide a theoretical framework for the combined use of deflation and the shifted Laplacian along with some numerical experiments.
  - In the sixth chapter, Calandra, Gratton, and Vasseur describe a two-level technique in which the coarse-level operator is approximated by the shifted Laplacian.

- **Part III Implementations and Industrial Applications**

  - In the seventh chapter, Plessix compares the time and frequency domain approach for the inversion of seismic data.
  - In the eight chapter, Antoine and Geuzaine review the Schwarz domain decomposition methods for the scalar and vector Helmholtz equation.
  - In the ninth and final chapter, Betcke, van 't Wout, and Gelat present a boundary element approach and give details on the discretization, preconditioning, and fast evaluation of the involved operators.

We wish to sincerely thank a number of people and organizations who made the TU Delft workshop and this book possible. We thank the workshop speakers for accepting our invitation to take the stage and contribute to this book. We thank the workshop sponsors for their financial contribution. We are grateful to Prof. Berkhout for his foreword. We much appreciate the effort that authors took to write their individual contributions. Both the seminar and the book would not have been possible without the help of the support staff at the TU Delft and Springer. We hope the reader will find this book pleasant and inspiring to read.

Delft, The Netherlands Domenico Lahaye
Delft, The Netherlands Jok Tang
Delft, The Netherlands Kees Vuik
2016

Reference

1. Erlangga, Y.A., Oosterlee, C.W., and Vuik, C., *A novel multigrid based preconditioner for heterogeneous Helmholtz problems*, SIAM Journal on Scientific Computing, 27(4), pp. 1471–1492, 2006.

Part I
Algorithms: New Developments and Analysis

In this part we foresee the description and analysis of new numerical solvers. Some of them may lead to nice insights and implementation for practical use.

Recent Results on Domain Decomposition Preconditioning for the High-Frequency Helmholtz Equation Using Absorption

Ivan G. Graham, Euan A. Spence, and Eero Vainikko

Abstract In this paper we present an overview of recent progress on the development and analysis of domain decomposition preconditioners for discretised Helmholtz problems, where the preconditioner is constructed from the corresponding problem with added absorption. Our preconditioners incorporate local subproblems that can have various boundary conditions, and include the possibility of a global coarse mesh. While the rigorous analysis describes preconditioners for the Helmholtz problem with added absorption, this theory also informs the development of efficient multilevel solvers for the "pure" Helmholtz problem without absorption. For this case, 2D experiments for problems containing up to about 50 wavelengths are presented. The experiments show iteration counts of order about $\mathscr{O}(n^{0.2})$ and times (on a serial machine) of order about $\mathscr{O}(n^{\alpha})$, with $\alpha \in [1.3, 1.4]$ for solving systems of dimension n. This holds both in the pollution-free case corresponding to meshes with grid size $\mathscr{O}(k^{-3/2})$ (as the wavenumber k increases), and also for discretisations with a fixed number of grid points per wavelength, commonly used in applications. Parallelisation of the algorithms is also briefly discussed.

1 Introduction

In this paper we describe recent work on the theory and implementation of domain decomposition methods for iterative solution of discretisations of the Helmholtz equation:

$$-(\Delta + k^2)u = f\,, \quad \text{in a domain} \quad \Omega, \tag{1}$$

I.G. Graham (✉) • E.A. Spence
Department of Mathematical Sciences, University of Bath, Bath BA2 7AY, UK
e-mail: I.G.Graham@bath.ac.uk; E.A.Spence@bath.ac.uk

E. Vainikko
Institute of Computer Science, University of Tartu, Tartu 50409, Estonia
e-mail: eero.vainikko@ut.ee

D. Lahaye et al. (eds.), *Modern Solvers for Helmholtz Problems*, Geosystems Mathematics, DOI 10.1007/978-3-319-28832-1_1

where $k(x) = \omega/c(x)$, with ω denoting frequency and c denoting the speed of acoustic waves in Ω. Our motivation originates from applications in seismic imaging, but the methods developed are applicable more generally, e.g. to earthquake modelling or medical imaging. While practical imaging problems often involve the frequency domain reduction of the elastic wave equation or Maxwell's equations, the scalar Helmholtz equation (1) is still an extremely relevant model problem which encapsulates many of the key difficulties of more complex problems.

We will focus here on solving (3) on a bounded domain Ω, subject to the first order absorbing (impedance) boundary condition:

$$\frac{\partial u}{\partial n} - iku = g \quad \text{on} \quad \Gamma = \partial\Omega \, , \tag{2}$$

although the methods presented are more general.

The theoretical part of this paper is restricted to the case of k constant. However the methods proposed can be used in the variable k case, and preliminary experiments are done on this case in Sect. 5.3.

Important background for our investigation is the large body of work on "shifted Laplace" preconditioning for this problem, starting from [12] and including, for example [11] and recent work on deflation [28]. (A fuller survey is given in [17, 21] and elsewhere in this volume.) In those papers (multigrid) approximations of the solution operator for the perturbed problem

$$-(\Delta + (k^2 + \mathrm{i}\varepsilon))u = f \, , \quad \text{with} \quad \frac{\partial u}{\partial n} - iku = g \quad \text{on} \quad \Gamma \, , \tag{3}$$

(suitably discretised and with carefully tuned "absorption" parameter $\varepsilon > 0$), were used as preconditioners for the iterative solution of (1). When k is variable, a slightly different shift strategy is appropriate (see Sect. 5.3).

One can see immediately the benefit of introducing ε in (3): When k is constant the fundamental solution $G_{k,\varepsilon}$ of the operator in (3) (for example in 3D) satisfies, for fixed $x \neq y$ with $k|x-y| = \mathscr{O}(1)$ and $\varepsilon \ll k^2$,

$$G_{k,\varepsilon}(x,y) = G_{k,0}(x,y) \exp\left(-\frac{\varepsilon}{2k}|x-y|\right)\left(1 + \mathscr{O}\left(\left(\frac{\varepsilon}{k^2}\right)^2 k|x-y|\right)\right) \, , \quad \text{as} \quad k \to \infty.$$

Thus, the effect of introducing ε is to exponentially damp the oscillations in the fundamental solution of problem (3), with the amount of damping proportional to ε/k. With slightly more analysis one can show that the weak form of problem (3) enjoys a coercivity property (with coercivity constant of order $\mathscr{O}(\varepsilon/k^2)$ in the energy norm (7) [21, Lemma 2.4]). This has the useful ramification that any finite element method for (3) is always well-posed (independent of mesh size) and enjoys a corresponding (albeit ε- and k-dependent) quasioptimality property. Therefore preconditioners constructed by applying local and coarse mesh solves applied to (3) are always well-defined; this is not true when $\varepsilon = 0$.

A natural question is then, how should one choose ε? To begin to investigate this question, we first introduce some notation. Let A_ε denote the finite element approximation of (3) and write $A = A_0$. Then A is the system matrix for problem (1), (2), which we want to solve.

Suppose an approximate inverse B_ε^{-1} for A_ε is constructed. Then a sufficient condition for B_ε^{-1} to be a good preconditioner for A is that $I - B_\varepsilon^{-1}A$ should be sufficiently small. Writing

$$I - B_\varepsilon^{-1}A = (I - B_\varepsilon^{-1}A_\varepsilon) + B_\varepsilon^{-1}A_\varepsilon(I - A_\varepsilon^{-1}A),$$

we see that a sufficient condition for the smallness of the term on the left-hand side is that

(i) $I - A_\varepsilon^{-1}A$ should be sufficiently small, and
(ii) $I - B_\varepsilon^{-1}A_\varepsilon$ should be sufficiently small.

At this stage, one might already guess that achieving both (i) and (ii) imposes somewhat contradictory requirements on ε. Indeed, on the one hand, (i) requires ε to be sufficiently small (since the ideal preconditioner for A is $A^{-1} = A_0^{-1}$). On the other hand, the larger ε is, the less oscillatory the shifted problem is, and the easier it should be to construct a good approximation to A_ε^{-1} for (ii).

Regarding (i): The spectral analysis in [14] of a 1-d finite-difference discretisation concluded that one needs $\varepsilon < k$ for the eigenvalues to be clustered around 1 (which partially achieves (i)). The analysis in [17] showed that, in both 2- and 3-d for a range of geometries and finite element discretisations, (i) is guaranteed if $\varepsilon/k \leq C_1$ for a small enough positive constant C_1, with numerical experiments indicating that this condition is sharp. Somewhat different investigations are contained in [11–13]. These performed spectral analyses that essentially aim to achieve (i) on a continuous level, and explored the best preconditioner of the form (3) for (1) in the 1D case with Dirichlet boundary conditions, based on the ansatz $k^2 + \mathrm{i}\varepsilon = k^2(a + \mathrm{i}b)$, where a, b are to be chosen; related more general results are in [30]. (For more detail, see, e.g., the summary in [17] and other articles in this volume.)

Regarding (ii): several authors have considered the question of when multigrid converges (in a k-independent number of steps) when applied to the shifted problem A_ε, with the conclusion that one needs $\varepsilon \sim k^2$ [2, 8, 14]. Note that this question of convergence is not quite the same question as whether a multigrid approximation to A_ε^{-1} is a good preconditioner for A_ε (property (ii)) or for A_0 (the original problem), but these questions are investigated numerically in [8]. For classical Additive Schwarz domain decomposition preconditioners, it was shown in [21] that (ii) is guaranteed (under certain conditions on the coarse grid diameter) if $\varepsilon \sim k^2$ (resonating with the multigrid results). In fact [21] also provides ε-explicit estimates of the rate of GMRES convergence when A_ε is preconditioned by the Schwarz algorithm. Although these estimates degrade sharply when ε is chosen less than k^2, numerical experiments in [21] indicate that improved estimates may be possible in the range $k \lesssim \varepsilon \lesssim k^2$.

The contradictory requirements that (i) requires ε/k to be sufficiently small, and (ii) requires $\varepsilon \sim k^2$ (at least for classical Additive Schwarz domain decomposition preconditioners) motivate the question of whether new choices of B_ε^{-1} can be devised that operate best when ε is chosen in the range $k \lesssim \varepsilon \lesssim k^2$. Such choices should necessarily use components that are more suitable for "wave-like" problems, rather than the essentially "elliptic" technology of classical multigrid or classical domain decomposition. In fact our numerical experiments below indicate that, for the preconditioners studied here, the best choice of ε varies, but is generally in the range $[k, k^{1.6}]$.

Domain decomposition methods offer the attractive feature that their coarse grid and local problems can be adapted to allow for "wave-like" behaviour. There is indeed a large literature on this (e.g. [4, 16, 18]), but methods that combine many subdomains and coarse grids and include a convergence analysis are still missing. The paper [21] provides the first such rigorous analysis in the many subdomain case, and current work is focused on extending this to the case when wave-like components are inserted, such as using (optimised) impedance or PML conditions on the local solves.

Another class of preconditioners for Helmholtz problems of great recent interest is the "sweeping" preconditioner [10] and its related variants—e.g. [7, 29, 31]. In principle these methods require the direct solution of Helmholtz subproblems on strips of the domain. A method of expediting these inner solves with an additional domain decomposition and off-line computation of local inverses is presented in [32]. Related domain decomposition methods for these inner solves, using tuned absorption, and with applications to industrial problems, are explored in [1, 27].

Finally it should be acknowledged that, while the reduction of the complicated question of the performance of B_ε^{-1} as a preconditioner for A into two digestible subproblems ((i) and (ii) above) is theoretically convenient, this approach is also very crude in several ways: Firstly the splitting of the problem into (i) and (ii) may not be optimal and secondly the overarching requirement that $\|I - B_\varepsilon^{-1}A\|$ should be small is far from necessary when assessing B_ε^{-1} as a preconditioner for A: for example good GMRES convergence is still assured if the field of values of $B_\varepsilon^{-1}A$ is bounded away from the origin in the complex plane (in a suitable inner product) and that $B_\varepsilon^{-1}A$ is bounded from above in the corresponding norm. We use this in the theory below.

2 Domain Decomposition

To start, we denote the nodes of the finite element mesh as $\{x_j : j \in \mathscr{I}^h\}$, for a suitable index set $\mathscr{I}^h$. These include nodes on the boundary Γ of Ω. The continuous piecewise linear finite element hat function basis is denoted $\{\phi_j : j \in \mathscr{I}^h\}$. To define preconditioners, we choose a collection of N non-empty relatively open subsets Ω_ℓ of $\overline{\Omega}$, which form an overlapping cover of $\overline{\Omega}$. Each $\overline{\Omega_\ell}$ is assumed to consist of a union of elements of the finite element mesh, and the corresponding nodes on Ω_ℓ are denoted $\{x_j : j \in \mathscr{I}^h(\Omega_\ell)\}$.

Now, for any $j \in \mathscr{I}^h(\Omega_\ell)$ and $j' \in \mathscr{I}^h$, we define the restriction matrix $(R_\ell)_{j,j'} := \delta_{j,j'}$. The matrix

$$A_{\varepsilon,\ell} := R_\ell A_\varepsilon R_\ell^T$$

is then just the minor of A_ε corresponding to rows and columns taken from $\mathscr{I}^h(\Omega_\ell)$. This matrix corresponds to a discretisation (on the fine mesh) of the original problem (3) restricted to the local domain Ω_ℓ, with a homogeneous Dirichlet condition at the interior boundary $\partial\Omega_\ell\backslash\Gamma$ and impedance condition at the outer boundary $\partial\Omega_\ell \cap \Gamma$ (when this is non-empty).

One-level domain decomposition methods are constructed from the inverses $A_{\varepsilon,\ell}^{-1}$. More precisely,

$$B_{\varepsilon,AS,local}^{-1} := \sum_\ell R_\ell^T A_{\varepsilon,\ell}^{-1} R_\ell, \tag{4}$$

is the classical one-level Additive Schwarz approximation of A_ε^{-1} with the subscript "*local*" indicating that the solves are on local subdomains Ω_ℓ.

The overlapping subdomains are required to satisfy certain technical conditions concerning their shape and the size and uniformity of the overlap. Moreover, each point in the domain is allowed to lie only in a bounded number of overlapping subdomains as the mesh is refined. We do not repeat these conditions here but refer the interested reader to [21, Sect. 3]. The theorems presented in Sect. 3 require these assumptions for their proof, as well as a quasi-uniformity assumption on the coarse mesh which is introduced next.

Two-level methods are obtained by adding a global coarse solve. We introduce a family of coarse simplicial meshes with nodes $\{x_j^H, j \in \mathscr{I}^H\}$, where each coarse element is also assumed to consist of the union of a set of fine grid elements. The basis functions are taken to be the continuous P_1 hat functions on the coarse mesh, which we denote $\{\Phi_p^H,\ p \in \mathscr{I}^H\}$. Then, introducing the fine-to-coarse restriction matrix $(R_0)_{pj} := \Phi_p^H(x_j^h)$, $j \in \mathscr{I}^h$, $p \in \mathscr{I}^H$, we can define the corresponding coarse mesh matrix $A_{\varepsilon,0} := R_0 A_\varepsilon R_0^T$. Note that, due to the coercivity property for problem (3), both $A_{\varepsilon,0}$ and $A_{\varepsilon,\ell}$ are invertible for all mesh sizes h, H and all choices of $\epsilon \neq 0$.

The classical Additive Schwarz preconditioner is then

$$B_{\varepsilon,AS}^{-1} := R_0^T A_{\varepsilon,0}^{-1} R_0 + B_{\varepsilon,AS,local}^{-1}, \tag{5}$$

(i.e. the sum of coarse solve and local solves) with $B_{\varepsilon,AS,local}^{-1}$ defined in (4).

The theoretical results outlined in the next section concern the properties of $B_{\varepsilon,AS}^{-1}$ as a preconditioner for A_ε (i.e. criterion (ii) in Sect. 1). The hypotheses for the theory involve conditions on k, ε and H (the coarse mesh diameter) as well as H_{sub} (the maximum of the diameters of the local subdomains Ω_ℓ). This theory is verified by some of the numerical experiments in [21] and we do not repeat those here. Instead, in Sect. 5 below we focus in detail on the performance of (variants of) $B_{\varepsilon,AS}^{-1}$ when used as preconditioners for the pure Helmholtz matrix A (hence aiming to satisfy criteria (i) and (ii) of Sect. 1 simultaneously). The variants of (5) which we will consider include the Restricted, Hybrid and local impedance preconditioners. These are defined in Sect. 4.

First we give a summary of the theoretical results for (5). These are taken from [21]. The proofs are based on an analysis of projection operators onto subspaces with respect to the sesquilinear form which underlies the shifted problem (3). This type of analysis is well-known for coercive elliptic problems, but [21] was the first to devise such a theory for the high-frequency Helmholtz equation.

3 Main Theoretical Results

Here we describe the main results from [21], namely Theorems 5.6 and 5.8 in that reference.

Since the systems arising from the discretisation of (3) are not Hermitian we need to use a general purpose solver. Here we used GMRES. Estimates of the condition number of the preconditioned matrix are not then enough to predict the convergence rate of GMRES. Instead one has to estimate either (i) the condition of the basis of eigenvectors of the system matrix, or (ii) bounds on its field of values. Here we take the second approach, making use of the classical theory of Eisenstat et al. [9] (see also [3]). A brief summary of this theory is as follows.

Consider a nonsingular linear system $C\mathbf{x} = \mathbf{d}$ in $\mathbb{C}^n$. Choose an initial guess $\mathbf{x}^0$ for $\mathbf{x}$, then introduce the residual $\mathbf{r}^0 = \mathbf{d} - C\mathbf{x}^0$ and the usual Krylov spaces: $\mathscr{K}^m(C, \mathbf{r}^0) := \mathrm{span}\{C^j\mathbf{r}^0 : j = 0, \ldots, m-1\}$. Introduce a Hermitian positive definite matrix D and the corresponding inner product on $\mathbb{C}^n$: $\langle \mathbf{V}, \mathbf{W}\rangle_D := \mathbf{W}^* D \mathbf{V}$, and let $\|\cdot\|_D$ denote the corresponding induced norm.

For $m \geq 1$, define $\mathbf{x}^m$ to be the unique element of $\mathscr{K}^m$ satisfying the minimal residual property:

$$\|\mathbf{r}^m\|_D := \|\mathbf{d} - C\mathbf{x}^m\|_D = \min_{\mathbf{x}\in\mathscr{K}^m(C,\mathbf{r}^0)} \|\mathbf{d} - C\mathbf{x}\|_D,$$

When $D = I$ this is just the usual GMRES algorithm, and we write $\|\cdot\| = \|\cdot\|_I$, but for more general D it is the weighted GMRES method [15] in which case its implementation requires the application of the weighted Arnoldi process [22]. The reason for including weighted GMRES in the discussion will become clear later in this section.

The following theorem is then a simple generalisation of the classical convergence result stated (for $D = I$) in [3]. A proof is given in [21].

Theorem 1 *Suppose* $0 \notin W_D(C)$. *Then*

$$\frac{\|\mathbf{r}^m\|_D}{\|\mathbf{r}^0\|_D} \leq \sin^m(\beta)\,, \quad \textit{where} \quad \cos(\beta) := \frac{\mathrm{dist}(0, W_D(C))}{\|C\|_D}\,, \tag{6}$$

where $W_D(C)$ *denotes the* field of values *(also called the* numerical range *of C) with respect to the inner product induced by D, i.e.*

$$W_D(C) \;=\; \{\langle \mathbf{x}, C\mathbf{x}\rangle_D : \mathbf{x} \in \mathbb{C}^n, \|\mathbf{x}\|_D = 1\}.$$

This theorem shows that if the preconditioned matrix has a bounded norm, and has field of values bounded away from the origin, then GMRES will converge independently of all parameters which are not present in the bounds.

With this criterion for robust convergence in mind, the following results were proved in [21]. These results use the notation $A \lesssim B$ (equivalently $B \gtrsim A$) to mean that A/B is bounded above by a constant independent of k, ε, and mesh diameters h, H_{sub}, H. We write $A \sim B$ when $A \lesssim B$ an $B \lesssim A$. In all the theoretical results below k is assumed constant.

In Theorem 2 and Corollary 1 below, the matrix D_k which appears is the stiffness matrix arising from discretising the energy inner product for the Helmholtz equation using the finite element basis. More precisely, the Helmholtz energy inner product and associated norm are defined by

$$(v, w)_{1,k} := \int_\Omega \left(\nabla v . \nabla \overline{w} + k^2 v \overline{w}\right) \mathrm{d}x, \qquad \text{and} \quad \|v\|_{1,k} = (v, w)_{1,k}^{1/2}\,. \tag{7}$$

For star-shaped Lipschitz domains, the norm $\|u\|_{1,k}$ of the solution u of the Helmholtz boundary-value problems (1), (2) (or alternatively (3) in the case of absorption) can be estimated in terms of the data f and g (measured in suitable norms) with a constant that is independent of k and ε (provided ε grows no faster than $\mathscr{O}(k^2)$). This fact is the starting point (and a crucial building block) for the theory in [21]. If ϕ_ℓ are the basis functions for the finite element space on the fine mesh, then the matrix D_k is defined by

$$(D_k)_{\ell,m} = (\phi_\ell, \phi_m)_{1,k} \quad \text{for all} \quad \ell, m.$$

The matrix D_k^{-1} appears as a weight in the result for right preconditioning in Theorem 3. These weights appear as artefacts of the method of analysis of the domain decomposition method which makes crucial use of the analysis of the Helmholtz equation in the energy norm. Fortunately, in practice, standard GMRES performs just as well as weighted GMRES (and is more efficient)—see Remark 1 below for more details.

Theorem 2 (Left Preconditioning)

$$(i) \qquad \|B_{\varepsilon,AS}^{-1}A_\varepsilon\|_{D_k} \;\lesssim\; \left(\frac{k^2}{\varepsilon}\right) \quad \textit{for all} \quad H, H_{\text{sub}}.$$

Furthermore, there exists a constant C_1 such that

$$(ii) \qquad |\langle \mathbf{V}, B_{\varepsilon,AS}^{-1}A_\varepsilon \mathbf{V}\rangle_{D_k}| \;\gtrsim\; \left(\frac{\varepsilon}{k^2}\right)^2 \|\mathbf{V}\|_{D_k}^2, \quad \textit{for all} \quad \mathbf{V} \in \mathbb{C}^n,$$

when

$$\max\left\{kH_{\text{sub}},\; kH\left(\frac{k^2}{\varepsilon}\right)^2\right\} \;\le\; C_1\left(\frac{\varepsilon}{k^2}\right). \tag{8}$$

This result contains a lot of information. In particular, if $\varepsilon \sim k^2$ and kH, kH_{sub} are uniformly bounded, then (weighted) left-preconditioned GMRES applied to systems with matrix A_ε will converge in a parameter-independent way. However when $\varepsilon/k^2 \to 0$, the bounds degrade. Nevertheless, numerical experiments in [21] (in the regime $H \sim H_{\text{sub}}$) suggest there is some room to sharpen the theory: In particular, if $\varepsilon \sim k^2$ the convergence of GMRES is parameter-independent even when $kH \to \infty$ quite quickly (that is much coarser coarse meshes than those predicted by the theory are possible). However if $\varepsilon \sim k$ then there appears not to be much scope to further reduce the coarse mesh diameter H.

Combining Theorems 1 and 2 we obtain:

Corollary 1 (GMRES Convergence for Left Preconditioning) *Consider the weighted GMRES method where the residual is minimised in the norm induced by D_k. Let $\mathbf{r}^m$ denote the mth iterate of GMRES applied to the system A_ε, left preconditioned with $B_{\varepsilon,AS}^{-1}$. Then*

$$\frac{\|\mathbf{r}^m\|_{D_k}}{\|\mathbf{r}^0\|_{D_k}} \;\lesssim\; \left(1 - \left(\frac{\varepsilon}{k^2}\right)^6\right)^{m/2}, \tag{9}$$

provided condition (8) *holds.*

Nowadays both left- and right-preconditioning play important roles in system solvers, and, in particular, right preconditioning is necessary if one wants to use Flexible GMRES (FGMRES) [26]. Fortunately Theorem 2 can be adapted to the case of right preconditioning as follows.

The first observation is that, for any $n \times n$ complex matrix C (and working in the inner product $\langle\cdot,\cdot\rangle_D$ induced by some SPD matrix D), we have, for any $\mathbf{v} \in \mathbb{C}^n$ and $\mathbf{w} := D\mathbf{v}$,

$$\frac{\langle \mathbf{v}, C\mathbf{v}\rangle_D}{\langle \mathbf{v}, \mathbf{v}\rangle_D} = \frac{\overline{\langle \mathbf{w}, C^*\mathbf{w}\rangle_{D^{-1}}}}{\langle \mathbf{w}, \mathbf{w}\rangle_{D^{-1}}}, \tag{10}$$

where $C^* = \overline{C}^\top$ denotes the Hermitian transpose of C. Thus estimates for the distance of the field of values of C from the origin with respect to $\langle\cdot,\cdot\rangle_D$ are equivalent to analogous estimates for the field of values of C^* with respect to $\langle\cdot,\cdot\rangle_{D^{-1}}$.

The second observation is that Theorem 2 also holds for the adjoint of problem (3). In the adjoint case, the sign of ε is reversed in the PDE and the boundary condition is replaced by $\partial u/\partial n + \mathrm{i}ku = g$. In this case the estimates in Theorem 2 continue to hold, but with ε replaced by $|\varepsilon|$. This is also proved in [21].

To handle the right-preconditioning case, we consider the field of values of the matrix $A_\varepsilon B_{\varepsilon,AS}^{-1}$ in the inner product induced by D_k^{-1}. By (10) these are provided by estimates of the field of values of $B_{\varepsilon,AS}^{-*}A_\varepsilon^*$ in the inner product induced by D_k. The latter are provided directly by the (the extended version of) Theorem 2. The required estimates for the norm of $A_\varepsilon B_{\varepsilon,AS}^{-1}$ are obtained by a similar argument.

The result (from [21]) is as follows.

Theorem 3 (Right Preconditioning) *With the same notation as in Theorem 2, we have*

$$(i) \qquad \|A_\varepsilon B_{\varepsilon,AS}^{-1}\|_{D_k^{-1}} \lesssim \left(\frac{k^2}{\varepsilon}\right) \quad \textit{for all} \quad H, H_{\text{sub}}.$$

Furthermore, provided condition (8) *holds,*

$$(ii) \qquad |\langle \mathbf{V}, A_\varepsilon B_{\varepsilon,AS}^{-1}\mathbf{V}\rangle_{D_k^{-1}}| \gtrsim \left(\frac{\varepsilon}{k^2}\right)^2 \|\mathbf{V}\|_{D_k^{-1}}^2, \quad \textit{for all} \quad \mathbf{V} \in \mathbb{C}^n.$$

Remark 1 As described earlier, the estimates above are in the weighted inner products induced by D_k and D_k^{-1}. It would be inconvenient to have to implement GMRES with these weights, especially the second one. It is thus an interesting question whether the use of weighted GMRES is necessary in practice for these problems. We investigated both standard and weighted GMRES (in the case of left preconditioning and with weight D_k) for a range of problems (some covered by the theory, some not). In practice there was little difference between the two methods. Therefore, the numerical experiments reported here use standard GMRES.

Remark 2 The theorems in [21] also allowed general parameter $\delta > 0$ which described the amount of overlap between subdomains, and included the dependence on δ explicitly in the estimates. We suppressed this here in order to make the exposition simpler.

4 Variants of the Preconditioners

In this section we describe the variants of the classical Additive Schwarz method defined in (5) which are investigated in the numerical experiments which follow.

The first variant which we consider is the *Restrictive Additive Schwarz* (RAS) preconditioner, which is well-known in the literature [5, 23]. Here, to define the local operator, for each $j \in \mathscr{I}^h$, choose a single $\ell = \ell(j)$ with the property that $x_j \in \Omega^{\ell(j)}$. Then the action of the local contribution, for each vector of fine grid freedoms $\mathbf{v}$, is:

$$(B^{-1}_{\varepsilon,RAS,local}\mathbf{v})_j \;=\; \left(R^T_{\ell(j)}A^{-1}_{\varepsilon,\ell(j)}R_{\ell(j)}\mathbf{v}\right)_j\,, \quad \text{for each} \quad j \in \mathscr{I}^h\,. \tag{11}$$

We denote this one level preconditioner as RAS1. (We shall in fact use a slight variation on this—as described precisely in Sect. 5.)

From this we could build the RAS preconditioner (in analogy to the standard Additive Schwarz method):

$$B^{-1}_{\varepsilon,RAS} \;=\; R_0^T A^{-1}_{\varepsilon,0}R_0 \;+\; B^{-1}_{\varepsilon,RAS,local}\,. \tag{12}$$

However we shall not use this directly in the following. Rather, instead of doing all the local and coarse grid problems independently (and thus potentially in parallel), we first do a coarse solve and then perform the local solves on the residual of the coarse solve. This was first introduced in [24]. As described in [20], this method is closely related to the deflation method [25], which has been used recently to good effect in the context of shifted Laplacian combined with multigrid [28]. The Hybrid RAS (HRAS) preconditioner then takes the form

$$B^{-1}_{\varepsilon,HRAS} := R_0^T A^{-1}_{\varepsilon,0}R_0 + P_0^T\left(B^{-1}_{\varepsilon,RAS,local}\right)P_0\,, \tag{13}$$

where

$$P_0 \;=\; I - AR_0^T A^{-1}_{\varepsilon,0}R_0\,.$$

Remembering that the local solves in $B^{-1}_{\varepsilon,RAS,local}$ are solutions of local problems with a Dirichlet condition on interior boundaries of subdomains, and noting that these are not expected to perform well for genuine wave propagation (i.e. ε small and k large), we also consider the use of impedance boundary conditions on the local solves. Let $A_{\varepsilon,Imp,\ell}$ be the stiffness matrix arising from the solution of (3) restricted to Ω_ℓ, where the impedance condition $\partial u/\partial n - \mathrm{i}ku$ is imposed on the boundary $\partial\Omega_\ell$, and dealt with in the finite element method as a natural boundary condition. This can be used as a local operator in the `HRAS` operator (13). The one-level variant is

$$(B^{-1}_{\varepsilon,Imp,RAS,local}\mathbf{v})_j \;=\; \left(\widetilde{R}^T_{\ell(j)}A^{-1}_{\varepsilon,Imp,\ell(j)}\widetilde{R}_{\ell(j)}\mathbf{v}\right)_j\,, \quad \text{for each} \quad j \in \mathscr{I}^h\,, \tag{14}$$

Here (noting that the local impedance condition is handled as a natural boundary condition on Ω_ℓ), $\widetilde{R}_\ell$ denotes the restriction operator $(\tilde{R}_\ell)_{j,j'} = \underline{\delta_{j,j'}}$, (as before) j' ranges over all $\mathscr{I}^h$, but now j runs over all indices such that $x_j \in \overline{\Omega}_\ell$.

The hybrid two-level variant is

$$B^{-1}_{\varepsilon,Imp,HRAS} := R_0^T A^{-1}_{\varepsilon,0} R_0 + P_0^T \left(B^{-1}_{\varepsilon,Imp,RAS,local}\right) P_0 \ . \tag{15}$$

We refer to these as the one- and two-level ImpHRAS preconditioners.

In the following section we will concentrate on illustrating the use of the four preconditioners defined in (11), (13), (14) and (15) for solving various problems with system matrix A (i.e. the discretisation of (3) with $\varepsilon = 0$). In our discussion and in the tables below we will use the following notation for the preconditioners:

$$(11) = \text{RAS1}, \qquad (13) = \text{HRAS}, \qquad (14) = \text{ImpRAS1}, \qquad (15) = \text{ImpHRAS} \ . \tag{16}$$

5 Numerical Experiments

Our numerical experiments concern the solution of (3) on the unit square, with $\eta = k$ and $\varepsilon = 0$, discretised by the continuous piecewise linear finite element method on a uniform triangular mesh. Thus, the problem being solved here is the "pure Helmholtz" problem without absorption and can be completely specified by the *fine mesh diameter*, here denoted h_{prob}. In [21] we also computed iteration numbers for solving (3) with $\varepsilon > 0$, thus an additional parameter $\varepsilon_{\text{prob}}$ was needed to specify the problem being solved. Here we restrict to the case $\varepsilon_{\text{prob}} = 0$. For the solver we shall use domain decomposition preconditioners built from various approximate inverses for (3). The choice of $\varepsilon > 0$ which is used to build the preconditioner is denoted $\varepsilon_{\text{prec}}$.

The experiments in Sect. 5.1 will be concerned with the case when the fine grid diameter is $h_{\text{prob}} \sim k^{-3/2}$. This is the discretisation level generally believed to be necessary to remove the pollution effect: roughly speaking the relative error obtained with this choice of h_{prob} is not expected to grow as $k \to \infty$. (However there is no proof of this except in the 1D case: See, e.g., the literature reviews in [17, Remark 4.2] and [19, Sect. 1.2.2].)

However the case of a fixed number of grid points per wavelength ($h_{\text{prob}} \sim k^{-1}$) is also frequently used in practice (especially in 3D) and provides sufficient accuracy in a limited frequency range. This regime is often studied in papers about Helmholtz solvers and so we include a substantial subsection (Sect. 5.2) on results for this case, which was not specifically discussed in [21]. Nevertheless the question of preconditioning the problem defined by $h_{\text{prob}} \sim k^{-1}$ and $\varepsilon_{\text{prob}} \sim k$ did arise in [21], as an "inner problem" in the multilevel solution of the problem with $h_{\text{prob}} \sim k^{-3/2}$, $\varepsilon_{\text{prob}} = 0$. (This is discussed again in Sect. 5.1 below.)

Interestingly, it turns out that the asymptotics (as k increases) of the solvers in each of the two cases $h_{\text{prob}} \sim k^{-3/2}$ and $h_{\text{prob}} \sim k^{-1}$ (both with $\varepsilon_{\text{prob}} = 0$) are somewhat different from each other and the best methods for one case are not necessarily the best for the other.

In the general theory given in Sect. 3, coarse grid size H and subdomain size H_{sub} are permitted to be unrelated. In our experiments here we construct local subdomains by first choosing a coarse grid and then taking each of the elements of the coarse grid and extending them to obtain an overlapping cover of subdomains with overlap parameter δ. This is chosen as large as possible, but with the restriction no two extended subdomains can touch unless they came from touching elements of the original coarse grid. In the literature this is called *generous overlap* and $H_{\text{sub}} \sim H$. Thus our preconditioners are completely determined by specifying the values of H and ε. In the case of constant k, we denote these by

$$H_{\text{prec}} \quad \text{and} \quad \varepsilon_{\text{prec}} \ . \tag{17}$$

We also have to specify how the RAS subdomains (recall (11)) are defined. Actually in our implementation involves a slight variation on (11) as follows. Our RAS subdomains are the original elements of the coarse grid (before extension). These overlap, but only at the edges of the coarse grid. Each node of the fine grid lies in a unique RAS subdomain except for nodes on the coarse grid edges. At these nodes the RAS operator (11) is extended so that it performs averaging of the contributions from all relevant subdomains at all such edge nodes.

When designing good domain decomposition methods we should be aware of cost. In the classical context (which we adopt here) where coarse grid and local problems are linked, a large-sized coarse grid problem will imply small-sized local problems and vice-versa. Coarse grids which are very fine and very coarse can both lead to very good methods in terms of iteration numbers, but not necessarily optimal in terms of time.

An "ideal" situation may be when all sub-problems are "load balanced". Let h_{prob} be the fine grid diameter and let H_{prec} be the coarse grid diameter, so that in $\mathbb{R}^d$, the dimension of the coarse grid problem is $\mathscr{O}(H_{\text{prec}}^{-d})$, while the dimension of the local problems are $\mathscr{O}((H_{\text{prec}}/h_{\text{prob}})^d)$. Then the classical domain decomposition method is load-balanced when $H_{\text{prec}} \sim h_{\text{prob}}^{1/2}$. If generous overlap is used, then a slightly smaller H_{prec} will give us load balancing. For example, in the pollution-free case $h_{\text{prob}} = k^{-3/2}$, the domain decomposition will be load-balanced at about $H_{\text{prec}} = k^{-0.8}$. While load balancing occurs at about $H_{\text{prec}} \sim k^{-0.6}$ when we are taking a fixed number of points per wavelength ($h_{\text{prob}} \sim k^{-1}$). We use these estimates as a guide in the experiments below.

In all the experiments below the stopping tolerance for GMRES was that the relative residual should be reduced by 10^{-6}.

In the experiments below, the system being solved is always the pure Helmholtz system $A\mathbf{u} = \mathbf{f}$. In the results given in Tables 1, 2, and 3 the right hand side vector $\mathbf{f}$ was chosen so that the finite element solution is an approximation of a plane wave (see [21, Sect. 6.2]). For the rest of the experiments $\mathbf{f} = \mathbf{1}$ was used.

Table 1 Comparison of HRAS and ImpHRAS for the problem with $h_{\text{prob}} \sim k^{-3/2}$, $\varepsilon_{\text{prob}} = 0$, using various choices of H_{prec} and $\varepsilon_{\text{prec}}$

$H_{\text{prec}} \sim k^{-1}$, $\varepsilon_{\text{prec}} = k$		
k	# HRAS	# ImpHRAS
20	12_{92}	17_{105}
40	18_*	21_*
60	25_*	27_*
80	33_*	35_*
100	43_*	45_*

$H_{\text{prec}} \sim k^{-1}$, $\varepsilon_{\text{prec}} = k^{1.2}$		
k	# HRAS	# ImpHRAS
20	13_{92}	18_{105}
40	18_*	21_*
60	25_*	27_*
80	32_*	34_*
100	42_*	43_*

$H_{\text{prec}} \sim k^{-1}$, $\varepsilon_{\text{prec}} = k^2$		
k	# HRAS	# ImpHRAS
20	37_{93}	34_{113}
40	63_*	56_*
60	86_*	78_*
80	110_*	101_*
100	136_*	123_*

$H_{\text{prec}} \sim k^{-0.6}$, $\varepsilon_{\text{prec}} = k$		
k	# HRAS	# ImpHRAS
20	51_{63}	26_{31}
40	125_{133}	50_{51}
60	$*_*$	69_{71}
80	$*_*$	74_{84}
100	$*_*$	84_{97}

$H_{\text{prec}} \sim k^{-0.6}$, $\varepsilon_{\text{prec}} = k^{1.2}$		
k	# HRAS	# ImpHRAS
20	48_{58}	26_{32}
40	114_{125}	48_{51}
60	$*_*$	69_{70}
80	$*_*$	74_{83}
100	$*_*$	84_{95}

$H_{\text{prec}} \sim k^{-0.6}$, $\varepsilon_{\text{prec}} = k^2$		
k	# HRAS	# ImpHRAS
20	39_{43}	36_{42}
40	81_6	73_{66}
60	113_{102}	104_{91}
80	135_{121}	126_{111}
100	156_{141}	148_{131}

Table 2 Iteration numbers for ImpHRAS with $\varepsilon_{\text{prob}} = k^{1.2} = \varepsilon_{\text{prec}}$, $h_{\text{prob}} = \pi/5k$ and $H_{\text{prec}} \sim k^{-1/2}$

k	#ImpHRAS
20	14_{16}
40	21_{23}
60	28_{30}
80	32_{31}
100	36_{34}
120	39_{38}
140	43_{41}

Table 3 GMRES iteration counts and timings for the inner-outer algorithm with $\varepsilon_{\text{prob}} = 0$, $h_{\text{prob}} = k^{-3/2}$, $H_{\text{prec}} = k^{-1}$ in the outer iteration, $H_{\text{prec}} = k^{-1/2}$ in the inner iteration and $\varepsilon_{\text{prec}} = k^{\beta}$ in both inner and outer iterations

	β						
k	0	0.4	0.8	1	1.2	1.6	2.0
20	**19(2)** 3.86 [0.08]	**19(2)** 3.72 [0.08]	**19(2)** 3.72 [0.08]	**19(2)** 3.68 [0.08]	**19(2)** 3.66 [0.08]	**25(1)** 4.00 [0.07]	**36(1)** 4.96 [0.07]
40	**22(4)** 54.8 [0.73]	**22(4)** 54.9 [0.73]	**22(4)** 54.8 [0.72]	**22(3)** 54.7 [0.71]	**22(3)** 54.8 [0.71]	**28(2)** 58.0 [0.69]	**61(1)** 80.4 [0.68]
60	**28(5)** 370 [3.20]	**28(5)** 371 [3.20]	**28(5)** 372 [3.19]	**28(5)** 370 [3.16]	**28(4)** 369 [3.11]	**35(2)** 383 [3.00]	**82(1)** 539 [3.10]
80	**36(6)** 1288 [8.62]	**36(6)** 1375 [8.69]	**36(6)** 1300 [8.59]	**36(5)** 1316 [8.51]	**35(5)** 1273 [8.38]	**42(2)** 1323 [8.08]	**104(1)** 1909 [8.19]
100	**46(8)** 3533 [16.5]	**46(8)** 3678 [16.01]	**46(7)** 3586 [16.4]	**45(7)** 3471 [15.9]	**44(6)** 3483 [16.2]	**49(2)** 3503 [15.5]	**126(1)** 4832 [16.4]

5.1 Pollution-Free Systems ($h_{\mathrm{prob}} \sim k^{-3/2}$)

The timings given in Tables 1, 2, and 3 below were for implementation on a serial workstation with Intel Xeon E5-2630L CPUs with 48 GB RAM. The later experiments were on a multiprocessor, described in Sect. 5.2.

The performance of GMRES for this case is investigated in detail in [21]. There we first studied the performance of domain decomposition preconditioners for systems with absorption (i.e. we set $\varepsilon_{\mathrm{prob}} = \varepsilon > 0$ and we studied the performance of B_ε^{-1} as a preconditioner for A_ε). With respect to that question we found that:

(i) the performance of the solvers reflected the theory given in Sect. 3;
(ii) There was little difference between left- and right-preconditioning;
(iii) There was little difference between the performance of standard GMRES and GMRES which minimised the residual in the weighted norm (in the case of left preconditioning) induced by D_k (see Remark 1 at the end of Sect. 3);
(iv) There was a marked superiority for HRAS over several other variants of Additive Schwarz;
(v) If H_{prec} is small enough ($H_{\mathrm{prec}} \sim k^{-1}$ is sufficient), then B_ε^{-1} is a good preconditioner for A_ε even for rather small ε (in fact, even $\varepsilon = 1$ gives acceptable results for HRAS);
(vi) If H_{prec} is small enough then it makes little difference whether the local problems have Dirichlet or impedance boundary conditions;
(vii) For larger H_{prec}, Dirichlet local problems perform very badly, while impedance local problems work well for large enough H_{prec}. In this case the coarse grid solver can be switched off without degrading the convergence of GMRES.

Based on these observations, the discussion in [21] then turned to the more important question of the solution of problems without absorption (i.e. $\varepsilon_{\mathrm{prob}} = 0$). The discussion in the rest of this subsection is an expansion of the discussion in [21].

We compare HRAS (Hybrid Restricted Additive Schwarz with Dirichlet local problems), as defined in (13) with ImpHRAS (Hybrid RAS with Impedance local problems), as defined in (15). In these experiments, $h_{\mathrm{prob}} = k^{-3/2}$ and in Table 1 below we give the number of GMRES iterations (with # denoting iteration count) for each of these two methods for various choices of H_{prec} and $\varepsilon_{\mathrm{prec}}$. In Table 1, the headline figure for each case is the iteration number for the Hybrid method (13) or (15), while as a subscript we give the iteration count for the corresponding one level methods (omitting the coarse grid solve), given respectively by (11) and (14). We include iteration numbers for the three cases $\varepsilon_{\mathrm{prec}} = k, k^{1.2}, k^2$. The optimal choice turns out to be around $\varepsilon_{\mathrm{prec}} \in [k, k^{1.2}]$, while $\varepsilon_{\mathrm{prec}} = k^2$ is provided for comparison. Data for a larger range of $\varepsilon_{\mathrm{prec}}$ and H_{prec} is given in [21]. A $*$ in the tables means the iteration did not converge after 200 iterations.

Based on the results in Table 1, we can make the following observations:

(i) When $H_{\text{prec}} \sim k^{-1}$, the coarse grid is sufficiently fine and does a good job. Using the data for $H_{\text{prec}} \sim k^{-1}$ and $\varepsilon_{\text{prec}} \sim k^{1.2}$ we observe that we have $\#HRAS \sim k^{0.71}$. Since we are here solving problems of size $n \sim k^3$, this is equivalent to $\#HRAS \sim n^{0.24}$. (Throughout the paper, rates of growth are obtained by linear least squares fits to the relevant log-log data.) Note that when $H_{\text{prec}} \sim k^{-1}$, there is little difference between HRAS and ImpHRAS, i.e. it does not matter here whether the local problems have Dirichlet or Impedance condition. This preconditioner has a competitive performance as n increases, but it incorporates an expensive coarse grid solve of size $H_{\text{prec}}^{-2} \sim k^2$ and it does not work without the coarse solve.

(ii) When $H_{\text{prec}} \sim k^{-0.6}$ the local problems are rather large (size $\sim k^{9/5}$) the ImpHRAS method works reasonably well with a slow growth of iteration count with respect to k (although higher actual iterations), while HRAS is not usable. Moreover in the case of ImpHRAS, the coarse grid solve has almost no effect and can be neglected.

(iii) In all cases the best choice of absorption parameter $\varepsilon_{\text{prec}}$ seems to be about $\varepsilon_{\text{prec}} \sim k^{\beta}$ with β close to 1.2. We note that the choice $\varepsilon_{\text{prec}} \sim k^2$ is remarkably inferior. A more extensive study of the variation of iteration numbers with respect to $\varepsilon_{\text{prec}}$ and H_{prec} is given in [21].

These observations led to the formulation of an inner-outer strategy for problems with $h_{\text{prob}} \sim k^{-3/2}$, with the outer iteration having preconditioner specified by $H_{\text{prec}} = k^{-1}$ and $\varepsilon_{\text{prec}} = k^{1.2}$. This "outer preconditioner" is a discretisation of (3) with $h_{\text{prob}} \sim k^{-1}$ and $\varepsilon_{\text{prob}} \sim k^{1.2}$, which is to be solved by a preconditioned inner iteration. So, as a precursor to formulating the inner-outer method, we study iteration counts for typical instances of this inner iteration. Here are some sample results with $h_{\text{prob}} = \pi/5k \sim k^{-1}$, $\varepsilon_{\text{prob}} = k^{1.2}$ using ImpHRAS as a preconditioner, with $H_{\text{prec}} \sim k^{-1/2}$ and $\varepsilon_{\text{prec}} = k^{1.2}$.

We see from Table 2 that, even without the coarse solve, the iteration numbers grow slowly, and even seem to be slowing down as k increases. Extrapolation using the last five entries of Table 2 (without the coarse solve) indicates that #ImpHRAS grows with approximately $\mathscr{O}(k^{0.38}) = \mathscr{O}(n^{0.19})$, where n is the size of the systems being solved in Table 2.

Therefore in [21] we proposed an inner-outer FGMRES iteration using (as the outer solver) HRAS with $H_{\text{prec}} = k^{-1}$ and (as the inner solver) ImpRAS1 with $H_{\text{prec}} = k^{-1/2}$. This method solves a system of dimension $\mathscr{O}(k^3)$ by solving $\mathscr{O}(k^2 + k)$ independent subdomain problems of dimension $\mathscr{O}(k^{1/2} \times k^{1/2}) = \mathscr{O}(k)$ and was found to have competitive properties.

In particular the subproblems are sufficiently small as to be very efficiently solved by a sparse direct solver. (Here we use `umfpack` included in the scipy sparse matrix package.) In this regard, an interesting observation is that, while positive definite systems coming from 2D finite element approximations of elliptic problems are often reported to be solvable by sparse direct solvers in optimal time ($\mathscr{O}(n)$, for dimension n up to about 10^5), this appears not to be the case for the indefinite systems encountered here. In our experience the computation time for the sub-systems encountered here grows slightly faster than linearly with respect to dimension n.

The following table gives some sample results for the composite inner/outer algorithm with $\varepsilon_{\text{prec}} = k^{\beta}$ (for both inner and outer iterations, for various β) and an inner tolerance $\tau = 0.5$ (found in [21] to be empirically best). The numbers in bold font denote the number of outer (respectively inner) iterations, while the smaller font numbers underneath denote the total time in seconds [with an average time for each outer iteration in square brackets]. (Other choices of inner tolerance are explored in [21]. Recall that the outer tolerance is 10^{-6}.) The best results occur with $\varepsilon_{\text{prec}} = k^{\beta}$ with $\beta \in [1, 1.2]$. Using the data in the column headed $\beta = 1$ (and remembering that we are here solving systems of dimension $n = k^3$), the outer iteration count grows with about $\mathscr{O}(k^{0.53}) \approx \mathscr{O}(n^{0.18})$, while the time per iteration is about $\mathscr{O}(n^{1.11})$ and the total time is $\mathscr{O}(n^{1.43})$. To give an idea of the size of the systems being solved, when $k = 100, n = 1,002,001$.

An interesting observation in Table 3 is the relative insensitivity of the results to the choice of β in the range $\beta \in [0, 1.6]$, and the very poor performance of $\beta = 2$. Thus for this method the choice of absorption $\varepsilon_{\text{prec}} = k^2$ is a relatively poor one, while in fact the choice $\varepsilon_{\text{prec}} = 1 = k^0$ is quite competitive. This is quite different to the experience reported using multigrid shifted Laplacian preconditioners. Note also that the number of inner iterations decreases as we read the rows of Table 3 from left to right, because increasing β means putting more absorption into the preconditioner and hence makes the inner problem easier to solve.

The remainder of the experiments in the paper were done on a linux cluster of 130 nodes. Each node consists of 2 CPUs (Intel Xeon E5-2660 v2 @ 2.20 GHz) with 10 cores: in total 20 cores and 64 GB RAM on each node. The nodes are connected with 4x QDR Infiniband networks. This cluster was used in serial mode except for the modest parallel experiment in Table 5, in which up to 10 of the 130 nodes were used.

5.2 10 *Grid-Points Per Wavelength* ($h \sim k^{-1}$)

5.2.1 Experiments with ImpRAS1 and ImpHRAS

In this section we consider the discretisation of (3) with $\varepsilon = 0$ and $h = \pi/5k$ (i.e. 10 grid points per wavelength). In this case the domain decomposition is load-balanced at about $H = k^{-0.6}$ and so we investigated the performance of preconditioned GMRES only for $H = k^{-\alpha}$, with α in the range $[0.4, 0.8]$. We found, for all choices of α, the method HRAS not to be effective (with or without coarse grid solve), and so we focused attention on ImpHRAS and its one-level variant ImpRAS1.

Sample results for ImpRAS1 (top) and ImpHRAS (bottom) are given in Table 4. Here T denotes the timing for the total solve process, while T_{it} denotes the time per iteration. Here the cost of the coarse grid solve is relatively small and the time per iteration for ImpHRAS is almost the same as that for ImpRAS1. Overall ImpRAS1 is slightly quicker than ImpHRAS: Using the last six entries of each column for

Table 4 Performance of ImpRAS1 (top) and ImpHRAS (bottom) with $\varepsilon_{\text{prob}} = 0$, $\varepsilon_{\text{prec}} = k$ and $h = \pi/5k$, for $H_{\text{prec}} = k^{-0.5}$, $k^{-0.4}$

		ImpRAS1					
		$H = k^{-0.5}$			$H = k^{-0.4}$		
k	n	*#GMRES*	T	T_{it}	*#GMRES*	T	T_{it}
60	9409	35	6.83	0.15	20	4.67	0.16
80	16,129	39	13.01	0.27	23	9.21	0.30
100	25,921	43	24.21	0.47	25	18.8	0.59
120	35,344	45	37.10	0.69	29	29.50	0.83
140	52,441	49	63.85	1.12	28	43.31	1.27
160	68,121	51	84.65	1.43	33	67.15	1.73
180	82,369	54	113.86	1.85	32	91.01	2.43
200	104,329	57	159.67	2.47	30	114.27	3.26
220	119,716	59	190.50	2.86	34	160.46	4.11
240	141,376	61	249.48	3.64	35	203.30	5.12
260	173,889	66	323.79	4.43	35	262.77	6.67
280	196,249	70	390.81	5.07	39	354.60	8.17
300	227,529	68	459.72	6.13	38	420.12	9.98

		ImpHRAS					
		$H = k^{-0.5}$			$H = k^{-0.4}$		
k	n	*#GMRES*	T	T_{it}	*#GMRES*	T	T_{it}
60	9409	33	5.09	0.11	21	4.36	0.14
80	16,129	40	10.87	0.22	25	9.18	0.29
100	25,921	43	20.80	0.40	24	17.11	0.57
120	35,344	47	34.08	0.61	29	27.99	0.79
140	52,441	52	61.25	1.01	27	40.45	1.24
160	68,121	55	82.11	1.28	32	63.16	1.67
180	82,369	53	103.99	1.69	32	88.40	2.37
200	104,329	56	147.72	2.29	31	115.10	3.20
220	119,716	59	180.19	2.66	35	161.90	4.05
240	141,376	60	233.24	3.42	35	198.54	4.99
260	173,889	64	295.08	4.05	34	252.30	6.56
280	196,249	69	361.86	4.63	37	332.66	8.01
300	227,529	67	430.55	5.69	37	403.04	9.76

ImpHRAS with $H = k^{-0.4}$, #GMRES is growing with order $\mathcal{O}(n^{0.18})$, while the total time is growing with order $\mathcal{O}(n^{1.5})$.

In Table 5 we give preliminary timing results for a parallel implementation of the ImpRAS1 method. The implementation is in python and is based on numpy and scipy with the mpi4py library used for message passing. The problem is run on $P = M^2$ processes, where M^2 is the number of subdomains in the preconditioner. Processes are mapped onto M cluster nodes with M processes running on each node. The column labelled P is the number of processors, which coincides with

Table 5 Parallel performance of ImpRAS1 with $\varepsilon_{prob}=0$, $\varepsilon_{\text{prec}}=k$ and $h=\pi/5k$, for $H_{\text{prec}}=k^{-0.4}$

k	$P=M^2$	n_{loc}	#*GMRES*	T	T_{par}	S
60	25	1444	20	4.67	0.38	12.25
80	36	1764	23	9.21	0.51	17.97
100	36	2916	25	18.8	1.02	18.54
120	49	2916	29	29.50	1.15	25.62
140	49	3969	28	43.31	1.62	26.66
160	64	3969	33	67.15	1.93	34.76
180	64	5041	32	91.01	2.37	38.43
200	64	6241	30	114.27	3.05	37.43
220	81	6084	34	160.46	3.24	49.53
240	81	6889	35	203.30	4.14	49.11
260	81	8281	35	262.77	5.34	49.23
280	100	8100	39	354.60	5.71	62.15
300	100	9025	38	420.12	6.73	62.43

Relative speedup S is shown for comparison of total time T_{par} on P processes with serial implementation time T

the number of subdomains. The column labelled n_{loc} gives the dimension of the local problem being solved on each processor. Note that n_{loc} grows with about $k^{1.2}$ while P grows with about $k^{0.8}$ in this implementation. T is the serial time, T_{par} is the parallel time and $S = T/T_{par}$. Based on the last six entries of the column T_{par}, the parallel solve time is growing with about $\mathscr{O}(k^{2.1}) = \mathscr{O}(n^{1.05})$ where n is the system dimension.

5.2.2 A Multilevel Version of ImpRAS1

From Table 4 we see that the case $H = k^{-0.4}$ provides a solver with remarkably stable iteration counts, having almost no growth with respect to k. However (although the coarse grid component of the preconditioner can be neglected), the local systems to be solved at each iteration are relatively large, being of dimension $\mathscr{O}((k^{0.6})^2) = \mathscr{O}(k^{1.2})$. We therefore consider inner-outer iterative methods where these large local problems are resolved by an inner GMRES preconditioned with an ImpRAS1 preconditioner based on decomposition of the local domains of diameter $k^{-0.4}$ into much smaller domains of diameter $(k^{-0.4})^2 = k^{-0.8}$. (Such inner-outer methods are also investigated in different ways in [32].) The local problems to be solved then are of dimension $\mathscr{O}((k^{0.2})^2) = \mathscr{O}(k^{0.4})$ and there are $\mathscr{O}(k^{1.6})$ of them to solve at each iteration.

The inclusion of this method in the present paper is rather tentative, because (for the range of k considered), breaking up the local problems of size $\mathscr{O}(k^{1.2})$ into smaller subproblems is not competitive time-wise with the direct solver in 2D. The times of this multilevel variant are far inferior to those reported in Table 4. However even though the inner tolerance is set quite large at 0.5, the (outer) iteration numbers

Table 6 Sample iteration counts for the inner-outer ImpRAS1 preconditioner $\varepsilon_{\text{prob}} = 0$, $h_{\text{prob}} = \pi/5k$, $\varepsilon_{\text{prec}} = k^{\beta}$, $H_{\text{prec}} = k^{-0.4}$ (for the outer iteration) and $H_{\text{prec}} = k^{-0.8}$ (for the inner iteration)

	β	
k	1.2	1.6
100	26(6)	31(4)
120	31(6)	36(4)
140	29(6)	35(4)
160	33(7)	39(5)
180	33(7)	38(5)
200	32(7)	39(5)
220	35(8)	42(5)
240	35(8)	42(5)
260	34(8)	42(5)
280	39(9)	45(6)
300	39(9)	45(6)

are remarkably unaffected (sample results are given in Table 6). In this table the outer tolerance is (as before) relative residual reduction of 10^{-6}. Similar results (although slightly inferior) are obtained with $\varepsilon_{\text{prec}} = k$, in which case the inner iterations are also almost identical with those reported in Table 4 for ImpRAS1 in the case $H_{\text{prec}} = k^{-0.4}$.

Since the action of this preconditioner involves the solution of $\mathscr{O}(k^{1.6})$ independent local systems of dimension only $\mathscr{O}(k^{0.4})$, this method has strong parallel potential and is also worth investigating in 3D, where the direct solvers are less competitive.

5.3 Variable Wave Speed ($h \sim \omega^{-3/2}$)

In this subsection we give some initial results on the performance of our algorithms when applied to problems with variable wave speed. A more detailed investigation of this problem is one of our next priorities and the discussion here should be regarded as somewhat preliminary.

Domain decomposition methods have the advantage that the subdomains (and possibly the coarse mesh) can be chosen to resolve jumps in the wave speed, if the wave speed is geometrically simple enough. At present the variable speed case is not covered by any theory, so this section is necessarily experimental.

We consider the analogue of the problem (3) with $k = \omega/c$ where ω is the angular frequency and $c = c(x)$ is the spatially dependent wave speed. For the preconditioners we consider approximate inverses of problems with variable absorption of the form:

$$- \Delta u - (1 + \mathrm{i}\rho) \left(\frac{\omega}{c}\right)^2 u = f \,, \quad \text{on} \quad \Omega, \tag{18}$$

on a bounded domain Ω with impedance boundary condition

$$\frac{\partial u}{\partial n} - i\left(\frac{\omega}{c}\right)u = g \quad \text{on} \quad \Gamma \tag{19}$$

where $\rho = \rho_{\text{prec}} \geq 0$ is a parameter to be chosen. Thus when c is constant, and $k := \omega/c$, the perturbed wavenumber is $k^2 + \mathrm{i}\rho k^2$ and so the choice $\varepsilon = k^\beta$ in (3) corresponds to the choice $\rho = k^{\beta-2}$ in (18). On the other hand when c is variable, the amount of absorption added is proportional to $(\omega/c)^2$ so more absorption is effectively added where c is relatively small and less is added when c is relatively large. We do not insert any absorption into the boundary condition (19).

We consider a test problem where Ω is the unit square. An internal square Ω_1 of side length $1/3$ is placed inside Ω and the wave speed is taken to have value c^* in the inner square and value 1 in $\Omega_2 := \Omega\backslash\Omega_1$. The square Ω_1 is either placed in the centre of Ω (this is the case "discontinuity resolved", where the coarse grid described below will resolve the interface) or at a position a few fine grid elements to the north and west of centre, with the distance moved in the directions north and west equal to the size of the overlap of the subdomains. In the latter case the coarse grid passes through the interface (and this is called "discontinuity unresolved" below). We perform experiments with c^* both bigger than 1 and less than 1 with the latter case expected to be hardest.

The problem is discretised by a uniform fine grid with $h_{\text{prob}} \sim \omega^{-3/2}$ and with the fine grid resolving the interface $\Gamma_{1,2}$ between Ω_1 and Ω_2. No absorption is added to the problem to be solved, i.e. $\rho_{\text{prob}} = 0$.

We apply the inner-outer algorithm as described in Sect. 5.1 (see Table 3) for this problem. The outer solver is HRAS with $H_{\text{prec}} \sim k^{-1}$ while the inner solver is ImpRAS1 with $H_{\text{prec}} \sim k^{-1/2}$. For both inner and outer solvers we set $\rho_{prec} = \omega^{\beta-2}$. In all cases generous overlap is used and the RAS domains are determined by the coarse grid as described in the introductory paragraphs to this section.

The coarse grid for the outer solve consists of uniform triangles of diameter $\sim k^{-1}$ which are chosen to resolve the square Ω_1 when it is placed in the centre, and do not resolve it when the square is moved. Numerical results, comparing the cases $c^* = 1.5, 1, 0.66$ are given in Tables 7, 8, and 9. In each row, for each value of β, the three figures indicate the number of outer HRAS iterations, the number of inner ImpRAS1 iterations (in brackets) and the total time on a serial machine. The outer tolerance is set at 10^{-6} while the inner tolerance is set at 0.5.

The times for $\beta = 1.6$ grow with about $\mathscr{O}(n^{1.4})$ in the case $c^* = 1.5$ and $c^* = 1$ (rather similar to the performance observed in Table 3). The actual times in the case $c^* = 0.66$ are considerably worse (which is to be expected as smaller $c*$ implies larger effective frequency on that domain). But the rate of growth of time with n is not affected very much, being about $\mathscr{O}(n^{1.5})$ in Table 9. The case $c^* = 1.5$ seems a little easier to solve than the case $c^* = 1$. There is not much difference in any case between the resolved and the unresolved cases.

Table 7 Performance of the inner-outer algorithm described in Sect. 5.3

	β							
ω	1.0		1.2		1.6		1.8	
$c^* = 1.5$, *discontinuity resolved*								
10	19(1)	0.71	19(1)	0.55	20(1)	0.53	21(1)	0.54
20	20(2)	3.25	20(2)	3.22	27(1)	3.65	30(1)	3.84
40	22(3)	50.09	23(3)	50.55	29(2)	54.04	44(1)	62.99
60	25(4)	356.71	26(4)	358.10	35(2)	381.06	57(1)	445.19
80	29(5)	1244.13	29(4)	1240.80	40(2)	1394.72	66(1)	1606.64
100	35(6)	3479.95	35(5)	3697.02	45(2)	3820.97	78(1)	4309.29
$c^* = 1.5$, *discontinuity unresolved*								
10	18(1)	0.70	19(1)	0.56	20(1)	0.53	21(1)	0.54
20	20(2)	3.26	20(2)	3.25	27(1)	3.65	30(1)	3.87
40	22(3)	50.80	23(3)	51.30	29(2)	54.56	44(1)	63.76
60	25(4)	363.19	26(4)	364.96	35(2)	387.40	58(1)	454.04
80	30(5)	1273.11	30(4)	1347.66	40(2)	1417.61	66(1)	1623.74
100	35(6)	3545.44	35(5)	3541.62	45(2)	3660.37	78(1)	4042.62

Discontinuous wave speed, $c* = 1.5$

Table 8 Performance of the inner-outer algorithm described in Sect. 5.3

	β							
ω	1.0		1.2		1.6		1.8	
$c^* = 1.0$								
10	18(1)	0.70	18(1)	0.54	19(1)	0.51	21(1)	0.54
20	19(2)	3.12	19(2)	3.16	25(1)	3.43	29(1)	3.73
40	22(3)	48.76	22(3)	48.58	28(2)	51.79	45(1)	62.22
60	28(5)	353.26	28(4)	352.74	35(2)	368.99	56(1)	429.53
80	36(5)	1253.44	35(5)	1244.01	42(2)	1361.78	66(1)	1476.53
100	45(7)	3487.02	44(6)	3693.13	49(2)	3728.06	79(1)	4179.60

Continuous wave speed $c* = 1$

6 Summary

In this paper we considered the construction of preconditioners for the Helmholtz equation (without or with absorption) by using domain decomposition methods applied to the corresponding problem with absorption.

These methods are related to the shifted Laplacian multigrid methods, but the relative simplicity of the method considered here permits rigorous analysis of the convergence of GMRES through estimates of the field of values of the preconditioned problem. The flexibility of the domain decomposition approach also allows for the insertion of sub-solvers which are appropriate for high frequency Helmholtz problems, such as replacing Dirichlet local problems with impedance (or PML) local problems.

Table 9 Performance of the inner-outer algorithm described in Sect. 5.3

	β							
ω	1.0		1.2		1.6		1.8	
$c^* = 0.66$, *discontinuity resolved*								
10	19(1)	0.73	20(1)	0.58	21(1)	0.54	23(1)	0.58
20	22(2)	3.38	22(2)	3.43	28(1)	3.68	33(1)	4.03
40	31(4)	55.03	32(3)	55.22	37(2)	57.46	54(1)	68.06
60	48(5)	418.78	48(4)	415.63	54(2)	426.52	79(1)	502.58
80	85(7)	1709.73	78(5)	1628.70	74(2)	1630.55	108(1)	1925.38
100	124(8)	4881.62	115(7)	4853.22	93(2)	4448.73	134(1)	5151.77
$c^* = 0.66$, *discontinuity unresolved*								
10	19(1)	0.72	19(1)	0.60	21(1)	0.54	23(1)	0.58
20	23(2)	3.54	23(2)	3.55	29(1)	3.84	34(1)	4.19
40	32(4)	58.89	32(3)	58.45	38(2)	61.33	55(1)	71.74
60	49(5)	450.16	49(4)	448.57	55(2)	458.60	80(1)	533.56
80	85(7)	1820.58	79(5)	1826.07	77(2)	1767.87	109(1)	2041.53
100	123(8)	5076.60	116(6)	5016.11	96(2)	4567.90	135(1)	5323.54

Discontinuous wave speed, $c* = 0.66$

For the analysis, two theoretical subproblems are identified: (i) What range of absorption *is permitted*, so that the problem with absorption remains an optimal preconditioner for the problem without absorption? and (ii) What range of absorption *is needed* so that the domain decomposition method performs optimally as a preconditioner for the problem with absorption?

The ranges that result from studying problems (i) and (ii) separately have been analysed, and this analysis is reviewed in the paper (Sects. 1 and 3). Since these ranges are disjoint, the best methods are obtained by using a combination of insight provided by the rigorous analysis and by numerical experimentation. The best methods involve careful tuning of the absorption parameter, the choice of coarse grid and the choice of boundary condition on the subdomains.

Of those methods studied, the best (in terms of computation time on a serial machine) differ, depending on the level of resolution of the underlying finite element grid. For problems with constant wave speed and with mesh diameter $h \sim k^{-3/2}$ (so chosen to resolve the pollution effect), a multilevel method with serial time complexity $\mathscr{O}(n^{\alpha})$ with $\alpha \in [1.3, 1.4]$ is presented, where $n \sim k^3$ is the dimension of the system being solved (Sect. 5.1). In this method a two level preconditioner with a fairly fine coarse grid is used, and the coarse grid problem is resolved by an inner iteration with a further one-level preconditioner with impedance local solves.

For discretisations involving a fixed number of grid points per wavelength, similar time complexity is achieved by highly parallelisable one-level methods using impedance local solves on relatively large subdomains.

We also illustrate the method when it is applied to a model problem with jumping wave speed (Sect. 5.3). A preliminary parallel experiment is also given.

References

1. P.N. Childs, I.G. Graham, and J.D. Shanks. Hybrid sweeping preconditioners for the Helmholtz equation. *Proc. 11th conference on mathematical and numerical aspects of wave propagation (Gammarth, Tunisia, June 2013)*, pages 285–286, 2013.
2. S. Cools and W. Vanroose. Local Fourier Analysis of the complex shifted Laplacian preconditioner for Helmholtz problems. *Numerical Linear Algebra with Applications*, 20:575–597, 2013.
3. B. Beckermann, S. A. Goreinov, and E. E. Tyrtyshnikov. Some remarks on the Elman estimate for GMRES. *SIAM journal on Matrix Analysis and Applications*, 27(3):772–778, 2006.
4. J-D. Benamou and B. Després. A domain decomposition method for the Helmholtz equation and related optimal control problems. *Journal of Computational Physics*, 136(1):68–82, 1997.
5. X-C. Cai and M. Sarkis. A restricted additive Schwarz preconditioner for general sparse linear systems. *SIAM Journal on Scientific Computing*, 21(2):792–797, 1999.
6. X-C. Cai and J. Zou. Some observations on the l^2 convergence of the additive Schwarz preconditioned GMRES method. *Numerical linear algebra with applications*, 9(5):379–397, 2002.
7. Z. Chen and X.Xiang. A source transfer domain decomposition method for Helmholtz equations in unbounded domains. *SIAM J. Numer. Anal*, 51(4):pp.2331–3356, 2013.
8. P-H. Cocquet and M. Gander. Analysis of multigrid performance for finite element discretizations of the shifted Helmholtz equation. *preprint*, 2014.
9. S. C. Eisenstat, H. C. Elman, and M. H. Schultz. Variational iterative methods for nonsymmetric systems of linear equations. *SIAM Journal on Numerical Analysis*, pages 345–357, 1983.
10. B. Engquist and L. Ying. Sweeping preconditioner for the Helmholtz equation: moving perfectly matched layers. *SIAM Multiscale Modelling and Simulation*, 9(2):686–710, 2010.
11. Y. A. Erlangga, C. W. Oosterlee, and C. Vuik. A novel multigrid based preconditioner for heterogeneous Helmholtz problems. *SIAM J. Sci. Comp.*, 27:1471–1492, 2006.
12. Y. A. Erlangga, C. Vuik, and C. W. Oosterlee. On a class of preconditioners for solving the Helmholtz equation. *Applied Numerical Mathematics*, 50(3):409–425, 2004.
13. Y. A. Erlangga. Advances in iterative methods and preconditioners for the Helmholtz equation. *Archives of Computational Methods in Engineering*, 15(1):37–66, 2008.
14. O. G. Ernst and M. J. Gander. Why it is difficult to solve Helmholtz problems with classical iterative methods. In I. G. Graham, T. Y. Hou, O. Lakkis, and R. Scheichl, editors, *Numerical Analysis of Multiscale Problems*, volume 83 of *Lecture Notes in Computational Science and Engineering*, pages 325–363. Springer, 2012.
15. A. Essai Weighted FOM and GMRES for solving nonsymmetric linear systems. *Numerical Algorithms*, 18(3-4):277–292, 1998.
16. C. Farhat, A. Macedo, M. Lesoinne, A two-level domain decomposition method for the iterative solution of high frequency exterior Helmholtz problems, *Numer. Math.* 85:283–308 (2000).
17. M.J. Gander, I.G. Graham, and E.A. Spence Applying GMRES to the Helmholtz equation with shifted Laplacian preconditioning: What is the largest shift for which wavenumber-independent convergence is guaranteed? Numer Math 131: 567–614.
18. M.J. Gander, F. Magoulès, and F. Nataf. Optimized Schwarz methods without overlap for the Helmholtz equation. *SIAM Journal on Scientific Computing*, 24(1):38–60, 2002.
19. I.G.Graham, M. Löhndorf, J.M. Melenk and E.A. Spence When is the error in the h-BEM for solving the Helmholtz equation bounded independently of k? *BIT Num. Math.*, 55(1): 171–214, 2015.
20. I. G. Graham and R. Scheichl. Robust domain decomposition algorithms for multiscale PDEs. *Numerical Methods for Partial Differential Equations*, 23(4):859–878, 2007.
21. I.G. Graham, E.A. Spence and E. Vainikko, Domain decomposition preconditioning for high-frequency Helmholtz problems with absorption. *Math. Comp.*, to appear, 2017.

22. S. Güttel and J. Pestana. Some observations on weighted GMRES. *Numerical Algorithms*, 67(4):733–752, 2014.
23. J-H. Kimn and M. Sarkis. Shifted Laplacian RAS solvers for the Helmholtz equation. In *Domain Decomposition Methods in Science and Engineering XX*, pages 151–158. Springer, 2013.
24. J. Mandel and M. Brezina. Balancing domain decomposition for problems with large jumps in coefficients. *Mathematics of Computation of the American Mathematical Society*, 65(216):1387–1401, 1996.
25. R. Nabben and C. Vuik. A comparison of deflation and coarse grid correction applied to porous media flow. *SIAM Journal on Numerical Analysis*, 42(4):1631–1647, 2004.
26. Y. Saad, A Flexible Inner-Outer Preconditioned GMRES Algorithm. *SIAM J. Sci. Comput.*, 14(2): 461–469, 1983.
27. J.D. Shanks, Robust Solvers for Large Indefinite Systems in Seismic Inversion, PhD Thesis, University of Bath, 2015.
28. A. H. Sheikh, D. Lahaye, and C. Vuik. On the convergence of shifted Laplace preconditioner combined with multilevel deflation. *Numerical Linear Algebra with Applications*, 20:645–662, 2013.
29. C. Stolk. A rapidly converging domain decomposition method for the Helmholtz equation. *Journal of Computational Physics*, 241:240–252, 2013.
30. M. B. Van Gijzen, Y. A. Erlangga, and C. Vuik. Spectral analysis of the discrete Helmholtz operator preconditioned with a shifted Laplacian. *SIAM Journal on Scientific Computing*, 29(5):1942–1958, 2007.
31. L. Zepeda-Núñez, L. Demanet, The method of polarized traces for the 2D Helmholtz equation, MIT Preprint 2014, to appear in *J. Comp. Phys.*
32. L. Zepeda-Núñez, L. Demanet, Nested domain decomposition with polarized traces for the 2D Helmholtz equation, MIT Preprint 2015.

High Order Transparent Boundary Conditions for the Helmholtz Equation

Lothar Nannen

Abstract We consider finite element simulations of the Helmholtz equation in unbounded domains. For computational purposes, these domains are truncated to bounded domains using transparent boundary conditions at the artificial boundaries. We present here two numerical realizations of transparent boundary conditions: the complex scaling or perfectly matched layer method and the Hardy space infinite element method. Both methods are Galerkin methods, but their variational framework differs. Proofs of convergence of the methods are given in detail for one dimensional problems. In higher dimensions radial as well as Cartesian constructions are introduced with references to the known theory.

1 Introduction

We consider finite element simulations of time-harmonic, scalar waves in open systems. Since standard mesh based methods like finite element or finite difference methods are restricted to bounded domains, for these methods unbounded domains of propagation have to be truncated to a bounded computational domain. Typically, such a truncation results in artificial reflections at the truncation boundary. Due to the non-locality of the waves, the reflections may pollute the solution in the whole computational domain.

The purpose of this chapter is to present some high order transparent boundary conditions such that artificial reflections are minimized. Thereby we restrict ourselves to finite element based transparent boundary conditions. For boundary element methods we refer to the corresponding chapter in this book.

The simplest transparent boundary condition is the so-called first order absorbing boundary condition. It has no extra costs, but the computational domain typically has to be quite large in order to minimize artificial reflections. For a review of higher order local absorbing boundary conditions we refer to [15, 24]. For these transparent boundary conditions, as for all subsequent ones, additional unknowns are needed.

L. Nannen (✉)
Technische Universität Wien, Wien, Austria
e-mail: lnannen@tuwien.ac.at

D. Lahaye et al. (eds.), *Modern Solvers for Helmholtz Problems*, Geosystems Mathematics, DOI 10.1007/978-3-319-28832-1_2

Since the construction and the theoretical framework are quite complicated, we will not present them in this chapter.

The so-called complex scaling or perfectly matched layer (PML) method (see e.g. [2, 8, 19, 31]) fits very well into the variational framework of finite element methods. It surrounds the computational domain with an artificial, anisotropic damping layer. It is very flexible and allows to reduce artificial reflections as much as necessary. A downside is, that it can be difficult to find optimal method parameters, since it depends on the damping profile, the thickness of the layer and on the finite element discretization in the absorbing layer.

For infinite elements no artificial truncation is needed. The unbounded domain outside of the computational domain is discretized with special basis and test functions. For classical infinite element methods (see [12, 13]) these functions fulfill the Sommerfeld radiation condition. Since the infinite elements are defined on an unbounded domain, integration over these basis and test functions needs to be done carefully. Moreover, the discretization matrices typically have large condition numbers.

Hardy space infinite elements (see [20, 32, 33]) also discretize the whole unbounded domain, but the basis functions are completely different to the classical ones. The basis functions are constructed using the pole condition [22, 34] as radiation condition. Roughly speaking this radiation condition characterizes outgoing waves by the poles of their Laplace transforms. These Laplace transforms belong to a certain class of Hardy spaces. The Hardy space infinite element method is a Galerkin method applied to a variational problem in a space which is built using a Hardy space. Just as the PML method, the Hardy space infinite element method allows arbitrary small discretization errors. It is even more flexible as the PML method and can be applied to time harmonic wave equations with phase and group velocities of different signs, where PML methods fail (see [17, 18]).

For the Helmholtz scattering problems given in Sect. 2 we present the PML (Sect. 3) and the Hardy space infinite element method (Sect. 4). To explain the basic ideas, we start for both methods with a one dimensional model problem, even though in one dimension there exists an easy to use exact transparent boundary condition. These ideas are then generalized to higher dimensions using radial, as well as Cartesian coordinates. In Sect. 5 we compare the two methods in terms of efficiency and programming effort.

2 Helmholtz Scattering Problems

In this section we start with the problem setting and the most popular radiation conditions in order to control the behavior of solutions u to the Helmholtz equation for large arguments.

2.1 Problem Setting

Let u be a solution to the Helmholtz equation

$$-\Delta u(x) - \omega^2(1 + p(x))u(x) = 0, \qquad x \in \Omega,$$

for an unbounded Lipschitz domain $\Omega \subset \mathbb{R}^d$, $d = 1, 2, 3$, with angular frequency $\omega > 0$. p is a coefficient function with compact support $\operatorname{supp}(p) := \{x \in \mathbb{R}^d : p(x) \neq 0\}$ in an open ball $B_R := \{x \in \mathbb{R}^d : |x| < R\}$ of radius $R > 0$. $|x| := \sqrt{\sum_{j=1}^{d} |x_j|^2}$ denotes the standard Euclidean norm.

Moreover, let the boundary $\partial\Omega$ be contained in B_R and let u fulfill the boundary condition

$$\frac{\partial u}{\partial \mathbf{n}} + \alpha u = g, \qquad \text{for } x \in \partial\Omega$$

with given functions α and g and the unit normal vector $\mathbf{n}$ pointing to the exterior of Ω. We refrain from Dirichlet boundary conditions in order to simplify the presentation.

Since problems on unbounded domains Ω cannot be discretized with standard finite elements, we introduce a bounded and star shaped Lipschitz domain $D \subset \mathbb{R}^d$ such that $\partial\Omega \subset D$ and $p \equiv 1$ in $\mathbb{R}^d \setminus D$. Then Ω is the disjoint union of the bounded interior domain $\Omega_{\text{int}} := \Omega \cap D$, the unbounded exterior domain $\Omega_{\text{ext}} := \mathbb{R}^d \setminus \overline{D}$ and the interface $\Gamma := \partial\Omega_{\text{int}} \cap \partial\Omega_{\text{ext}}$. For example, one could choose $D = B_R$.

In Ω_{ext}, we are looking for solutions u of the homogeneous problem

$$-\Delta u - \omega^2 u = 0, \qquad \text{in } \Omega_{\text{ext}}, \tag{1a}$$

$$u = u_0, \qquad \text{on } \Gamma, \tag{1b}$$

$$u \quad \text{is outgoing for } |x| \to \infty. \tag{1c}$$

The radiation condition (1c) ensures that (1) is uniquely solvable for all Dirichlet data $u_0 \in H^{1/2}(\Gamma)$ and all $\omega > 0$, and that these solutions are physically meaningful. For such a unique solution u_{u_0} we define the so-called Dirichlet-to-Neumann operator $\text{DtN} : H^{1/2}(\Gamma) \to H^{-1/2}(\Gamma)$ by

$$\text{DtN}\, u_0 := \frac{\partial u_{u_0}}{\partial \mathbf{n}}.$$

Here, the unit normal vector $\mathbf{n}$ on Γ points to the interior of Ω_{ext}. The interior problem for $u \in H^1(\Omega_{\text{int}})$ in weak form is given by

$$\int_{\Omega_{\text{int}}} \left(\nabla u \cdot \nabla v - \omega^2(1 + p)u\, v\right) dx + \int_{\partial\Omega} \alpha u v \, ds + \int_{\Gamma} (\text{DtN}\, u|_{\Gamma})\, v \, ds = \int_{\partial\Omega} g v \, ds \tag{2}$$

for all test functions $v \in H^1(\Omega_{\text{int}})$. Representation formulas of the Dirichlet-to-Neumann operator for $d = 2, 3$ will be the subject of the following subsections.

In one dimension solutions to (1a) are given by linear combinations of $x \mapsto \exp(i\omega|x|)$ and $x \mapsto \exp(-i\omega|x|)$. Using the standard convention $\exp(-i\omega t)$ for the time-harmonic ansatz, $x \mapsto \exp(i\omega|x|)$ is a radiating solution. Hence, the Dirichlet-to-Neumann operator in one dimension is simply given by DtN $u_0 = i\omega u_0$.

2.2 Sommerfeld Radiation Condition

Following [28] for time-harmonic waves of the form $\Re\left(u(x)\exp\left(-i\omega t\right)\right)$, the averaged outward energy flux through the interface Γ is given by

$$J_\Gamma(u) := -\frac{1}{2\omega}\Im\left\{\int_\Gamma u\,\frac{\partial\overline{u}}{\partial\mathbf{n}}ds\right\}.$$

Using Green's first identity in a domain $B_R \cap \Omega_{\text{ext}}$, it can be shown that

$$J_\Gamma(u) = \frac{1}{4\omega^2}\lim_{R\to\infty}\left(-\int_{\partial B_R}\left|\frac{\partial u}{\partial\mathbf{n}} \mp i\omega u\right|^2 ds \pm \int_{\partial B_R}\left(\left|\frac{\partial u}{\partial\mathbf{n}}\right|^2 + \omega^2|u|^2\right)ds\right) \tag{3}$$

for all solutions $u \in H^2_{\text{loc}}(\Omega_{\text{ext}})$[1] to (1a). Using (3) with the minus sign in the first integral, $J_\Gamma(u)$ is non-negative for solutions to (1a), if u fulfills the Sommerfeld radiation condition

$$\lim_{|x|\to\infty}|x|^{(d-1)/2}\left(\frac{\partial u(x)}{\partial|x|} - i\omega u(x)\right) = 0 \quad \text{uniformly for all directions } \frac{x}{|x|}. \tag{4}$$

Moreover, using (4) as radiation condition the problem (1) is uniquely solvable (see e.g. [35, Sect. 9, Theorem 1.3]). So the Sommerfeld radiation condition leads to a well defined Dirichlet-to-Neumann operator.

It can also be used to construct an approximation to the exact Dirichlet-to-Neumann operator: If the interface is a sphere of radius $R > 0$, then the so-called first order absorbing boundary condition is given by $u_0 \mapsto i\omega u_0$. This Robin boundary condition is only the exact DtN operator for $d = 1$. But since for a numerical realization no extra costs are needed, it is widely used in practice. Typically, R has to be quite large in order to guarantee, that the artificial reflections at Γ are negligible.

[1] $H^r_{\text{loc}}(\Omega)$ denotes the space of functions, which belong to $H^r(\hat{\Omega})$ for each compact $\hat{\Omega} \subset \Omega$.

2.3 Dirichlet-to-Neumann Operator

For $x \in \Omega_{\text{ext}} = \mathbb{R}^d \setminus \overline{B_R}$ we can use polar coordinates $x = r\hat{x}$ with $r := |x| > 0$ and $\hat{x} = x/r \in \partial B_1$ in order to construct a representation formula for solutions u to the exterior problem (1). In polar coordinates the Helmholtz equation (1a) is given by

$$-\frac{\partial^2 u(r\hat{x})}{\partial^2 r} - \frac{d-1}{r}\frac{\partial u(r\hat{x})}{\partial r} - \frac{1}{r^2}\Delta_{\hat{x}}\, u(r\hat{x}) - \omega^2 u(r\hat{x}) = 0, \qquad r > R, \hat{x} \in \partial B_1.$$

$-\Delta_{\hat{x}}$ is the negative Laplace-Beltrami operator. As it is hermitian and positive semi-definite, all eigenvalues are non-negative. For example, in [11] it is shown, that for $d = 3$ the eigenvalues are given by $\lambda_\nu := \nu(\nu+1)$ with multiplicities $M_\nu := 2\nu+1$, $\nu \in \mathbb{N}_0$.[2] For $d = 2$ the eigenvalues are $\lambda_\nu := \nu^2$ with multiplicities $M_\nu := 2$ for $\nu \in \mathbb{N}$ and $M_0 := 1$ for $\nu = 0$. The corresponding eigenfunctions, the spherical harmonics $Y_\nu^{(\mu)}$, build a complete orthonormal set of $L^2(\partial B_1)$. Hence, there holds

$$u(r\hat{x}) = \sum_{\nu=0}^{\infty}\sum_{\mu=1}^{M_\nu} u_{\nu,\mu}(r) Y_\nu^{(\mu)}(\hat{x}), \qquad r > R, \quad \hat{x} \in \partial B_1 \tag{5}$$

with $u_{\nu,\mu}(r) := \int_{\partial B_1} u(r\hat{x})\overline{Y_\nu^{(\mu)}(\hat{x})}d\hat{x}$. The series converges for each $r > R$ in the $L^2(\partial B_1)$ sense. If u is a sufficiently smooth solution to (1a), we can differentiate under the integral and deduce that $u_{\nu,\mu}$ is a solution to the (spherical) Bessel equation

$$-u_\nu''(r) - \frac{d-1}{r}u_\nu'(r) + \left(\frac{\lambda_\nu}{r^2} - \omega^2\right) u_\nu(r) = 0, \qquad r > R. \tag{6}$$

Solutions to (6) with $\omega = 1$ are linear combinations of the (spherical) Hankel functions of the first and second kind. We denote the Hankel functions ($d = 2$) and the spherical Hankel functions ($d = 3$) of the first and second kind by $\mathscr{H}_\nu^{(1,2)}$. Their asymptotic behavior is given by

$$\mathscr{H}_\nu^{(1,2)}(t) = \tfrac{C_d}{t^{(d-1)/2}} \exp\left(\pm i\left(t - \tfrac{\nu\pi}{2}\right)\right)\left(1 + \mathscr{O}\left(\tfrac{1}{t}\right)\right), \qquad t \to \infty, \tag{7a}$$

$$\mathscr{H}_\nu^{(1,2)'}(t) = \tfrac{\pm i C_d}{t^{(d-1)/2}} \exp\left(\pm i\left(t - \tfrac{\nu\pi}{2}\right)\right)\left(1 + \mathscr{O}\left(\tfrac{1}{t}\right)\right), \qquad t \to \infty, \tag{7b}$$

with $C_2 := \sqrt{2/\pi}\exp(\mp i\pi/4)$ and $C_3 := \exp(\mp i\pi/2)$. Hence, there holds

$$\lim_{r\to\infty} r^{(d-1)/2}\left(\mathscr{H}_\nu^{(1,2)'} \mp i\mathscr{H}_\nu^{(1,2)}(r)\right) = 0.$$

[2] $\mathbb{N}$ denotes the set of all positive natural numbers and $\mathbb{N}_0 := \{0\} \cup \mathbb{N}$.

In particular, the functions $u(r\hat{x}) := \mathscr{H}_\nu^{(1)}(\omega r)Y_\nu^{(\mu)}(\hat{x})$ solve the Helmholtz equation (1a) and satisfy the Sommerfeld radiation condition (4). Moreover, using (3) we compute

$$J_\Gamma(u) = \begin{cases} \frac{1}{\omega}, & d = 2 \\ \frac{2\pi}{3\omega^2}, & d = 3 \end{cases}, \tag{8}$$

i.e. the outward energy flux is positive and independent of ν and μ. So these functions radiate energy to infinity and are therefore physically meaningful.

Remark 2.1 A second way of motivating the choice of outgoing solutions is the limiting absorption principle (see e.g. [35, Sect. 9]). Similar to the idea of shifted Laplace preconditioners, we replace the positive frequency ω in the Helmholtz equation by $\omega(1 + \epsilon i)$ with $\epsilon > 0$ adding artificial absorption to the problem. Since the solutions to the perturbed problem should be bounded for $r \to \infty$, these solutions are given by $u_\epsilon(r\hat{x}) := \mathscr{H}_\nu^{(1)}(\omega(1 + i\epsilon)r)Y_\nu^{(\mu)}(\hat{x})$. Passing ϵ to the limit 0 leads again to the Hankel functions of the first kind.

Using the Hankel functions of the first kind in (5) and incorporating the boundary condition (1b) leads to the series representation

$$u(r\hat{x}) = \sum_{\nu=0}^{\infty}\sum_{\mu=0}^{M_\nu} \frac{\int_{\partial B_1} u_0(R\hat{x})\overline{Y_\nu^{(\mu)}(\hat{x})}d\hat{x}}{\mathscr{H}_\nu^{(1)}(\omega R)} \mathscr{H}_\nu^{(1)}(\omega r)Y_\nu^{(\mu)}(\hat{x}), \quad r > R, \quad \hat{x} \in \partial B_1. \tag{9}$$

For $u_0 \in L^2(\partial B_R)$ it is shown in [11, Theorem 2.14], that this series as well as the series of the term by term derivatives converges absolutely and uniformly on compact subsets of $\Omega_{\text{ext}} = \mathbb{R}^3 \setminus \overline{B_R}$. The results holds true for $\Omega_{\text{ext}} = \mathbb{R}^2 \setminus \overline{B_R}$ as well. Moreover, it is indeed a solution to (1) with the Sommerfeld radiation condition and each solution to (1) satisfying the Sommerfeld radiation condition is given by (9). Hence, (9) can be used to construct the Dirichlet-to-Neumann operator on spheres of radius R by

$$\text{DtN}\, u_0 := \sum_{\nu=0}^{\infty}\sum_{\mu=0}^{M_\nu} \left(\int_{\partial B_1} u_0(R\hat{x})\overline{Y_\nu^{(\mu)}(\hat{x})}d\hat{x}\right) \frac{\omega \mathscr{H}_\nu^{(1)'}(\omega R)}{\mathscr{H}_\nu^{(1)}(\omega R)} Y_\nu^{(\mu)}(\hat{x}). \tag{10}$$

A second representation formula for solutions to (1) can be deduced using the fundamental solution of the Helmholtz equation

$$\Phi(x, y) := \begin{cases} \frac{i}{4}\mathscr{H}_0^{(1)}(\omega|x - y|), & d = 2 \\ \frac{\omega}{4\pi}\mathscr{H}_0^{(1)}(\omega|x - y|), & d = 3 \end{cases}.$$

In [11, 35]) is shown, that for smooth boundary Γ a solution u of the exterior problem (1) combined with the Sommerfeld radiation condition has the integral representation

$$u(x) = \int_{\Gamma} \left(u(y) \frac{\partial \Phi(x,y)}{\partial \mathbf{n}(y)} - \frac{\partial u}{\partial \mathbf{n}}(y) \Phi(x,y) \right) ds(y), \qquad x \in \Omega_{\text{ext}}. \tag{11}$$

This representation can be used to construct a Dirichlet-to-Neumann operator for arbitrary smooth boundaries Γ.

Remark 2.2 The representation formulas (9) and (11) can also be used as radiation conditions. Since the (spherical) Hankel functions are holomorphic in $\{z \in \mathbb{C} : \Re(z) > 0\}$, the solutions u to (1) using these radiation conditions are holomorphic with respect to complex frequencies ω with $\Re(\omega) > 0$. This is not the case, if the Sommerfeld radiation condition is used, since for ω with $\Re(\omega) > 0$ and $\Im(\omega) < 0$ the Hankel functions of the second kind fulfill the Sommerfeld radiation condition. So for resonance problems, where the frequency is the sought complex resonance, the Sommerfeld radiation condition is not useful.

3 Complex Scaling Method

For test functions $v \in H^1(\Omega_{\text{ext}})$ with compact support in $\Omega_{\text{ext}} \cup \Gamma$, the variational form of (1) is given by

$$\int_{\Omega_{\text{ext}}} \left(\nabla u \cdot \nabla v - \omega^2 u \, v \right) dx = \int_{\Gamma} \text{DtN} \, u_0 \, v \, ds. \tag{12}$$

In the complex scaling or perfectly matched layer method the left hand side of this equation is first reformulated such that the solution u and the integrand is exponentially decaying for $|x| \to \infty$. Then a truncation of the unbounded domain Ω_{ext} to a bounded layer leads to an approximation of the Dirichlet-to-Neumann operator on the right hand side. As we will see, this approximation converges exponentially to the correct Dirichlet-to-Neumann operator with respect to the thickness of the layer.

3.1 One Dimensional PML

For simplicity we start with a one dimensional problem. Let $u \in H^1_{\text{loc}}(\mathbb{R}_+)$ be an outgoing solution to

$$\int_0^\infty \left(u'(x)v'(x) - \omega^2(1 + p(x))u(x)v(x) \right) dx = -u_0' v(0) \tag{13}$$

for all test functions $v \in H^1(\mathbb{R}_+)$ with compact support in $\mathbb{R}_{\geq 0}$. $u_0' \in \mathbb{C}$ denotes a given Neumann boundary value of $u'(0)$. If $p \in L^2(\mathbb{R}_+)$ with $\operatorname{supp}(p) \subset [0,R)$, u is given by

$$u(x) = \begin{cases} u_{\text{int}}(x), & x \in \Omega_{\text{int}} := (0,R) \\ u_{\text{int}}(R)\exp(i\omega(x-R)), & x \in \Omega_{\text{ext}} := (R,\infty) \end{cases} \tag{14}$$

where $u_{\text{int}} \in H^1(\Omega_{\text{int}})$ is a solution to the interior problem.

For the complex scaling we use a continuously differentiable function $\tau : \mathbb{R}_{\geq 0} \to \mathbb{R}_{\geq 0}$ with $\tau(0) = 0$ and $\tau(t) \geq Ct$, $C > 0$, for sufficiently large t. One might use simply the identity. For $\sigma \in \mathbb{C}$ and $R > 0$ the complex scaling function is defined by

$$\gamma_{\sigma,R}(r) := \begin{cases} r, & r \leq R \\ (\sigma-1)\tau(r-R) + r, & r > R \end{cases}. \tag{15}$$

$\gamma_{\sigma,R}$ is continuous and at least continuously differentiable for all $r \neq R$. For the monomials $\tau(t) = t^k$, $\gamma_{\sigma,R}$ is $k-1$ times continuously differentiable at $r = R$ and arbitrary smooth elsewhere.

Since outgoing solutions u are given by (14), the complex scaled function $u_{\sigma,R} := u \circ \gamma_{\sigma,R}$ is exponentially decaying for $x \to \infty$ if and only if $\Im(\sigma) > 0$. As $u|_{\Omega_{\text{ext}}}$ has a holomorphic extension, $u_{\sigma,R}$ solves the complex scaled Helmholtz equation

$$\frac{\partial}{\partial x}\left(\frac{\tilde{u}_{\sigma,R}'(x)}{\gamma_{\sigma,R}'(x)}\right) - \omega^2\gamma_{\sigma,R}'(x)\tilde{u}_{\sigma,R}(x) = 0, \qquad x > R. \tag{16}$$

Hence,

$$\int_{\Omega_{\text{ext}}} \left(u'v' - \omega^2 uv\right) dx = \int_{\Omega_{\text{ext}}} \left(\frac{u_{\sigma,R}'v_{\sigma,R}'}{(\gamma_{\sigma,R}')^2} - \omega^2 u_{\sigma,R}v_{\sigma,R}\right)\gamma_{\sigma,R}'\,dx$$

follows with partial integration on both sides for all test functions $v \in H^1(\Omega_{\text{ext}})$ with compact support in $[R,\infty)$ and all $v_{\sigma,R} \in H^1(\Omega_{\text{ext}})$ with $v_{\sigma,R}(R) = v(R)$.

Theorem 3.1 *Let $\omega, R > 0$, $p \in L^2(\mathbb{R}_+)$ with $\operatorname{supp}(p) \subset [0,R)$, and $\sigma \in \mathbb{C}$ with positive imaginary part. Moreover, let the assumptions on τ be fulfilled.*

Then $u \in H^1_{\text{loc}}(\mathbb{R}_+)$ is an outgoing solution to (13) if and only if $u_{\sigma,R} := u \circ \gamma_{\sigma,R} \in H^1(\mathbb{R}_+)$ is a solution to

$$\int_0^\infty \left(\frac{u_{\sigma,R}'v_{\sigma,R}'}{\gamma_{\sigma,R}'} - \omega^2(1+p))\gamma_{\sigma,R}'u_{\sigma,R}v_{\sigma,R}\right) dx = -u_0'v(0), \quad v_{\sigma,R} \in H^1(\mathbb{R}_+). \tag{17}$$

Proof The proof is a variant of the one in [10, Theorem 1]. We have already shown the first direction. Vice versa, let $\tilde{u}$ be a solution to (17). Using test functions $v_{\sigma,R}$ with compact support in (R,∞) and elliptic regularity results, $\tilde{u}_{\text{ext}} := \tilde{u}|_{(R,\infty)} \in H^2((R,\infty))$ solves (16). Hence, $\tilde{u}_{\text{ext}}$ is a linear combination of $x \mapsto \exp(\pm i\omega\gamma_{\sigma,R}(x))$. Since $\Re(\pm i\omega\gamma_{\sigma,R}(x)) = \mp\omega\Im(\sigma)\tau(x-R)$ and $\tilde{u}_{\text{ext}} \in H^2((R,\infty))$, we have $\tilde{u}_{\text{ext}}(x) = \tilde{u}_{\text{ext}}(R)\exp(i\omega(\gamma_{\sigma,R}(x) - R))$. Plugging this into (17) and using partial integration in $[R,\infty)$ leads to

$$\int_0^R \left(\tilde{u}'v' - \omega^2(1+p)\tilde{u}v\right) dx = i\omega\tilde{u}(R)v(R) - u_0'v(0), \quad v \in H^1(\Omega_{\text{int}}),$$

i.e. to the correct Dirichlet-to-Neumann operator at $x = R$. Thus, u defined by (14) with $u_{\text{int}} := \tilde{u}|_{\Omega_{\text{int}}}$ is outgoing and solves (13). □

Corollary 3.2 *Let u be a solution to (17). Then $u|_{\Omega_{\text{int}}}$ is independent of the damping function $\gamma_{\sigma,R}$.*

Of course, (17) is still posed on an unbounded domain $\mathbb{R}_+$ and cannot be discretized directly using standard finite element methods. But since the integrand is exponentially decaying, $\mathbb{R}_+$ is typically truncated to a bounded domain $(0, R+L)$ with $L > 0$ sufficiently large. Then, the truncated problem on $H^1((0, R+L))$ is discretized using standard finite element methods.

3.2 Convergence of a One Dimensional PML

Proving convergence of a truncated and discretized PML is typically done in the following way (see e.g. [1, 23]): Similar to the last proof, the problem in the perfectly matched layer $(R, R+L)$ is solved analytically at first. This results into a perturbed Dirichlet-to-Neumann operator at the interface $x = R$. Typically the error to the correct Dirichlet-to-Neumann operator is bounded by the complex scaled function at the truncation boundary $R+L$, i.e. the truncation error decays exponentially with increasing layer thickness L. For sufficiently large $L > 0$ it is then shown, that the truncated problem is uniquely solvable if the untruncated problem is uniquely solvable.

Once this is established, compact perturbation arguments of strictly coercive operators can be used to show, that the discrete problem on the truncated domain is uniquely solvable for sufficiently fine discretization. Moreover, using the generalized Céa Lemma the discretization error can be bounded by the approximation error.

Here, we will use an approach, where truncation and discretization error are treated simultaneously. For scalar waveguides this approach was proposed in [21]. For simplicity, let us assume that τ is the identity. Then $\gamma_{\sigma,R}'(x) \equiv \sigma$ for $x > R$ and there exists a rotation $\theta \in \{z \in \mathbb{C} : |z| = 1, \Re(z) > 0\}$ and a constant $\alpha_1 > 0$ such that

$$\Re\left(\theta \int_R^\infty \left(\frac{1}{\sigma}\left|u'\right|^2 - \omega^2\sigma\left|u\right|^2\right) dx\right) > \alpha_1 \|u\|^2_{H^1((R,\infty))}, \qquad u \in H^1((R,\infty)).$$

Since $\Re(\theta) > 0$, the Gårding inequality

$$\Re\left(\theta \int_0^\infty \left(\frac{|u'|^2}{\gamma_{\sigma,R}} - \omega^2(1+p)\gamma_{\sigma,R}\left|u\right|^2\right) dx + C\int_0^R |u|^2 dx\right) > \alpha\|u\|^2_{H^1(\mathbb{R}_+)} \tag{18}$$

holds for $u \in H^1(\mathbb{R}_+)$ with constants $\alpha := \min\{\alpha_1, \Re(\theta)\} > 0$ and $C > 0$ sufficiently large. Since $L^2((0,R))$ is compactly embedded in $H^1((0,R))$, a Fredholm operator of the form $A_{\sigma,R} + K_{\sigma,R} : H^1(\mathbb{R}_+) \to H^1(\mathbb{R}_+)$ can be associated to (17). $A_{\sigma,R}$ is continuous and strictly coercive and $K_{\sigma,R}$ is compact. Hence, Riesz-Fredholm theory can be used to show convergence of the truncated and discretized problem with homogeneous Dirichlet boundary condition at the truncation boundary. For the more complicated case of an acoustic waveguide, the following theorem was proven in [21, Theorem 5.6].

Theorem 3.3 *Let $V_{h,L} \subset \{f \in H^1((0,R+L)) : f(R+L) = 0\}$ be a usual finite element discretization of the truncated domain, such that for all $v \in H^1((0,R+L))$ with $v(R+L) = 0$ the orthogonal projection converges point wise, i.e.*

$$\lim_{h\to 0} \inf_{v_h \in V_{h,L}} \|v - v_{h,L}\|_{H^1((0,R+L))} = 0. \tag{19}$$

If (17) is uniquely solvable with solution $u_{\sigma,R} \in H^1(\mathbb{R}_+)$, then there exists for all sufficiently small $h > 0$ and all sufficiently large $L > 0$ a unique solution $u_{h,L} \in V_{h,L}$ to

$$\int_0^{R+L} \left(\frac{u'_{\sigma,R}v'_{\sigma,R}}{\gamma'_{\sigma,R}} - \omega^2(1+p))\gamma'_{\sigma,R}u_{\sigma,R}v_{\sigma,R}\right) dx = -u'_0 v(0), \quad v \in V_{h,L}. \tag{20}$$

Moreover, there exists a constant $C > 0$ independent of h and L such that

$$\|u_{\sigma,R} - u_{h,L}\|_{H^1((0,R+L))} \leq C\left(\inf_{v_{h,L}\in V_{h,L}} \|u_{\sigma,R} - v_{h,L}\|_{H^1((0,R+L))} + \|u_{\sigma,R}\|_{H^1((R+L,\infty))}\right). \tag{21}$$

Proof We define a finite dimensional subspace of $H^1(\mathbb{R}_+)$ by

$$\tilde{V}_{h,L} := \{f \in H^1(\mathbb{R}_+) : f|_{(0,R+L)} \in V_{h,L}, f|_{[R+L,\infty)} \equiv 0\}.$$

Since functions with compact support are dense in $H^1(\mathbb{R}_+)$, the orthogonal projection onto $\tilde{V}_{h,L}$ converges point-wise for $h \to 0$ and $L \to \infty$. So the first part of the theorem follows with [27, Theorem 13.7], since (20) is the projection of (17) to $\tilde{V}_{h,L}$. The error estimation is a consequence of Céa's Lemma (see e.g. [27, Theorem 13.6]). □

Equation (21) includes truncation and discretization error. Since

$$|u_{\sigma,R}(x)| = |u_{\sigma,R}(R)| \exp(-\omega \Im(\sigma)\tau(x-R)), \qquad x > R,$$

the second term of (21) decays exponentially with respect to L. For the first term we introduce for fixed $\epsilon > 0$ and $k \in \mathbb{N}$ the functions

$$g_{\epsilon,k}(x) := \begin{cases} 1, & x \leq R+L-\epsilon \\ 1 - \left(\frac{x+\epsilon-R-L}{\epsilon}\right)^k, & x \in (R+L-\epsilon, R+L) \end{cases} \tag{22}$$

such that $x \mapsto u_{\sigma,R}(x) g_{\epsilon,k}(x)$ belongs to $H^k((0, R+L))$ and vanishes at $R+L$. Hence, it can be approximated by functions $v_h \in V_{h,L}$ using (19). The remaining $H^1((0, R+L))$-error of $x \mapsto u_{\sigma,R}(x)(1 - g_{\epsilon,k}(x))$ again decays exponentially with respect to L for fixed ϵ and k.

Remark 3.4 For functions $u \in H^{k+1}(\Omega)$ the approximation error of finite element discretizations typically is bounded by

$$\inf_{v_h \in V_h} \|u - v_h\|_{H^1(\Omega)} \leq C h^k \|u\|_{H^{k+1}(\Omega)}. \tag{23}$$

The constant $C > 0$ is independent of the mesh size h, but depends amongst others on the order $k \in \mathbb{N}$ of the used polynomials. See e.g. [6, Sect. 4.4] or [9, Theorem 3.2.1] for sufficient conditions on finite elements such that (23) holds.

For those (19) is satisfied by density of $H^2(\Omega)$ in $H^1(\Omega)$. Moreover, (23) can be used to bound the approximation error of $u_{\sigma,R} g_{\epsilon,k+1}$.

3.3 Radial Complex Scaling

For problems in higher dimensions we may use radial complex scaling. Let us assume, that the interface Γ between the interior and the exterior domain is piecewise smooth, i.e. there exists a parametrization of Γ which is piecewise k times continuously differentiable with $k \in \mathbb{N}$. Moreover, we require that for all $x \in \Gamma$ with normal vector $\mathbf{n}_\Gamma(x)$ the scalar product $x \cdot \mathbf{n}_\Gamma(x)$ does not vanish and that Γ is the boundary of a domain D, which is star shaped with respect to the origin. Most often, Γ is just a sphere, but e.g. convex polyhedrons are also possible.

Using the complex scaling function of Sect. 3.1 we define for all $x \in \Omega \setminus \{0\}$ the complex scaled variable

$$x_\sigma(x) := \frac{\gamma_{\sigma,R(x)}(|x|)}{|x|} x \quad \text{with } R(x) := \sup\{r \in \mathbb{R}_+ : r\frac{x}{|x|} \in \Omega_{\text{int}}\}. \tag{24}$$

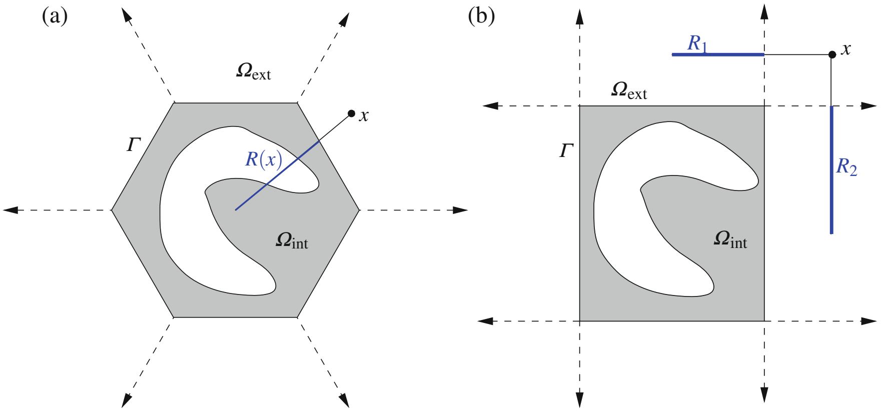


Fig. 1 Sketch of a complex scaling. The *dotted lines* indicate possible discontinuities of the Jacobian $D_x x_\sigma$ in Ω_{ext}. (**a**) Radial; (**b**) Cartesian

If 0 is contained in Ω, we define $x_\sigma(0) = 0$. See Fig. 1a for a sketch of the radial complex scaling.

For a spherical complex scaling, i.e. $\Gamma = \partial B_R$, $R(x) = R$ becomes constant. It is straightforward to see, that $x_\sigma(x) = x$ for all $x \in \Omega_{\text{int}}$ and that $\Im(x_\sigma(x)) = \Im(\sigma)\frac{\tau(|x|-R(x))}{|x|}x$. Since τ increases at least linearly for sufficiently large arguments, for $\Im(\sigma) > 0$ the imaginary part of the Cartesian components of $x_\sigma(x)$ increase at least linearly with respect to the distance of x to the interface Γ.

Lemma 3.5 *Let $\tilde{\Gamma} \subset \Gamma$ be parametrized by a k times continuously differentiable function η and let the function τ in the definition of the complex scaling function $\gamma_{\sigma,R(x)}$ be also k times continuously differentiable. Then x_σ is k times continuously differentiable in the pyramidal frustum $\{r\hat{x} \in \mathbb{R}^d : r > 1, \hat{x} \in \tilde{\Gamma}\}$.*

On the interfaces between the pyramidal frustums and to the interior domain Ω_{int}, x_σ is at least continuous.

Proof For $x \in \Omega_{\text{ext}} \cup \Gamma$ there exists at least one intersection point of the rays $\{rx \in \mathbb{R}^d : r > 0\}$ with Γ, since Γ is the boundary of a domain containing the origin and not containing x. This intersection point is unique, since otherwise the bounded domain would not be star shaped or there would be an $\hat{x} \in \Gamma$ with $x \cdot \mathbf{n}_\Gamma(\hat{x}) = 0$. Clearly, this intersection point depends continuously on x. Hence, $R(x)$ in the definition (24) of the complex scaling depends continuously on x, since it is the Euclidean norm of this intersection point. Since the complex scaling function $\gamma_{\sigma,R(x)}$ is continuous with respect to the argument and to R, $x_\sigma(x)$ is continuous in Ω.

Now, let x be in the interior of one pyramidal frustum $\{r\hat{x} \in \mathbb{R}^d : r > 1, \hat{x} \in \tilde{\Gamma}\}$ and let $\tilde{\Gamma} = \eta(S)$ with $S \subset \mathbb{R}^{d-1}$. We have to show, that for $x = r(x)\eta(\varphi(x))$, $\varphi \in S$, the function r is k times continuously differentiable. Since $R(x)$ in the definition of the complex scaling is given by $R(x) = |x|/r(x)$, this proves the claim.

So we define $F : \Omega_{\text{ext}} \times (\mathbb{R}_+ \times S) \to \mathbb{R}^d$ by $F(x, (r, \varphi)) := x - r\eta(\varphi)$. Since the Jacobian $D_{r,\varphi}F(x, (r, \varphi)) = (-\eta(\varphi), -rD_\varphi \eta_\varphi)$ is always invertible due to the assumption $\hat{x} \cdot \mathbf{n}(\hat{x}) \neq 0$ with $\hat{x} = \eta(\varphi)$ and since F is k times continuously differentiable, the implicit function theorem guarantees the smoothness of r. □

Explicit forms of the Jacobian $J_\sigma(x) = D_x x_\sigma(x)$ are complicated, but for the most common situation $\Gamma = \partial B_R$ it is straightforward to compute

$$J_\sigma(x) = \frac{\gamma_{\sigma,R}(|x|)}{|x|}\text{Id}_d + \frac{\gamma'_{\sigma,R}(|x|)|x| - \gamma_{\sigma,R}(|x|)}{|x|^3}xx^\top, \qquad x \in \Omega \setminus \{0\}. \tag{25}$$

$xx^\top \in \mathbb{R}^{d\times d}$ denotes the dyadic product and $\text{Id}_d \in \mathbb{R}^{d\times d}$ the identity matrix. In the following we restrict ourselves to spherical interfaces in order to simplify the proof.

Theorem 3.6 *Let Γ be a sphere of radius R and let $u \in H^1_{\text{loc}}(\Omega_{\text{ext}})$ be a radiating solution to (12). If $\Im(\sigma) > 0$, then $u_\sigma = u \circ x_\sigma \in H^1(\Omega_{\text{ext}})$ decays exponentially and there holds*

$$\int_{\Omega_{\text{ext}}} \left(J_\sigma^{-T}\nabla u_\sigma \cdot J_\sigma^{-T}\nabla v - \omega^2 u_\sigma\, v\right)\det(J_\sigma)dx = \int_\Gamma \text{DtN}\, u_0\, v\, ds \tag{26}$$

for all $v \in H^1(\Omega_{\text{ext}})$.

Vice versa, if $\Im(\sigma) > 0$ and if $\tilde{u} \in \{f \in H^1(\Omega_{\text{ext}}) : f|_\Gamma = u_0\}$ is a solution to (26) for all $v \in H^1_0(\Omega_{\text{ext}})$, then (26) holds true for all $v \in H^1(\Omega_{\text{ext}})$.

Proof The series representation (9) of a solution u to (12) converges absolutely and uniformly on compact subsets of Ω_{ext}. The same holds true for the series of the term by term derivatives. Moreover, the spherical Hankel functions are holomorphic in $\mathbb{C} \setminus \{0\}$ and the Hankel functions are holomorphic in $\mathbb{C} \setminus \mathbb{R}_{\leq 0}$. Hence, the series representation has a holomorphic extension from $x = r\hat{x} \in \Omega_{\text{ext}}$ with $r = |x| > R$ and $\hat{x} = x/r$ to complex variables $\tilde{x} = \tilde{r}\hat{x}$ with complex radius $\tilde{r} \in \mathbb{C} \setminus \mathbb{R}_{\leq 0}$.

So u_σ is well defined and the last lemma guarantees, that $u_\sigma \in H^1_{\text{loc}}(\Omega_{\text{ext}})$. Based on the integral representation (11) for a sphere in the interior of Ω_{ext}, it can be shown that u_σ decays exponentially. So the first part of the theorem follows with the chain rule for the transformation of the gradients.

The second part can be shown using a separation into (spherical) Bessel problems. For the details see [10, Theorem 1]. □

For a different kind of complex scaling it is shown in [4], that there holds a Gårding inequality for a complex scaled bilinear form, which is similar to the one in (26) with Ω instead of Ω_{ext}. Hence, for this modified complex scaling the same approach as for the one dimensional problem in Sect. 3.2 can be used. The error induced by truncation and finite element discretization is again bounded by an exponentially decaying truncation error and the usual finite element approximation error.

3.4 Cartesian Complex Scaling

If Γ is the boundary of a rectangle ($d = 2$) or a cuboid ($d = 3$), usually a Cartesian complex scaling is used. In the radial complex scaling (24) basically the absolute value of $x \in \Omega_{\text{ext}}$ is scaled. Hence, all Cartesian components are scaled simultaneously with the same scaling function. In Cartesian complex scaling, each Cartesian component can be scaled individually.

W.l.o.g. we assume, that $\Omega_{\text{int}} = \Omega \cap \bigotimes_{j=1}^{d}(-R_j, R_j)$ with $R_j > 0, j = 1, \ldots, d$. It is possible to choose $2d$ different functions $\tau_j^{(1,2)}$ in the complex scaling function, but we confine ourselves to one function τ. Then for $\sigma_j^{(1,2)} \in \mathbb{C}, j = 1, \ldots, d$, and $x = (x_1, \ldots, x_d)^\top \in \Omega$ we define the complex scaled variable $x_\sigma = ((x_\sigma)_1, \ldots, (x_\sigma)_d)^\top$ by

$$(x_\sigma)_j := \begin{cases} \gamma_{\sigma_j^{(1)}, R_j}(x_j), & x_j > R_j \\ x_j, & x_j \in [-R_j, R_j] \\ -\gamma_{\sigma_j^{(2)}, R_j}(-x_j), & x_j < -R_j \end{cases}, \qquad j = 1, \ldots, d. \tag{27}$$

In Fig. 1b a sketch of the Cartesian complex scaling is given. Since $\tau(0) = 0$, x_σ is continuous everywhere. The regularity of τ carries over to the regularity of x_σ in $\{x \in \Omega_{\text{ext}} : x_j \neq \pm R_j, j = 1, \ldots, d\}$. The Jacobian $J_\sigma(x)$ for a Cartesian scaling is a diagonal matrix, where the diagonal entries are the derivatives of the scaling functions. Therefore a Cartesian complex scaling is typically much easier to implement than a radial complex scaling.

In contrast to the radial PML, the convergence theory is more involved, see e.g. [5, 26]. In [5, Theorem 5.8] it is shown, that for $\tau(t) = t$ and with some constraints on σ the truncation error decays exponentially with respect to the thickness of the layer.

3.5 Choice of Complex Scalings and Bibliographical Remarks

Cartesian complex scaling typically is easier to implement than a radial one. But if the most popular linear complex scaling $\tau(t) = t$ is used, one has to take into account the discontinuities of the Jacobian J_σ shown in Fig. 1. Since the solution u_σ in this case is only in $H^1(\Omega)$, a high order finite element method would suffer a lot (confer with Remark 3.4). This can be avoided, if the finite element mesh is chosen such that the discontinuities of J_σ are part of the skeleton of the mesh. Hence, u_σ is smooth in the interior of each finite element, which guarantees the standard approximation error estimates of high order methods. Of course, choosing more regular damping functions also solves this issue.

The choice of the thickness of the complex scaling layer, the damping function and of the mesh in the layer is delicate. For a linear complex scaling one might use a priori error estimators for the truncation error of the form $\exp(-\omega\Im(\sigma)L)$ with L being a measure for the layer thickness. Afterwards, for the truncated problem standard mesh refinement strategies can be used (see e.g. [7]).

There is a vast amount of literature using the complex scaling method. In comparison theoretical results are rare. Without claiming to be exhaustive, we mention the following references, where unique solvability and exponential convergence of the truncated complex scaling problem is shown for Helmholtz problems in free space. Note, that in most cases for the truncated problems standard finite element results can be used.

The results in [29, 30] include spherical complex scaling as in (24) with $\Re(\sigma) = 1$ and some additional assumptions on τ. In particular, due to an assumption $\tau''(t) > 0$, linear complex scaling is not covered from the theory there.

Spherical complex scaling is also studied in [23] with one main difference: In this work, γ has to be at least two times continuously differentiable in a bounded transition zone $(R, \tilde{R})$. Moreover, for all $r > \tilde{R}$ the complex scaling is purely linear, i.e. $\gamma(r) = \sigma r$. So in contrast to (15) with $\tau(t) = t$, there is a translation by $(\sigma - 1)R$. A similar scaling is used in [4], where this translation is crucial for the existence of a Gårding inequality.

In [3] spherical complex scalings are used with scaling functions of the form $\gamma(r) = r + i\sigma(r)$ with $\sigma(r) = \log(\hat{R} - 1) - \log(\hat{R} - r)$ for $r \in (1, \hat{R})$ with $\hat{R} > 1$. For this kind of complex scaling no truncation is needed, but the coefficients in the complex scaled variational formulation become singular.

In [19] the spectral properties of untruncated radial complex scalings are investigated for $\Re(\sigma) = 1$ and two times continuously differentiable scaling functions. This work was extended in [25] with studies on truncated radial complex scalings.

Cartesian complex scalings were studied e.g. in a series of papers by Joseph E. Pasciak and coauthors [5, 26]. The last includes convergence results for linear complex scaling under some constraints on σ.

4 Hardy Space Infinite Element Method

Classical infinite element methods (see [12, 13]) directly discretize the exterior variational formulation (12) with special test and basis functions for $|x| \to \infty$. These basis functions have to satisfy the Sommerfeld radiation condition (4). Hardy space infinite element methods use the same idea, but they are based on the pole condition. This is another kind of radiation condition, which is to some extent equivalent to the classical ones.

4.1 One Dimensional Pole Condition

We start with this pole condition for one dimensional problems of the form (13). The details can be found in [20, Sect. 2].

Arbitrary solutions u to (13), which do not fulfill a radiation condition, are given for $x \geq R$ by $u(x) = Cu_{\text{out}}(x-R) + Du_{\text{inc}}(x-R)$ with complex constants $C, D \in \mathbb{C}$ and

$$u_{\text{out}}(r) := \exp(i\omega r), \quad u_{\text{inc}}(r) := \exp(-i\omega r), \qquad r \geq 0.$$

In the following we will use the Laplace transform $(\mathscr{L}v)(s) := \int_0^\infty \exp(-sr)v(r)dr$, $\Re(s) > 0$, and for a complex constant $\kappa_0 \in \mathbb{C} \setminus \{0\}$ a Möbius transform

$$(\mathscr{M}_{\kappa_0}\hat{v})(z) := \frac{1}{z-1}\hat{v}\left(i\kappa_0\frac{z+1}{z-1}\right), \qquad z \neq 1. \tag{28}$$

The constant κ_0 will be the main parameter of the Hardy space infinite element method. It is somehow equivalent to the complex scaling parameter σ for a linear complex scaling.

Since

$$(\mathscr{M}_{\kappa_0}\mathscr{L}u_{\text{out}})(z) = \frac{1}{i(\kappa_0-\omega)z+i(\kappa_0+\omega)}, \qquad z \in \mathbb{C}, \tag{29a}$$

$$(\mathscr{M}_{\kappa_0}\mathscr{L}u_{\text{inc}})(z) = \frac{1}{i(\kappa_0+\omega)z+i(\kappa_0-\omega)}, \qquad z \in \mathbb{C}, \tag{29b}$$

$\mathscr{M}_{\kappa_0}\mathscr{L}\{u(\bullet+R)\}$ is a meromorphic function with poles at $\left(\frac{\omega+\kappa_0}{\omega-\kappa_0}\right)^{\pm 1}$. For $\omega > 0$ and $\Re(\kappa_0) > 0$, the pole of $\mathscr{M}_{\kappa_0}\mathscr{L}u_{\text{out}}$ has absolute value larger than 1. So $\mathscr{M}_{\kappa_0}\mathscr{L}u_{\text{out}}$ can be expanded into the Taylor series

$$(\mathscr{M}_{\kappa_0}\mathscr{L}u_{\text{out}})(z) = \frac{1}{i(\kappa_0+\omega)}\sum_{j=0}^{\infty}\left(\frac{\omega-\kappa_0}{\omega+\kappa_0}\right)^j z^j, \qquad z \in \mathbb{C},$$

which converges for all $|z| \leq 1$. In particular, $\mathscr{M}_{\kappa_0}\mathscr{L}u_{\text{out}}$ is holomorphic in the complex unit disk and belongs to the Hardy space[3] $H^+(S^1)$ of the complex unit sphere $S^1 := \{z \in \mathbb{C} : |z| = 1\}$.

If $\Re(\kappa_0), \omega > 0$, $\mathscr{M}_{\kappa_0}\mathscr{L}u_{\text{inc}} \notin H^+(S^1)$, since it has a pole with absolute value smaller than 1. So we can use Hardy spaces in order to ensure, that a solution u to (13) only contains the outgoing solution u_{out}.

[3] $H^+(S^1) \subset L^2(S^1)$ consists of functions of the form $\sum_{j=0}^{\infty}\alpha_j z^j$, $z \in S^1$, with a square summable series (α_j). These functions are boundary values of some functions, which are holomorphic in the complex unit disk. Equipped with the $L^2(S^1)$ scalar product, $H^+(S^1)$ is a Hilbert space. For more details to Hardy spaces we refer to [14].

Definition 4.1 (Pole Condition) Let $H^+(S^1)$ denote the Hardy space of the complex unit sphere S^1 and let $\kappa_0 \in \mathbb{C}$ with positive real part be fixed.

Then a function $u \in L^2_{\mathrm{loc}}((R,\infty))$ is outgoing, if $\mathscr{M}_{\kappa_0}\mathscr{L}u(\bullet + R)$ is well defined and belongs to $H^+(S^1)$.

4.2 *Hardy Space Variational Formulation in One Dimension*

In order to be able to use this radiation condition, we have to reformulate (13). First, we define interior functions $u_{\mathrm{int}} := u|_{(0,R)}$, $v_{\mathrm{int}} := v|_{(0,R)}$ and shifted exterior functions $u_{\mathrm{ext}}(r) := u(r+R)$, $v_{\mathrm{ext}}(r) := v(r+R)$ for $r > 0$. For these functions (13) is split into

$$-u_0' v(0) = b_{\mathrm{int}}(u_{\mathrm{int}}, v_{\mathrm{int}}) + b_{\mathrm{ext}}(u_{\mathrm{ext}}, v_{\mathrm{ext}}),$$

with interior and exterior bilinear forms

$$b_{\mathrm{int}}(u_{\mathrm{int}}, v_{\mathrm{int}}) := \int_0^R \left(u_{\mathrm{int}}'(x)v_{\mathrm{int}}'(x) - \omega^2(1+p(x))u_{\mathrm{int}}(x)v_{\mathrm{int}}(x)\right) dx,$$

$$b_{\mathrm{ext}}(u_{\mathrm{ext}}, v_{\mathrm{ext}}) := \int_0^\infty \left(u_{\mathrm{ext}}'(r)v_{\mathrm{ext}}'(r) - \omega^2 u_{\mathrm{ext}}(r)v_{\mathrm{ext}}(r)\right) dr.$$

Using test functions of the form $v_{\mathrm{ext}}(r) = v_{\mathrm{int}}(R)\exp(i\lambda r)$ with $\Im(\lambda) > 0$ and $\Re(\lambda/\kappa_0) > 0$, we can use the identity in [20, Lemma A.1] to show

$$b_{\mathrm{ext}}(u_{\mathrm{ext}}, v_{\mathrm{ext}}) = q_{\kappa_0}(\mathscr{M}_{\kappa_0}\mathscr{L}u_{ext}', \mathscr{M}_{\kappa_0}\mathscr{L}v_{ext}') - \omega^2 q_{\kappa_0}(\mathscr{M}_{\kappa_0}\mathscr{L}u_{ext}, \mathscr{M}_{\kappa_0}\mathscr{L}v_{ext}), \tag{30}$$

with the bilinear form $q : H^+(S^1) \times H^+(S^1) \to \mathbb{C}$ defined by

$$q_{\kappa_0}(U,V) := \frac{-2i\kappa_0}{2\pi}\int_0^{2\pi} U(\exp(i\varphi))V(\exp(-i\varphi))d\varphi, \qquad U,V \in H^+(S^1). \tag{31}$$

q_{κ_0} is almost the $L^2(S^1)$ scalar product: Let $z \mapsto \overline{z}$ denote the standard complex conjugation and let $\mathscr{C} : H^+(S^1) \to H^+(S^1)$ denote the involution defined by $(\mathscr{C}V)(z) := \overline{V(\overline{z})}$, $z \in S^1$. Then $q_{\kappa_0}(U, \mathscr{C}V) = \frac{-2i\kappa_0}{2\pi}(U,V)_{L^2(S^1)}$. Moreover, the monomials $z \mapsto z^j, j \in \mathbb{N}_0$, are orthogonal with respect to the bilinear form q_{κ_0}.

There are two main difficulties in (30). First we have to ensure, that our basis and test functions are continuous at the interface $x = R$. Due to (29a), e.g. for the test functions there holds $v_{\mathrm{int}}(R) = \frac{1}{2i\kappa_0}(\mathscr{M}_{\kappa_0}\mathscr{L}v_{ext})(1)$. The right hand side would not be well defined for an arbitrary function $V \in H^+(S^1) \subset L^2(S^1)$. The second challenge is the terms $\mathscr{M}_{\kappa_0}\mathscr{L}u_{ext}'$ and $\mathscr{M}_{\kappa_0}\mathscr{L}v_{ext}'$, which have to be computed if test functions for $\mathscr{M}_{\kappa_0}\mathscr{L}u_{ext}$ and $\mathscr{M}_{\kappa_0}\mathscr{L}v_{ext}$ are used.

Both issues can be solved with one modification. We define the operators $\mathscr{T}_\pm : \mathbb{C} \times H^+(S^1) \to H^+(S^1)$ by

$$\mathscr{T}_\pm(v_0, V)(z) := \frac{1}{2}\left(v_0 + (z \pm 1)V(z)\right), \quad z \in S^1, \qquad (v_0, V) \in \mathbb{C} \times H^+(S^1). \tag{32}$$

Lemma 4.2 *Let $v \in H^1_{\mathrm{loc}}(\mathbb{R}_+) \cap C(\mathbb{R}_{\geq 0})$ be such that the Möbius and Laplace transformed function $\mathscr{M}_{\kappa_0}\mathscr{L}v$ is well defined. Moreover, we assume that $\mathscr{M}_{\kappa_0}\mathscr{L}v \in \mathscr{T}_-(\mathbb{C} \times H^+(S^1))$, i.e. there exists $(v_0, V) \in \mathbb{C} \times H^+(S^1)$ such that $\mathscr{M}_{\kappa_0}\mathscr{L}v = \frac{1}{i\kappa_0}\mathscr{T}_-(v_0, V)$.*

Then $v_0 = v(0)$ and $\mathscr{M}_{\kappa_0}\mathscr{L}v' = \mathscr{T}_+(v_0, V)$.

Proof By a limit theorem of the Laplace transform, there holds

$$v(0) = \lim_{r\to 0} v(r) = \lim_{s\to\infty} s(\mathscr{L}v)(s) = \lim_{z\to 1} i\kappa_0\left((z+1)\left(\mathscr{M}_{\kappa_0}\mathscr{L}v\right)(z)\right).$$

The limit of the right hand side exists, since by assumption $\left(\mathscr{M}_{\kappa_0}\mathscr{L}v\right)(z) = 1/(2i\kappa_0)(v_0 + (z-1)V(z))$ with $V \in L^2(S^1)$. Hence, $v_0 = v(0)$.

The second assertion follows from direct calculations with $(\mathscr{L}v')(s) = s(\mathscr{L}v)(s) - v(0)$. □

Using this lemma, the exterior bilinear form becomes

$$\begin{aligned} b_{\mathrm{ext},\kappa_0}\left((u_0, U), (v_0, V)\right) :=& q_{\kappa_0}\left(\mathscr{T}_+(u_0, U), \mathscr{T}_+(v_0, V)\right) \\ &- \omega^2 q_{\kappa_0}\left(\frac{1}{i\kappa_0}\mathscr{T}_-(u_0, U), \frac{1}{i\kappa_0}\mathscr{T}_-(v_0, V)\right), \end{aligned} \tag{33}$$

with $(u_0, U), (v_0, V) \in \mathbb{C} \times H^+(S^1)$. u_0 and v_0 represent the Dirichlet values of $u_{\mathrm{ext}}(0) = u_{\mathrm{int}}(R)$ and $v_{\mathrm{ext}}(0) = v_{\mathrm{int}}(R)$ respectively. This allows a continuous coupling of classical finite elements for v_{int} with infinite elements for $\mathscr{M}_{\kappa_0}\mathscr{L}v_{\mathrm{ext}}$.

Lemma 4.3 *For $\Re(\kappa_0) > 0$ there exists a rotation $\theta \in \{z \in \mathbb{C} : |z| = 1, \Re(z) > 0\}$ and a constant $\alpha > 0$ such that for all $(v_0, V) \in \mathbb{C} \times H^+(S^1)$*

$$\Re\left(\theta b_{\mathrm{ext},\kappa_0}\left((v_0, V), (\overline{v_0}, \mathscr{C}V)\right)\right) \geq \alpha \|(v_0, V)\|^2_{\mathbb{C}\times H^+(S^1)}. \tag{34a}$$

Moreover, $b_{\mathrm{ext},\kappa_0}$ is continuous, i.e. there exists a constant $C > 0$ such that for all $(u_0, U), (v_0, V) \in \mathbb{C} \times H^+(S^1)$

$$|b_{\mathrm{ext},\kappa_0}\left((u_0, U), (v_0, V)\right)| \leq C\|(u_0, U)\|_{\mathbb{C}\times H^+(S^1)}\|(v_0, V)\|_{\mathbb{C}\times H^+(S^1)}. \tag{34b}$$

The norm on $\mathbb{C} \times H^+(S^1)$ is thereby defined as

$$\|(v_0, V)\|_{\mathbb{C}\times H^+(S^1)} := \sqrt{|v_0|^2 + \|V\|^2_{L^2(S^1)}}, \qquad (v_0, V) \in \mathbb{C} \times H^+(S^1).$$

Proof The continuity of $b_{\text{ext},»_0}$ follows from the continuity of q_{κ_0} and of the operators $\mathscr{T}_\pm$. Since $2\mathscr{T}_\pm(v_0, V)(z) = v_0 + zV(z) \pm V(z)$, the parallelogram identity leads to

$$\begin{aligned}\|\mathscr{T}_-(v_0, V)\|^2_{L^2(S^1)} + \|\mathscr{T}_+(v_0, V)\|^2_{L^2(S^1)} &= \frac{1}{2}\|v_0 + \bullet V(\bullet)\|^2_{L^2(S^1)} + \frac{1}{2}\|V\|^2_{L^2(S^1)} \\ &= \frac{1}{2}|v_0|^2 + \|V\|^2_{L^2(S^1)} \geq \frac{1}{2}\|(v_0, V)\|^2_{\mathbb{C}\times L^2(S^1)}.\end{aligned}$$

The last identity yields by orthogonality of the monomials $z \mapsto z^j, j \in \mathbb{N}_0$, in $L^2(S^1)$. Choosing θ with $\Re(\theta)$ such that

$$\Re\left(\frac{-2i\kappa_0}{2\pi}\theta\right) = \frac{1}{\pi}\Im(\kappa_0\theta) \text{ and } \Re\left(\frac{(-2i\kappa_0)(-\omega^2)}{2\pi(i\kappa_0)^2}\theta\right) = \frac{\omega^2}{\pi|\kappa_0|^2}\Im(\overline{\kappa_0}\theta)$$

are positive, yields the claim. □

The last proof as well as the next two ones are simplifications of those in [21, Sect. 6] and [20, Theorem 2.4].

Theorem 4.4 *Let $\omega, R > 0$, $p \in L^2(\mathbb{R}_+)$ with* $\operatorname{supp}(p) \subset [0, R)$*, and $\kappa_0 \in \mathbb{C}$ with positive real part. If u is an outgoing solution to (13) and $u_{\text{int}} := u|_{(0,R)}$, then there exists a function $U \in H^+(S^1)$ such that $(u_{\text{int}}, U) \in H^1((0, R)) \times H^+(S^1)$ solves*

$$-u_0'v(0) = b_{\text{int}}(u_{\text{int}}, v_{\text{int}}) + b_{\text{ext},\kappa_0}((u_{\text{int}}(R), U), (v_{\text{int}}(R), V)) \tag{35}$$

for all test functions $(v_{\text{int}}, v) \in H^1((0, R)) \times H^+(S^1)$. Vice versa, if $(u_{\text{int}}, U) \in H^1((0, R)) \times H^+(S^1)$ is a solution to (35), then u_{int} is the restriction to $(0, R)$ of an outgoing solution $u \in H^1_{\text{loc}}(\mathbb{R}_+)$ to (13).

Proof For a radiating solution u to (13), $\mathscr{M}_{\kappa_0}\mathscr{L}u(\bullet + R) = \frac{1}{i\kappa_0}\mathscr{T}_-(u(R), U)$ with

$$U(z) = \frac{(\omega - \kappa_0)u(R)}{(\kappa_0 - \omega)z + (\kappa_0 + \omega)}, \qquad z \in S^1. \tag{36}$$

We have already shown, that $(u_{\text{int}}, U) \in H^1((0, R)) \times H^+(S^1)$ solves (35) for a special kind of test functions. Since these test functions are dense in $H^+(S^1)$ (see [20, Lemma A.2]) and since the bilinear form in (35) is continuous, (35) holds true for all test functions in $H^1((0, R)) \times H^+(S^1)$.

Conversely, let $(u_{\text{int}}, U) \in H^1((0, R)) \times H^+(S^1)$ be a solution to (35). As in the proof of Theorem 3.1, we start with test functions $(v_{\text{int}}, V) \in H^1((0, R)) \times H^+(S^1)$ with $v_{\text{int}} \equiv 0$. (35) reduces to the exterior bilinear form with $v_{\text{int}}(R) = 0$, which is coercive due to the last lemma. Hence, U is unique and due to the first part of the proof given by (36). Plugging U into (33) with arbitrary test functions $(v_{\text{int}}, V) \in H^1((0, R)) \times H^+(S^1)$ leads to $b_{\text{ext},\kappa_0}((u_{\text{int}}(R), U), (v_{\text{int}}(R), V)) = i\omega u_{\text{int}}(R)v_{\text{int}}(R)$, i.e. the correct Dirichlet-to-Neumann operator. □

Corollary 4.5 *Let $(u_{\mathrm{int}}, U) \in H^1((0,R)) \times H^+(S^1)$ be a solution to (35). Then u_{int} is independent of the parameter κ_0.*

4.3 Hardy Space Infinite Elements in One Dimension

Theorem 4.6 *With the same assumptions as in Theorem 4.4 let $V_{\mathrm{int,h}} \subset H^1((0,R))$ be a standard finite element space such that the orthogonal projection onto $V_{\mathrm{int,h}}$ converges point wise for all $v \in H^1((0,R))$. Moreover, let $\Pi_N \subset H^+(S^1)$ denote the set of polynomials of maximal order $N \in \mathbb{N}_0$ and*

$$V_{h,N} := V_{\mathrm{int,h}} \times \Pi_N \subset H^1((0,R)) \times H^+(S^1). \tag{37}$$

If (36) is uniquely solvable with solution $(u_{\mathrm{int}}, U) \in H^1((0,R)) \times H^+(S^1)$, then for sufficiently small h and sufficiently large N there exists a unique solution $(u_{\mathrm{int,h}}, U_N) \in V_{h,N}$ to (36) with test functions only in $V_{h,N}$. Moreover, there exist constants $C, c > 0$ independent of h and L such that

$$\|u_{\mathrm{int,h}} - u_{\mathrm{int}}\|_{H^1((0,R))} \leq C \left(\inf_{v_{\mathrm{int,h}} \in V_{\mathrm{int,h}}} \|u_{\mathrm{int}} - v_{\mathrm{int,h}}\|_{H^1((0,R))} + \exp(-cN) \right). \tag{38}$$

Proof Similarly to the linear complex scaling in Sect. 3.2 there holds a Gårding inequality in $H^1((0,R)) \times H^+(S^1)$ and the theorem is a consequence of the projection method applied to a compact perturbation of a coercive operator [27, Theorems 13.6 and 13.7]. For $\kappa_0 = \omega$ there is no approximation error in the Hardy space. Otherwise, U has a pole at $p := \left(\frac{\omega+\kappa_0}{\omega-\kappa_0}\right)$, which has absolute value larger than one. Since $U(z) = u_{\mathrm{int}}(R) \sum_{j=0}^{\infty} p^{-(j+1)} z^j$, $\inf_{V_N \in \Pi_N} \|V_N - U\|_{L^2(S^1)}$ converges exponentially with $p^{-(N+1)}$. □

In the one dimensional case the choice of the parameter κ_0 is obvious: If $\kappa_0 = \omega$, we have $U \equiv 0$ and the Hardy space method reduces to the correct Dirichlet-to-Neumann operator. In higher dimensions this is no longer the case, but typically $\kappa_0 \approx \omega$ remains a good choice.

In contrast to the complex scaling method, no truncation error occurs and no mesh in the exterior domain is needed. Moreover, we have exponential convergence with respect to the number of unknowns in the Hardy space. But we have to implement a new bilinear form and a new infinite element.

In the one dimensional case this is extremely easy. As basis functions for $(v_{\mathrm{int,h}}(R), V_N) \in \mathbb{C} \times \Pi_N$ we use $\Phi_{-1}(z) := (1,0)$ and the monomials $\Phi_j(z) := (0, z^j)$, $j = 0, \ldots, N$. The operators $\mathcal{T}_{\pm,N} : \mathbb{C} \times \Pi_N \to \Pi_{N+1} = \operatorname{span}\{z^0, \ldots, z^{N+1}\}$

in this basis are given by the bidiagonal matrices

$$T_{\pm,N} = \frac{1}{2}\begin{pmatrix} 1 & \pm 1 & & \\ & \ddots & \ddots & \\ & & 1 & \pm 1 \\ & & & 1 \end{pmatrix} \in \mathbb{R}^{(N+2)\times(N+2)}.$$

Since q_{κ_0} is orthogonal with respect to the monomials, we have

$$S_N := \left(q_{\kappa_0}\left(\mathscr{T}_+\Phi_j, \mathscr{T}_+\Phi_k\right)\right)_{j,k=-1}^{N} = (-2i\kappa_0)T_{+,N}^{\top}T_{+,N}, \tag{39a}$$

$$M_N := \left(q_{\kappa_0}\left(1/(i\kappa_0)\mathscr{T}_-\Phi_j, 1/(i\kappa_0)\mathscr{T}_-\Phi_k\right)\right)_{j,k=-1}^{N} = \frac{2i}{\kappa_0}T_{-,N}^{\top}T_{-,N} \tag{39b}$$

and finally

$$\left(b_{\text{ext},\kappa_0}\left(\Phi_j, \Phi_k\right)\right)_{j,k=-1}^{N} = S_N - \omega^2 M_N.$$

Only this matrix has to be implemented for Hardy space infinite elements in one dimension. The first row and the first column belong to $v_{\text{int}}(R)$ and $u_{\text{int}}(R)$ respectively. Hence, they have to be coupled with the corresponding degrees of freedoms of $V_{\text{int,h}} \subset H^1((0,R))$.

4.4 Radial Hardy Space Infinite Elements

As for the complex scaling method there exists different ways to generalize one dimensional infinite elements to two or three dimensions. For generalized Cartesian Hardy space infinite elements in two dimensions we refer to [33, Sect. 2.3.1.]. Here, we only use radial infinite elements. Since the correct mathematical framework is rather involved, we restrict ourselves to the presentation of the numerical method. For a mathematically correct construction of the method we refer to [20] and for proof of convergence to [16].

We use the same assumptions on the interface $\Gamma = \overline{\Omega_{\text{int}}} \cap \overline{\Omega_{\text{ext}}}$ as for the radial complex scaling in Sect. 3.3. For a parametrization $\eta : S \subset \mathbb{R}^{d-1} \to \Gamma$ of the interface, we parametrize the exterior domain by $F : \mathbb{R}_+ \times S \to \Omega_{\text{ext}}$ with

$$F(r,\varphi) := (1+r)\eta(\varphi), \qquad r > 0, \varphi \in S. \tag{40}$$

If η is piecewise smooth, due to Lemma 3.5 F is piecewise smooth in each segment of the exterior domain (see Fig. 1a) and at least continuous everywhere. Hence, the exterior bilinear form in (12) is given by

$$\int_{\mathbb{R}_+\times S}\left(\begin{pmatrix}\partial_r u_{\text{ext}}\\ \nabla_\varphi u_{\text{ext}}\end{pmatrix}^\top J^{-1}J^{-\top}\begin{pmatrix}\partial_r v_{\text{ext}}\\ \nabla_\varphi v_{\text{ext}}\end{pmatrix}-\omega^2 u_{\text{ext}}v_{\text{ext}}\right)\det(J)\,d(r,\varphi),$$

with $u_{\text{ext}} := u\circ F$, $v_{\text{ext}} := v\circ F$ and Jacobian $J(r,\varphi) = (\eta(\varphi),(1+r)D_\varphi\eta(\varphi)) \in \mathbb{R}^{d\times d}$. Since $J(r,\varphi) = \hat{J}(\varphi)\begin{pmatrix}1 & 0\\ 0 & (1+r)\mathrm{Id}_{d-1}\end{pmatrix}$ with $\hat{J}(\varphi) := (\eta(\varphi), D_\varphi\eta(\varphi))$, we define

$$\det(\hat{J}(\varphi))\hat{J}(\varphi)^{-1}\hat{J}(\varphi)^{-\top} =: \begin{pmatrix}G_{11}(\varphi) & G_{21}(\varphi)^\top\\ G_{21}(\varphi) & G_{22}(\varphi)\end{pmatrix},$$

with $G_{11}\in\mathbb{R}$, $G_{21}\in\mathbb{R}^{d-1}$ and $G_{22}\in\mathbb{R}^{(d-1)\times(d-1)}$. For the exterior bilinear form we have to discretize the two integrals

$$\begin{aligned}\int_{\Omega_{\text{ext}}}\nabla u\cdot\nabla v dx = \int_S\int_0^\infty \Big(&(1+r)^{d-1}\partial_r u_{\text{ext}}(r,\varphi)G_{11}(\varphi)\partial_r v_{\text{ext}}(r,\varphi)+\\ &(1+r)^{d-2}\partial_r u_{\text{ext}}(r,\varphi)G_{21}(\varphi)^\top\nabla_\varphi v_{\text{ext}}(r,\varphi)+\\ &(1+r)^{d-2}\nabla_\varphi u_{\text{ext}}(r,\varphi)^\top G_{21}(\varphi)\partial_r v_{\text{ext}}(r,\varphi)+\\ &(1+r)^{d-3}\nabla_\varphi u_{\text{ext}}(r,\varphi)^\top G_{22}(\varphi)\nabla_\varphi v_{\text{ext}}(r,\varphi)\Big)dr\,d\varphi\end{aligned} \tag{41a}$$

and

$$\int_{\Omega_{\text{ext}}} u\,v\,dx = \int_S\int_0^\infty u_{\text{ext}}(r,\varphi)v_{\text{ext}}(r,\varphi)(1+r)^{d-1}\det(\hat{J}(\varphi))\,dr\,d\varphi. \tag{41b}$$

Similar to Definition 4.1 we formulate the radiation condition in terms of the Möbius and Laplace transformed function: u_{ext} is outgoing if $\mathscr{M}_{\kappa_0}\mathscr{L}u_{\text{ext}}(\bullet,\varphi)$ exists for all $\varphi\in S$ and belongs to the Hardy space $H^+(S^1)$. In order to use this radiation condition, we transform the integrals $\int_0^\infty(\dots)dr$ in radial direction as in the one dimensional case into the bilinear form (31) using the identity [20, Lemma A.1]. Special attention has to be paid to the factors $(1+r)^{\pm1}$.

In order to treat these, we first study the Möbius and Laplace transformation of a multiplication operator. If $\mathscr{M}_{\kappa_0}\mathscr{L}v$ and $(\mathscr{M}_{\kappa_0}\mathscr{L}v)'$ belong to the Hardy space $H^+(S^1)$, then $\mathscr{M}_{\kappa_0}\mathscr{L}\{r\mapsto rv(r)\} = \frac{-1}{2i\kappa_0}\mathscr{D}\mathscr{M}_{\kappa_0}\mathscr{L}v$ with

$$(\mathscr{D}V)(z) := (z-1)^2V'(z) + (z-1)V(z), \qquad V\in H^+(S^1).$$

Hence, we deduce

$$\mathscr{M}_{\kappa_0}\mathscr{L}\left\{r\mapsto(1+r)^{\pm1}v(r)\right\} = \left(\mathscr{I}-\frac{1}{2i\kappa_0}\mathscr{D}\right)^{\pm1}\mathscr{M}_{\kappa_0}\mathscr{L}v, \tag{42}$$

with the identity operator $\mathscr{I} : H^+(S^1) \to H^+(S^1)$. For implementation the orthogonal projection onto Π_{N+1} of $\mathscr{I} - \frac{1}{2i\kappa_0}\mathscr{D} : \Pi_{N+1} \to \Pi_{N+2}$ is useful. In the monomial basis it is given by

$$D_{\kappa_0} := \mathrm{Id}_{N+2} - \frac{1}{2i\kappa_0}\begin{pmatrix} -1 & 1 & & & & & \\ 1 & -3 & 2 & & & & \\ & 2 & -5 & 3 & & & \\ & & \ddots & \ddots & \ddots & & \\ & & & N & -2N-1 & N & \\ & & & & N+1 & -2N-3 \end{pmatrix} \in \mathbb{C}^{(N+2)\times(N+2)}.$$

Note, that it is symmetric, i.e. $D_{\kappa_0}^\top = D_{\kappa_0}$. For the inverse operator $(\mathscr{I} - \frac{1}{2i\kappa_0}\mathscr{D})^{-1}$ we use the inverse $D_{\kappa_0}^{-1}$ as approximation.

We are now ready for a computation of the radial Hardy space infinite elements: Let $V_h \subset H^1(\Omega_{\text{int}})$ be a standard finite element discretization of $H^1(\Omega_{\text{int}})$ and let $\Psi_1, \dots, \Psi_{N_\Gamma}$ with $N_\Gamma \in \mathbb{N}$ denote the non vanishing traces of the finite element basis functions in $H^1(\Omega_{\text{int}})$. We need standard finite element matrices on the interface Γ

$$M_\Gamma := \left(\int_S \det(\hat{J}(\varphi))\, \Psi_j(\varphi)\Psi_k(\varphi)\, d\varphi\right)_{j,k=1}^{N_\Gamma}, \tag{43a}$$

$$S_\Gamma^{00} := \left(\int_S \Psi_j(\varphi) G_{11}(\varphi)\Psi_k(\varphi)\, d\varphi\right)_{j,k=1}^{N_\Gamma}, \tag{43b}$$

$$S_\Gamma^{10} := \left(\int_S \left(\nabla_\varphi \Psi_j(\varphi)\right)^\top G_{21}(\varphi)\, \Psi_k(\varphi)\, d\varphi\right)_{j,k=1}^{N_\Gamma}, \tag{43c}$$

$$S_\Gamma^{11} := \left(\int_S \left(\nabla_\varphi \Psi_j(\varphi)\right)^\top G_{22}(\varphi) \nabla_\varphi \Psi_k(\varphi)\, d\varphi\right)_{j,k=1}^{N_\Gamma}, \tag{43d}$$

and for the basis functions $\Phi_{-1}, \dots, \Phi_{N_r+1} \in \mathbb{C} \times \Pi_{N_r}$ defined in the last subsection the non-standard Hardy space infinite elements

$$M_r := \frac{2i}{\kappa_0} T_{-,N_r}^\top D_{\kappa_0}^{d-1} T_{-,N_r}, \tag{44a}$$

$$S_r^{00} := \frac{2i}{\kappa_0} T_{-,N_r}^\top D_{\kappa_0}^{d-3} T_{-,N_r}, \tag{44b}$$

$$S_r^{10} := -2 T_{+,N_r}^\top D_{\kappa_0}^{d-2} T_{-,N_r}, \tag{44c}$$

$$S_r^{11} := (-2i\kappa_0) T_{+,N_r}^\top D_{\kappa_0}^{d-1} T_{+,N_r}. \tag{44d}$$

For the tensor product basis functions $\Phi_j \otimes \Psi_k$ the discretization of $\int_{\Omega_{\text{ext}}} \nabla u \cdot \nabla v\, dx$ is due to (41) given by

$$S := S_r^{11} \otimes S_\Gamma^{00} + S_r^{10} \otimes S_\Gamma^{01\,\top} + S_r^{10\,\top} \otimes S_\Gamma^{10} + S_r^{00} \otimes S_\Gamma^{11}. \tag{45a}$$

The infinite element matrix for $\int_{\Omega_{\mathrm{ext}}} uv\, dx$ is given by

$$M := M_r \otimes M_\Gamma . \tag{45b}$$

As in the one dimensional case the basis functions $\Phi_{-1} \otimes \Psi_k$, $k = 1, \ldots, N_\Gamma$, have to be coupled to the corresponding basis functions in $H^1(\Omega_{\mathrm{int}})$ in order to ensure continuity at the interface Γ.

Remark 4.7 If u_{int} is a solution to (2), $\tilde{u}_{\mathrm{ext}}$ a solution to (1), $u_{\mathrm{ext}} = \tilde{u}_{\mathrm{ext}} \circ F$ and

$$U(z,\varphi) := \frac{2i\kappa_0 \mathscr{M}_{\kappa_0} \mathscr{L}\{u_{\mathrm{ext}}(\bullet,\varphi)\}(z) - u_{\mathrm{int}}(\eta(\varphi))}{z-1},$$

then (u_{int}, U) belongs to a subspace of $H^1(\Omega_{\mathrm{int}}) \times H^+(S^1) \otimes L^2(S)$. This subspace is constructed such that the bilinear forms in (41) are continuous (see [20, Eq. (3.7) and Lemma A.3] for a slightly different bilinear form or [16, Eq. (3.7)]). The Hardy space infinite element method is in this space a Galerkin method with tensor product elements. In [16] a Gårding inequality is shown leading to super-algebraic convergence with respect to the number of unknowns in the Hardy space.

The numerical results in [20, 32, 33] confirm this result. The main parameters of the method are κ_0, the number of unknowns in radial direction and the choice of the interface Γ. The numerical results indicate, that $\kappa_0 \approx \omega$ is recommendable. The interface Γ has to be chosen carefully.

Of course, in order to minimize the computational effort in Ω_{int}, one would like to choose Ω_{int} as small as possible. On the other hand the numerical results show, that the number of radial unknowns has to be increased, if the distance of Γ to a source of the scattered wave becomes smaller. For a distance of one or two wavelengths typically less than 10 radial unknowns are needed to ensure, that the error of the infinite elements is negligible. As mentioned in [16, Remark 3.3], highly anisotropic interfaces Γ should be avoided, when radial infinite elements are used. For such interfaces Cartesian infinite elements as in [33, Sect. 2.3.1.] are preferable.

5 Summary

We have presented PML and Hardy space infinite element methods for Helmholtz problems in open systems. Both methods are Galerkin methods and for both methods convergence can be shown. However, the type of convergence is different.

PML methods converge exponentially with increasing layer thickness. The convergence with respect to the finite element discretization of the perfectly matched layer depends on the used finite elements and typically is h^k for polynomials of order k. Hardy space infinite element methods converge super-algebraically with respect to the number of unknowns in radial direction and with the usual finite element convergence order for the interface unknowns. In a comparison in [33] the Hardy space infinite element method was superior to a complex scaling method for a two dimensional problem with inhomogeneous exterior domain. Of course, this might change in a different situation.

For the Hardy space infinite element method the programming effort typically is noticeable larger than for a standard PML. A non-standard infinite element with non-standard discretization matrix has to be implemented. The implementation of the matrix itself is very easy and does not require a remarkable effort. On the other hand a standard PML will not converge, if the layer thickness or the damping is not increased. Realizing this in a given finite element code is not an easy task neither.

One big advantage of both methods is the flexibility. In this chapter we have only used Helmholtz problems in free space, but the methods can be used for waveguides [1, 21] and inhomogeneous exterior domains [7, 33] as well. Moreover, they are not restricted to scalar problems in frequency domain.

Acknowledgements Support from the Austrian Science Fund (FWF) through grant P26252 is gratefully acknowledged.

References

1. É. Bécache, A.-S. Bonnet-BenDhia, and G. Legendre, *Perfectly matched layers for the convected Helmholtz equation*, SIAM Journal on Numerical Analysis, 42 (2004), pp. 409–433.
2. J.-P. Berenger, *A perfectly matched layer for the absorption of electromagnetic waves*, J. Comput. Phys., 114 (1994), pp. 185–200.
3. A. Bermúdez, L. Hervella-Nieto, A. Prieto, and R. Rodríguez, *An exact bounded perfectly matched layer for time-harmonic scattering problems*, SIAM J. Sci. Comput., 30 (2007/08), pp. 312–338.
4. J. H. Bramble and J. E. Pasciak, *Analysis of a finite PML approximation for the three dimensional time-harmonic Maxwell and acoustic scattering problems*, Math. Comp., 76 (2007), pp. 597–614 (electronic).
5. ———, *Analysis of a Cartesian PML approximation to acoustic scattering problems in $\mathbb{R}^2$ and $\mathbb{R}^3$*, J. Comput. Appl. Math., 247 (2013), pp. 209–230.
6. S. C. Brenner and L. R. Scott, *The mathematical theory of finite element methods*, vol. 15 of Texts in Applied Mathematics, Springer, New York, third ed., 2008.
7. Z. Chen, C. Liang, and X. Xiang, *An anisotropic perfectly matched layer method for Helmholtz scattering problems with discontinuous wave number*, Inverse Problems and Imaging, 7 (2013), pp. 663–678.
8. W. C. Chew and W. H. Weedon, *A 3d perfectly matched medium from modified Maxwell's equations with stretched coordinates*, Microwave Optical Tech. Letters, 7 (1994), pp. 590–604.
9. P. G. Ciarlet, *The finite element method for elliptic problems*, North-Holland Publishing Co., Amsterdam-New York-Oxford, 1978. Studies in Mathematics and its Applications, Vol. 4.
10. F. Collino and P. Monk, *The perfectly matched layer in curvilinear coordinates*, SIAM J. Sci. Comput., 19 (1998), pp. 2061–2090 (electronic).
11. D. Colton and R. Kress, *Inverse acoustic and electromagnetic scattering theory*, vol. 93 of Applied Mathematical Sciences, Springer-Verlag, Berlin, second ed., 1998.
12. L. Demkowicz and K. Gerdes, *Convergence of the infinite element methods for the Helmholtz equation in separable domains*, Numer. Math., 79 (1998), pp. 11–42.
13. L. Demkowicz and F. Ihlenburg, *Analysis of a coupled finite-infinite element method for exterior Helmholtz problems*, Numer. Math., 88 (2001), pp. 43–73.

14. P. L. DUREN, *Theory of H^p spaces*, Pure and Applied Mathematics, Vol. 38, Academic Press, New York, 1970.
15. D. GIVOLI, *High-order local non-reflecting boundary conditions: a review*, Wave Motion, 39 (2004), pp. 319–326.
16. M. HALLA, *Convergence of Hardy space infinite elements for Helmholtz scattering and resonance problems*, Preprint 10/2015, Institute for Analysis and Scientific Computing; TU Wien, 2013, ISBN: 978-3-902627-05-6, 2015.
17. M. HALLA, T. HOHAGE, L. NANNEN, AND J. SCHÖBERL, *Hardy space infinite elements for time harmonic wave equations with phase and group velocities of different signs*, Numerische Mathematik, (2015), pp. 1–37.
18. M. HALLA AND L. NANNEN, *Hardy space infinite elements for time-harmonic two-dimensional elastic waveguide problems*, Wave Motion, 59 (2015), pp. 94 – 110.
19. P. D. HISLOP AND I. M. SIGAL, *Introduction to spectral theory*, vol. 113 of Applied Mathematical Sciences, Springer-Verlag, New York, 1996. With applications to Schrödinger operators.
20. T. HOHAGE AND L. NANNEN, *Hardy space infinite elements for scattering and resonance problems*, SIAM J. Numer. Anal., 47 (2009), pp. 972–996.
21. ———, *Convergence of infinite element methods for scalar waveguide problems*, BIT Numerical Mathematics, 55 (2014), pp. 215–254.
22. T. HOHAGE, F. SCHMIDT, AND L. ZSCHIEDRICH, *Solving time-harmonic scattering problems based on the pole condition. I. Theory*, SIAM J. Math. Anal., 35 (2003), pp. 183–210.
23. ———, *Solving time-harmonic scattering problems based on the pole condition. II. Convergence of the PML method*, SIAM J. Math. Anal., 35 (2003), pp. 547–560.
24. F. IHLENBURG, *Finite element analysis of acoustic scattering*, vol. 132 of Applied Mathematical Sciences, Springer-Verlag, New York, 1998.
25. S. KIM AND J. E. PASCIAK, *The computation of resonances in open systems using a perfectly matched layer*, Math. Comp., 78 (2009), pp. 1375–1398.
26. ———, *Analysis of a Cartesian PML approximation to acoustic scattering problems in $\mathbb{R}^2$*, J. Math. Anal. Appl., 370 (2010), pp. 168–186.
27. R. KRESS, *Linear integral equations*, vol. 82 of Applied Mathematical Sciences, Springer-Verlag, New York, second ed., 1999.
28. R. KRESS, *Chapter 1.2.1 - specific theoretical tools*, in Scattering, R. P. Sabatier, ed., Academic Press, London, 2002, pp. 37 – 51.
29. M. LASSAS AND E. SOMERSALO, *On the existence and the convergence of the solution of the PML equations*, Computing, 60 (1998), pp. 229–241.
30. M. LASSAS AND E. SOMERSALO, *Analysis of the PML equations in general convex geometry*, Proc. Roy. Soc. Edinburgh Sect. A, 131 (2001), pp. 1183–1207.
31. N. MOISEYEV, *Quantum theory of resonances: Calculating energies, width and cross-sections by complex scaling*, Physics reports, 302 (1998), pp. 211–293.
32. L. NANNEN, T. HOHAGE, A. SCHÄDLE, AND J. SCHÖBERL, *Exact Sequences of High Order Hardy Space Infinite Elements for Exterior Maxwell Problems*, SIAM J. Sci. Comput., 35 (2013), pp. A1024–A1048.
33. L. NANNEN AND A. SCHÄDLE, *Hardy space infinite elements for Helmholtz-type problems with unbounded inhomogeneities*, Wave Motion, 48 (2010), pp. 116–129.
34. F. SCHMIDT AND P. DEUFLHARD, *Discrete transparent boundary conditions for the numerical solution of Fresnel's equation*, Computers Math. Appl., 29 (1995), pp. 53–76.
35. M. E. TAYLOR, *Partial differential equations. II*, vol. 116 of Applied Mathematical Sciences, Springer-Verlag, New York, 1996. Qualitative studies of linear equations.

On the Optimality of Shifted Laplacian in a Class of Polynomial Preconditioners for the Helmholtz Equation

Siegfried Cools and Wim Vanroose

Abstract This paper introduces and explores a class of polynomial preconditioners for the Helmholtz equation, denoted as expansion preconditioners $EX(m)$, that form a direct generalization to the classical complex shifted Laplace (CSL) preconditioner. The construction of the $EX(m)$ preconditioner is based on a truncated Taylor series expansion of the original Helmholtz operator inverse. The expansion preconditioner is shown to significantly improve Krylov solver convergence rates for growing values of the number of series terms m. However, the addition of multiple terms in the expansion also increases the computational cost of applying the preconditioner. A thorough cost-benefit analysis of the addition of extra terms in the $EX(m)$ preconditioner proves that the CSL or $EX(1)$ preconditioner is the most efficient member of the expansion preconditioner class for general practical and solver problem settings. Additionally, possible extensions to the expansion preconditioner class that further increase preconditioner efficiency are suggested, and numerical experiments in 1D and 2D are presented to validate the theoretical results.

1 Introduction

1.1 Overview of Recent Developments

The propagation of waves through any material is often mathematically modeled by the Helmholtz equation, which represents the time-independent waveforms in the frequency domain. For high wavenumbers, i.e. high spatial frequencies, the sparse linear system that results from the discretization of this PDE is distinctly indefinite,

S. Cools (✉) • W. Vanroose
Department of Mathematics and Computer Science, University of Antwerp, Middelheimlaan 1, 2020 Antwerp, Belgium
e-mail: siegfried.cools@uantwerp.be; wim.vanroose@uantwerp.be

D. Lahaye et al. (eds.), *Modern Solvers for Helmholtz Problems*, Geosystems Mathematics, DOI 10.1007/978-3-319-28832-1_3

causing most of the classic direct and iterative solution methods to perform poorly. Over the past few years, many different Helmholtz solution methods have been proposed, an overview of which can be found in [18]. Krylov subspace methods like GMRES [31] or BiCGStab [40] are known for their robustness and are hence frequently used for the solution of Helmholtz problems [3, 22, 27, 36]. However, due to the indefinite nature of the problem, Krylov methods are generally not efficient as Helmholtz solvers without the inclusion of a suitable preconditioner.

In this chapter we focus on the class of so-called shifted Laplace preconditioners, which were introduced in [25] and [15], and further analyzed in [16, 17]. It was shown in the literature that contrary to the original discretized Helmholtz system, the complex shifted Laplace (CSL) system (or damped Helmholtz equation) can be solved efficiently using iterative methods [18, 29]. Originally introduced by Fedorenko in [19], multigrid methods [7–9, 13, 37, 39] have been proposed as scalable solution methods for the shifted Laplace system in the literature [17]. Typically only one multigrid V-cycle on the CSL system yields a sufficiently good approximate inverse, which can then be used as a preconditioner to the original Helmholtz system [14, 16, 28, 32]. The main concept behind the shifted Laplace preconditioner is deceivingly simple: by shifting the spectrum of the Helmholtz operator down into the complex plane, close-to-zero eigenvalues (leading to near-singularity) that possibly destroy the iterative solver convergence are avoided, as illustrated in Fig. 1. Nevertheless, the results in this work will show that this apparent simplicity is exactly what makes the CSL preconditioner into a powerful tool for the iterative solution of the Helmholtz equation.

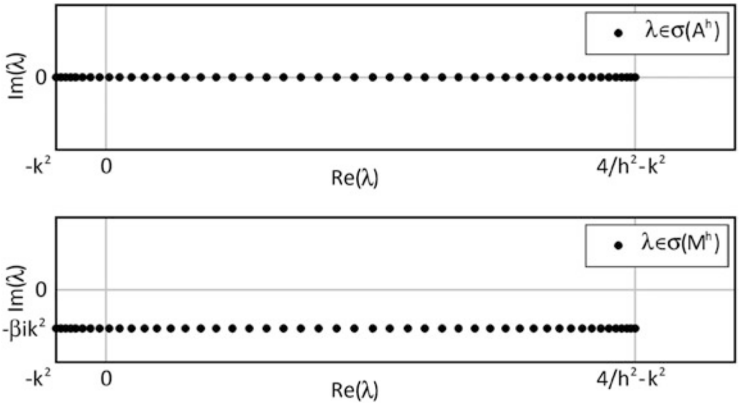


Fig. 1 *Top*: spectrum of the 1D Helmholtz operator discretized using second order finite differences on a 48 × 48 equidistant grid with standard homogeneous Dirichlet boundary conditions. *Bottom*: spectrum of the corresponding Complex Shifted Laplacian (CSL) operator

1.2 Outline of This Chapter

This study presents a generalization of the class of shifted Laplace preconditioners, which is based on a Taylor series expansion [2, 6, 26] of the original Helmholtz operator inverse around a complex shifted Laplacian operator. This formulation relates the original Helmholtz inverse to an infinite sum of shifted Laplace problems. By truncating the series we are able to define a class of so-called expansion preconditioners, denoted by $EX(m)$, where the number of terms m in the expansion is a parameter of the method. The expansion preconditioner directly generalizes the classic complex shifted Laplace preconditioner, since the CSL preconditioner appears as the operator $EX(1)$, i.e. the first term in the Taylor expansion.

Using a spectral analysis [12, 16, 41], the incorporation of additional series terms in the $EX(m)$ preconditioning polynomial is shown to greatly improve the spectral properties of the preconditioned system. When used as a preconditioner the $EX(m)$ operator hence allows for a significant reduction of the number of outer Krylov steps required to solve the Helmholtz problem for growing values of m. However, the addition of multiple terms in the expansion also increases the computational cost of applying the preconditioner, since each additional series term comes at the cost of one extra shifted Laplace operator inversion. The performance trade-off between the reduction of the number of outer Krylov iterations and the cost of additional terms (CSL inversions) in the preconditioner polynomial is analyzed in-depth. Furthermore, several theoretical extensions to the expansion preconditioner are introduced to improve preconditioner efficiency. These extensions provide the reader with supplementary insights into Helmholtz preconditioning. The proposed methods show similarities to the research on flexible Krylov methods [30, 35] and more specifically the work on multi-preconditioned GMRES [21].

A variety of numerical experiments are performed to validate the $EX(m)$ preconditioner and illustrate the influence of the number of series terms m on convergence. Performance and computational cost of CSL- and $EX(m)$-preconditioned BiCGStab [40] are compared for one- and two-dimensional Helmholtz model problems with absorbing boundary conditions. These absorbing boundaries are implemented using Exterior Complex Scaling (ECS) [1, 28, 34], a technique which has been related to Perfectly Matched Layers (PMLs) [5] by Chew and Weedon [11].

The remainder of this chapter is organized as follows. In Sect. 2 we outline the theoretical framework for this work and we introduce the expansion preconditioner class $EX(m)$. Following its formal definition, an overview of the theoretical, numerical and computational properties of the expansion preconditioner is given. Section 3 presents several possible extensions to the proposed $EX(m)$ preconditioner, which are shown to improve preconditioner efficiency even further. The new preconditioner class is validated in Sect. 4, where it is applied to a 1D and 2D Helmholtz benchmark problem. A spectral analysis confirms the asymptotic exactness of the expansion preconditioner as the number of terms m grows towards infinity. Additionally, experiments are performed to compare the efficiency and computational cost of the $EX(m)$ preconditioner for various values of m. Conclusions and a short discussion on the results are formulated in Sect. 5.

2 The Expansion Preconditioner

In this section we introduce the general framework for the construction of the expansion preconditioner. Starting from the notion of the classic complex shifted Laplace operator, we define the class of expansion preconditioners based on a Taylor expansion of the original Helmholtz operator inverse around a shifted Laplace problem with an arbitrary shift parameter. The definition of the expansion preconditioner is followed by an overview of the fundamental analytical and computational properties of the new preconditioner class.

2.1 The Complex Shifted Laplacian Preconditioner

In this work we aim to construct an efficient solution method for the d-dimensional Helmholtz equation

$$(-\Delta - k^2(\mathbf{x}))\, u(\mathbf{x}) = f(\mathbf{x}), \quad \mathbf{x} \in \Omega \subset \mathbb{R}^d, \tag{1}$$

with outgoing wave boundary conditions

$$u = \text{outgoing on } \partial\Omega, \tag{2}$$

where $-k(\mathbf{x})^2$ is a distinctly negative shift. Here $k \in \mathbb{R}$ designates the wavenumber, which will be assumed to be a spatially independent constant throughout most of this work for simplicity. However, note that the definitions in this section do not depend on this assumption. The above equation is discretized using a finite difference, finite element or finite volume scheme of choice, yielding a system of linear equations of the general form

$$Au = f, \tag{3}$$

where the matrix operator A represents a discretization of the Helmholtz operator $A \overset{d}{=} (-\Delta - k^2)$. It has been shown in the literature that iterative methods in general, and multigrid in particular, fail at efficiently solving the discretized Helmholtz system (3) due to the indefiniteness of the operator A [13, 18]. However, the addition of a complex shift in the Helmholtz system induces a damping. This allows for a more efficient solution of the resulting system, which is known as the complex shifted Laplacian (CSL)

$$(-\Delta - (1 + \beta i)k^2(\mathbf{x}))\, u(\mathbf{x}) = f(\mathbf{x}), \quad \mathbf{x} \in \Omega \subset \mathbb{R}^d, \tag{4}$$

where $\beta \in \mathbb{R}^+$ is the complex shift (or damping) parameter that is conventionally chosen to be positive [15, 17]. The discretized shifted Laplace system is denoted by

$$Mu = f, \tag{5}$$

where M is the discretization of the complex shifted Helmholtz operator $\mathcal{M} \overset{d}{=} (-\Delta - (1 + \beta i)k^2)$.

It is well-known that this system can be solved using multigrid when the shift parameter β is sufficiently large [14–16]. Furthermore, it has been shown in the literature that the complex shift parameter β should be chosen at least as large as the critical value $\beta_{\min}$. For a multigrid V-cycle with the standard linear interpolation and full weighting restriction operators and a traditional ω-Jacobi or Gauss-Seidel type smoothing scheme, the rule-of-thumb value for $\beta_{\min}$ was shown to be 0.5 for a V(1,0)-cycle [16], and lies roughly around 0.6 when solving the CSL system using a V(1,1) multigrid cycle [12]. Note that the latter value for the shift will be commonly used throughout this chapter.

The exact solution to (4) is a damped waveform that is fundamentally different from the solution to the original Helmholtz system (1). However, the inverse of the shifted matrix operator M can be used as a preconditioner to the original system. This preconditioning technique is known as the complex shifted Laplace preconditioner, and was shown to perform well as a preconditioner for Helmholtz problems, see [14, 16, 41].

2.2 The Class of Expansion Preconditioners

2.2.1 General Criteria for Preconditioner Efficiency

The aim of this work is to extend the existing class of shifted Laplace preconditioners to obtain a more efficient preconditioner to the original Helmholtz system (3). Moreover, we would like to construct a preconditioner M such that $M \approx A$ or, as an equivalent measure, we require that the eigenvalues of the preconditioned operator $M^{-1}A$ are concentrated around one, i.e.

$$\mathrm{spec}(M^{-1}A) \approx 1. \tag{6}$$

In the context of an efficient iterative solution, the condition number $\kappa = \kappa(M^{-1}A)$ of the preconditioned system is often related to the number of Krylov iterations [23]. Although this relation is somewhat heuristic in the context of non-normal matrices [38], we believe that it provides an insightful intuition on preconditioner efficiency. The above requirement (6) can hence be broadened by requiring that the condition number κ approximately equals one, i.e.

$$\kappa(M^{-1}A) \approx 1. \tag{7}$$

We trivially note that by the above characterizations the best preconditioner to the Helmholtz system (3) is (a good approximation to) the original operator A itself. However, the discrete operator A is generally close to singular and hence cannot be easily inverted in practice. On the other hand, given a sufficiently large complex shift, the CSL system (4) can be approximately inverted using e.g. a (series of) multigrid V-cycle(s). Note that the complex shifted Laplacian is generally not a very precise approximation to the original Helmholtz operator unless the shift parameter β is very small, which in practice is not achievable due to stability requirements on the preconditioner inversion, see [12].

Following the general idea of preconditioning, the optimal Helmholtz preconditioner M_{opt} thus ideally satisfies the following two key properties, which are inspired by analogous conditions that were formulated in [20]:

(P1) M_{opt}^{-1} is a good approximation to the exact inverse of the original Helmholtz operator A^{-1}, such that condition (6) is satisfied, i.e. the spectrum of the preconditioned operator $M_{\text{opt}}^{-1}A$ is clustered around one,

(P2) for any given vector v, $M_{\text{opt}}^{-1}v$ can be efficiently computed iteratively. This implies a good approximation to $M_{\text{opt}}^{-1}v$ is found after a 'moderate' number of iterations of the chosen method, with a manageable cost per iteration.

In the context of this chapter condition (P2) is satisfied if M_{opt}^{-1} can be formulated in terms of shifted Laplace operator inverses with a shift parameter β that is sufficiently large to ensure a stable iterative solution of the shifted Laplace inverses. Indeed, given that the shift parameter β is sufficiently large, a good approximation to the CSL inverse can be computed using e.g. only one multigrid V-cycle, see [16].

Note that due to the strong indefiniteness of the Helmholtz operator conditions (P1) and (P2) are generally incompatible. The classic CSL preconditioner trivially satisfies the second condition given that the shift parameter β is large enough, however, the first condition is typically violated when β is large. In the following we aim at constructing a preconditioning scheme based upon the shifted Laplace preconditioner, which effectively satisfies *both* of the above conditions.

2.2.2 Taylor Series Expansion of the Inverse Helmholtz Operator

The complex shifted Laplace preconditioning operator M can be written more generally as

$$M(\beta) \overset{d}{=} -\Delta - k^2(\mathbf{x}) + P(\beta, \mathbf{x}), \tag{8}$$

where $P(\beta, \mathbf{x})$ is a possibly spatially dependent linear operator in the shift parameter $\beta \in \mathbb{R}^+$, satisfying $P(0, \mathbf{x}) = 0$, such that $M(0) = A$. The above formulation

(8) characterizes, apart from the complex shifted Laplace (CSL) operator, also the concept of complex stretched grid (CSG), where the underlying grid is rotated into the complex plane. This results in a damped Helmholtz problem that is equivalent to complex shifted Laplacian, see [29]. For the remainder of this text we however assume $P(\beta, \mathbf{x}) = -\beta i k^2(\mathbf{x})$ as suggested by (4).

We define an operator functional f based on the general shifted Laplacian operator M as follows:

$$f(\beta) := M(\beta)^{-1} = (-\Delta - (1 + \beta i)k^2)^{-1}. \tag{9}$$

Choosing $\beta \equiv 0$ in the above expression results in the inverse of the original Helmholtz operator $A = M(0)$, whereas choosing $\beta > 0$ yields the inverse of the shifted Laplace operator $M(\beta)$. The derivatives of the functional f are given by

$$f^{(n)}(\beta) = n!\,(k^2 i)^n\,(-\Delta - (1 + \beta i)k^2)^{-(n+1)}, \tag{10}$$

for any $n \in \mathbb{N}$. Constructing a Taylor series expansion [2, 6, 26] of $f(\beta)$ around a fixed shift $\beta_0 \in \mathbb{R}^+$ leads now to the following expression

$$f(\beta) = \sum_{n=0}^{\infty} \frac{f^{(n)}(\beta_0)}{n!}(\beta - \beta_0)^n, \tag{11}$$

where the derivatives $f^{(n)}(\beta_0)$ are defined by (10). Note that the derivatives of f in (11) are negative powers of the complex shifted Laplace operator $M(\beta_0)$. By evaluating the functional $f(\beta)$ in $\beta = 0$ and by choosing a sufficiently large positive value for β_0, Eq. (11) yields an approximation of the original Helmholtz operator inverse in terms of CSL operator inverses, i.e.

$$\begin{aligned} f(0) = M(0)^{-1} &= \sum_{n=0}^{\infty} (-\beta_0)^n \frac{f^{(n)}(\beta_0)}{n!} \\ &= \sum_{n=0}^{\infty} (-\beta_0 k^2 i)^n\,(-\Delta - (1 + \beta_0 i)k^2)^{-(n+1)} \\ &= \sum_{n=0}^{\infty} (-\beta_0 k^2 i)^n\,M(\beta_0)^{-(n+1)}. \end{aligned} \tag{12}$$

Hence, the computation of the infinite series of easy-to-compute inverse CSL operators $M(\beta_0)$ with an arbitrary shift parameter β_0 asymptotically results in the exact inversion of the original Helmholtz operator $M(0) = A$.

2.2.3 Definition of the Expansion Preconditioner

By truncating the expansion in (12), we can now define a new class of polynomial Helmholtz preconditioners. For a given m, each particular member of this preconditioner class is denoted as the *expansion preconditioner* of degree m,

$$EX(m) := \sum_{n=0}^{m-1} \alpha_n \, (-\Delta - (1+\beta_0 i)k^2)^{-(n+1)}, \tag{13}$$

where the coefficients $\alpha_0, \ldots, \alpha_{m-1}$ are defined as

$$\alpha_n = (-\beta_0 k^2 i)^n, \qquad (0 \leq n \leq m-1), \tag{14}$$

by the Taylor series expansion (10)–(11). The expansion preconditioner $EX(m)$ is hence a degree m polynomial in the inverse complex shifted Laplace operator $M(\beta_0)^{-1} = (-\Delta - (1+\beta_0 i)k^2)^{-1}$. The above Taylor series approach appears quite natural. However, other series approximations to the Helmholtz operator inverse may be constructed using alternative choices for the series coefficients. We refer to Sect. 3 for a more elaborate discussion on the choice of the series coefficients.

2.2.4 Properties of the Expansion Preconditioner

Following the formal definition (13), we formulate some essential properties of the $EX(m)$ class of preconditioners in this section. Firstly, one trivially observes that the classic CSL preconditioner is a member of the class of expansion preconditioners. Indeed, the complex shifted Laplace inverse $M(\beta_0)^{-1}$ is the first order term in the Taylor expansion (12), and hence we have $M(\beta_0)^{-1} = EX(1)$.

By including additional terms in the preconditioning polynomial (i.e. for $m \to \infty$), the $EX(m)$ preconditioner becomes an increasingly accurate approximation to the original Helmholtz operator inverse A^{-1}. Hence, the class of $EX(m)$ preconditioners is *asymptotically exact*, since

$$\lim_{m\to\infty} EX(m) = \sum_{n=0}^{\infty} \alpha_n \, (-\Delta - (1+\beta_0 i)k^2)^{-(n+1)} = M(0)^{-1}. \tag{15}$$

This implies that, if we assume that the computational cost of computing the inverse matrix powers in (15) is manageable, $EX(m)$ satisfies *both* conditions (P1) and (P2) for efficient Helmholtz preconditioning suggested in Sect. 2.2.1. It should be stressed that (P1) in fact holds asymptotically, and is thus in practice only satisfied when a large number of series terms is taken into account. The $EX(m)$ preconditioner is thus expected to be increasingly more efficient for growing m, which suggests a significant reduction in the number of outer Krylov iterations. On the other hand, condition (P2) is satisfied when m is not too large. This creates a trade-off for the value of m, which is commented on in Sect. 2.2.5.

While the approximation precision of the $EX(m)$ preconditioner clearly benefits from the addition of multiple terms in the expansion, note that the accuracy of the m-term $EX(m)$ approximation is governed by the truncation error of the series (11). This truncation error is of order $\mathscr{O}({\beta_0}^m)$ for any $EX(m)$ preconditioner (with $m > 0$), i.e.

$$M(0)^{-1} = EX(m) + \mathscr{O}({\beta_0}^m). \tag{16}$$

The efficiency of the $EX(m)$ preconditioner is hence also intrinsically dependent on the value of the shift parameter β_0. However, it is well-known that β_0 cannot be chosen below a critical value for iterative (multigrid) solver stability, which typically lies around 0.5 or 0.6, see [12, 16]. Consequently, it is clear from (16) that convergence of the Taylor series (11) is slow, being in the order of ${\beta_0}^m$. This indicates that a large number of terms has to be taken into account in the $EX(m)$ polynomial to obtain a high-precision approximation to the original Helmholtz operator inverse.

2.2.5 Computational Cost of the Expansion Preconditioner

The inclusion of additional series terms yields a higher-order $EX(m)$ preconditioner polynomial, which is expected to improve preconditioning efficiency as derived above. Therefore, if the computational cost of the CSL inversions would be negligible compared to the cost of applying one Krylov iteration, there would theoretically be no restriction on the number of terms that should be included in $EX(m)$. Unfortunately, even when approximating each CSL inversion by one V-cycle, the computational cost of the CSL inversions is the main bottleneck for the global cost of the solver in practice. Indeed, while the addition of multiple series terms improves performance, it also increases the computational cost of applying the preconditioner. In this section we briefly expound on the computational cost of the $EX(m)$ preconditioner using a simple theoretical cost model.

We model the computational cost of the $EX(m)$ preconditioner by assuming that its cost is directly proportional to the number of CSL operator inversions that need to be performed when solving the preconditioning system. Each additional term in the series (11) requires exactly one extra shifted Laplace system to be inverted, since the $EX(m)$ polynomial can be constructed as follows:

$$EX(1)w = \alpha_0 \underbrace{M(\beta_0)^{-1}w}_{:=v_0},$$

$$EX(2)w = \alpha_0 v_0 + \alpha_1 \underbrace{M(\beta_0)^{-1}v_0}_{:=v_1},$$

$$\vdots$$

$$EX(m)w = \sum_{n=0}^{m-2} \alpha_n v_n + \alpha_{m-1} \underbrace{M(\beta_0)^{-1} v_{m-2}}_{:= v_{m-1}}, \qquad (17)$$

where $w \in \mathbb{R}^N$ is a given vector of size N, i.e. the number of unknowns. Note that all complex shifted Laplace systems in (17) feature the same shift parameter β_0. Hence, an additional CSL system of the form

$$M(\beta_0)\, v_i = v_{i-1}, \quad \text{with } v_{-1} := w, \quad (0 \le i \le m-1), \qquad (18)$$

has to be solved for each term in the expansion preconditioner, resulting in a total of m inversions to be performed. We again stress that in practice the shifted Laplace inverse $M(\beta_0)^{-1}$ is never calculated explicitly, but the approximate solution to (18) is rather computed iteratively by a multigrid V-cycle.

The question rises whether the reduction in outer Krylov iterations when using the multi-term $EX(m)$ preconditioning polynomial compensates for the rising cost of the additional (approximate) CSL inversions. Let the computational cost of one approximate CSL inversion be denoted as one work unit (1 WU), and let the total computational cost of the $EX(m)$ preconditioner in a complete $EX(m)$-preconditioned Krylov solve be denoted by $\mathscr{C}_{tot}$. If the number of Krylov iterations until convergence (up to a fixed tolerance `tol`) is $p(m)$, then $\mathscr{C}_{tot} = m \cdot p(m)$ WU. Hence, it should hold that $p(m) < C/m$ for some moderate constant C for the cost of the $EX(m)$ preconditioner to support the inclusion of multiple series terms. Numerical experiments in Sect. 4 of this work will show that this is generally not the case for the Taylor expansion polynomial in many practical applications, and the classic CSL preconditioner is hence the optimal choice for a preconditioner in the $EX(m)$ class. In the next section we propose several extensions to the $EX(m)$ preconditioner class to further improve its performance.

3 Extensions and Further Analysis

In this section we propose two theoretical extensions to the Taylor series representation (11) for the inverse Helmholtz operator. These extensions provide essential insights into the expansion preconditioners and aim at further improving preconditioner efficiency. The primary goal is to improve the performance of the expansion preconditioner class, resulting in a more cost-efficient preconditioner with respect to the number of terms m. The theoretical results obtained in this section are supported by various numerical experiments in Sect. 4 that substantiate the analysis and illustrate the efficiency of the extended expansion preconditioner.

3.1 The Expansion Preconditioner As a Stationary Iterate

We first consider an extension of the $EX(m)$ preconditioner class that allows manual optimization of the series coefficients for each degree m. To this aim, we illustrate how the $EX(m)$ preconditioner can be interpreted as the m-th iterate of a specific fixed-point iteration. Consequently, a class of extended expansion preconditioners is defined by optimizing the fixed-point iteration.

3.1.1 Taylor Series-Based Polynomial Preconditioners As Fixed-Point Iterates

Recall that the foundation for the Taylor-based $EX(m)$ preconditioner class presented in Sect. 2 is the reformulation of the original Helmholtz operator inverse as a Taylor series. Equation (12) can alternatively be reformulated as

$$\begin{aligned} f(0) = M(0)^{-1} &= M(\beta_0)^{-1} \sum_{n=0}^{\infty} (-\beta_0 k^2 i)^n M(\beta_0)^{-n} \\ &= M(\beta_0)^{-1} \left(I + \beta_0 k^2 i M(\beta_0)^{-1}\right)^{-1}, \end{aligned} \tag{19}$$

where the last equation follows from the limit expression

$$\sum_{n=0}^{\infty} x^n = \frac{1}{1-x} \quad \text{for } |x| \leq 1. \tag{20}$$

For general (matrix) operators, this series is known in the literature as a Neumann series [42]. Note that for the matrix equation (19) the requirement $|x| \leq 1$ is met if $\rho(-\beta_0 k^2 i M(\beta_0)^{-1}) \leq 1$, which is trivially satisfied since $|\lambda_j(M(\beta_0))| \geq |-\beta_0 k^2 i|$ for all $j = 1, \dots, N$, where we assume $M(\beta_0) \in \mathbb{C}^{N\times N}$. For notational convenience, let us denote

$$L := -\beta_0 k^2 i M(\beta_0)^{-1}. \tag{21}$$

so that the last line in (19) reads

$$M(0)^{-1} = M(\beta_0)^{-1} (I - L)^{-1}. \tag{22}$$

Equation (22) shows that inverting the indefinite Helmholtz operator $M(0)$ is equivalent to subsequently inverting the operator $M(\beta_0)$ followed by the inversion of the operator $(I - L)$. The first operator is simply the inverse of a CSL operator and can easily be solved iteratively. However, the second inversion is non-trivial, as it requires the solution u to the system

$$(I-L)u=b, \tag{23}$$

given a right-hand side $b \in \mathbb{R}^N$. The linear system (23) can alternatively be formulated as a fixed-point iteration (or stationary iterative method)

$$u^{(m+1)}=Lu^{(m)}+b, \qquad m>0. \tag{24}$$

One observes that by setting $u^{(0)}=0$, the m-th iterate of this fixed-point method generates the $EX(m)$ preconditioner, since

$$u^{(m)}=\left(\sum_{n=0}^{m-1} L^n\right) b \approx (I-L)^{-1}\, b, \tag{25}$$

which implies

$$M(0)^{-1}\, b \approx M(\beta_0)^{-1}\, u^{(m)}=\left(M(\beta_0)^{-1}\sum_{n=0}^{m-1} L^n\right) b = EX(m)\, b. \tag{26}$$

The fixed-point iteration (24) thus asymptotically generates the Taylor series expansion (11).

The truncation analysis in Sect. 2.2.4 indicated that the Taylor series displays a slow convergence. Alternatively, convergence behavior can now be analyzed by studying the convergence of the fixed-point iteration (24), which is governed by the spectral radius of the iteration matrix

$$\rho(L)=\rho\left(-\beta_0 i k^2\left(-\Delta-(1+\beta_0 i)k^2\right)^{-1}\right). \tag{27}$$

For $\beta_0>0$ this spectral radius tends to be relatively close to one, since

$$\rho(L)=\max_{1\le j\le N}\left|\frac{-\beta_0 i k^2}{\lambda_j-(1+\beta_0 i)k^2}\right| = \left(\min_{1\le j\le N}\left|1-\frac{\lambda_j-k^2}{\beta_0 i k^2}\right|\right)^{-1}\approx 1, \tag{28}$$

where λ_j $(1\le j\le N)$ are the eigenvalues of the negative Laplacian. Hence, the slow convergence of the Taylor series is apparent from the spectral properties of the fixed-point iteration.

3.1.2 Weighted Fixed-Point Iteration to Improve Convergence

To obtain a convergence speed-up, the fixed-point iteration (24) can be substituted by a more general weighted stationary iteration

$$u^{(m+1)}=(1-\omega)u^{(m)}+\omega L u^{(m)}+\omega b, \qquad \omega\in[0,2], \quad m>0, \tag{29}$$

for $b \in \mathbb{R}^N$, which can alternatively be written as

$$u^{(m+1)} = \tilde{L}u^{(m)} + \omega b, \qquad \omega \in [0, 2], \quad m > 0, \tag{30}$$

using the notation $\tilde{L} := (1 - \omega)I + \omega L$. Setting $u^{(0)} = 0$ as the initial guess, this iteration constructs a different class of polynomial expansion preconditioners for any choice of $\omega \in [0, 2]$, as follows;

$$EX_\omega(m)\, b := M(\beta_0)^{-1}\, u^{(m)}, \quad m > 0, \tag{31}$$

where $u^{(m)}$ is given by (30). We call this class of preconditioners the *extended expansion preconditioner* of degree m, and denote them by $EX_\omega(m)$ to indicate their dependency on ω.

The parameter ω allows us to modify the coefficients of the series expansion to obtain a more suitable truncated series approximation to the original Helmholtz inverse. Note that the Taylor expansion preconditioner $EX(m)$ can be constructed from iteration (30) by setting the parameter $\omega = 1$. Additionally, note that choosing $\omega = 0$ trivially yields the CSL preconditioner $M(\beta_0)^{-1} = EX_0(m)$ for all $m > 0$.

The careful choice of the parameter $\omega \in [0, 2]$ in (31) possibly results in a series that converges faster than the Taylor series generated by (24). Indeed, the parameter ω can be chosen to modify the polynomial coefficients such that

$$\rho(\tilde{L}) < \rho(L), \tag{32}$$

yielding a series that converges faster than the Taylor series (11). Consequently, for the right choice of ω, the $EX_\omega(m)$ truncated series results in a more efficient preconditioner than the original $EX(m)$ polynomial of the same degree. We refer to the numerical results in Sect. 4 to support this claim.

3.2 GMRES-Based Construction of the Expansion Polynomial

The optimization of the series coefficients through the choice of the parameter ω in the $EX_\omega(m)$ operator can be generalized even further by letting the coefficients of the series expansion vary freely. Moreover, the coefficients can be optimized depending on the degree m of the preconditioning polynomial. By replacing the stationary fixed-point iterations in (24)–(30) by a more advanced Krylov solution method, an optimal degree m polynomial approximation to the original Helmholtz operator can be constructed.

3.2.1 Optimization of the Expansion Preconditioner

For the fixed-point method (24) we essentially constructed the Taylor polynomial $EX(m)$ using a fixed linear combination of the following basis polynomials

$$\mathscr{R}(m) = \left\{L,\, L\,(I+L),\, L\left(I+L+L^2\right),\, \ldots\right\}. \tag{33}$$

Alternatively, the extended expansion preconditioner $EX_\omega(m)$ was formed as a fixed linear combination of the basis polynomials

$$\mathscr{S}(m) = \left\{\omega L,\, (2\omega-\omega^2)L + \omega^2 L^2,\, \ldots\right\}, \tag{34}$$

for the weighted fixed-point iteration (30). Note that in this section L designates the inverse of the CSL operator as before, up to scaling by a scalar constant, i.e.

$$L := M(\beta_0)^{-1}. \tag{35}$$

As a direct generalization of the above constructions, we now consider the coefficients in each step of the iterative procedure to be variable. This boils down to constructing the preconditioning polynomial from the monomial basis

$$\mathscr{T}(m) = \text{span}\left\{L,\, L^2,\, L^3,\, \ldots,\, L^m\right\}. \tag{36}$$

Since the preconditioning polynomial asymptotically results in the exact Helmholtz operator inverse, we can alternatively solve the preconditioning system

$$L\,v = g, \qquad \text{with } \; v = Au \; \text{ and } \; g = Lf, \tag{37}$$

using m steps of GMRES [31], which results in construction of an m-term polynomial from the Krylov basis

$$\mathscr{K}_m(L,\, A\,r_0) = \text{span}\left\{A\,r_0,\, L\,A\,r_0,\, L^2\,A\,r_0,\, \ldots,\, L^{m-1}\,A\,r_0\right\}. \tag{38}$$

After an additional multiplication with the operator L, the m-th Krylov subspace exactly generates a preconditioning polynomial from the basis $\mathscr{T}(m)$ (36). Hence, a generalized expansion preconditioner can be constructed by applying m steps of GMRES on the system $Lv = g$, which allows for a free choice of the polynomial coefficients. Moreover, since GMRES minimizes the residual over the m-th Krylov subspace, the resulting m-term preconditioner is the optimal polynomial approximation of degree m to the exact Helmholtz inverse. These concepts resemble the principles of polynomial smoothing by a GMRES(m)-based construction, see [10].

3.2.2 Simultaneous Construction of Preconditioner and Krylov Solver Basis

Further extending the above methodology, we outline the theoretical framework for an integrated construction of the preconditioner polynomial in the Krylov subspace construction at the solver runtime level. The key notions in this section show some similarities to the work on multi-preconditioned GMRES in [21]. We additionally refer to the closely related literature on flexible Krylov solvers [30, 35].

Consider the Krylov method solution to the $EX(m)$-preconditioned Helmholtz system

$$EX(m)\,Au = EX(m)f, \tag{39}$$

for a fixed polynomial degree m. Using GMRES to solve this system, the k-th residual $r_k = EX(m)(f - Au^{(k)})$ is minimized over the Krylov subspace

$$\mathscr{K}_k(EX(m)\,A, r_0) = \text{span}\left\{r_0,\; EX(m)\,A\,r_0,\; \ldots,\; (EX(m)\,A)^{k-1}\,r_0\right\}. \tag{40}$$

Note that this basis spans an entirely different subspace compared to the Krylov subspace (38). Indeed, for the basis terms in (40) the preconditioning polynomial degree is fixed at m while the power of the Helmholtz operator A is variable. To form the generalized $EX(m)$ polynomial in (38) on the other hand, the powers of the CSL inverse L vary but the power of A is fixed at one.

A combination of the two principles characterized by (38) and (40) can be made by embedding the iterative procedure for the construction of the expansion preconditioner polynomial $EX(m)$ in the governing Krylov solver. This results in a mixed basis consisting of a structured mixture of powers of the inverse CSL operator L and the original system matrix A applied to the initial residual r_0. We denote the mixed basis corresponding to the $EX(m)$ preconditioner by

$$\mathscr{K}_k^{EX(m)}(A, r_0) = \text{span}\left\{L^{i\cdot j}A^j\,r_0\;:\, 1 \leq i < m,\; 0 \leq j < k\right\}, \tag{41}$$

for any $m \geq 1$. One trivially observes from the definition (41) that

$$\mathscr{K}_k^{EX(m-1)}(A, r_0) \subset \mathscr{K}_k^{EX(m)}(A, r_0), \tag{42}$$

which generalizes the embedding of the $EX(1)$ or CSL preconditioner in the class of $EX(m)$ expansion operators to the mixed basis setting. The subspace spanned by (41) theoretically allows for the simultaneous construction of the preconditioning polynomial and the solution of the preconditioned system. However, the mixed basis $\mathscr{K}_k^{EX(m)}(A, r_0)$ generally does not span a Krylov subspace for any $m > 1$, making its practical construction non-trivial.

As mentioned earlier, the addition of extra terms in the $EX(m)$ polynomial improves the polynomial approximation to the exact Helmholtz operator inverse, resulting in faster convergence in terms of outer Krylov iterations. This implies lower powers of the Helmholtz operator A in the mixed basis (41). However, the addition of extra terms in the polynomial $EX(m)$ also increases the number of vectors constituting the mixed basis, and hence gives rise to a higher computational cost of the total method. This trade-off between preconditioner approximation precision and computational cost in function of the number of terms m is apparent from (41).

As a final remark, note that the extensions proposed in this section are mainly intended as an insightful theoretical framework. In practice the GMRES-based extended preconditioner proposed in Sect. 3.2.1 is unlikely to perform significantly better than the extended $EX_\omega(m)$ preconditioner proposed in the previous section. This is a consequence of the fact that, given a sufficiently large shift parameter β, the convergence speed of any monomial-based series of this type is slow, as was already pointed out in Sect. 2.2.4.

4 Numerical Results

In this section we present experimental results that illustrate the practical application of the class of expansion preconditioners to enhance the Krylov convergence on a 1D and 2D Helmholtz benchmark problem. The primary aim is to validate the $EX(m)$ expansion preconditioner for degrees $m > 1$ and illustrate the asymptotic behavior of the $EX(m)$ preconditioner as $m \to \infty$. Initial numerical experiments in Sects. 4.1–4.3 will use exact inverses of the complex shifted Laplace operators appearing in the polynomial preconditioner. We then introduce a multigrid V(1,1)-cycle as an approximate solver for the CSL systems in the expansion. The performance of the $EX(m)$ preconditioner is consequently compared to that of the classic complex shifted Laplace or $EX(1)$ preconditioner. Additionally, the extensions to the class of generalized $EX_\omega(m)$ preconditioners and the combination of the preconditioner polynomial and Krylov basis construction (see Sect. 3) are shown to display the potential to improve the preconditioner's efficiency.

4.1 *Problem Setting: A 1D Constant Wavenumber Helmholtz Problem with Absorbing Boundary Conditions*

Consider the one-dimensional constant wavenumber Helmholtz model problem on the unit domain

$$(-\Delta - k^2)\, u(x) = f(x), \quad x \in \Omega = [0, 1], \tag{43}$$

where the right-hand side $f(x)$ represents a unit source in the domain center. The wavenumber is chosen to be $k^2 = 2 \times 10^4$. The equation is discretized using a standard Shortley-Weller finite difference discretization [33], required to treat the absorbing boundary layers (see below). The unit domain Ω is represented by an $N + 1 = 257$ equidistant point grid, defined as $\Omega^h = \{x_j = jh, 0 \le j \le N\}$, respecting the physical wavenumber criterion $kh = 0.5524 < 0.625$ for a minimum of 10 grid points per wavelength [4].

To simulate outgoing waves near the edges of the numerical domains, we use *exterior complex scaling* [11, 34], or ECS for short, adding absorbing layers to both sides of the numerical domain. The absorbing layers are implemented by the addition of two artificial complex-valued extensions to the left and right of the domain Ω, defined by the complex grid points $\{z_j = \exp(i\theta_{ECS})\,x_j\}$, where $x_j = jh$, for $-N/4 \leq j < 0$ and $N < j \leq 5/4N$. The ECS complex scaling angle that determines the inclination of the extensions in the complex plane is chosen as $\theta_{ECS} = \pi/6$. The two complex-valued extensions feature $N/4$ grid points each, implying the discretized Helmholtz equation takes the form of an extended linear system

$$Au = f, \tag{44}$$

where $A \in \mathbb{C}^{\frac{3}{2}N\times\frac{3}{2}N}$. The discretized right-hand side $f = (f_j) \in \mathbb{C}^{\frac{3}{2}N}$ is defined as

$$f_j = f(x_j) = \begin{cases} 1 \text{ for } j = N/2, \\ 0 \text{ elsewhere,} \end{cases} \tag{45}$$

representing a unit source located in the center of the domain for this example.

The Helmholtz model problem (43) is solved using $EX(m)$-preconditioned BiCGStab [40] up to a relative residual tolerance $\|r_p\|/\|r_0\| < \texttt{tol} = 1\text{e}{-8}$. The CSL operator inverses in the $EX(m)$ polynomial are either computed exactly using LU factorization for the purpose of analysis, or approximated using one multigrid V-cycle as is common in realistic applications. Note that the complex shift parameter β in the CSL operators is chosen as $\beta = 0.6$, which guarantees multigrid V(1,1)-cycle stability [12].

4.2 Spectral Analysis of the Expansion Preconditioner

To analyze the efficiency of the $EX(m)$ Helmholtz preconditioner we perform a classic eigenvalue analysis of the $EX(m)$-preconditioned Helmholtz operator. For convenience of analysis, the CSL inversions in the $EX(m)$ preconditioning polynomial are solved using a direct method in this section.

The typical pitchfork shaped spectrum of the indefinite Helmholtz operator with ECS boundary conditions is shown in the left panel of Fig. 2. The leftmost eigenvalue is located near $-k^2 = -2 \times 10^4$, while the rightmost eigenvalue is close to $4/h^2 - k^2 \approx 2.4 \times 10^5$, cf. [28]. The right panel of Fig. 2 shows the spectrum of the $EX(m)$-preconditioned Helmholtz operator for various numbers of terms in the Taylor polynomial $EX(m)$. Note how the spectra become more clustered around 1 when additional series terms are taken into account, illustrating the asymptotic exactness of the $EX(m)$ preconditioner class.

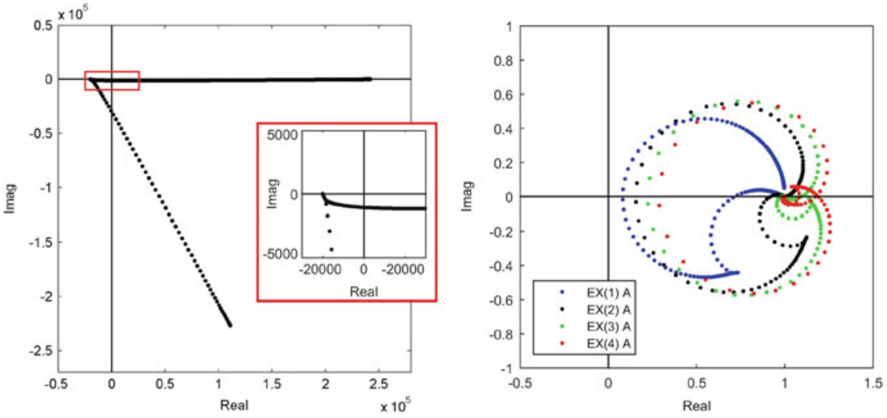


Fig. 2 Spectral analysis of the discretized 1D Helmholtz model problem (43). Exact preconditioner inversion. *Left*: spectrum of the Helmholtz operator A with ECS absorbing boundary conditions. *Right*: spectrum of the polynomial preconditioned operator $EX(m)A$ for various values of m. The spectrum becomes more clustered around 1 for increasing values of m

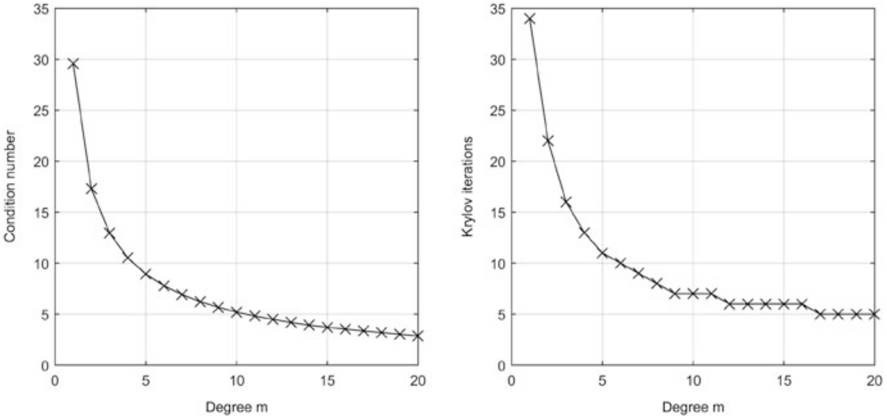


Fig. 3 Conditioning and $EX(m)$-BiCGStab performance on the discretized 1D Helmholtz model problem (43). Exact preconditioner inversion. *Left*: condition number of the preconditioned operator $EX(m)A$ as a function of m. *Right*: number of $EX(m)$-BiCGStab iterations required to solve the Helmholtz system (44) as a function of m

The condition number of the preconditioned operator $\kappa(EX(m)A)$ is displayed in the left panel of Fig. 3 as a function of m. One observes that conditioning improves significantly by the addition of extra terms in the polynomial $EX(m)$. This observation is reflected in the number of Krylov iterations required to solve the problem, which is displayed in Fig. 3 (right panel) for a range of values of m. The number of Krylov iterations (right panel) appears to be directly proportional to the condition number of the preconditioned system (left panel).

4.3 Performance Analysis of the Expansion Preconditioner

It is clear from the spectral analysis that the addition of multiple series terms improves the conditioning of the preconditioned system. In this section, the performance of the $EX(m)$ preconditioner is analyzed using the simple theoretical cost model introduced in Sect. 2.2.5.

The experimentally measured performance of the $EX(m)$-BiCGStab solver on the model problem (43) is displayed in Table 1. The table shows the number of Krylov iterations required to solve system (44) until convergence up to `tol` = 1e−8 (first column), the iteration ratio compared to the standard CSL preconditioner (second column), and the effective number of work units (CSL inversions) for the entire run of the method (third column). The $EX(m)$ preconditioner becomes increasingly more efficient in reducing the number of Krylov iterations in function of larger m. Comparing e.g. the $EX(3)$ preconditioner to the classic $EX(1)$ (CSL) scheme, one observes that the number of Krylov iterations is slightly more than halved. The largest improvement is obtained by adding the first few terms, which is a consequence of the slow Taylor convergence. Note that the computational cost of the Krylov solver itself is not incorporated into this cost model.

Although higher-order series approximations clearly result in a qualitatively better preconditioner, the number of Krylov iterations is not reduced sufficiently to compensate for the cost of the extra inversions. Indeed, while the addition of multiple series terms in the $EX(m)$ preconditioner improves the spectral properties of the preconditioned system, the increased computational cost of the extra CSL inversions appears to be a bottleneck for performance. Hence, one observes that standard CSL preconditioning—which takes only one series term into account—is the most cost-efficient, requiring a minimum of 34 WU for the entire solve.

Table 1 Performance of $EX(m)$-BiCGStab for different values of m on the discretized 1D Helmholtz model problem (43)

m	$p(m)$	$\frac{p(m)}{p(1)}$	$m \cdot p(m)$
1	34	1.00	34 WU
2	22	0.65	44 WU
3	16	0.47	48 WU
4	13	0.38	52 WU
5	11	0.32	55 WU

Exact preconditioner inversion. Column 1: number of BiCGStab iterations $p(m)$ required to solve the system (44). Col. 2: iteration ratio compared to classic CSL. Col. 3: preconditioner computational cost based on $p(m)$

Table 2 Performance of $EX(m)$-BiCGStab for different values of m on the discretized 1D Helmholtz model problem (43)

m	$p(m)$	$\frac{p(m)}{p(1)}$	CPU time (s)
1	49	1.00	0.57
2	39	0.80	0.68
3	34	0.69	0.80
4	31	0.63	0.88
5	30	0.61	1.02

V(1,1)-cycle approximate preconditioner inversion. Column 1: number of BiCGStab iterations $p(m)$ required to solve the system (44). Col. 2: iteration ratio compared to CSL. Col. 3: CPU time until convergence (in seconds)

4.4 Multigrid Inversion of the Expansion Preconditioner

For convenience of analysis a direct inversion of the preconditioning scheme was used in the previous sections. However, in realistic large-scale applications the terms of the $EX(m)$ preconditioner often cannot be computed directly. Instead, the CSL systems comprising the $EX(m)$ polynomial are approximately solved using some iterative method. In this section we use one geometric multigrid V(1,1)-cycle to approximately solve the CSL systems, which is a standard approach in the Helmholtz literature [14, 16]. The V(1,1)-cycle features the traditional linear interpolation and full weighting restriction as intergrid operators, and applies one weighted Jacobi iteration (with parameter 2/3) as a pre- and post-smoother. The choice of the damping parameter $\beta = 0.6$ guarantees stability of the multigrid solver for the inversion of the CSL operators, see [12].

Table 2 summarizes the number of Krylov iterations and CPU time[1] required to solve system (44) using $EX(m)$-BiCGStab for different values of m. The application of the $EX(m)$ operator is approximately computed using a total of m V(1,1)-cycles, see Sect. 2.2.5. The corresponding convergence histories are shown in Fig. 4. The addition of terms in the $EX(m)$ preconditioner reduces the number of Krylov iterations as expected, although the improvement is less pronounced compared to the results in Table 1 due to non-exact inversion of the CSL operators. However, the increased cost to (approximately) compute the additional series terms for larger values of m is clearly reflected in the timings. Hence, in terms of preconditioner computational cost, the classic CSL or $EX(1)$ preconditioner is the most cost-efficient member of the $EX(m)$ preconditioner class for this benchmark problem.

[1] Hardware specifications: Intel Core i7-2720QM 2.20 GHz CPU, 6MB Cache, 8 GB RAM. Software specifications: Windows 7 64-bit OS, experiments implemented in MATLAB R2015a.

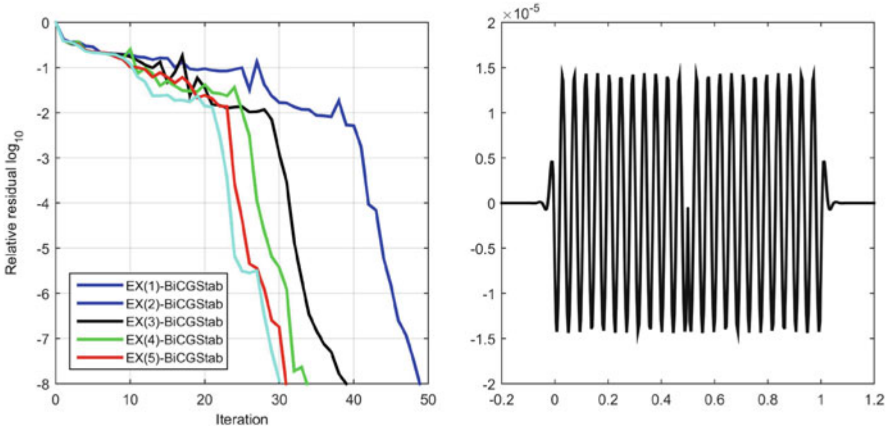


Fig. 4 Convergence of $EX(m)$-BiCGStab and solution of the discretized 1D Helmholtz model problem (43). V(1,1)-cycle approximate preconditioner inversion. *Left*: $EX(m)$-BiCGStab relative residual history $\|r_p\|/\|r_0\|$ for various values of m. Vertical axis in log scale. *Right*: numerical solution $u(x)$ to Eq. (43) up to the relative residual tolerance `tol` $=$ 1e$-$8. The ECS absorbing boundary layer rapidly damps the solution outside the unit domain $\Omega = [0, 1]$

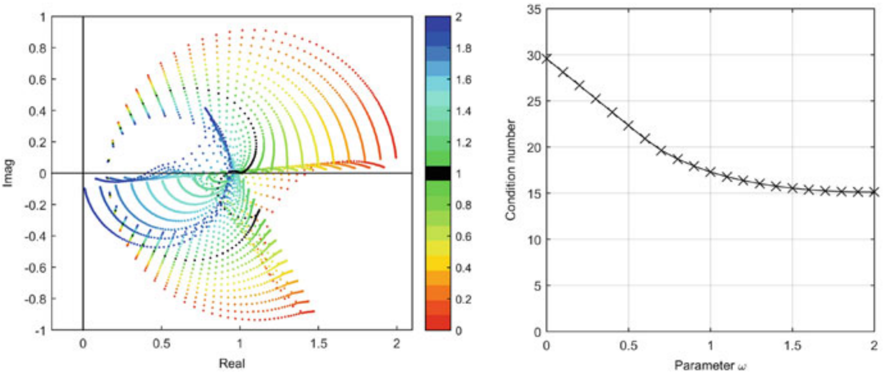


Fig. 5 Spectral analysis of the discretized 1D Helmholtz model problem (43). Exact preconditioner inversion. *Left*: spectrum of the preconditioned operator $EX_\omega(2)A$ for different values of the parameter ω. The spectrum for $EX(m)A$ with $\omega = 1$ is indicated in *black*, see also Fig. 2. *Right*: condition number of the preconditioned operator $EX_\omega(2)A$ as a function of the weight ω

4.5 Validation of the Extended Expansion Preconditioner

In this section we validate the generalizations to the expansion preconditioner proposed in Sect. 3 on the 1D Helmholtz model problem (43).

The weighted fixed-point iteration (30) generates the generalized class of expansion preconditioners $EX_\omega(m)$. Figure 5 shows the spectrum (left panel) and condition number (right panel) of the Helmholtz operator preconditioned by the

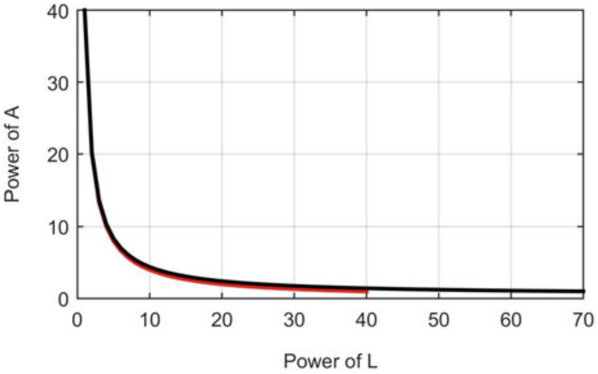


Fig. 6 Performance of the $EX(m)$ preconditioner formed by the construction of the mixed basis (41) on the discretized 1D Helmholtz model problem (43). Maximum power of L, i.e. the polynomial preconditioner degree m, vs. maximum power of A, i.e. the number of Krylov iterations $p(m)$. *Red curve*: theoretical upper bound required for cost-efficiency. *Black curve*: experimentally measured results

two-term $EX_\omega(2)$ polynomial for different values of the parameter ω. Note that the condition number of the standard $EX(2)$ preconditioner ($\omega = 1$) is $\kappa(EX(2)A) = 17.29$. A small improvement in conditioning is achievable through the right choice of the parameter ω, reducing the condition number to $\kappa(EX_\omega(2)A) = 15.13$ for parameter choices around $\omega = 2$. With the optimal choice for ω, the condition number of the classic CSL preconditioner $EX_0(2)$ is halved when using $EX_\omega(2)$, suggesting a halving of the number of Krylov iterations may be achievable by using $EX_\omega(2)$ instead of the classic CSL preconditioner. The smaller condition number implies an increase in performance compared to the $EX(2)$ preconditioner, making the addition of extra terms in $EX_\omega(m)$ theoretically cost-efficient.

A first step towards a simultaneous construction of the $EX(m)$ preconditioning polynomial and the outer Krylov basis resulting in the mixed basis (41) is illustrated in Fig. 6. The black curve shows the experimentally determined maximum power of A (Krylov iterations) versus the maximum power of L (terms in the preconditioning polynomial) required to solve the Helmholtz benchmark problem (43) up to a relative residual tolerance $\texttt{tol} = 1\text{e}{-8}$. Subject to this tolerance, a solution is found either after 40 $EX(1)$-BiCGStab iterations or alternatively after one $EX(70)$-BiCGStab iteration. Indeed, the incorporation of 70 terms in the $EX(m)$ expansion preconditioner effectively reduces the number of outer Krylov iterations to 1. However, note that to be cost-efficient with respect to the number of CSL inversions, the same solution should be found using the $EX(40)$ (degree 40) polynomial preconditioner. The red curve represents a constant number of CSL inversions for the total run of the method. To ensure cost-efficiency of the class of $EX(m)$ preconditioners for $m > 1$, the experimental black curve should fall below the theoretical red curve, which is not the case. Hence, the most simple case of the $EX(1)$ or CSL preconditioner can again be considered optimal w.r.t. cost-efficiency.

4.6 Problem Setting: A 2D Constant Wavenumber Helmholtz Problem with Absorbing Boundary Conditions

To conclude this work we extend the above 1D model problem (43) to a two dimensional Helmholtz problem. Numerical results for solution using $EX(m)$-preconditioned BiCGStab and GMRES are provided, and we comment on the scalability of the expansion preconditioner functionality to higher spatial dimensions.

Consider the two-dimensional constant wavenumber Helmholtz model problem

$$(-\Delta - k^2)\, u(x,y) = f(x,y), \quad (x,y) \in \Omega = [0,1]^2, \tag{46}$$

where the right-hand side $f(x,y)$ again represents a unit source in the domain center and outgoing wave boundary conditions are implemented using Exterior Complex Scaling with $\theta_{ECS} = \pi/6$. We consider two different wavenumbers, namely $k^2 = 5\text{e}{+}3$ and $k^2 = 2\text{e}{+}4$, corresponding to a moderate- and high-energetic wave respectively. Equation (46) is discretized using $n_x = n_y = 128$ (for $k^2 = 5\text{e}{+}3$) and $n_x = n_y = 256$ (for $k^2 = 2\text{e}{+}4$) real-valued grid points in each spatial dimension, respecting the wavenumber criterion $kh < 0.625$ [4] in every direction. Note that the discretized 2D Helmholtz operator with ECS boundary conditions can be constructed from the 1D Helmholtz operator using Kronecker products, i.e. $A^{2D} = A_x^{1D} \otimes I_y + I_x \otimes A_y^{1D}$, where $I_x \in \mathbb{C}^{n_x \times n_x}$ and $I_y \in \mathbb{C}^{n_y \times n_y}$ are identity matrices.

Figure 7 shows the $EX(m)$-BiCGStab solver convergence history for various values of m (left) and the solution $u(x,y)$ (right) to the 2D model problem (46) for different wavenumbers and corresponding discretizations. The $EX(m)$ preconditioner is approximately inverted using m multigrid V(1,1)-cycles with a weighted Jacobi smoother (weighting parameter 4/5). The corresponding number of BiCGStab iterations until convergence up to the relative residual tolerance `tol`$=$ 1e$-$8 are displayed in Table 3. The observations from the 1D spectral analysis extend directly to the 2D setting, as the table shows that the use of the $EX(m)$ preconditioner results in a significant reduction of the number of outer Krylov iterations for growing values of m.

Table 3 additionally features the CPU timings for the wavenumbers $k^2 = 5\text{e}{+}3$ and 2e+4. Note that although the number of Krylov iterations is reduced as a function of the preconditioner degree m, the CPU timings are rising in function of m. The computational cost of performing m multigrid V-cycles (compared to just one V-cycle for the CSL preconditioner) has a clear impact on the CPU timings. As a result, it is often advisable in view of cost-efficiency to restrict the expansion to the first term only (CSL preconditioner), where only one V-cycle is required to obtain an (approximate) preconditioner inverse. These observations are comparable to the 1D results from Sect. 4.4.

In Table 4 results for solving the same system using $EX(m)$-preconditioned GMRES are shown. Note that contrary to the BiCGStab results in Table 3, the

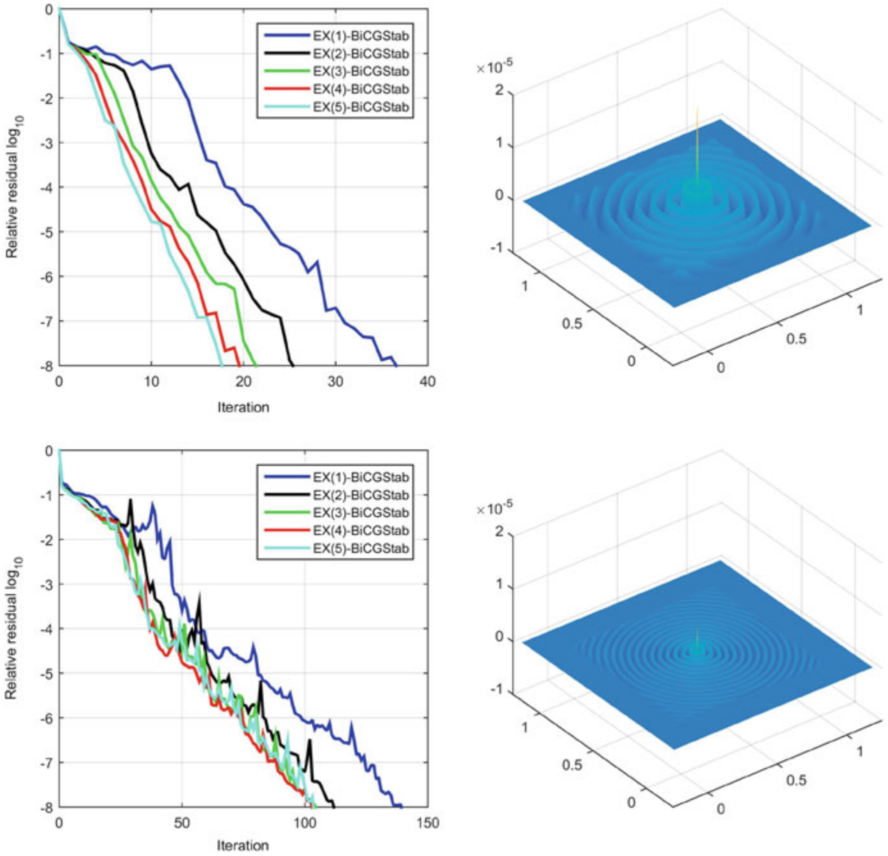


Fig. 7 Convergence of $EX(m)$-BiCGStab and solutions of the discretized 2D Helmholtz model problem (46). *Top*: wavenumber $k^2 = 5 \times 10^3$ and $N = n_x \times n_y = 128 \times 128$ unknowns. *Bottom*: wavenumber $k^2 = 2 \times 10^4$ and $N = 256 \times 256$ unknowns. V(1,1)-cycle approximate preconditioner inversion. *Left*: $EX(m)$-BiCGStab relative residual history $\|r_p\|/\|r_0\|$ for various values of m. Vertical axis in log scale. *Right*: numerical solution $u(x, y)$ to Eq. (46) up to the relative residual tolerance $\texttt{tol} = 1\text{e}{-8}$

$EX(m)$ preconditioner does appear to be cost-efficient for the 2D Helmholtz problem with wavenumber $k^2 = 2\text{e}{+}4$ for values of $m > 1$. Indeed, in this case the optimal preconditioner with respect to CPU time is the second-order polynomial $EX(2)$, which reduces the number of outer Krylov iterations to 191 and minimizes the CPU time to 477.3 s, compared to 540.8 s for $EX(1)$-GMRES. The main cause for this phenomenon is the relatively high per-iteration computational cost of the GMRES algorithm, which is caused by the orthogonalization procedure with respect to the Krylov subspace basis vectors. This cost is especially pronounced for larger iteration numbers. Hence, the good approximation properties of the $EX(m)$ preconditioner for higher values of m may prove useful when the Krylov iteration

Table 3 Performance of $EX(m)$-BiCGStab for different values of m on the discretized 2D Helmholtz model problem (46)

	$n_x \times n_y = 128 \times 128$ $k^2 = 5e+3$		$n_x \times n_y = 256 \times 256$ $k^2 = 2e+4$	
m	$p(m)$	CPU time (s)	$p(m)$	CPU time (s)
1	37	14.7	140	157.0
2	26	19.0	112	210.8
3	22	23.1	105	277.9
4	20	26.3	104	351.0
5	18	30.5	103	436.2
	EX(m)-preconditioned BiCGStab			

V(1,1)-cycle approximate preconditioner inversion. Columns 1 and 3: number of $EX(m)$-BiCGStab iterations $p(m)$. Cols. 2 and 4: total CPU time until convergence (in seconds)

Table 4 Performance of $EX(m)$-GMRES for different values of m on the discretized 2D Helmholtz model problem (46)

	$n_x \times n_y = 128 \times 128$ $k^2 = 5e+3$		$n_x \times n_y = 256 \times 256$ $k^2 = 2e+4$	
m	$p(m)$	CPU time (s)	$p(m)$	CPU time (s)
1	67	19.0	233	540.8
2	50	22.9	191	477.3
3	41	26.8	175	497.5
4	37	31.8	168	547.8
5	34	35.9	165	611.3
	EX(m)-preconditioned GMRES			

V(1,1)-cycle approximate preconditioner inversion. Columns 1 and 3: number of $EX(m)$-GMRES iterations $p(m)$. Cols. 2 and 4: total CPU time until convergence (in seconds)

cost is non-marginal compared to the cost of applying the preconditioner. The idea of polynomial preconditioners for GMRES has recently been proposed in the literature, see e.g. [24].

5 Conclusions

In this work we have proposed a theoretical framework that generalizes the classic shifted Laplacian preconditioner by introducing the class of polynomial expansion preconditioners $EX(m)$. This concept extends the one-term CSL preconditioner to an m-term Taylor polynomial in the inverse complex shifted Laplace operator. The outer iteration for solving the preconditioned system used in this work is a traditional Krylov iteration such as BiCGStab or GMRES.

Key properties of the $EX(m)$ preconditioner class are its structure as a finite m-term series of powers of CSL inverses (Neumann series), and its resulting asymptotic exactness, meaning $EX(m)$ approaches A^{-1} in the limit for m going to infinity. The polynomial structure of the $EX(m)$ preconditioner makes it easy to compute an iterative approximation to this polynomial, using e.g. m multigrid V(1,1)-cycles (one for each term), which are guaranteed to converge given that the complex shift is chosen to be sufficiently large.

The preconditioning efficiency of the $EX(m)$ preconditioner is validated using a classic eigenvalue analysis. The addition of extra terms in the preconditioning polynomial clusters the spectrum around 1, which reduces the condition number of the preconditioned Helmholtz operator and suggests a significant reduction in the number of outer Krylov iterations for large m.

Numerical results on 1D and 2D Helmholtz benchmark problems support the theoretical results. The number of outer Krylov iterations is reduced significantly by the higher degree expansion preconditioners. Unfortunately, the computational cost of applying the $EX(m)$ preconditioner is directly proportionate to the number of terms m, since an extra (approximate) CSL inversion is required for each additional term in the polynomial. The use of a large number of series terms is thus not necessarily guaranteed to result in a more cost-efficient preconditioner.

Following the numerical results of the 1D and 2D experiments, these conclusions are expected to be directly generalizable to higher spatial dimensions. Moreover, the constant wavenumber experiments performed in this paper are extensible to Helmholtz problems with heterogeneous and/or discontinuous wavenumbers, provided that the complex shift variable in the $EX(m)$ polynomial is large enough to allow for a stable numerical solution of the shifted Laplace systems for all wavenumber regimes occurring in the problem. Hence, from a practical preconditioning interest, the simple one-term shifted Laplace preconditioner appears to be the optimal member of the $EX(m)$ class for many applications and problem configurations.

Furthermore, two generalizations to the class of expansion preconditioners were presented and analyzed. These generalizations primarily prove to be insightful from a theoretical point of view. It is shown that the Taylor expansion preconditioner can be substituted by an optimal m-term polynomial which is theoretically cost-efficient. However, in practical applications the reduction of the number of outer Krylov iterations due to $EX(m)$ preconditioning often does not pay off to the cost of the extra approximate multigrid inversions.

For systems in which the cost of applying the outer Krylov step becomes significant relative to the cost of the preconditioner application, the use of multiple terms in the $EX(m)$ expansion could result in a more cost-efficient solver. Possible scenarios for this include the application of non-restarted GMRES as the outer Krylov solver, the use of higher order discretization schemes for the original Laplace operator (while maintaining second order discretization for the preconditioner), etc.

Additionally, the treatment of extremely large-scale HPC systems on massively parallel hardware may warrant the need for higher-order polynomial preconditioners. Since Krylov methods are typically communication (or bandwidth) bound instead of compute bound in this context, polynomial preconditioning directly reduces the number of communication bottlenecks (dot-products) by reducing the number of Krylov iterations, while simultaneously improving the arithmetic intensity of the solver. A detailed analysis of these individual scenarios is however well beyond the scope of this text, and is left for future work.

The generalization to the shifted Laplace preconditioner for Helmholtz problems proposed in this work is particularly valuable from a theoretical viewpoint, providing fundamental insights into the concept of shifted Laplace preconditioning by situating the classic complex shifted Laplace operator in a broader theoretical context, and proving the classic CSL preconditioner to be the most cost-efficient member of the $EX(m)$ preconditioner class for the most common practical problems.

Acknowledgements This research in this chapter was funded by the Research Council of the University of Antwerp under BOF grant number 41-FA070300-FFB5342. The scientific responsibility rests with its authors. The authors would like to cordially thank Dr. Bram Reps for fruitful discussions on the subject.

References

1. J. Aguilar and J.M. Combes. A class of analytic perturbations for one-body Schrödinger Hamiltonians. *Communications in Mathematical Physics*, 22(4):269–279, 1971.
2. G.B. Arfken and H.J. Weber. *Mathematical methods for physicists: A comprehensive guide*. Academic press, 2011.
3. A. Bayliss, C.I. Goldstein, and E. Turkel. An iterative method for the Helmholtz equation. *Journal of Computational Physics*, 49(3):443–457, 1983.
4. A. Bayliss, C.I. Goldstein, and E. Turkel. On accuracy conditions for the numerical computation of waves. *Journal of Computational Physics*, 59(3):396–404, 1985.
5. J.P. Berenger. A perfectly matched layer for the absorption of electromagnetic waves. *Journal of computational physics*, 114(2):185–200, 1994.
6. R. Bhatia. *Matrix analysis*. Springer, 1997.
7. A. Brandt. Multi-level adaptive solutions to boundary-value problems. *Mathematics of Computation*, 31(138):333–390, 1977.
8. A. Brandt and S. Ta'asan. Multigrid method for nearly singular and slightly indefinite problems. *Multigrid Methods II, Lecture Notes in Math.*, 1228:99–121, 1986.
9. W.L. Briggs, V.E. Henson, and S.F. McCormick. *A Multigrid Tutorial*. Society for Industrial Mathematics, Philadelphia, 2000.
10. H. Calandra, S. Gratton, R. Lago, X. Pinel, and X. Vasseur. Two-level preconditioned Krylov subspace methods for the solution of three-dimensional heterogeneous Helmholtz problems in seismics. *Numerical Analysis and Applications*, 5(2):175–181, 2012.
11. W.C. Chew and W.H. Weedon. A 3D perfectly matched medium from modified Maxwell's equations with stretched coordinates. *Microwave and optical technology letters*, 7(13):599–604, 1994.
12. S. Cools and W. Vanroose. Local Fourier analysis of the complex shifted Laplacian preconditioner for Helmholtz problems. *Numerical Linear Algebra with Applications*, 20(4):575–597, 2013.
13. H.C. Elman, O.G. Ernst, and D.P. O'Leary. A multigrid method enhanced by Krylov subspace iteration for discrete Helmholtz equations. *SIAM Journal on scientific computing*, 23(4):1291–1315, 2002.

14. Y.A. Erlangga and R. Nabben. On a multilevel Krylov method for the Helmholtz equation preconditioned by shifted Laplacian. *Electronic Transactions on Numerical Analysis*, 31(403–424):3, 2008.
15. Y.A. Erlangga, C.W. Oosterlee, and C. Vuik. On a class of preconditioners for solving the Helmholtz equation. *Applied Numerical Mathematics*, 50(3–4):409–425, 2004.
16. Y.A. Erlangga, C.W. Oosterlee, and C. Vuik. A novel multigrid based preconditioner for heterogeneous Helmholtz problems. *SIAM Journal on Scientific Computing*, 27(4):1471–1492, 2006.
17. Y.A. Erlangga, C. Vuik, and C.W. Oosterlee. Comparison of multigrid and incomplete LU shifted-Laplace preconditioners for the inhomogeneous Helmholtz equation. *Applied Numerical Mathematics*, 56(5):648–666, 2006.
18. O.G. Ernst and M.J. Gander. Why it is difficult to solve Helmholtz problems with classical iterative methods. *Lecture Notes in Computational Science and Engineering*, 83:325–363, 2012.
19. R.P. Fedorenko. The speed of convergence of one iterative process. *Zhurnal Vychislitel'noi Matematiki i Matematicheskoi Fiziki*, 4(3):559–564, 1964.
20. M.J. Gander, I.G. Graham, and E.A. Spence. Applying GMRES to the Helmholtz equation with shifted Laplacian preconditioning: what is the largest shift for which wavenumber-independent convergence is guaranteed? *Numerische Mathematik*, pages 1–48, 2015.
21. C. Greif, T. Rees, and D.B. Szyld. Multi-preconditioned GMRES. *Technical report - UBC CS TR-2011-12*, 2011.
22. A. Laird and M. Giles. Preconditioned iterative solution of the 2D Helmholtz equation. Technical report, NA-02/12, Comp. Lab. Oxford University UK, 2002.
23. J. Liesen and Z. Strakos. *Krylov subspace methods: principles and analysis*. Oxford University Press, 2012.
24. Q. Liu, R.B. Morgan, and W. Wilcox. Polynomial preconditioned GMRES and GMRES-DR. *SIAM Journal on Scientific Computing*, 37(5):S407–S428, 2015.
25. M.M.M. Made. Incomplete factorization-based preconditionings for solving the Helmholtz equation. *Int. J. Numer. Meth. Eng.*, 50(5):1077–1101, 2001.
26. P.M. Morse and H. Feshbach. Methods of theoretical physics, International series in pure and applied physics. *New York: McGraw-Hill*, 1(1953):29, 1953.
27. D. Osei-Kuffuor and Y. Saad. Preconditioning Helmholtz linear systems. *Applied Numerical Mathematics*, 60(4):420–431, 2010.
28. B. Reps and W. Vanroose. Analyzing the wave number dependency of the convergence rate of a multigrid preconditioned Krylov method for the Helmholtz equation with an absorbing layer. *Numerical Linear Algebra with Applications*, 19(2):232–252, 2012.
29. B. Reps, W. Vanroose, and H. Zubair. On the indefinite Helmholtz equation: Complex stretched absorbing boundary layers, iterative analysis, and preconditioning. *Journal of Computational Physics*, 229(22):8384–8405, 2010.
30. Y. Saad. FGMRES: A flexible inner-outer preconditioned GMRES algorithm. *SIAM Journal on Scientific Computing*, 14(3):461–469, 1993.
31. Y. Saad and M.H. Schultz. GMRES: A generalized minimal residual algorithm for solving nonsymmetric linear systems. *SIAM Journal on Scientific and Statistical Computing*, 7:856–869, 1986.
32. A.H. Sheikh, D. Lahaye, and C. Vuik. On the convergence of shifted Laplace preconditioner combined with multilevel deflation. *Numerical Linear Algebra with Applications*, 20(4):645–662, 2013.
33. G.H. Shortley and R. Weller. The numerical solution of Laplace's equation. *Journal of Applied Physics*, 9(5):334–348, 1938.
34. B. Simon. The definition of molecular resonance curves by the method of Exterior Complex Scaling. *Physics Letters A*, 71(2):211–214, 1979.
35. V. Simoncini and D.B. Szyld. Flexible inner-outer Krylov subspace methods. *SIAM Journal on Numerical Analysis*, pages 2219–2239, 2003.

36. V. Simoncini and D.B. Szyld. Recent computational developments in Krylov subspace methods for linear systems. *Numerical Linear Algebra with Applications*, 14(1):1–59, 2007.
37. K. Stüben and U. Trottenberg. Multigrid methods: Fundamental algorithms, model problem analysis and applications. *Multigrid methods, Lecture Notes in Math.*, 960:1–176, 1982.
38. L.N. Trefethen and M. Embree. *Spectra and pseudospectra: the behavior of nonnormal matrices and operators*. Princeton University Press, 2005.
39. U. Trottenberg, C.W. Oosterlee, and A. Schüller. *Multigrid.* Academic Press, New York, 2001.
40. H.A. Van der Vorst. BiCGSTAB: A fast and smoothly converging variant of BiCG for the solution of nonsymmetric linear systems. *SIAM Journal on Scientific Computing*, 13(2):631–644, 1992.
41. H.A. van Gijzen, Y. Erlangga, and C. Vuik. Spectral analysis of the discrete Helmholtz operator preconditioned with a shifted Laplacian. *SIAM Journal on Scientific Computing*, 29(5):1942–1985, 2007.
42. D. Werner. *Funktionalanalysis*. Springer, 2006.

Part II
Algorithms: Practical Methods and Implementations

In this part (parallel) algorithms will be described that are currently used in real applications.

How to Choose the Shift in the Shifted Laplace Preconditioner for the Helmholtz Equation Combined with Deflation

D. Lahaye and C. Vuik

Abstract In recent work we showed that the performance of the complex shifted Laplace preconditioner for the discretized Helmholtz equation can be significantly improved by combining it multiplicatively with a deflation procedure that employs multigrid vectors. In this chapter we argue that in this combination the preconditioner improves the convergence of the outer Krylov acceleration through a new mechanism. This mechanism allows for a much larger damping and facilitates the approximate solve with the preconditioner. The convergence of the outer Krylov acceleration is not significantly delayed and occasionally even accelerated. To provide a basis for these claims, we analyze for a one-dimensional problem a two-level variant of the method in which the preconditioner is applied after deflation and in which both the preconditioner and the coarse grid problem are inverted exactly. We show that in case that the mesh is sufficiently fine to resolve the wave length, the spectrum after deflation consists of a cluster surrounded by two tails that extend in both directions along the real axis. The action of the inverse of the preconditioner is to shrink the length of the tails while at the same time rotating them and shifting the center of the cluster towards the origin. A much larger damping parameter than in algorithms without deflation can be used.

1 Introduction

The Helmholtz equation is a classical model equation for the propagation of waves. Examples of its use in various branches of science and engineering are given in the references cited. Fast and scalable methods to solve the linear system that arise after discretization are urgently needed.

The advent of the complex shifted Laplacian in [1, 2] led to a breakthrough in solver capabilities. The basis of this work was laid in [3] and [4]. A work in which similar ideas are proposed albeit with a different perspective is [5]. The complex shifted Laplacian was reconsidered in [6–10] and led to a boost in tackling various

D. Lahaye (✉) • C. Vuik
DIAM, TU Delft, Mekelweg 4, 2611 CN Delft, The Netherlands
e-mail: d.j.p.lahaye@tudelft.nl; c.vuik@tudelft.nl

D. Lahaye et al. (eds.), *Modern Solvers for Helmholtz Problems*, Geosystems Mathematics, DOI 10.1007/978-3-319-28832-1_4

industrial applications as documented in [11–20]. For a survey we refer to [21]. Recent publications on various solution approaches include [22–31].

The convergence of the complex shifted Laplace preconditioners is analyzed in [10, 32, 33]. The preconditioner introduces damping and shifts small eigenvalues of the preconditioned system away from the origin such that the outer Krylov will be faster to converge. As the wavenumber increases while the number of grid points per wavelength is kept constant however, the number of small eigenvalues becomes too large for the preconditioner to handle effectively, and the required number of outer Krylov iterations increases linearly with the wavenumber. This motivated the development in [34] of a deflation approach aiming at removing small eigenvalues using a projection procedure. The paper [34] considers deflating the preconditioned operator using the columns of the coarse to fine grid interpolation operator as deflation vectors. The multilevel extension of the method requires a Krylov subspace acceleration at each level. The method is therefore called a multilevel Krylov method. Some form of approximation is required to avoid the explicit construction of the preconditioned operator and to render the method computationally feasible. By the approximation proposed in [34], the projection property of the deflation operator is lost. This renders the results of a model analysis of the method using Fourier modes more tedious to interpret.

The method we developed in [35] borrows ideas from [34]. However, instead of deflating the preconditioned operator, we instead deflate the original Helmholtz operator. We subsequently combined the deflation and complex shifted Laplacian multiplicatively. We thus avoid having to approximate a computationally expensive operator and preserve the projection property of the deflation operator. This construction allows to

- add a term to the deflation operator to shift a set of eigenvalues away from zero without significantly disturbing the non-zero eigenvalues. This in turn allows to extend the deflation method to multiple levels in a multigrid hierarchy. This multilevel extension can be interpreted as a multigrid method in which at least formally the complex shifted Laplacian acts as a smoother. As in [34], the method requires a Krylov acceleration at each level of the multigrid hierarchy;
- deduce the algebraic multiplicity of the zero eigenvalue of the deflated operator in a model problem analysis. This facilitates the computation of the non-zero eigenvalues;
- re-use implementations of the multigrid approximate inversion of the complex shifted Laplacian to code the operation with the deflation operator. In this re-use one has to construct the coarser grid operators by Galerkin coarsening, to provide a Krylov acceleration on the intermediate coarse levels and to provide a flexible Krylov method on the finest level. This can be done with for instance the PETSc software library [36].

In our model problem analysis we employ 10 or 20 grid points per wavelength on the finest level. We also assume that Dirichlet boundary conditions are employed. These conditions render the boundaries reflective for outward traveling waves and act as a worst case in terms of convergence for situations in which Sommerfeld or other types of absorbing boundary conditions are imposed. The spectrum of the operator after applying deflation without any preconditioning is real-valued and consists of a tight cluster surrounded by two tails. These tails spread in opposite directions along the real axis as the wavenumber increases. Elements in each tail correspond to the elements in the near-kernel of the Helmholtz operator on either side of zero. The role of the preconditioner is to scale and rotate the eigenvalues of the deflated operator. The spectrum of the operator after applying both deflation and preconditioning is complex-valued and consists of a cluster surrounded by two tails. These tails spread along a line in opposite directions in the complex plane away from the cluster with increasing wavenumber. The abscissa and slope of this line as well as the spread of the eigenvalues along this line are functions of the damping parameter in the preconditioner. Our results convincingly show that the use of deflation allows to significantly increase the damping parameter. Results in [35] give evidence for the fact the use of deflation results in a reduction of outer Krylov iterations. Results in [37] illustrate how the reduction of iterations leads to a significant speed-up of the computations.

This paper is structured as follows: in Sect. 2 we present the one-dimensional problem we intend to solve. In Sect. 3 we discuss the eigenvalues distribution of the complex shifted Laplace preconditioned matrix for the case of Dirichlet and Sommerfeld boundary conditions. In Sect. 4 we combine the preconditioner multiplicatively with a deflation operator that employs multigrid vectors. In Sect. 5 we derive closed form expressions for the eigenvalues of the preconditioned deflated system matrix through a model problem analysis. In Sect. 6 we present numerical results. Finally, we draw conclusions in Sect. 7.

2 Problem Formulation

The Helmholtz equation for the unknown field $u(x)$ on the one-dimensional domain $\Omega = (0, 1)$ reads

$$-\Delta u - (1 - \alpha\iota)k^2 u = g \text{ on } \Omega\,, \tag{1}$$

where $\iota, \alpha \in \mathbb{R}^+$, $k(x, y)$ and $g(x, y)$ are the imaginary unit, the damping parameter, the wave number and the source function, respectively. Here we are primarily be interested in solving the hard case without damping, i.e., the case in which $\alpha = 0$. We use the case with damping to illustrate various arguments. The wave number k, the frequency f and the angular frequency $\omega = 2\pi f$, the speed of propagation $c(x, y)$ and the wavelength $\lambda = \frac{c(x,y)}{f}$ are related by

$$k = \frac{2\pi}{\lambda} = \frac{\omega}{c}\,. \tag{2}$$

On the boundary $\partial\Omega$ we impose either homogeneous Dirichlet or first order Sommerfeld radiation boundary conditions. The latter are given by

$$\frac{\partial u}{\partial n} - \iota k u = 0 \text{ on } \partial\Omega \,. \tag{3}$$

This condition renders the end points of the one-dimensional domain transparent for outgoing waves. The spectrum of the coefficient matrix is such that the problem with Dirichlet boundary conditions is harder to solve than the problem with Sommerfeld boundary conditions as observed in e.g. [32]. This statement remains valid even if deflation is deployed.

2.1 Finite Difference Discretization

The finite difference discretization of the above problems on a uniform mesh with mesh width h using the stencil

$$[A_h] = \frac{1}{h^2}\left[-1 \;\; 2-(1-\alpha\iota)\kappa^2 \;\; -1\right] \text{ where } \kappa = k\,h \,, \tag{4}$$

results after elimination of the boundary conditions in the linear system

$$A_h x_h = b_h \,, \tag{5}$$

where

$$A_h = -\Delta_h - (1-\alpha\iota)k^2 I_h \in \mathbb{C}^{(n-1)\times(n-1)} \,. \tag{6}$$

The discretization requires due care to avoid the pollution error [38, 39]. This can be done by either imposing a minimum number of grid points per wavelength or by imposing the more stringent condition that $k^2\,h^3$ remains constant.

2.2 Eigenvalues of the Discrete Helmholtz Operator

The linear system matrix A_h is sparse and symmetric. In the case of no damping and a sufficiently high wave-number (and thus a sufficiently fine mesh), A_h is indefinite and has a non-trivial near-null space. In case that the constant wave number one-dimensional problem is supplied with Dirichlet boundary conditions and is discretized using a uniform mesh with mesh width $h = 1/n$, the eigenvalues of A_h are easy to compute. As other types of boundary conditions introduce some form of damping, the resulting spectrum is generally more favorable for the convergence of the outer Krylov acceleration. This implies that the use of Dirichlet boundary conditions acts as a worst-case scenario in the analysis of the convergence of Krylov methods via the spectrum. With these assumptions, the eigenvalues of A_h are the negatively shifted eigenvalues of the discrete Poisson operator and are given by Sheikh et al. [35] and Ernst and Gander [40]

$$\lambda^{\ell}(A_h) = \frac{1}{h^2}(2 - 2c_{\ell} - \kappa^2)\,, \tag{7}$$

for $1 \leq \ell \leq n-1$, where

$$c_{\ell} = \cos(\ell\,\pi\,h)\,. \tag{8}$$

The corresponding eigenvectors are the orthogonal set of discrete sine modes denoted by ϕ^{ℓ} where $1 \leq \ell \leq n-1$. Each ϕ^{ℓ} is thus a vector with $n-1$ components indexed by i and given by

$$\phi_i^{\ell} = \sin(i\,h\,\ell\,\pi) \text{ for } 1 \leq i \leq n-1\,. \tag{9}$$

The mutual orthogonality of the ϕ^{ℓ}'s implies that the matrix A_h is normal and that theory for the convergence of GMRES applies. In case that damping is included in the Helmholtz equation, a purely imaginary contribution is added to (7), shifting the eigenvalues away from the origin. This increase in distance away from the origin makes the damped version easier to solve. From (7), it follows that the eigenvalues of the h^2-scaled operator $h^2\,A_h$ vary continuously between

$$\lambda^{1}(h^2\,A_h) \approx -\kappa^2 \text{ and } \lambda^{n-1}(h^2\,A_h) \approx 4 - \kappa^2 \tag{10}$$

where $c_1 \approx 1$ and $c_{n-1} \approx -1$, respectively.

In case that damping is added in the Helmholtz equation (1) by setting the damping parameter $\alpha > 0$, the imaginary component $\iota\alpha k^2$ is added to the eigenvalue expression (7). The eigenvectors remain unaltered. The eigenvalues are shifted upwards in the complex plane, and the problem becomes easier to solve iteratively.

In case that the Dirichlet boundary conditions are replaced by the Sommerfeld boundary conditions, both the eigenvalues and the eigenvectors change. An analytical computation of the spectrum in the limit of large k can be found in [41]. For the undamped case and for $k = 100$ and 10 grid points per wavelength, we computed the spectrum of the h^2-scaled matrix $h^2\,A_h$ numerically. We plotted the sorted real and imaginary part of the eigenvalues as a function of the index ℓ in Fig. 1a and b, respectively. The sorting is such that different orderings are used in both figures. The real part closely resembles the expression for the Dirichlet case given by (7). The presence of a non-zero imaginary part in the eigenvalues render the use of Sommerfeld boundary conditions similar to the case of damping with Dirichlet boundary conditions with damping. The imaginary contribution shifts the eigenvalues away from the origin and renders the problem easier to solve

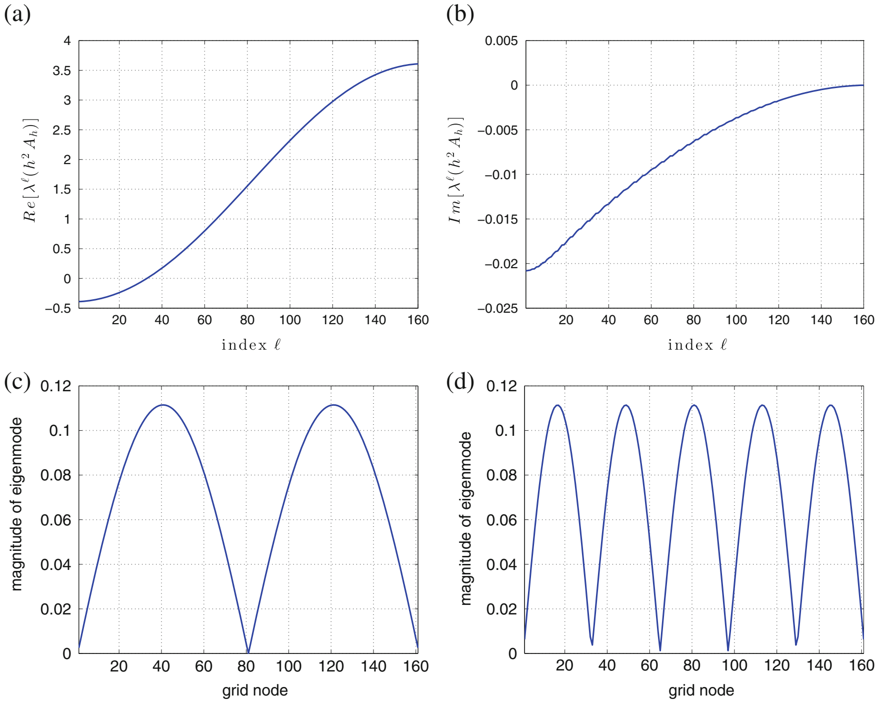


Fig. 1 Eigenvalues and magnitude of the second and fifth eigenmode of the discrete Helmholtz operator with Sommerfeld boundary conditions for $k = 100$ using 10 grid points per wavelength. (**a**) Sorted real part of eigenvalues. (**b**) Sorted imaginary part of eigenvalues. (**c**) Magnitude of second eigenmode. (**d**) Magnitude of fifth eigenmode

numerically. The magnitude of the second and fifth eigenvector is shown as a function of the grid index in Fig. 1c and d, respectively.

2.3 Multigrid Considerations

In the previous paragraph we assumed the mesh to be sufficiently fine to represent the wavelength. In this paper we will however consider an approach in which the Helmholtz equation without damping is discretized on a hierarchy of increasingly coarser meshes. This is the essential difference with CSLP precondition in previous work [1, 2] in which the original Helmholtz equation is discretized on the finest mesh only and in which the Helmholtz equation with damping only is coarsened.

The discretization of the undamped Helmholtz equation on the multigrid hierarchy motivates looking into properties of $h^2 A_h$ on the different levels of this hierarchy. We will derive bounds on the eigenvalues and a measure for the diagonal dominance of $h^2 A_h$ into account. For a fixed value of the wavenumber, each level

Table 1 Eigenvalue bounds and diagonal dominance measure of the h^2-scaled discretized Helmholtz operator $h^2 A_h$ for fixed wavenumber and for various values of the number of grid points per wavelength (gpw) on a multigrid hierarchy with five levels

κ	gpw	$\lambda^1(h^2 A_h) = -\kappa^2$	$\lambda^{n-1}(h^2 A_h) = 4 - \kappa^2$	$\lvert 2 - \kappa^2 \rvert$
0.3125	20	−0.0977	3.9023	1.9023
0.625	10	− 0.3906	3.6094	1.6094
1.25	5	−1.5625	2.4375	0.4375
2.5	2.5	−6.25	−2.25	4.2500
5	1.25	−25	−21	23

of the hierarchy corresponds to a number of grid points per wavelength, and thus to a value of κ. Here we will consider the case in which the Helmholtz operator on the coarser levels is obtained via rediscretization and leave the case of Galerkin coarsening to later in the paper. To obtain bounds for the eigenvalues of the discrete Helmholtz operator on each level of the hierarchy in case of rediscretization, it suffices to substitute the corresponding value for κ into the bounds (10). As a measure for the diagonal dominance, we will adopt the absolute value of the diagonal element $|2-\kappa^2|$. For a multigrid hierarchy consisting of five levels obtained by standard $h \to 2h$ coarsening each level except for the coarsest, the eigenvalue bounds (10) and the value of $|2 - \kappa^2|$ are tabulated in Table 1. Motivating this measure for diagonal dominance is the fact that the weight of the off-diagonal elements does not change in traversing the hierarchy. The middle columns of Table 1 shows that in traversing the multigrid hierarchy from finest to coarsest level, the spectrum shifts in the negative direction and that on a sufficiently coarse level (here at 2.5 grid points per wavelength) even the largest eigenvalue becomes negative. From that level onward, the matrix ceases to be indefinite. The right-most column of Table 1 shows that the measure for the diagonal dominance initially decreases and increases again starting at a sufficiently coarse level (here again at 2.5 grid points per wavelength). At this coarsest level, the problem can be easily solved using the method of Jacobi for instance. For $k = 1000$ and for 10 grid points per wavelength for instance, the problem becomes definite and diagonally dominant starting from the third coarsest level onward. On these levels the use of complex solution algorithms such as the CSLP preconditioner is unnecessary. Similar ideas have already been discussed in [42]. We will return to this observation when discussing how to choose the damping parameter in the complex shifted Laplace preconditioner on the intermediate coarser levels.

3 Complex Shifted Laplace Preconditioner

In this section we introduce the complex shifted Laplace preconditioner [1, 2] and derive closed form expressions for the eigenvalues of the preconditioner and the preconditioned operator. We will in particular look into the effect of choosing a very large damping parameter in the preconditioner. The information collected in this section will serve as a reference for our model problem analysis in Sect. 5.

Denoting by β_2 the strictly positive damping parameter, the complex shifted Laplace (CSLP) preconditioner can be written as

$$M_{h,\beta_2} = -\Delta_h - (1 - \iota\beta_2)\kappa^2 I_h \text{ where } \beta_2 \in \mathbb{R}^+ \setminus 0 \,. \tag{11}$$

The value of β_2 needs to balance the quality of the preconditioner (favoring a small value) with the ease of approximately inverting it (favoring a large value). We assume that the boundary conditions implemented in A_h are imposed on M_{h,β_2} as well. In both the case of Dirichlet and Sommerfeld boundary conditions, the submatrices of M_{h,β_2} and A_h corresponding to the interior nodes differ by a scalar multiple of the purely imaginary constant diagonal matrix $\iota\kappa^2 I_h$. In the absence and presence of damping, the scalar involved is equal to β_2 and $\beta_2 - \alpha$, respectively.

3.1 Eigenvalues of the CSLP Preconditioner

Given that the matrices A_h and M_{h,β_2} differ by a purely imaginary constant diagonal contribution, the eigenvalues of M_{h,β_2} are the eigenvalues of A_h shifted along the imaginary axis. In both the case of Dirichlet and Sommerfeld boundary conditions, the eigenvectors of M_{h,β_2} and A_h are the same. In the one-dimensional problem with Dirichlet boundary conditions, we have that the eigenvalues of the h^2-scaled preconditioner $h^2 M_{h,\beta_2}$ for $1 \leq \ell \leq n-1$ are given by

$$\lambda^\ell(h^2 M_{h,\beta_2}) = 2 - 2\,\mathrm{c}_\ell - \kappa^2(1 - \iota\beta_2) \,. \tag{12}$$

Let $\mu^\ell(h^2 M_{h,\beta_2})$ denote the inverse of this eigenvalue. Separating this inverse into a real and imaginary part, we obtain

$$\begin{aligned} \mu^\ell(h^2 M_{h,\beta_2}) &= \frac{1}{2 - 2\,\mathrm{c}_\ell - \kappa^2(1 - \iota\beta_2)} \\ &= \frac{2 - 2\,\mathrm{c}_\ell - \kappa^2}{[2 - 2\,\mathrm{c}_\ell - \kappa^2]^2 + \kappa^4\beta_2^2} - \iota\frac{\kappa^2\beta_2}{[2 - 2\,\mathrm{c}_\ell - \kappa^2]^2 + \kappa^4\beta_2^2} \\ &= \frac{2 - 2\,\mathrm{c}_\ell - \kappa^2}{|\lambda^\ell(h^2 M_{h,\beta_2})|} - \iota\frac{\kappa^2\beta_2}{|\lambda^\ell(h^2 M_{h,\beta_2})|} \\ &= \mathrm{Re}[\mu^\ell(h^2 M_{h,\beta_2})] + \iota\mathrm{Im}[\mu^\ell(h^2 M_{h,\beta_2})] \,. \end{aligned} \tag{13}$$

From these expressions we conclude that for $1 \leq \ell \leq n-1$

$$0 < \mathrm{Re}[\mu^\ell(h^2 M_{h,\beta_2})] < 1 \quad \forall \beta_2 > 0, \tag{14}$$

$$-1 < \mathrm{Im}[\mu^\ell(h^2 M_{h,\beta_2})] < 0 \quad \forall \beta_2 > 0 \,, \tag{15}$$

and that in the limit for strong damping that

$$\mathrm{Re}[\mu^{\ell}(h^2 M_{h,\beta_2})] \to 0 \text{ as } \beta_2 \to +\infty, \tag{16}$$

$$\mathrm{Im}[\mu^{\ell}(h^2 M_{h,\beta_2})] \to 0 \text{ as } \beta_2 \to +\infty\,. \tag{17}$$

These results will be used to derive expressions for the eigenvalues of the preconditioned operator and the deflated preconditioned operator in the next paragraph and the next section, respectively.

3.2 Eigenvalues of the CSLP Preconditioned Operator

In deriving the eigenvalues of the preconditioned operator, we will assume the preconditioner to be inverted exactly. We will consider both the case of Dirichlet and Sommerfeld boundary conditions. In the former case, M_{h,β_2}^{-1} and A_h share the set of discrete sine modes given by (9). In the absence of damping, the eigenvalues of the preconditioned operator $M_{h,\beta_2}^{-1} A_h$ are the scaled and rotated eigenvalues of A_h given by

$$\begin{aligned} \lambda^{\ell}(M_{h,\beta_2}^{-1} A_h) &= \mu^{\ell}(M_{h,\beta_2})\,\lambda^{\ell}(A_h) \qquad (18)\\ &= \mathrm{Re}[\mu^{\ell}(h^2 M_{h,\beta_2})]\,\lambda^{\ell}(A_h) + \iota\,\mathrm{Im}[\mu^{\ell}(h^2 M_{h,\beta_2})]\,\lambda^{\ell}(A_h)\\ &= \frac{\lambda^{\ell}(A_h)(2-2\,\mathrm{c}_{\ell}-\kappa^2)}{[2-2\,\mathrm{c}_{\ell}-\kappa^2]^2+\kappa^4\beta_2^2} - \iota\frac{\lambda^{\ell}(A_h)\kappa^2\beta_2}{[2-2\,\mathrm{c}_{\ell}-\kappa^2]^2+\kappa^4\beta_2^2}\,. \end{aligned}$$

This computation can be generalized to include non-zero damping (i.e., $\alpha = 0$) in the Helmholtz equation. In the case of Sommerfeld boundary conditions, we will resort to the numerical computations of the eigenvalues.

In Fig. 2 we plotted the eigenvalues $\lambda^{\ell}(M_{h,\beta_2}^{-1} A_h)$ for $1 \leq \ell \leq n-1$ in the complex plane for $k = 1000$ and 10 grid points per wavelength for four cases. In all four cases we highlighted a small region around the origin with a circle. In Fig. 2a we consider the case of Dirichlet boundary conditions without damping using $\beta_2 = 0.5$. We used shaded and non-shaded symbols to distinguish the eigenvalues that correspond to the index ℓ for which $\lambda^{\ell}(A_h)$ is negative and positive, respectively. Clearly both the real and imaginary part of $M_{h,\beta_2}^{-1} A_h$ are small for those values of ℓ for which $\lambda^{\ell}(A_h)$ shows a change of sign, i.e., for the values of ℓ that correspond to the near-kernel of A_h. These small eigenvalues hamper the convergence of the outer Krylov acceleration.

In Fig. 2b we consider again the case of Dirichlet boundary conditions without damping, this time using the larger value $\beta_2 = 1$. Comparing this figure with Fig. 2a confirms that for larger values of β_2 the eigenvalues $\mu^{\ell}(M_{h,\beta_2})$ and therefore the eigenvalues $\lambda^{\ell}(M_{h,\beta_2}^{-1} A_h)$ shift towards the origin. This causes the quality of

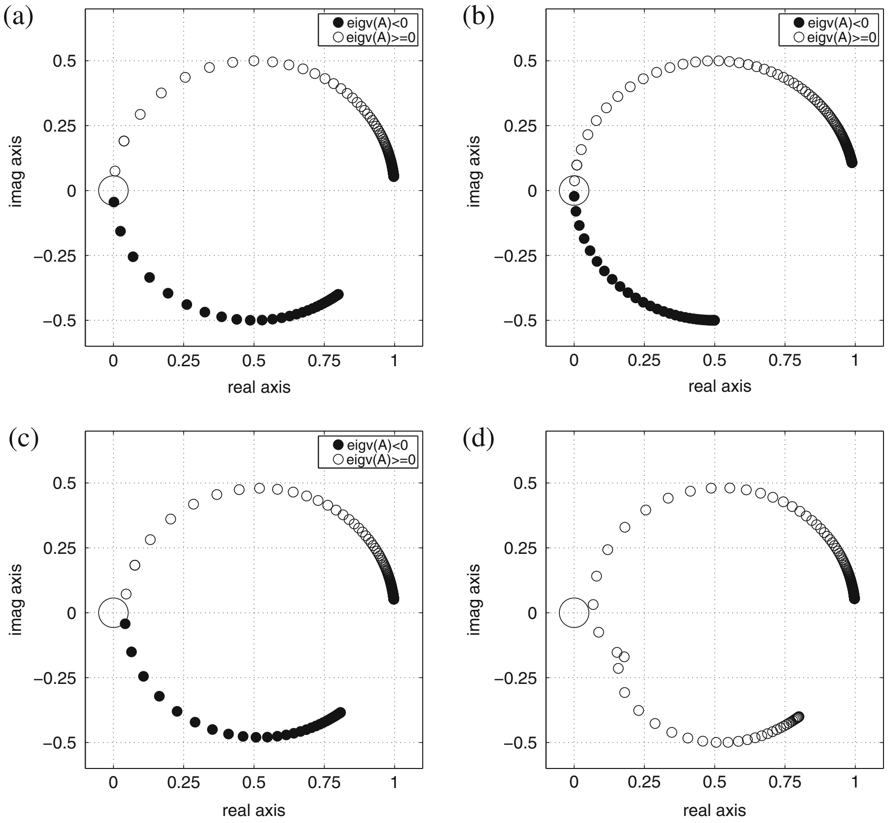


Fig. 2 Eigenvalues of the CSLP preconditioned operator for various preconditioning strategies for $k = 1000$ using 10 grid points per wavelength. (**a**) No damping and $\beta_2 = 0.5$. (**b**) No damping and $\beta_2 = 1$. (**c**) $\alpha = 0.02$ and $\beta_2 = 0.5$. (**d**) Sommerfeld and $\beta_2 = 0.5$

the preconditioner to degrade. The analysis in [32] shows that a similar shift of eigenvalues towards the origin occurs as k increases while β_2 and κ is kept constant.

In Fig. 2c we consider once more the case of Dirichlet boundary conditions, this time using a damping coefficient $\alpha = 0.02$. Comparing this figure with Fig. 2a shows that by introducing damping in the Helmholtz equation, the eigenvalues close to the origin shift towards the right in the complex plane. The increase of the magnitude of the eigenvalues that are small in size renders the preconditioned systems easier to solve.

In Fig. 2d finally we consider the case of Sommerfeld boundary conditions. This figure closely resembles to Fig. 2c. The introduction of the Sommerfeld boundary conditions is seen to introduce damping that causes a shift of small eigenvalues away from the origin. The preconditioned system again becomes easier to solve.

4 Combining Deflation and Precondition Multiplicatively

In this section we describe how we combine the complex shifted Laplacian preconditioner (CSLP) with a deflation technique. This approach is motivated by the fact that the convergence of the CSLP preconditioned Krylov acceleration is hampered by a few eigenvalues that are small in size. This is especially a problem in cases without damping. The objective of deflation is to remove these undesirable eigenvalues by a projection procedure. We describe the deflation technique on two levels, its extension to multiple levels, and the multiplicative combination of the preconditioner and the deflation technique.

4.1 Deflation by Two-Grid Vectors

Assuming p to be a non-zero natural number, we discretize the computational domain $\Omega = (0, 1)$ by a uniform mesh with $n = 2^p$ elements and mesh width $h = 1/n$ resulting in a fine mesh Ω^h. The discretization of the Helmholtz equation results after elimination of the boundary nodes in a discrete operator $A_h \in \mathbb{C}^{(n-1)\times(n-1)}$. Standard $h \to H = 2h$ coarsening of the fine mesh Ω^h results in a coarse mesh Ω^H with $n/2 - 1$ internal nodes. We denote by $Z_{h,H} \in \mathbb{R}^{(n-1)\times(n/2-1)}$ the coarse-to-fine grid interpolation operator. We employ a linear interpolation operator that, for fine grid points not belonging to the coarse grid, has the stencil

$$[Z_{h,H}] = \frac{1}{4}\left[1\ 2\ 1\right]_H^h . \tag{19}$$

The columns of $Z_{h,H}$ are referred to as the deflation vectors. A deflation technique that uses these vectors is referred as two-grid deflation. The restriction operator is set equal to the full-weighting restriction operator. With this choice the restriction is equal to the transpose of the interpolation. This construction fits the theoretical framework of deflation methods.

The coarse grid operator E_H is constructed by Galerkin coarsening

$$E_H = Z_{h,H}^T A_h Z_{h,H} \in \mathbb{C}^{(n/2-1)\times(n/2-1)} . \tag{20}$$

The spectral properties of E_H will be discussed in the next section. We then define the coarse grid solve operator $Q_{h,H}$ as

$$Q_{h,H} = Z_{h,H} E_H^{-1} Z_{h,H}^T \in \mathbb{C}^{(n-1)\times(n-1)} , \tag{21}$$

and the deflation operator $P_{h,H}$ as

$$P_{h,H} = I_h - A_h Q_{h,H} \in \mathbb{C}^{(n-1)\times(n-1)} . \tag{22}$$

The construction of E_H by Galerkin coarsening is such that $P_{h,H}$ satisfies the relation $P_{h,H}^2 = P_{h,H}$. $P_{h,H}$ is thus a projection and has eigenvalues 0 and 1. The matrix $P_{h,H}$ corresponds to the residual propagation matrix in a basic iterative solution method based on the splitting $A_h = Q_{h,H} - (Q_{h,H} - A_h)$ for the linear system (5).

By applying deflation, the linear system (5) is transformed into

$$P_{h,H}\, A_h x_h = P_{h,H}\, b_h \, . \tag{23}$$

The columns of $Z_{h,H}$ lie in the kernel of the deflated operator [43], i.e.,

$$P_{h,H}\, A_h\, Z_{h,H} = 0_{(n-1)\times(n/2-1)} \, . \tag{24}$$

The matrix $P_{h,H}\, A_h$ is thus singular and has a zero eigenvalue with multiplicity $n/2 - 1$. The computation of the remaining $n/2$ eigenvalues will be shown in the next section. The solution of the linear system (23) is defined up to a component in the kernel of $P_{h,H}\, A_h$. Such a solution can be found by a Krylov subspace method on the condition that in the application of $P_{h,H}$ the coarse linear system with E_H is solved to full precision at each iteration. What this condition implies and how it can be alleviated will be discussed in the next paragraph.

4.2 Multilevel Extension

For large problems in two or three dimensions, the exact inversion of the coarser grid matrix E_H is impractical and one has to resort to approximate coarser grid solves. Without proper care, this will however lead to the zero eigenvalue of $P_{h,H}\, A_h$ to be replaced by a cluster of near-zero eigenvalues. Such a cluster impedes the fast convergence of the outer Krylov acceleration. This can be avoided introducing a shift over a distance γ with $Q_{h,H}$ in the deflation operator $P_{h,H}$ and to define $P_{h,H,\gamma}$ as

$$P_{h,H,\gamma} = P_{h,H} + \gamma\, Q_{h,H} = I_h - A_h\, Q_{h,H} + \gamma\, Q_{h,H} \, . \tag{25}$$

With this definition, the equivalent of (24) for $P_{h,H,\gamma}$ is

$$P_{h,H,\gamma}\, A_h\, Z_{h,H} = \gamma\, Z_{h,H} \, , \tag{26}$$

i.e., γ is an eigenvalue with multiplicity $n/2 - 1$ of deflated matrix $P_{h,H,\gamma}\, A_h$. The value of γ is chosen once a choice for the preconditioner is made. We will give details in the next paragraph. The shift away from zero of the eigenvalues of the deflated matrix allows to solve the coarse grid system with coefficient matrix E_H approximately for instance by a recursive application of the two-level algorithm described. The use of a Krylov subspace solver on the coarser level requires to resort to a flexible Krylov subspace solver on the fine level. The depth of the multigrid cycle can be limited in accordance to the discussion given in Sect. 2.

4.3 Multiplicative Combining of Preconditioning and Deflation

The CSLP preconditioner M_{h,β_2} and the deflation operator $P_{h,H,\gamma}$ including the shift with γ can be combined multiplicatively to construct a composite preconditioner. If the precondition is applied after the deflation, the linear system to be solved can be written as

$$B_{h,H,\beta_2,\gamma}\, x = (M_{h,\beta_2}^{-1}\, P_{h,H} + \gamma\, Q_{h,H})\, b\,, \tag{27}$$

where $B_{h,H,\beta_2,\gamma}$ is the preconditioned deflated operator

$$B_{h,H,\beta_2,\gamma} = (M_{h,\beta_2}^{-1}\, P_{h,H} + \gamma\, Q_{h,H})\, A_h\,. \tag{28}$$

In case that $\gamma = 1$, the matrix $I - B_{h,H,\beta_2,\gamma}$ is the error propagation matrix of a two-grid $V(0,1)$ cycle applied to the linear system (5) with Galerkin coarsening and with M_{h,β_2} assuming at least formally the role of the smoother. In case that $\gamma \neq 1$, the composite preconditioner can be implemented as the additive combination of previously described $V(0,1)$ cycle and a shift with $\gamma = 1$. Closed form expressions for the eigenvalues of $B_{h,H,\beta_2,\gamma}$ defined by (28) for $\gamma = 0$ and $\gamma \neq 0$ will be derived in the next section.

4.4 Comments on a Practical Implementation

An implementation of a multigrid approximate inversion CSLP as preconditioner can be easily extended to its combined use with the above described deflation technique. The multigrid components already in place can be recycled. A flexible Krylov acceleration on each level is required.

5 Model Problem Analysis

In this section we first derive closed form expressions for the eigenvalues of the Galerkin coarse grid operator E_H and the deflation operator $P_{h,H}$ defined by (20) and (22), respectively. Next we extend this analysis of the eigenvalues of the deflated operator $P_{h,H}\, A_h$ and the preconditioned deflated operator $M_{h,\beta_2}^{-1}\, P_{h,H}\, A_h$ given in the left-hand side of (23) and (27) with $\gamma = 0$, respectively. We consider the one-dimensional problem with Dirichlet boundary conditions with and without damping.

Based on the arguments on the resemblance of the eigenvalues in the problem with damping and with Sommerfeld boundary conditions in Sect. 2, we assume here that the problem with damping in the Helmholtz equation offers a good representation of the problem with Sommerfeld boundary conditions. We will derive expression for the eigenvalues by computing the action of these operators on the set of discrete sine modes defined by (9). This analysis is referred to a Rigorous Fourier analysis to distinguish it from a Local Fourier analysis in which the influence of the boundary conditions is not taken into account. Assuming Dirichlet boundary conditions, the set of sine modes ϕ_h^ℓ given by (9) forms a basis in which both the discrete operator A_h and the preconditioner M_{h,β_2} are diagonal. The analysis of the coarse operator E_H and the deflation operator $P_{h,H}$ requires care in handling the grid aliasing effect in the intergrid transfer operators $Z_{h,H}$ and $Z_{h,H}^T$. The eigenvalue expressions resulting from our analysis are fractions in which the eigenvalues of the coarse grid operator E_H appear in the denominator. These expressions form the basis for a subsequent analysis. The scattering of the eigenvalues along both sides of the real axis in case of 10 grid points per wavelength for instance can then be related to the near-kernel eigenvalues of the coarse grid operator E_H.

We assume the one-dimensional problem on $\Omega = (0, 1)$ with Dirichlet boundary conditions to be discretized by a uniform mesh with mesh width h. The coarse mesh obtained by standard coarsening then has a mesh width $H = 2h$. The use of Dirichlet boundary conditions was motivated in Sect. 2. We will perform a two-level analysis and assume that the Galerkin coarse grid operator E_H defined by (20) is inverted exactly. By reordering the eigenvectors of A_h defined by (9) in a standard way in $(\ell, n-\ell)$ pairs [44], we obtain the basis

$$V_h = \{(\phi_h^\ell, \phi_h^{n-\ell}) \mid \ell = 1, \ldots, n/2-1\} \cup \{\phi_h^{n/2}\}. \tag{29}$$

The modes ϕ_h^ℓ and $\phi_h^{n-\ell}$ form a pair by coarse grid aliasing. In the basis (29) first the deflation operator $P_{h,H}$, subsequently the deflated operator $P_{h,H} A_h$ and finally the preconditioned deflated operator $M_{h,\beta_2}^{-1} P_{h,H} A_h$ can be written in a block diagonal form. For a generic $(n-1) \times (n-1)$ matrix B, we will denote this block decomposition as

$$B = \left[(B)^\ell\right]_{1 \le \ell \le n/2}, \tag{30}$$

where for $1 \le \ell \le n/2 - 1$ the block $(B)^\ell$ is of size 2×2 and where for $\ell = n/2$ the block B^ℓ is a number. From this block diagonal form the eigenvalues of B can be computed with relative ease. For the restriction operator $Z_{h,H}^T$ and the coarse grid operator E_H that have size $(n/2-1) \times (n-1)$ and $(n/2-1) \times (n/2-1)$ the size of the diagonal blocks reduces to 1×2 and 1×1, respectively.

In the following we will subsequently compute the eigenvalues of the Galerkin coarse grid operator E_H, the deflation operator $P_{h,H}$, the deflated operator $P_{h,H} A_h$ and finally the preconditioned deflated operator without shift $M_{h,\beta_2}^{-1} P_{h,H} A_h$ and with shift $(M_{h,\beta_2}^{-1} P_{h,H} + \gamma\, Q_{h,H}) A_h$. As before, we will especially look into large values of the damping parameter β_2.

5.1 Eigenvalues of the Coarse Grid Operator E_H

The block diagonal representation of the interpolation operator $Z_{h,H}^T$ in the basis (29) can be obtained by a standard computation [44]. Using the fact that $c_{n-\ell} = -c_\ell$, one obtains for $1 \leq \ell \leq n/2 - 1$ the 1×2 blocks

$$(Z_{h,H}^T)^\ell = 1/2 \left[(1 + c_\ell) \; -(1 - c_\ell) \right] . \tag{31}$$

Given the $n/2$-th sine mode $\phi^{n/2}$ is equal to zero in all the coarse grid nodes and given the stencil (19), we have that

$$(Z_{h,H}^T)^{n/2} = 1/2 . \tag{32}$$

The diagonal block of the discrete operator A_h in the basis (29) are for $1 \leq \ell \leq n/2 - 1$ given by

$$(A_h)^\ell = \begin{pmatrix} \lambda^\ell(A_h) & 0 \\ 0 & \lambda^{n-\ell}(A_h) \end{pmatrix} = \frac{1}{h^2} \begin{pmatrix} 2 - 2c_\ell - \kappa^2 & 0 \\ 0 & 2 + 2c_\ell - \kappa^2 \end{pmatrix} , \tag{33}$$

and for $\ell = n/2$ by

$$(A_h)^{n/2} = (\kappa^2 - 2)/h^2 . \tag{34}$$

The 1×1 diagonal blocks $(E_H)^\ell$ of the Galerkin coarse grid operator are obtained by the Galerkin coarsening of the individual blocks and results for all coarse grid values of ℓ including $\ell = n/2$ in

$$(E_H)^\ell = (Z_{h,H}^T)^\ell (A_h)^\ell (Z_{H,h})^\ell = \frac{1}{2h^2} [2(1 - c_\ell^2) - \kappa^2(1 + c_\ell^2)] . \tag{35}$$

Given that in the basis (29) the operator E_H is diagonal, we have that the ℓ-th eigenvalue $\lambda^\ell(E_H)$ is equal to the ℓ-th diagonal block $(E_H)^\ell$. The eigenvalues of the H^2-scaled operator E_H are then for $1 \leq \ell \leq n/2$ given by

$$\lambda^\ell(H^2 \, E_H) = -(2 + 2c_\ell^2) \, \kappa^2 + 4 - 4c_\ell^2 . \tag{36}$$

In the end points of the range from $\ell = 1$ to $\ell = n/2$, these expressions reduce to

$$\lambda^1(H^2 \, E_H) \approx -4\kappa^2 < 0 \text{ and } \lambda^{n/2}(H^2 \, E_H) \approx 4 - 2\kappa^2 > 0 , \tag{37}$$

where $c_{\ell=1} \approx 1$ and $c_{\ell=n/2} = 0$, respectively. In the range of ℓ considered, close to zero eigenvalues of $\lambda^\ell(H^2 \, E_H)$ thus exist. The expressions (36) are the coarse grid equivalents of (10) and can be generalized to the case with damping by introducing a shift with $\iota \, \alpha \, \kappa^2$.

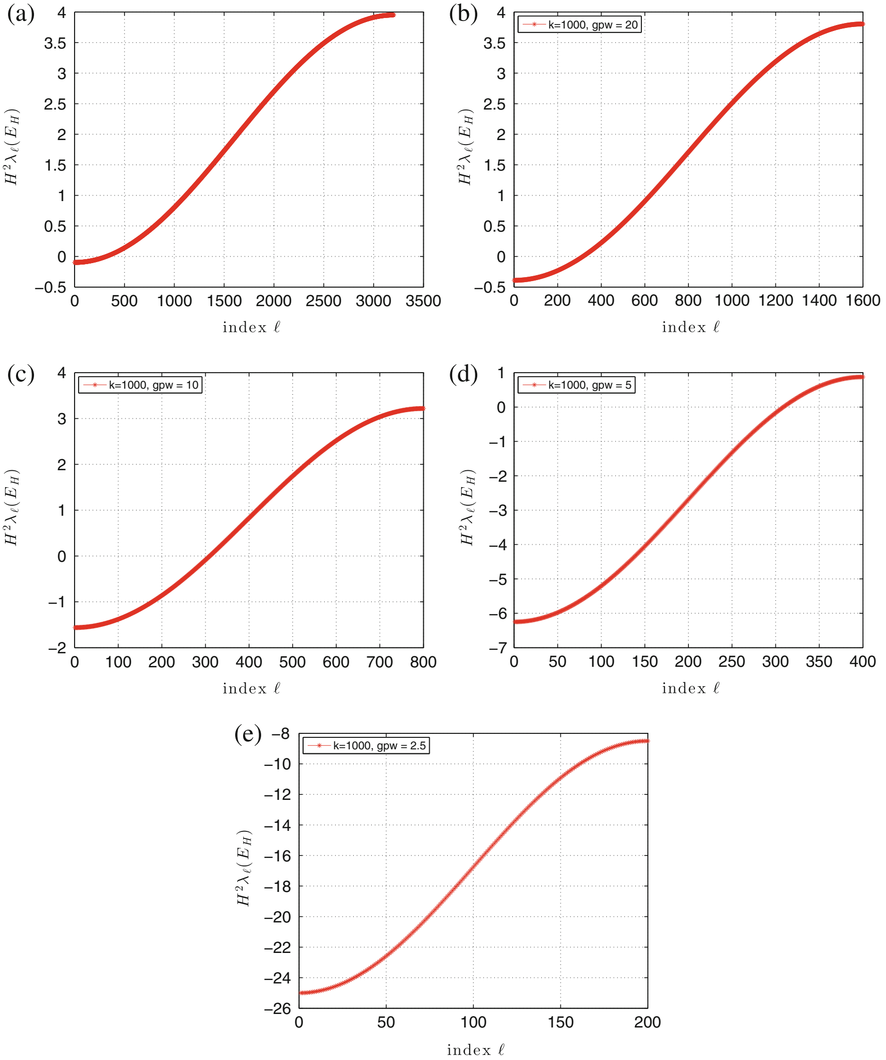


Fig. 3 Eigenvalues of the Helmholtz Galerkin coarse grid operator E_H for $k = 1000$ as a function of the index ℓ on a multigrid hierarchy consisting of five levels. On the finest level 40 grid points per wave length are employed. On each coarser level the number of grid points per wave length is halved. (**a**) 40 gpw; (**b**) 20 gpw; (**c**) 10 gpw; (**d**) 5 gpw; (**e**) 2.5 gpw

In Fig. 3 we plotted $\lambda^\ell(H^2 E_H)$ given by (37) as a function of ℓ for $k = 1000$ using various grid points per wavelength ranging from 40 (corresponding to $\kappa = 0.15625$) in the top left of the figure to 2.5 (corresponding to $\kappa = 2.5$) in the bottom of the figure. This figure clearly shows that in traversing the multigrid hierarchy from finest to coarser level (and thus increasing κ) the eigenvalues of the coarse grid operator

E_H shift towards the left on the real plane until all eigenvalues become negative and bounded away from zero on sufficiently coarse grids (sufficiently large values of κ). This is in accordance with the bounds (37).

The fact that in problems without damping the matrix E_H on fine and intermediate levels has several close to zero eigenvalues will play a central role in the subsequent analysis. By introducing damping, the issue of these small eigenvalues will be alleviated to some extent.

5.2 Eigenvalues of the Deflation Operator $P_{h,H}$

The diagonal blocks $(P_{h,H})^{\ell}$ of the deflation operator $P_{h,H}$ are for $1 \leq \ell \leq n/2-1$ given by

$$(P_{h,H})^{\ell} = I - (Z_{h,H})^{\ell} \, [(E_H)^{\ell}]^{-1} \, (Z_{h,H}^T)^{\ell} \, (A_h)^{\ell} \, , \tag{38}$$

and for $\ell = n/2$ by

$$(P_{h,H})^{n/2} = 1 \, . \tag{39}$$

As $P_{h,H}$ is a deflation operator, the individual 2×2 blocks $(P_{h,H})^{\ell}$ are projections as well and therefore have 0 and 1 as eigenvalue. Less immediate results will follow next.

5.3 Eigenvalues of the Deflated Operator $P_{h,H} A_h$

The diagonal blocks $(P_{h,H} A_h)^{\ell}$ of the deflated operator $P_{h,H} A_h$ are for $1 \leq \ell \leq n/2-1$ given by

$$\begin{aligned} (P_{h,H} A_h)^{\ell} &= (P_{h,H})^{\ell} \, (A_h)^{\ell} \\ &= (A_h)^{\ell} - (Z_{h,H})^{\ell} \, [(E_H)^{\ell}]^{-1} \, (Z_{h,H}^T)^{\ell} \, [(A_h)^{\ell}]^2 \, , \end{aligned} \tag{40}$$

and for $\ell = n/2$ by

$$(P_{h,H} A_h)^{n/2} = (2-\kappa^2)/h^2 \, . \tag{41}$$

Property (24) translates on the 2×2 block level to $(P_{h,H} A_h)^{\ell} (Z_{h,H})^{\ell} = 0_{2\times 1}$. Block $(P_{h,H} A_h)^{\ell}$ thus has a zero eigenvalue with multiplicity one. The remaining non-zero eigenvalue is then equal to the trace $\mathrm{Tr}[(P_{h,H} A_h)^{\ell}]$. Computations show that for

$1 \leq \ell \leq n/2 - 1$ the elements of the h^2-scaled block $h^2\,(P_{h,H}\,A_h)^\ell$ are given by

$$h^2\,(P_{h,H}\,A_h)^\ell = \frac{1}{\lambda^\ell(H^2\,E_H)} \begin{pmatrix} pa^\ell_{11,h,H} & pa^\ell_{12,h,H} \\ pa^\ell_{21,h,H} & pa^\ell_{22,h,H} \end{pmatrix}, \tag{42}$$

where the four matrix elements $pa^\ell_{ij,h,H}$ for $1 \leq i,j \leq 2$ are polynomials of second degree in κ^2. The diagonal elements $pa^\ell_{11,h,H}$ and $pa^\ell_{22,h,H}$ are more precisely given by

$$pa^\ell_{11,h,H} = (c_\ell - 1)^2(2c_\ell - 2 + \kappa^2)(2c_\ell + 2 - \kappa^2) \tag{43}$$

$$pa^\ell_{22,h,H} = (c_\ell + 1)^2(2c_\ell - 2 + \kappa^2)(2c_\ell + 2 - \kappa^2)\,. \tag{44}$$

Observe that these expressions only differ by the sign in the first factor. As the off-diagonal elements $pa^\ell_{12,h,H}$ and $pa^\ell_{21,h,H}$ are not required to compute the trace, their detailed expression is omitted here. The non-zero eigenvalue of the ℓ-th block $h^2\,(P_{h,H}\,A_h)^\ell$ is then given by

$$\begin{aligned} \lambda^\ell\left(h^2\,P_{h,H}\,A_h\right) &= \mathrm{Tr}[h^2\,(P_{h,H}\,A_h)^\ell] \\ &= \frac{1}{\lambda^\ell(H^2 E_H)}[pa^\ell_{11,h,H} + pa^\ell_{22,h,H}] \\ &= \frac{2}{\lambda^\ell(H^2 E_H)}(c_\ell^2 + 1)(2c_\ell - 2 + \kappa^2)(2c_\ell + 2 - \kappa^2)\,. \end{aligned} \tag{45}$$

Give that the deflated operator involves a coarse grid solve, it is natural that the eigenvalue $\lambda^\ell(H^2 E_H)$ of the coarse grid operator appears in the denominator. In the range from $\ell = 1$ to $\ell = n/2$, the eigenvalues (45) decrease from

$$\lambda^1\left(h^2\,P_{h,H}\,A_h\right) \approx 4 - \kappa^2 \text{ to } \lambda^{n/2}\left(h^2\,P_{h,H}\,A_h\right) = 2 - \kappa^2 \tag{46}$$

where $c_{\ell=1} \approx 1$ and $c_{\ell=n/2} = 0$, respectively. This decrease is however not monotone. Indeed, for those values of ℓ that corresponds to the near-kernel of $H^2\,E_H$, the numerator of (45) is finite and the denominator very small. The eigenvalues $\lambda^\ell\left(h^2\,P_{h,H}\,A_h\right)$ thus become very large for those values of ℓ. Stated differently, the closest-to-zero eigenvalue of $H^2 E_H$ causes of a vertical asymptote to appear in the plot of $\lambda^\ell\left(h^2\,P_{h,H}\,A_h\right)$ versus ℓ.

In Fig. 4 we plotted $\lambda^\ell\left(h^2\,P_{h,H}\,A_h\right)$ given by (45) as a function of ℓ for $k = 1000$. As in Fig. 3, we consider a sequence of five grids in which the number of grid points per wavelength ranges from 40 on the finest to 2.5 on the coarsest. On each level we consider a two-level construction of the deflation operator. For illustration purposes, we superimposed in each plot of $\lambda^\ell\left(h^2\,P_{h,H}\,A_h\right)$ a plot of $\lambda^\ell\left(H^2\,E_H\right)$. On the y-axis we labeled the extreme values $2 - \kappa^2$ and $4 - \kappa^2$. In the various subfigures of Fig. 4, the eigenvalues are seen to be bounded by $2 - \kappa^2$ and $4 - \kappa^2$, except for values close

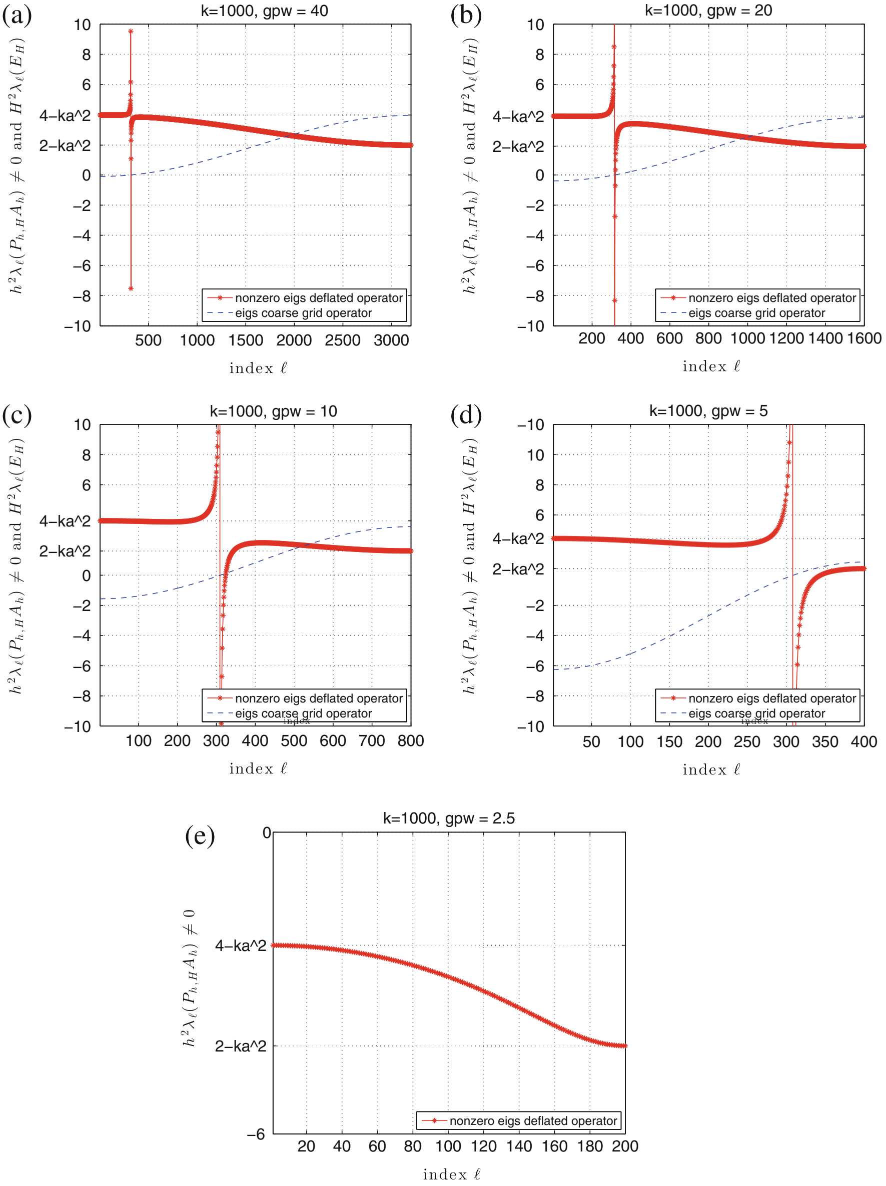


Fig. 4 Eigenvalues of the Helmholtz Galerkin coarse grid operator E_H (*dashed line*) and the two-grid deflated Helmholtz operator $P_{h,H}A_h$ (*solid line*) for $k = 1000$ as a function of the index ℓ on a multigrid hierarchy consisting of five levels. On the finest level 40 grid points per wave length are employed. On each coarser level the number of grid points per wave length is halved. (**a**) 40 gpw; (**b**) 20 gpw; (**c**) 10 gpw; (**d**) 5 gpw; (**e**) 2.5 gpw

to a vertical asymptote. The value of ℓ for which this asymptote occurs, is seen to coincide with the value of ℓ for which $\lambda^{\ell}\left(H^2\,E_H\right) \approx 0$. This value of ℓ shifts towards the right on coarser meshes until disappearing completely. This agrees with our discussion of $\lambda^{\ell}\left(H^2\,E_H\right)$ in the previous paragraph. The number of eigenvalues large in size of $h^2\,P_{h,H}\,A_h$ is proportional to the number of close-to-zero eigenvalues of $H^2\,E_H$. This number is small on the finest mesh considered in Fig. 4, increases on intermediate coarser meshes and is zero on the coarsest mesh.

The previous discussion implies that in a plot $\lambda^{\ell}\left(h^2\,P_{h,H}\,A_h\right)$ on the real axis (instead of versus ℓ as before), the spectrum appears clustered between $2-\kappa^2$ and $4-\kappa^2$, except for two tails that spread along both sides of the real axis. The spread of these tails is inversely proportional to the size of the smallest eigenvalues of $H^2 E_H$. The number of elements in these tails is proportional to the number close-to-zero eigenvalues of E_H. For a fixed value of the wavenumber, the spread and number of elements in the tail vary with the number of grid points per wavelength employed.

5.4 Eigenvalues of the Preconditioned Deflated Operator $M^{-1}_{h,\beta_2}\,P_{h,H}\,A_h$

The diagonal blocks of the preconditioned deflated operator $(M^{-1}_{h,\beta_2}\,P_{h,H}\,A_h)^{\ell}$ are for $1 \leq \ell \leq n/2-1$ given by

$$(M^{-1}_{h,\beta_2}\,P_{h,H}\,A_h)^{\ell} = (M^{-1}_{h,\beta_2})^{\ell}\,(P_{h,H}\,A_h)^{\ell}\,. \tag{47}$$

and for $\ell = n/2$ by

$$(M^{-1}_{h,\beta_2}\,P_{h,H}\,A_h)^{n/2} = \frac{2-\kappa^2}{2-\kappa^2(1-\iota\beta_2)}\,. \tag{48}$$

From the singularity of the block $(P_{h,H}\,A_h)^{\ell}$ and (47) follows that the ℓ-th diagonal block $(M^{-1}_{h,\beta_2}\,P_{h,H}\,A_h)^{\ell}$ is singular as well. Its non-zero eigenvalue can thus be computed by merely computing its trace. The diagonal blocks of the h^2-scaled preconditioner $h^2\,M_{h,\beta_2}$ in the basis (29) are for $1 \leq \ell \leq n/2-1$ given by

$$(h^2\,M_{h,\beta_2})^{\ell} = \begin{pmatrix} \lambda^{\ell}(h^2\,M_{h,\beta_2}) & 0 \\ 0 & \lambda^{n-\ell}(h^2\,M_{h,\beta_2}) \end{pmatrix} \tag{49}$$

Assuming the preconditioner to be inverted exactly, the diagonal blocks of the inverse of the preconditioner are given by

$$(h^{-2}\,M^{-1}_{h,\beta_2})^{\ell} = \begin{pmatrix} \mu^{\ell}(h^2\,M_{h,\beta_2}) & 0 \\ 0 & \mu^{n-\ell}(h^2\,M_{h,\beta_2}) \end{pmatrix}\,. \tag{50}$$

The non-zero eigenvalue of the ℓ-th diagonal block $(M_{h,\beta_2}^{-1}\, P_{h,H}\, A_h)^\ell$ is then given by

$$\begin{aligned}\lambda^\ell(M_{h,\beta_2}^{-1}\, P_{h,H}\, A_h) &= \mathrm{Tr}[(M_{h,\beta_2}^{-1}\, P_{h,H}\, A_h)^\ell] \\ &= \frac{1}{\lambda^\ell(H^2 E_H)} \big[\mu^\ell(h^2\, M_{h,\beta_2})\, pa_{11,h,H}^\ell + \\ &\quad \mu^{n-\ell}(h^2\, M_{h,\beta_2})\, pa_{22,h,H}^\ell\big]\,. \end{aligned} \tag{51}$$

Observe that the eigenvalues of the Galerkin coarse grid operator E_H appear in the denominator. The coefficients $pa_{11,h,H}^\ell$ and $pa_{22,h,H}^\ell$ are real-valued. It is thus easy to split the non-zero eigenvalue $\lambda^\ell(M_{h,\beta_2}^{-1}\, P_{h,H}\, A_h)$ is a real and imaginary part and obtain for $1 \le \ell \le n/2 - 1$

$$\begin{aligned}\mathrm{Re}\left[\lambda^\ell(M_{h,\beta_2}^{-1}\, P_{h,H}\, A_h)\right] = \frac{1}{\lambda^\ell(H^2\, E_H)} \Big[&\mathrm{Re}\left[\mu^\ell(h^2\, M_{h,\beta_2})\right] pa_{11,h,H}^\ell + \\ &\mathrm{Re}\left[\mu^{n-\ell}(h^2\, M_{h,\beta_2})\right] pa_{22,h,H}^\ell \Big]\,,\end{aligned} \tag{52}$$

and

$$\begin{aligned}\mathrm{Im}\left[\lambda^\ell(M_{h,\beta_2}^{-1}\, P_{h,H}\, A_h)\right] = \frac{1}{\lambda^\ell(H^2\, E_H)} \Big[&\mathrm{Im}\left[\mu^\ell(h^2\, M_{h,\beta_2})\right] pa_{11,h,H}^\ell + \\ &\mathrm{Im}\left[\mu^{n-\ell}(h^2\, M_{h,\beta_2})\right] pa_{22,h,H}^\ell \Big]\,.\end{aligned} \tag{53}$$

Next we will use the results derived in Sect. 3 to find upper bounds for this real and imaginary part. These bounds will allow us to argue how the preconditioner transforms the eigenvalues of the deflated operator and how in particular the value of the damping parameter β_2 affects this transformation.

We start by considering the real part (52). The inequality (14) states that both $\mathrm{Re}\left[\mu^\ell(h^2\, M_{h,\beta_2})\right]$ and $\mathrm{Re}\left[\mu^{n-\ell}(h^2\, M_{h,\beta_2})\right]$ are bounded above by 1. We thus have that

$$\mathrm{Re}\left[\lambda^\ell(M_{h,\beta_2}^{-1}\, P_{h,H}\, A_h)\right] \le \frac{1}{\lambda^\ell(H^2\, E_H)} \left[pa_{11,h,H}^\ell + pa_{22,h,H}^\ell\right] = \lambda^\ell(h^2\, P_{h,H}\, A_h)\,, \tag{54}$$

where we have used expression (45) for $\lambda^\ell(h^2\, P_{h,H}\, A_h)$. The distance between $\mathrm{Re}\left[\lambda^\ell(M_{h,\beta_2}^{-1}\, P_{h,H}\, A_h)\right]$ and $\lambda^\ell(h^2\, P_{h,H}\, A_h)$ can be increased by taking large values of β_2. This is particularly interesting for those values of ℓ for which $\lambda^\ell(h^2\, P_{h,H}\, A_h)$ is large in size, i.e., for those values of ℓ corresponding to the near-null space of E_H. By taking large values of β_2, these large values of $\lambda^\ell(h^2\, P_{h,H}\, A_h)$ can be reduced, i.e., brought back to the center of the cluster of the eigenvalues by the action of the preconditioner. Eigenvalues $\lambda^{\ell}(h^2 \, P_{h,H} \, A_h)$ that lie in the interval from $2-\kappa^2$ to $4-\kappa^2$ are mapped to eigenvalues with a real part in a bounded interval. The length of this interval shrinks and its midpoint shift to zero of β_2 increases. Despite of this shift to zero, a larger damping than in the case without deflation can be chosen.

Next we consider the imaginary part (53). We use the expression in (18) to rewrite the imaginary parts $\text{Im}\left[\mu^{\ell}(h^2 \, M_{h,\beta_2})\right]$ and $\text{Im}\left[\mu^{n-\ell}(h^2 \, M_{h,\beta_2})\right]$ to obtain that

$$\text{Im}\left[\lambda^{\ell}(M_{h,\beta_2}^{-1} \, P_{h,H} \, A_h)\right] = \frac{-\beta_2 \kappa^2}{\lambda^{\ell}(H^2 \, E_H)} \left[\frac{p a_{11,h,H}^{\ell}}{|\lambda^{\ell}(h^2 \, M_{h,\beta_2})|} + \frac{p a_{22,h,H}^{\ell}}{|\lambda^{n-\ell}(h^2 \, M_{h,\beta_2})|} \right]. \quad (55)$$

On meshes with a sufficient number of grid points per wavelength, κ^2 is a small number. Expression (55) thus yields a small imaginary part except for those values of ℓ for which $\lambda^{\ell}(H^2 \, E_H) \approx 0$ and thus also $\lambda^{\ell}(h^2 \, M_{h,\beta_2}) \approx 0$. Eigenvalues $\lambda^{\ell}(h^2 \, P_{h,H} \, A_h)$ inside and outside the interval from $2-\kappa^2$ to $4-\kappa^2$ are mapped to eigenvalues with an imaginary part that is small and that increases proportionally to β_2, respectively.

In Fig. 5 we plotted the non-zero eigenvalues of $M_{h,\beta_2}^{-1} \, P_{h,H} \, A_h$ in the complex plane for $k = 1000$ and $\beta_2 = 1$. In traversing the hierarchy from the finest to the coarsest level, the range in the real part of the eigenvalues is seen to first increases to subsequently decrease starting from the third coarsest level with five grid point per wavelength. This in accordance with our previous discussion.

We can summarize the discussion by stating the action of the preconditioner is to contract and rotate the eigenvalues of the deflated operator. This is illustrated in Fig. 6 in which the eigenvalues $\lambda^{\ell}(M_{h,\beta_2}^{-1} \, P_{h,H} \, A_h)$ are plotted in the complex plane for $k = 1000$ and ten grid point per wavelength.

6 Numerical Results

In this section we present numerical results for the one-dimensional problem on the unit interval and the two-dimensional problem on the unit square. For both problem we consider the problem without damping supplied with homogeneous Dirichlet boundary conditions discretized using either 10 or 20 grid points per wavelength. We adopt a two-level variant of the deflation operator and assume the both preconditioner on the finest level and the coarse grid operator to be inverted exactly. As outer Krylov we run full GMRES with a zero initial guess. We declare convergence at the k-th iteration if the relative residual norm $\|A_h x_h^k - b_h\|_2 / \|b_h\|_2$ drops below 10^{-6}. We compare the following five algorithmic variants. The first variant merely employs A-DEF1 (without CSLP) as a preconditioner. The second, third and fourth variant combine A-DEF1 and CSLP multiplicative with β_2 equal to 0.5, 1 and 10, respectively. The fifth variant employs $\beta_2 = 10$ and approximates the CSLP preconditioner by its diagonal. The required numbered GMRES iterations for the one and two-dimensional problem are given in Tables 2 and 3, respectively. For

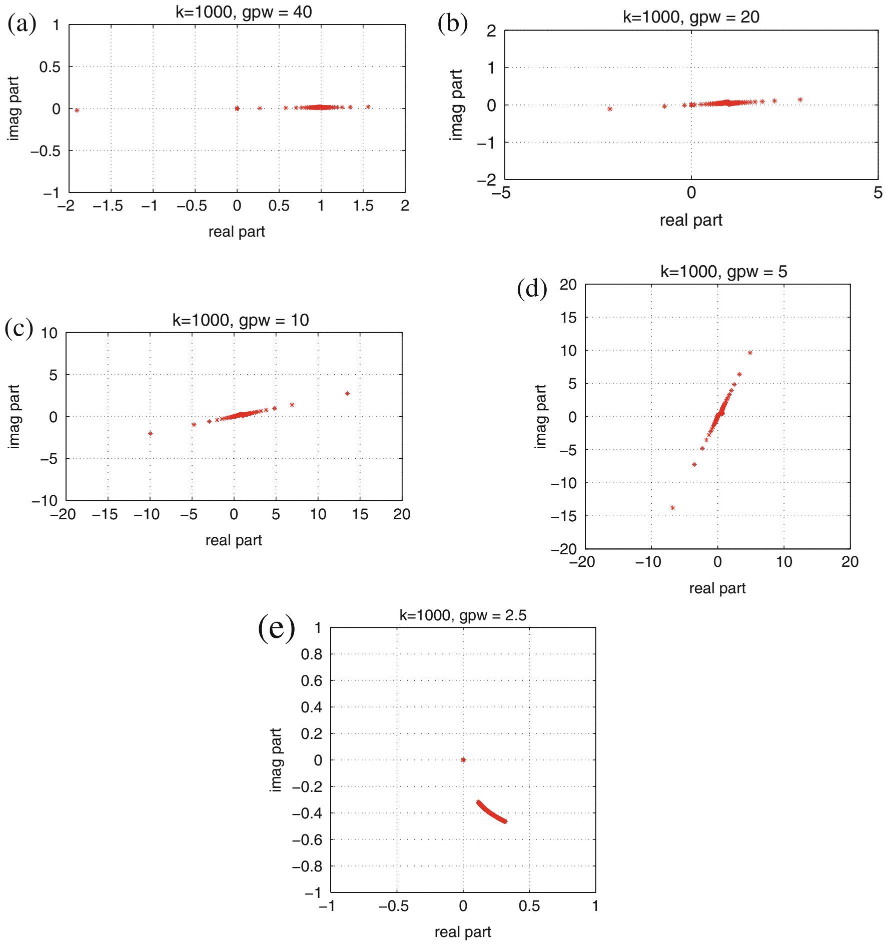


Fig. 5 Eigenvalues of the preconditioned deflated operator $M_{h,\beta_2}^{-1} P_{h,H} A_h$ in the complex plane for $k = 1000$ and $\beta_2 = 1$ on a multigrid hierarchy consisting of five levels. On the finest level 40 grid points per wave length are employed. On each coarser level the number of grid points per wave length is halved. (**a**) 40 gpw; (**b**) 20 gpw; (**c**) 10 gpw; (**d**) 5 gpw; (**e**) 2.5 gpw

the second and third variant we compare the multiplicative combination of A-DEF1 and CSLP with merely using CSLP as a preconditioner.

From Tables 2 and 3 we conclude that for the one and two-dimensional problem the combined use of A-DEF1 and CSLP

- results in a lower iteration count than either A-DEF1 or CSLP used separately. This reduction grows with the wave number;
- allows to use a large damping parameter β_2 without significantly increasing the iteration count;

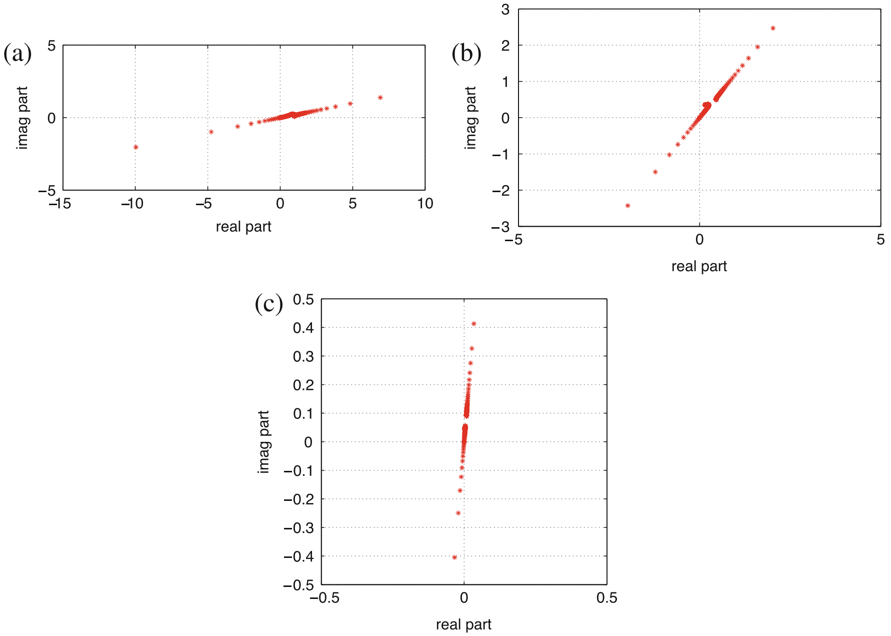


Fig. 6 Eigenvalues of the preconditioned deflated operator $M_{h,\beta_2}^{-1} P_{h,H} A_h$ on a fixed mesh for various values of the damping parameter β_2 using 10 grid points per wavelength for $k = 1000$. (**a**) $\beta_2 = 1$; (**b**) $\beta_2 = 10$; (**c**) $\beta_2 = 100$

- allows to set $\beta_2 = 10$ and to approximate the CSLP preconditioner by its diagonal without significantly increasing the iteration count.

7 Conclusions

In this paper we considered a solution method for the Helmholtz equation that combines the complex shifted Laplace preconditioner with a deflation technique that employs multigrid vectors. We derived closed form expressions for the eigenvalues of the deflated preconditioned operator through a model problem analysis. From this analysis we conclude that a much larger damping parameter can be used without adversely affecting the convergence of outer Krylov acceleration. Further research is required to tune the algorithmic to large scale applications.

Table 2 Iteration count for various methods for the 1D problem without damping

1D without damping					
		$\beta_2 = 0.5$	$\beta_2 = 1$	$\beta_2 = 10$	$\beta_2 = 10$
k	A-DEF1	CSLP/CSLP+A-DEF1	CSLP/CSLP+A-DEF1	CSLP+A-DEF1	JACOBI+A-DEF1
10 gpw					
10	5	7/3	8/4	5	5
20	9	10/5	12/6	7	7
40	15	16/8	20/8	9	9
80	15	23/8	33/9	9	9
160	20	36/13	55/14	14	12
320	30	61/19	97/20	19	19
640	45	108/33	179/33	34	32
20 gpw					
10	9	7/3	8/4	5	6
20	13	10/4	12/4	5	6
40	14	15/5	19/5	6	6
80	15	22/6	33/6	6	7
160	19	37/8	56/8	8	8
320	18	59/9	95/9	9	9
640	28	104/14	174/14	14	15
1280	36	190/23	328/23	23	23

Table 3 Iteration count for various methods for the 2D problem without damping

2D without damping					
		$\beta_2 = 0.5$	$\beta_2 = 1$	$\beta_2 = 10$	$\beta_2 = 10$
k	A-DEF1	CSLP/CSLP+A-DEF1	CSLP/CSLP+A-DEF1	CSLP+A-DEF1	JACOBI+A-DEF1
10 gpw					
10	18	9/5	11/5	9	11
20	24	17/7	22/8	10	11
40	36	45/16	64/16	19	21
80	68	130/43	210/41	45	46
20 gpw					
5	18	5/3	6/3	5	9
10	17	9/3	11/3	3	5
15	21	12/4	16/4	5	9
20	24	16/6	22/6	7	10
30	20	29/5	40/5	5	12

Acknowledgements The authors would like to sincerely thank Prof. Ira Livshits and Prof. Martin Gander for numerous discussions that indirectly give raise to material in this paper.

References

1. Y. A. Erlangga, C. Vuik, and C. W. Oosterlee. On a class of preconditioners for solving the Helmholtz equation. *Appl. Numer. Math.*, 50(3–4):409–425, 2004.
2. Y. A. Erlangga, C. W. Oosterlee, and C. Vuik. A novel multigrid based preconditioner for heterogeneous Helmholt problems. *SIAM J. Sci. Comput*, 27:1471–1492, 2006.
3. C. I. Goldstein A. Bayliss and E. Turkel. An iterative method for the Helmholtz equation. *Journal of Computational Physics*, 49:443 – 457, 1983.
4. L. A. Laird and M. B. Giles. Preconditioned iterative solution of the 2D Helmholtz equation. Technical report, Comp. Lab. Oxford University UK, 2002. NA-02/12.
5. M. M. M. Made. Incomplete factorization-based preconditionings for solving the Helmholtz equation. *International Journal for Numerical Methods in Engineering*, 50:1077–1101, 2001.
6. Y. A. Erlangga and R. Nabben. On a multilevel Krylov method for the Helmholtz equation preconditioned by shifted Laplacian. *Electronic Transactions on Numerical Analysis*, 31:403–424, 2008.
7. B. Reps, W. Vanroose, and H. Bin Zubair. On the indefinite Helmholtz equation: Complex stretched absorbing boundary layers, iterative analysis, and preconditioning. *J. Comput. Phys.*, 229:8384–8405, November 2010.
8. M. Bollhöfer, M. J. Grote, and O. Schenk. Algebraic multilevel preconditioner for the Helmholtz equation in heterogeneous media. *SIAM Journal on Scientific Computing*, 31:3781–3805, 2009.
9. T. Airaksinen, E. Heikkola, A. Pennanen, and J. Toivanen. An algebraic multigrid based shifted-Laplacian preconditioner for the Helmholtz equation. *Journal of Computational Physics*, 226:1196 – 1210, 2007.
10. M.J. Gander, I. G. Graham, and E. A. Spence. Applying GMRES to the Helmholtz equation with shifted Laplacian preconditioning: what is the largest shift for which wavenumber-independent convergence is guaranteed? *Numerische Mathematik*, pages 1–48, 2015.
11. J. Zhu, X. W. Ping, R. S. Chen, Z. H. Fan, and D. Z. Ding. An incomplete factorization preconditioner based on shifted Laplace operators for FEM analysis of microwave structures. *Microwave and Optical Technology Letters*, 52:1036–1042, 2010.
12. T. Airaksinen, A. Pennanen, and J. Toivanen. A damping preconditioner for time-harmonic wave equations in fluid and elastic material. *J. Comput. Phys.*, 228:1466–1479, March 2009.
13. C. D. Riyanti, A. Kononov, Y. A. Erlangga, C. Vuik, C. Oosterlee, R.E. Plessix, and W.A. Mulder. A parallel multigrid-based preconditioner for the 3D heterogeneous high-frequency Helmholtz equation. *Journal of Computational Physics*, 224:431–448, 2007.
14. R. E. Plessix. A Helmholtz iterative solver for 3D seismic-imaging problems. *Geophysics*, 72:SM185–SM194, 2007.
15. R. E. Plessix. Three-dimensional frequency-domain full-waveform inversion with an iterative solver. *Geophysics*, 74(6):149–157, 2009.
16. N. Umetani, S. P. MacLachlan, and C. W. Oosterlee. A multigrid-based shifted Laplacian preconditioner for a fourth-order Helmholtz discretization. *Numerical Linear Algebra with Applications*, 16:603–626, 2009.
17. T. Airaksinen and S. Mönkölä. Comparison between the shifted-Laplacian preconditioning and the controllability methods for computational acoustics. *J. Comput. Appl. Math.*, 234:1796–1802, July 2010.
18. L. Zepeda-Núnez and L. Demanet. The method of polarized traces for the 2D Helmholtz equation. *Journal of Computational Physics*, 308:347–388, 2016.
19. D. Osei-Kuffuor and Y. Saad. Preconditioning Helmholtz linear systems. *Appl. Numer. Math.*, 60:420–431, April 2010.
20. H. Calandra, S. Gratton, R. Lago, X. Pinel, and X. Vasseur. Two-level preconditioned Krylov subspace methods for the solution of three-dimensional heterogeneous Helmholtz problems in seismics. *Numerical Analysis and Applications*, 5:175–181, 2012.

21. Y.A. Erlangga. Advances in iterative methods and preconditioners for the Helmholtz equation. *Archives of Computational Methods in Engineering*, 15:37–66, 2008.
22. Huangxin Chen, Haijun Wu, and Xuejun Xu. Multilevel preconditioner with stable coarse grid corrections for the helmholtz equation. *SIAM Journal on Scientific Computing*, 37(1):A221–A244, 2015.
23. L. Conen, V. Dolean, Rolf R. Krause, and F. Nataf. A coarse space for heterogeneous helmholtz problems based on the Dirichlet-to-Neumann operator. *Journal of Computational and Applied Mathematics*, 271:83–99, 2014.
24. M. Ganesh and C. Morgenstern. An efficient multigrid algorithm for heterogeneous acoustic media sign-indefinite high-order FEM models. *Numerical Linear Algebra with Applications*, 2016.
25. I. Livshits. Multiple galerkin adaptive algebraic multigrid algorithm for the helmholtz equations. *SIAM Journal on Scientific Computing*, 37(5):S195–S215, 2015.
26. L. N. Olson and J. B. Schroder. Smoothed aggregation for helmholtz problems. *Numerical Linear Algebra with Applications*, 17(2-3):361–386, 2010.
27. C. C. Stolk. A dispersion minimizing scheme for the 3-d Helmholtz equation based on ray theory. *Journal of Computational Physics*, 314:618–646, 2016.
28. J. Poulson, B. Engquist, S. Li, and L. Ying. A parallel sweeping preconditioner for heterogeneous 3d Helmholtz equations. *SIAM Journal on Scientific Computing*, 35(3):C194–C212, 2013.
29. P. Tsuji, B. Engquist, and L. Ying. A sweeping preconditioner for time-harmonic Maxwell equations with finite elements. *Journal of Computational Physics*, 231(9):3770–3783, 2012.
30. A. Vion and C. Geuzaine. Double sweep preconditioner for optimized schwarz methods applied to the helmholtz problem. *Journal of Computational Physics*, 266:171–190, 2014.
31. AH Sheikh, D Lahaye, L Garcia Ramos, R Nabben, and C Vuik. Accelerating the shifted laplace preconditioner for the helmholtz equation by multilevel deflation. *Journal of Computational Physics*, 322:473–490, 2016.
32. M. B. van Gijzen, Y. A. Erlangga, and C. Vuik. Spectral analysis of the discrete Helmholtz operator preconditioned with a shifted Laplacian. *SIAM Journal on Scientific Computing*, 29:1942–1958, 2007.
33. S. Cools and W. Vanroose. Local Fourier analysis of the complex shifted Laplacian preconditioner for helmholtz problems. *Numerical Linear Algebra with Applications*, 20(4):575–597, 2013.
34. Y. A. Erlangga and R. Nabben. Deflation and balancing preconditioners for Krylov subspace methods applied to nonsymmetric matrices. *SIAM J. Matrix Anal. Appl.*, 30(2):684–699, 2008.
35. A. H. Sheikh, D. Lahaye, and C. Vuik. On the convergence of shifted Laplace preconditioner combined with multilevel deflation. *Numerical Linear Algebra with Applications*, 20(4):645–662, 2013.
36. S. Balay, J. Brown, K. Buschelman, V. Eijkhout, W. D. Gropp, D. Kaushik, M. G. Knepley, L. Curfman McInnes, B. F. Smith, and H. Zhang. PETSc users manual. Technical Report ANL-95/11 - Revision 3.4, Argonne National Laboratory, 2013.
37. A. H. Sheikh. *Development of the Helmholtz Solver Based On A Shifted Laplace Preconditioner and A Multilevel Deflation Technique*. PhD thesis, DIAM, TU Delft, 2014.
38. A. Bayliss, C.I. Goldstein, and E. Turkel. On accuracy conditions for the numerical computation of waves. *Journal of Computational Physics*, 59(3):396 – 404, 1985.
39. I. M. Babuska and S. A Sauter. Is the pollution effect of the FEM avoidable for the Helmholtz equation considering high wave numbers? *SIAM review*, 42(3):451–484, 2000.
40. O. G. Ernst and M. J. Gander. Why it is difficult to solve Helmholtz problems with classical iterative methods. In I. G. Graham, T. Y. Hou, O. Lakkis, and R. Scheichl, editors, *Numerical Analysis of Multiscale Problems*, volume 83 of *Lecture Notes in Computational Science and Engineering*, pages 325–363. Springer Berlin Heidelberg, 2012.

41. M. Gander. Fourier analysis of Helmholtz problems with Robin boundary conditions. Private Communications 2016.
42. I. Livshits. Use of shifted laplacian operators for solving indefinite helmholtz equations. *Numerical Mathematics: Theory, Methods and Applications*, 8(01):136–148, 2015.
43. J. M. Tang. *Two Level Preconditioned conjugate gradient methods with applications to bubbly flow problems.* PhD thesis, DIAM, TU Delft, 2008.
44. U. Trottenberg, C. W. Oosterlee, and A. Schüller. *Multigrid.* Academic Press, London, 2000.

The Multilevel Krylov-Multigrid Method for the Helmholtz Equation Preconditioned by the Shifted Laplacian

Yogi A. Erlangga, Luis García Ramos, and Reinhard Nabben

Abstract This chapter discusses a multilevel Krylov method (MK-method) for solving the Helmholtz equation preconditioned by the shifted Laplacian preconditioner, resulting in the so-called multilevel Krylov-multigrid (MKMG) method. This method was first presented in Erlangga and Nabben (E. Trans. Numer. Anal. 31:403–424, 2008). By combining the MK method with the shifted Laplacian preconditioner, it is expected that the issues related to indefiniteness and small eigenvalues of the Helmholtz matrix can be resolved simultaneously, leading to an equivalent system, whose matrix is spectrally favorable for fast convergence of a Krylov method. The eigenvalues of the preconditioned system preconditioned by the ideal MKMG operator lie on (or inside) the same circles known from the shifted Laplace preconditioning. But they are much better clustered. Here, we distinguish between the so-called ideal MKMG and the practical MKMG method. Numerical results for the practical MKMG presented here suggest that it is indeed possible to achieve an almost gridsize- and wavenumber-independent convergence, provided that the coarse-grid system in the ideal MK method is properly and accurately approximated.

1 Introduction

Fast iterative solvers for the Helmholtz equation have been an active research area in the last 15 years. One of the main drivers comes from the oil industry, where the Helmholtz solver is an important component in the frequency-domain Full Waveform Inversion (FWI) for oil prospecting. For other benchmark elliptic problems as the Poisson problem, there exist well-established fast solvers (e.g., multigrid) and the convergence can be made independent of the grid size with an

Y.A. Erlangga (✉)
Mathematics Department, Nazarbayev University, 53 Kabanbay Batyr Ave., 010000 Astana, Kazakhstan
e-mail: yogiae@gmail.com; yogi.erlangga@nu.edu.kz

L. García Ramos • R. Nabben
Institut für Mathematik, TU Berlin, MA 3-3, Strasse des 17. Juni 136, 10623 Berlin, Germany
e-mail: garcia@math.tu-berlin.de; nabben@math.tu-berlin.de

D. Lahaye et al. (eds.), *Modern Solvers for Helmholtz Problems*, Geosystems Mathematics, DOI 10.1007/978-3-319-28832-1_5

error reduction factor of, say, 0.1. In contrast, for the Helmholtz equation, in addition to grid-independent convergence we require a scalable solver that requires a number of iterations independent of the wavenumber k (or equivalently, the frequency f). It is evident that achieving an (h, k)-independent convergence is quite difficult and challenging, but recent developments have demonstrated that it may not be out of reach.

Some promising directions, for example, include wave-ray methods [8, 25] and Krylov methods preconditioned by the sweeping preconditioner [12, 30, 36]. The wave-ray method was developed in the context of multigrid methods and is based on the representation of components of multigrid errors, which are difficult to reduce, by a ray function. The associated errors can be made smooth enough such that the error reduction can be effectively performed (using the so-called ray cycle). The sweeping preconditioner benefits from the use of a perfectly matched layer [1, 2, 6, 7, 39] for the discretization of the Sommerfeld boundary condition. This highly accurate mathematical representation allows the wave propagation in the complete domain to be treated locally and rather accurately in smaller subdomains. This approach can be viewed as a domain-decomposition method with interface conditions that ensure continuous propagation of waves within subdomains.

Another direction was proposed by the authors, within the context of the complex shifted-Laplacian (CSL) preconditioner and projection-type methods [15]. The shifted Laplacian preconditioner, initially introduced in [4, 5] and further extended in [24] and [18, 19], leads to a preconditioned Helmholtz system, which is spectrally favorable for fast convergence of a Krylov method. With a proper choice of parameters involved in the preconditioning operator, the eigenvalues of the preconditioned Helmholtz matrix can be bounded above (in magnitude) by one, and this bound is independent of h and k.

The trouble comes from the smallest eigenvalues, which become too close to zero as k increases. This suggests convergence deterioration with an increase in k, which is confirmed by numerical experiments. The good news, nevertheless, is that the number of iterations to reach convergence exhibits a linear increase with respect to k, with a small constant, and for a fixed k, the convergence can be made independent of h.

In many cases, for the system

$$Au = b, \quad A \in \mathbb{C}^{n \times n}, \tag{1}$$

the convergence of a Krylov method can be measured by the spectrum of A. If A is Hermitian positive definite (HPD), the convergence rate of CG can be bounded in terms of the condition number of A, $\kappa(A)$ [32], which in this case is the ratio of the largest eigenvalue to the smallest one. For general matrices, convergence bounds are somewhat more difficult to establish and do not express a direct connection with the condition number of A. In either case, small condition numbers are desirable, which qualitatively means clustering of eigenvalues around a value far from zero. A good preconditioner is a (nonsingular, simple to invert) matrix M such that the eigenvalues of $M^{-1}A$ are more clustered and farther from zero than those of A.

There exist methods to deal with small eigenvalues of A. One of them is called *deflation*, proposed by Nicolaides [28] for conjugate gradient methods, which has been interpreted and used in various ways; see, for instance, [9, 21, 26]. The objective, generally speaking, is to (explicitly) shift small eigenvalues of A to 0. For an HPD matrix A, following [21], this process can be represented by the matrix

$$P_D = I - AZE^{-1}Z^T, \quad E = Z^TAZ, \tag{2}$$

which is therefore called the *deflation* matrix. In (2), $Z \in \mathbb{C}^{n\times r}$ is a full rank matrix, whose columns form a basis for the deflation subspace. It is shown in [21] that the spectrum of P_DA contains r zero eigenvalues; see also [13, 27]. Since the components of the residuals corresponding to the zero eigenvalues do not enter the iteration, the convergence rate is now bounded in terms of the *effective* condition number of P_DA, which for a HPD matrix A is the ratio between the largest eigenvalue and the smallest nonzero eigenvalue of P_DA.

For convergence acceleration, it is desirable to deflate as many small eigenvalues as possible to zero. This however makes the matrix $E = Z^TAZ \in \mathbb{C}^{r\times r}$ too large for an efficient inversion by a direct method. While it is still possible to carry out this inversion implicitly using an iterative method, care has to be taken in determining the termination criteria [37], since deflation is rather sensitive to an inaccurate computation of E^{-1} [27].

Originally, deflation methods were introduced with the columns of the matrix Z consisting of (approximate) eigenvectors, corresponding to the small eigenvalues of A. These eigenvalues are then deflated to zero. In the last decade, deflation was used and analyzed in combination with domain decomposition and multigrid methods [27, 37]. There, the rectangular matrices Z^T and Z are known as the restriction and the prolongation (interpolation) operator, respectively. The matrix $E = Z^TAZ$ is the Galerkin or coarse-grid matrix, which is typically still very large. Theoretical results established in [27, 37] were however based merely on the assumption that Z is full rank. For implementations, Z was a sparse matrix that corresponds to different interpolation techniques. For arbitrary matrices Z the nonzero eigenvalues of P_DA may differ from the eigenvalues of A. It is then important to quantify the difference between the nonzero spectrum of the deflated matrix P_DA and the spectrum of the original matrix, and to determine if the nonzero eigenvalues of P_DA are shifted near zero or grow increasingly large.

An alternative to deflating eigenvalues to zero is by shifting them to the largest eigenvalue [14]. To put this procedure in the framework of the shifted Laplacian preconditioner for the Helmholtz system, we introduce an equivalent linear system to (1):

$$\hat{A}\hat{u} = \hat{b}, \tag{3}$$

where $\hat{A} = M_1^{-1}AM_2^{-1}$, $\hat{u} = M_2u$, and $\hat{b} = M_1^{-1}b$. Here, M_1 and M_2 are nonsingular preconditioning matrices. The shift of r small eigenvalues to the largest eigenvalue of $\hat{A}$ is done via the action of the matrix[1]

$$P_{\hat{N}} = I - \hat{A}Z\hat{E}^{-1}Z^T + \lambda_n Z\hat{E}^{-1}Z^T, \quad \hat{E} = Z^T\hat{A}Z, \tag{4}$$

on the general system (3), with λ_n the largest eigenvalue (in magnitude) of $\hat{A}$.

With (4), we then solve the left preconditioned system

$$P_{\hat{N}}\hat{A}u = P_{\hat{N}}\hat{b}$$

with a Krylov method.

The right preconditioning version of (4) can also be defined using the shift matrix

$$Q_{\hat{N}} = I - Z\hat{E}^{-1}Z^T\hat{A} + \lambda_n Z\hat{E}^{-1}Z^T. \tag{5}$$

Given (5), we then solve the preconditioned system

$$\hat{A}Q_{\hat{N}}\widetilde{u} = \hat{b}, \hat{u} = Q_{\hat{N}}\widetilde{u} \tag{6}$$

with a Krylov method. Note that $P_{\hat{N}}$ and $Q_{\hat{N}}$ are nonsingular and an explicit expression for the inverses for $\lambda_n = 1$ is given in [17]. Moreover, we have $\sigma(P_{\hat{N}}\hat{A}) = \sigma(\hat{A}Q_{\hat{N}})$, with $\sigma(\cdot)$ the spectrum of the matrix in the argument.

Different from (2), $P_{\hat{N}}$ is insensitive to an inexact inversion of $\hat{E}$, which therefore allows us to use a large deflation subspace to shift as many small eigenvalues as possible. The Galerkin matrix $\hat{E}$ can now be inverted implicitly and a (inner) Krylov method with a less tight termination criterion can be used. That means that a few steps of an inner Krylov method need to be performed. The convergence rate of this inner iteration can be significantly improved if a shift operator similar to (4) is also applied to the Galerkin system. The action of this shift operator will require solving another Galerkin system, which will be carried out again by just a few steps of a Krylov method. By this construction, however, the preconditioner is no longer fixed, and hence a flexible Krylov method such as FGMRES needs to be employed. A recursive application of this process leads to the so-called *multilevel Krylov* (MK) method. To refer to the number of iterations on the second, third etc, level, we write, e.g., MK(8,4,1), which means 8 iterations on the second level, 4 on the third, and just one on all the other levels.

For the solution of the Helmholtz equation, in (3), we use $M_1 = I$ and $M_2 = M$ as the complex shifted Laplace (CSL) preconditioner, to be inverted approximately

[1] Since $\hat{A}$ is not necessarily symmetric, the more general form for the shift matrix is

$$P_{\hat{N}} = I - \hat{A}Z\hat{E}^{-1}Y^T + \lambda_n Z\hat{E}^{-1}Y^T, \quad \hat{E} = Y^T\hat{A}Z,$$

where $Y, Z \in \mathbb{C}^{n\times r}$ are full rank and such that $\hat{E} = Y^T\hat{A}Z$ is nonsingular. We however prefer to start with (4) because in the end we will set $Y = Z$.

by one multigrid (MG) iteration. The resulting method is called MKMG method, or, in more detail, e.g., MKMG(8,4,1). Here, we have to distinguish between the *ideal* (two-level) MKMG and the *practical* MKMG method. In the ideal version, the coarse grid system with system matrix $\hat{E} = Z^T\hat{A}Z = Z^TAM^{-1}Z$ is solved exactly. This will require the exact inversion of M and an exact computation of $\hat{E}$. In higher dimensions (2D or 3D) this approach is no longer feasible. Moreover, since in the end the inverse of M is approximately computed by one multigrid iteration, M^{-1} is not explicitly available. In this case we use some approximation of $\hat{E} = Z^TAM^{-1}Z$, resulting in the practical MKMG method.

Algorithm 1 describes the preconditioning part of the MKMG method, which can be a building block in a flexible Krylov method. As mentioned above, in the ideal two-level MKMG method the coarse grid system will be solved exactly. In the other cases the coarse grid systems will be approximated and solved by a few iterations of a flexible Krylov method preconditioned by the same type of preconditioner. Note that (5) can be written as

$$Q_{\hat{N}} = I + Z\hat{E}^{-1}Z^T(\lambda_n I - AM^{-1}).$$

In Algorithm 2 we present a more detailed pseudocode for the practical MKMG method incorporated in a flexible GMRES method.

```
Solve $Ax = b$ by a flexible Krylov method right preconditioned by $M^{-1}\hat{Q}_N$
Multiplication of the preconditioner with a vector $v$
    $w := M^{-1}v$ /* Multigrid solve of shifted Laplacian */
    $s := Aw$ /* Matrix vector multiplication */
    $t := \lambda_n v - s$ /* Adding the shift */
    $\hat{t} := Z^T t$ /* Restriction */
    $\hat{s} := \hat{E}^{-1}\hat{t}$ /* Coarse grid solve, recursively with the same */
              /* preconditioner or exactly */
    $s := Z\hat{s}$ /* Prolongation */
    $w := v + s$ /* Update */
    $z := M^{-1}w$ /* Multigrid solve of shifted Laplacian */
```

Algorithm 1: MKMG as a preconditioner

It is proven in [23] that the eigenvalues of $\hat{A}\hat{Q}_N = AM^{-1}\hat{Q}_N$ (the ideal MKMG operator) lie on exact the same circles as the CSL-preconditioned linear system for Dirichlet boundary conditions and are enclosed by the same circles for Sommerfeld boundary conditions. Moreover, they are better clustered. This holds for any dimension, any wavenumber and any choice of Z. In [34] exact formulas for the eigenvalues of $AM^{-1}\hat{Q}_N$ in the 1D case are given. These theoretical results help to explain the good performance of the MKMG method. Although the eigenvalue formulas show that for $k \to \infty$, the eigenvalues of $AM^{-1}\hat{Q}_N$ move to zero, this happens only for very high wavenumbers (see Fig. 6).

Numerical experiments with the ideal MKMG suggests that an h,k-independent fast convergence can be attained. For the practical MKMG method, our numerical results demonstrate a convergence, which is mildly h-dependent and k-dependent (Sect. 5). As $h \to 0$, the convergence can be made practically independent of h. Considering the fast convergence of the ideal MKMG method, improvement from the current practical MKMG can still be achieved by, e.g., having a better approximation of the coarse-grid matrix $\hat{E}$.

Remark 1 Instead of preconditioning the matrix first and then deflate, alternatively, one can first deflate and then precondition. This is done by the operator

$$B = M^{-1}P_D + ZE^{-1}Z^T.$$

Similarly to $\hat{Q}_N$, some eigenvalues are shifted to one if B is used as a preconditioner. The operator B is called the A-DEF1 operator in [37]. Motivated by the MK framework given in [15], a similar recursive structure using inner Krylov iterations to solve the coarse grid systems with E is used in [35]. The resulting method can be seen as a preconditioned MK method, see [34].

Remark 2 Elman et al. [10] used Krylov iterations (in this case, GMRES [33]) in a multilevel fashion for solving the Helmholtz equation. However, their approach is basically a multigrid method, specially adapted for the Helmholtz equation. While at the finest and coarsest level, standard smoothers still have good smoothing properties, at the intermediate levels GMRES is employed in place of standard smoothers. Since GMRES does not have a smoothing property, it plays a role in reducing the errors *but* not in smoothing them. A substantial number of GMRES iterations at the intermediate levels, however, is required to achieve a significant reduction of errors; see [29].

Remark 3 Even though the multilevel Krylov method uses a hierarchy of linear systems similar to multigrid, the way it treats each system and establishes a connection between systems differs from multigrid [14, 16]. In fact, the multilevel Krylov method is not by definition an instance of a multigrid method. With regard to the work in [10], we shall show numerically that the multilevel Krylov method can handle linear systems at the intermediate levels efficiently; i.e., a fast multilevel Krylov convergence can be achieved with only a few Krylov iterations at the intermediate levels.

We organize this chapter as follows. In Sect. 2, we first revisit the Helmholtz equation and our preconditioner of choice, the shifted Laplace preconditioner. In Sect. 3, some relevant theoretical results concerning our multilevel Krylov method are discussed. Practical implementations are explained in Sect. 4. Numerical results from 2D Helmholtz problems are presented in Sect. 5. Finally, in Sect. 6, we draw some conclusions and outlook. The material presented here is based on the presentation in [15].

2 The Helmholtz Equation and the Shifted Laplace Preconditioner

The 2D Helmholtz equation for heterogeneous media can be written as

$$\mathscr{A}u := -\left(\frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2} + k^2(x,y)\right) u(x,y) = g(x,y), \text{ in } \Omega \subset \mathbb{R}^2, \tag{7}$$

where $k(x,y)$ is the wavenumber, and g is the source term. Dirichlet, Neumann, or Sommerfeld (non-reflecting) conditions can be applied at the boundaries $\Gamma = \partial\Omega$; see, e.g., [11]. Discretization of (7) and the boundary conditions results in a large linear system with sparse and symmetric but indefinite matrix. For this kind of system, Krylov subspace methods with standard preconditioners, e.g., ILU, [20], converge unacceptably slowly to a solution.

The general shifted Laplacian is defined as follows:

$$\mathscr{M} := -\frac{\partial^2}{\partial x^2} - \frac{\partial^2}{\partial y^2} - (\alpha + \hat{\imath}\beta)k^2(x,y), \tag{8}$$

for some constants $\alpha, \beta \in \mathbb{R}$ and $\hat{\imath} = \sqrt{-1}$. The preconditioning matrix M is obtained from a discretization of (8), with the same boundary conditions as for (7). The solution u is computed from the (right) preconditioned system

$$AM^{-1}\hat{u} = b, \quad u = M^{-1}\hat{u}, \tag{9}$$

where A and M are the associated Helmholtz and preconditioner matrix respectively.

References [22, 40], e.g., discuss at length on how (α, β) has to be chosen. For our presentation, we shall only consider the pair $(\alpha, \beta) = (1, 0.5)$, which in [18] is shown to lead to an efficient and robust preconditioning operator. Since the convergence of Krylov methods is closely related to the spectrum of the given matrix, we shall give some insight on the spectrum of the preconditioned Helmholtz system (9) in the remainder of this section.

Theorem 1 *Let $A = L + \hat{\imath}C - K$ and $M = L + \hat{\imath}C - (1 - 0.5\hat{\imath})K$, with L, C and K matrices associated with a discretization of the negative Laplacian, the boundary conditions and the Helmholtz (k^2) term, respectively.*

1. *For Dirichlet boundary conditions, $C = 0$ and the eigenvalues of $M^{-1}A$ lie on the circle in the complex plane with center $c = (\frac{1}{2}, 0)$ and radius $R = \frac{1}{2}$.*
2. *For Sommerfeld boundary conditions, $C \neq 0$ and the eigenvalues of $M^{-1}A$ are enclosed by the circle with center $c = (\frac{1}{2}, 0)$ and radius $R = \frac{1}{2}$.*

Proof The proof for arbitrary (α, β) can be found in [40]. □

Since $\sigma(M^{-1}A) = \sigma(AM^{-1})$, Theorem 1 holds also for AM^{-1}. For the Helmholtz equation with Dirichlet boundary conditions some detailed information about the spectrum, e.g., the largest and smallest eigenvalues, can also be derived. We shall follow the approach used in [19], which was based on a continuous formulation of the problem. The results, however, also hold for the discrete formulation as indicated in [19]. For simplicity, we consider the 1D Helmholtz equation.

At the continuous level, the eigenvalue problem of the preconditioned system can be written as

$$-\left(\frac{d^2}{dx^2}-k^2\right)u=\lambda\left(-\frac{d^2}{dx^2}-(1-0.5\hat{j})k^2\right)u, \tag{10}$$

with λ the eigenvalue and u now the eigenfunction. By using the ansatz $u = \sin(i\pi x)$, $i \in \mathbb{N}$, from(10), we find that

$$\lambda_i = \frac{i^2\pi^2-k^2}{i^2\pi^2-(1-0.5\hat{j})k^2},$$

with

$$\begin{cases} \mathrm{Re}(\lambda_i) = \dfrac{(i^2\pi^2-k^2)^2}{(i^2\pi^2-k^2)^2+0.25k^4}, \\ \mathrm{Im}(\lambda_i) = \dfrac{-0.5(i^2\pi^2-k^2)k^2}{(i^2\pi^2-k^2)^2+0.25k^4}. \end{cases} \tag{11}$$

Notice that $0 < \mathrm{Re}(\lambda_i) < 1$, and therefore

$$\lim_{i\to\infty}\mathrm{Re}(\lambda_i) = \lim_{k\to\infty}\mathrm{Re}(\lambda_i) = 1.$$

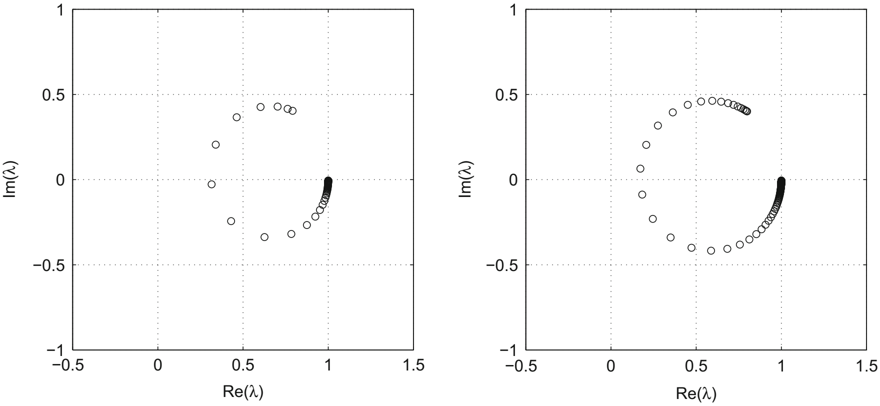
The real parts are close to zero if $i^2\pi^2$ are close to k^2. The sign of the imaginary parts depends on the mode i. Also, $\lim\limits_{k\to\infty}\mathrm{Im}(\lambda_i) = 0.5$ and $\lim\limits_{k\to\infty}\mathrm{Im}(\lambda_i) = -0.5$. Eliminating $i^2\pi^2$ in (11) yields

$$(\mathrm{Re}(\lambda_i)-0.5)^2+\mathrm{Im}(\lambda_i)^2 = 0.25.$$

Thus, λ_i lie on the circle with center $c = (\frac{1}{2}, 0)$ and radius $R = \frac{1}{2}$, as suggested by Theorem 1 (i). The largest $|\lambda_i|$ is approached as $i \to \infty$, where, in this case, $\mathrm{Re}(\lambda_i) \to 1$ and $\mathrm{Im}(\lambda_i) \to 0$. Thus, $\lim\limits_{i\to\infty}|\lambda_i| = 1$. This result holds for any choice of constant k.

Suppose now that for some i, $i^2\pi^2 - k^2 = \varepsilon$. For $\varepsilon \ll k$, $\mathrm{Re}(\lambda_i) \approx 4\varepsilon^2/k^4$ and $\mathrm{Im}(\lambda_i) \approx -2\varepsilon/k^2$, and hence

$$|\lambda_i|^2 = \mathrm{Re}(\lambda_i)^2+\mathrm{Im}(\lambda_i)^2 = \left(\frac{4\varepsilon^2}{k^4}\right)^2+\left(\frac{2\varepsilon}{k^2}\right)^2 \approx \frac{4\varepsilon^2}{k^4}.$$



$k = 20$

$k = 50$

Fig. 1 Spectrum of a typical 1D Helmholtz problem preconditioned with the shifted Laplacian. The wavenumbers k are 20 (*left*) and 50 (*right*). The Sommerfeld boundary condition is used

Therefore, while the spectrum of $M^{-1}A$ is more clustered than the spectrum of A, some eigenvalues still lie at a distance of order $\mathscr{O}(\varepsilon/k^2)$ from zero.

Figure 1 shows spectra associated with a 1D Helmholtz problem with $k = 20$ and 50 and the Sommerfeld condition at the boundaries. Clearly, the largest eigenvalue for both k's is essentially the same and close to one, but the smallest eigenvalue moves towards zero as k increases.

3 Multilevel Krylov Method

In the following we choose $M_1 = I$ and $M_2 = M$ as the shifted Laplacian in (8). As discussed in Sect. 1, one way to handle the small eigenvalues of $\hat{A} = AM^{-1}$ is by shifting them to the largest eigenvalue using either (4) or (5). Based on the result stated below Eq. (11), we set $\lambda_n = 1$.

For (5) the following spectral property holds.

Theorem 2 *Let $\hat{A}$ be normal, with eigenvalues $\lambda_1, \ldots, \lambda_n \in \sigma(\hat{A}) \subset \mathbb{C}$, ordered increasingly in magnitude. Let $Z \in \mathbb{C}^{n\times r}$, with $r < n$, be a full rank matrix whose columns are eigenvectors associated with the r smallest eigenvalues (in magnitude) of $\hat{A}$. Let $Q_{\hat{N}}$ be defined as in* (5)*. Then*

$$\sigma(\hat{A}Q_{\hat{N}}) = \{\lambda_n, \ldots, \lambda_n, \lambda_{r+1}, \ldots, \lambda_n\}.$$

Proof The proof requires the identity $P_{\hat{D}}\hat{A}Z = 0$, where $P_{\hat{D}} = I - \hat{A}Z\hat{E}^{-1}Z^T$, which is easily verified by a direct computation (see, e.g., [21]), and Theorem 3.5 of [14], which establishes the spectral equivalence $\sigma(P_{\hat{N}}\hat{A}) = \sigma(\hat{A}Q_{\hat{N}})$, with $P_{\hat{N}}$ as in (4).

First, for $i = 1, \dots, r$, we have $P_{\hat{N}}\hat{A}Z = P_{\hat{D}}\hat{A}Z + \lambda_n Z\hat{E}^{-1}Z^T\hat{A}Z = \lambda_n$. Next, for $r + 1 \leq i \leq n$, we have that

$$P_{\hat{N}}\hat{A}z_i = \hat{A}z_i - \hat{A}Z\hat{E}^{-1}Z^T\hat{A}z_i + \lambda_n Z\hat{E}^{-1}Z^T\hat{A}z_i = \lambda_i z_i,$$

due to orthogonality of eigenvectors. Finally, by using Theorem 3.5 of [14], $\sigma(P_{\hat{N}}\hat{A}) = \{\lambda_n, \dots, \lambda_n, \lambda_{+1}, \dots, \lambda_n\} = \sigma(\hat{A}Q_{\hat{N}})$. □

The above theorem can be generalized to the case where $\hat{A}$ is non-normal, with the term Z^T in (5) replaced by Y^T. Here the columns of Z and Y are, respectively, right and left eigenvectors of $\hat{A}$.

Thus, after applying $Q_{\hat{N}}$ to $\hat{A}$, r eigenvalues are no longer small and have been shifted to λ_n. The smallest eigenvalue (in magnitude) is now λ_{r+1}, and the rest of the spectrum remains untouched. If λ_{r+1} is of the same order of magnitude as λ_n, a Krylov subspace method is expected to converge faster.

The computation of eigenvectors, however, is very expensive for large linear systems and the use of eigenvectors for the columns of Z leads to a dense matrix.

In the following we consider the deflation and the shift operator under any full rank Z. We start with the deflation operator. Since

$$\hat{A}Q_{\hat{D}}Z = \hat{A}Z - \hat{A}Z\hat{E}^{-1}Z^T\hat{A}Z = \hat{A}Z - \hat{A}Z = 0,$$

with $Q_{\hat{D}} = I - Z\hat{E}^{-1}Z^T\hat{A}$, we obtain

$$\sigma(\hat{A}Q_{\hat{D}}) := \{0, \dots, 0, \mu_{r+1}, \dots, \mu_n\}.$$

Thus, $\hat{A}Q_{\hat{D}}$ has r zero eigenvalues for any matrix Z. In contrast to Theorem 2, the remaining eigenvalues $\mu_{r+1}, \dots, \mu_n$ are not, in general, eigenvalues of $\hat{A}$. Thus, r eigenvalues of $\hat{A}$ are shifted to zero, and the rest to μ_i.

The following theorem establishes a spectral relationship between deflation and the shift operator with any full rank Z.

Theorem 3 *Let $Z \in \mathbb{C}^{n\times r}$ be of rank r, $\hat{A}$ be nonsingular, and let $Q_{\hat{D}} = I - Z\hat{E}^{-1}Z^T\hat{A}$. If $Q_{\hat{N}}$ is defined as* (5)*, and Z is such that*

$$\sigma(\hat{A}Q_{\hat{D}}) := \{0, \dots, 0, \mu_{r+1}, \dots, \mu_n\},$$

then

$$\sigma(\hat{A}Q_{\hat{N}}) = \{\lambda_n, \dots, \lambda_n, \mu_{r+1}, \dots, \mu_n\}.$$

Proof Combine Theorems 3.4 and 3.5 of [14]. Note that the columns of Z are the left eigenvectors of $\hat{A}Q_{\hat{D}}$ corresponding to the eigenvalue equal to zero. We obtain

$$\hat{A}Q_{\hat{N}}Z = \lambda_n Z.$$

Theorem 3.5 of [14] gives

$$\sigma(\hat{A}Q_{\hat{N}}) = \sigma(P_{\hat{N}}\hat{A}).$$

Now, if

$$\hat{A}Q_{\hat{D}}x_i = \mu_i x_i,$$

for $r + 1 \leq i \leq n$ and some eigenvectors x_i, we easily obtain

$$P_{\hat{N}}\hat{A}(Q_{\hat{D}}x_i) = \mu_i(Q_{\hat{D}}x_i),$$

which completes the proof. □

Thus, while $Q_{\hat{D}}$ shifts r eigenvalues of $\hat{A}$ to zero, $Q_{\hat{N}}$ shifts r eigenvalues to λ_n. Under the arbitrariness of Z, the rest of the eigenvalues is also shifted to μ_i, $i = r + 1, \ldots, n$, but these eigenvalues are the same for both $\hat{A}Q_{\hat{D}}$ and $\hat{A}Q_{\hat{N}}$. Their exact values depend on the choice of Z. In general, $\mu_n \neq \lambda_n$. However, for any μ_n and λ_n, there exists a constant $\omega \in \mathbb{C}$ such that $\mu_n = \omega\lambda_n$. The constant ω is called the *shift scaling factor*. A shift correction can be incorporated in (5) by replacing λ_n with $\omega\lambda_n$. With this scaling, the spectrum of $\hat{A}Q_{\hat{D}}$ and $\hat{A}Q_{\hat{N}}$ differ only in the multiplicity of eigenvalue zero and μ_n. If the convergence can be measured merely by the ratio of the largest and smallest nonzero eigenvalues, a similar convergence for both methods can be expected.

As mentioned in Theorem 1, with Dirichlet conditions, the eigenvalues of the system preconditioned by the shifted Laplacian lie on a circle in the complex plane, and for Sommerfeld conditions inside this circle. For the ideal two-level MK method, where the coarse-grid (or, second-level) system $\hat{E} = Z^T AM^{-1}Z$ is solved exactly, the eigenvalues of $AM^{-1}Q_{\hat{N}}$ lie on the same circle as the eigenvalues of AM^{-1} for Dirichlet conditions. For Sommerfeld conditions the eigenvalues lie inside the same circle as the eigenvalues of AM^{-1}. Moreover, they are much better clustered. These surprising results are proven in [23]:

Theorem 4 *Let $Z \in \mathbb{C}^{n\times r}$ be of rank r. Let $Q_{\hat{N}}$ be defined as in* (5) *with $\lambda_n = 1$. Then, for Dirichlet boundary conditions the spectrum of $AM^{-1}Q_{\hat{N}}$ lies on the boundary of the same circle as the spectrum of AM^{-1}, i.e. the spectrum of $AM^{-1}Q_{\hat{N}}$ lies on the boundary of the circle with center $c = (\frac{1}{2}, 0)$ and radius $R = \frac{1}{2}$, for $(\alpha, \beta) = (1, 0.5)$. For Sommerfeld boundary conditions the spectrum of $AM^{-1}Q_{\hat{N}}$ is enclosed by the same circle. Moreover, in the case of Dirichlet boundary conditions,*

$$|\lambda_{\min}(AM^{-1})| \leq |\lambda_{\min}(AM^{-1}Q_{\hat{N}})|.$$

Proof The proof for arbitrary (α, β) is given in [23]. □

Note that this theorem holds for any dimension, any wavenumber, and any choice of full rank Z. Figure 2 illustrates the spectral result in Theorem 4 for a

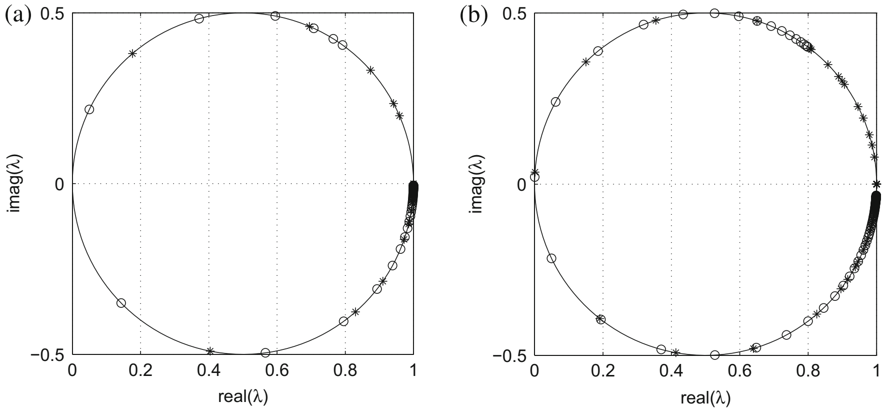


Fig. 2 Spectra of a preconditioned 1D Helmholtz problem with a Dirichlet boundary condition, and with $k = 20$ (*left*, **a**) and 50 (*right*, **b**). The number of grid points is $n = 100$ and 250, respectively. Eigenvalues of AM^{-1} are shown by "*open circle*", and of $AM^{-1}Q_{\hat{N}}$ by "*asterisk*". In $Q_{\hat{N}}$, Z is a random matrix of rank $r = n/2$

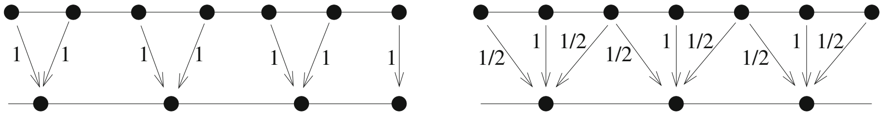


Fig. 3 Interpolation in 1D: piecewise-constant interpolation (*left*) and linear interpolation (*right*)

1D Helmholtz problem with a Dirichlet condition. To generate the spectrum, we used a full rank random matrix for Z in $Q_{\hat{N}}$. For $k = 20$, $|\lambda_{\min}(AM^{-1})| = 0.2230 < 0.4197 = |\lambda_{\min}(AM^{-1}Q_{\hat{N}})|$.

Using a dense matrix, like a random matrix or eigenvectors matrix, is not practical due to possibly excessive memory requirements, especially when r is quite large. Moreover, eigenvectors, in particular, are expensive to compute. For Z, we require that this matrix is sparse to avoid excessive memory requirements. A class of matrix satisfying this requirement is the multigrid prolongation/interpolation matrices. In this case, Z^T is a restriction operator, and $\hat{E} = Z^TAM^{-1}Z$ is the (Galerkin) coarse-grid approximation to A. Figure 3 shows two possible options for the transfer operator in a 1D setting: zeroth-order (piecewise-constant) interpolation and linear interpolation.

Spectra of $AM^{-1}Q_{\hat{N}}$ with Z based on the above-mentioned interpolations are shown in Fig. 4. Here, a Dirichlet condition and $k = 20$ or $k = 50$ are used to construct A and M. The spectral result of Theorem 4 is clearly satisfied: the eigenvalues lie on the circle, but now are much more clustered around one. Furthermore, using the linear interpolation yields the minimum eigenvalue (in magnitude), which is farther from the origin than that using the piecewise-constant interpolation.

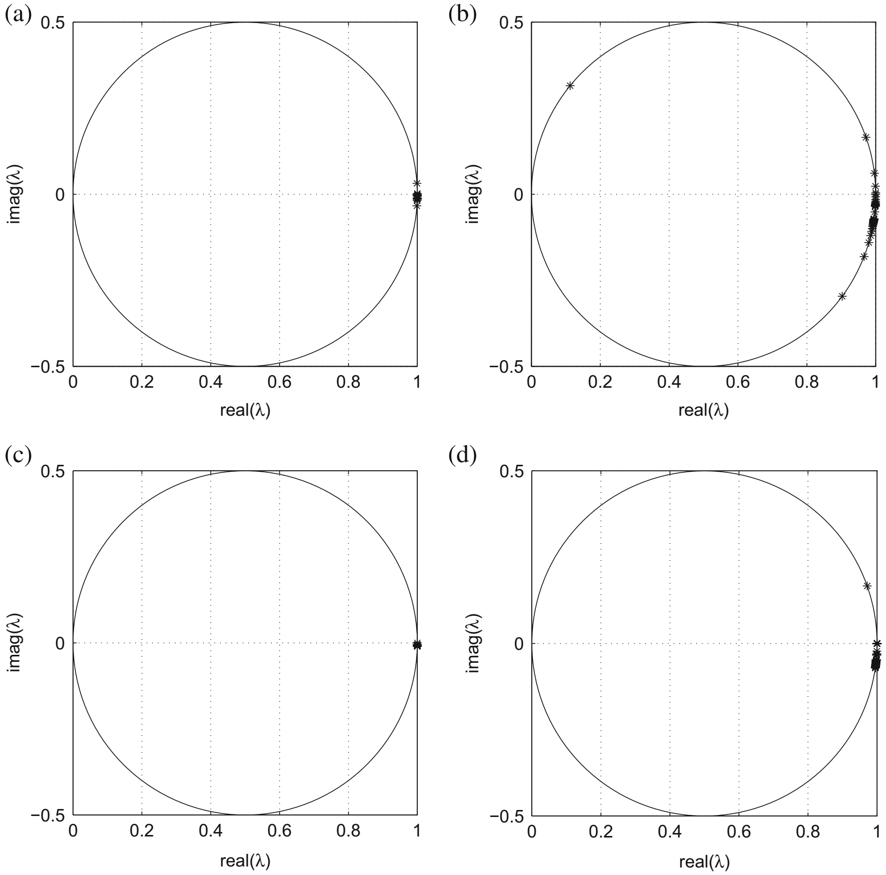


Fig. 4 Spectra of $AM^{-1}Q_{\hat{N}}$. A is associated with a 1D Helmholtz problem with a Dirichlet boundary condition, and with $k = 20, 50$. The number of grid points for each k is $n = 100$. In $Q_{\hat{N}}$, Z is of rank $r = 50$, and represents either the piecewise-constant interpolation or linear interpolation. (**a**) Piecewise-constant. $k = 20$. (**b**) Piecewise-constant. $k = 50$. (**c**) Linear interpolation. $k = 20$. (**d**) Linear interpolation. $k = 50$

Finally, we compute eigenvalues of $AM^{-1}Q_{\hat{N}}$, for the case where the Sommerfeld boundary condition is imposed, and with Z associated with the piecewise-constant and linear interpolation. The spectra are shown in Fig. 5, again for $k = 20$ and $k = 50$, which suggest a tight clustering of eigenvalues around 1. Furthermore, the eigenvalues are enclosed inside the circle.

We performed numerical experiments based on the 1D Helmholtz problem with constant wavenumber, with M and $\hat{E}$ inverted exactly (the ideal MKMG). We apply GMRES to (6) and measure the number of iterations needed to reduce the relative residual by six orders of magnitude. Convergence results are shown in Table 1, with $Z \in \mathbb{C}^{n \times r}$ based on either piecewise-constant interpolation or linear interpolation.

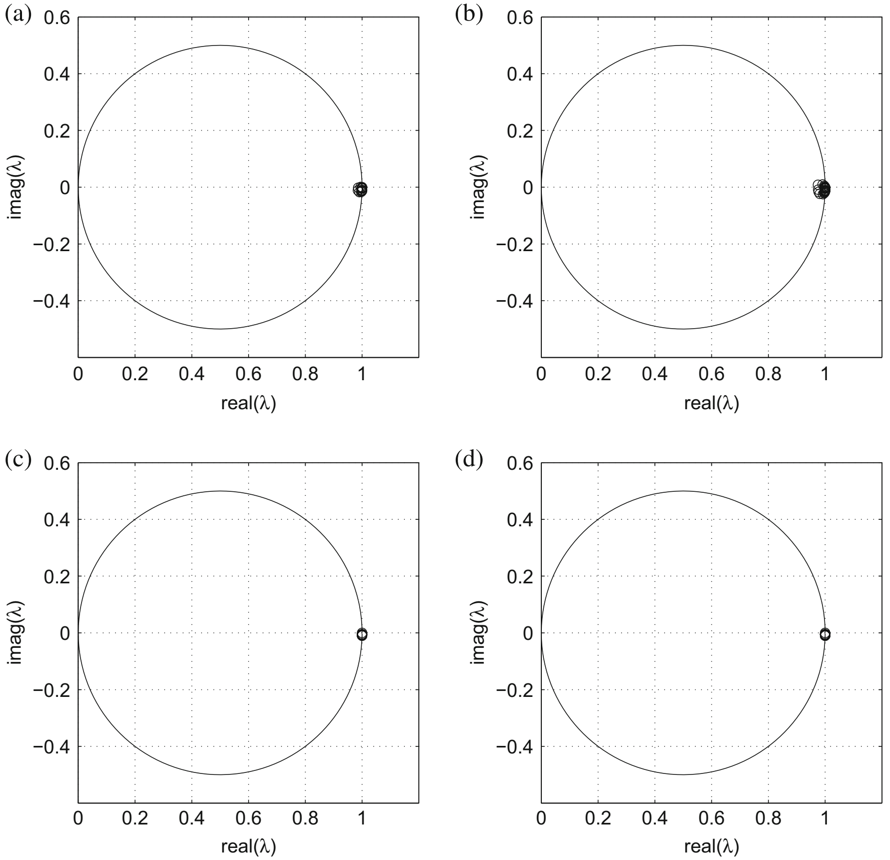


Fig. 5 Spectra of a preconditioned 1D Helmholtz problem with the Sommerfeld condition. The number of grid points for each k is $n = 100$ (for $k = 20$) and 250 (for $k = 50$), and $r = n/2$. (**a**) Piecewise-constant, $k = 20$. (**b**) Piecewise-constant, $k = 50$. (**c**) Linear interpolation, $k = 20$. (**d**) Linear interpolation, $k = 50$

In all cases, $r = n/2$, where $n = 1/h$ and h is the mesh size. The mesh size h decreases when the wavenumber k increases, so that the solutions are solved on grids equivalent to 30, 15 or 8 gridpoints per wavelength. (The use of 8 gridpoints per wavelength on the finest grid is, however, too coarse for a second-order finite-difference scheme used in this experiment, as the pollution error becomes dominant, see, e.g., [3, 5]. For a second-order scheme, the rule of thumb is to use at least 12 gridpoints per wavelength. For this reason, this is the only example where 8 gridpoints per wavelength are used.)

For the case without $Q_{\hat{N}}$, denoted by "standard", we observe convergence, which depends linearly on the wavenumber k. The convergence becomes less dependent on k if $Q_{\hat{N}}$ is incorporated. In particular, if Z is the linear interpolation matrix, the

Table 1 Number of preconditioned GMRES iterations for a 1D Helmholtz problem

	$k = 20$	$k = 50$	$k = 100$	$k = 200$	$k = 500$
Standard (no $Q_{\hat{N}}$)	14/15/15	24/25/26	39/40/42	65/68/78	142/146/157
$Q_{\hat{N}}$, piecewise-constant	4/5/7	4/6/10	5/7/14	6/10/20	7/15/37
$Q_{\hat{N}}$, linear interpolation	3/4/5	3/4/7	3/4/8	3/5/10	3/5/12

Equidistant grids equivalent to 30/15/8 gridpoints per wavelength are used, and $r = n/2$. The relative residual is reduced by six orders of magnitude

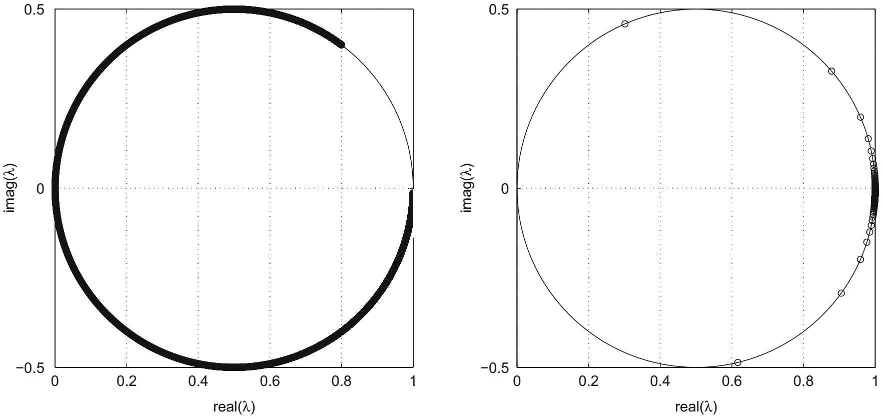


Fig. 6 Spectrum of $\hat{A}$ (*left*) and $\hat{A}Q_{\hat{N}}$ (*right*) for $\alpha = 1$, $\beta = 0.5$. $k = 10{,}000$ and $kh = 0.314$

convergence can be made almost independent of k, unless the grid is too coarse. Convergence deterioration is worse when piecewise-constant interpolation is used.

In [34] exact formulas for the eigenvalues of $AM^{-1}\hat{Q}_N$ in the 1D case are given, which explicitly indicate better clustering of eigenvalues than AM^{-1}. Although the eigenvalue formulas show that for $k \to \infty$, the eigenvalues of $AM^{-1}\hat{Q}_N$ move to zero, thus suggesting a convergence deterioration, this happens only for very high wavenumbers (see Fig. 6). Numerical results in Sect. 5 indicate, however, that for this *ideal* MKMG setting and for wavenumbers within the range of practical interests, the convergence is practically independent of k.

4 A Practical Multilevel Krylov Method for Helmholtz Systems

The *ideal* implementation of the multilevel Krylov method for the Helmholtz equation, where M and $\hat{E}$ are inverted explicitly as done in Sect. 3, is not at all practical, especially in higher dimensions (2D or 3D). Furthermore, since the inverse of M will be approximately computed by one multigrid iteration, M^{-1} is not explicitly available.

A *practical* implementation can be based on the following approximation to $\hat{E}$:

$$\begin{aligned}\hat{E} &:= Z^T \hat{A} Z = Z^T A M^{-1} Z \\ &\approx Z^T A Z (Z^T M Z)^{-1} Z^T Z = A_H M_H^{-1} B_H =: \hat{A}_H,\end{aligned} \tag{12}$$

where the products $A_H := Z^T A Z$, $M_H := Z^T M Z$, and $B_H := Z^T Z$ are the Galerkin matrices associated with A, M, and I respectively. The corresponding coarse-grid system then looks like this:

$$v_R' = A_H M_H^{-1} B_H v_R, \tag{13}$$

where the solution vector v_R is computed only approximately by using a Krylov subspace method. A fast convergence of a Krylov method for (13) can still be attained by applying a projection on (13). While the approximation (12) is effective for convergence acceleration (see Sect. 5), there exists no mathematical foundation that justifies this approach and, hence, no measure on its accuracy.

To construct a multilevel Krylov algorithm, we shall use notations which incorporate level identification. For example, for the two-level Krylov method discussed above, A, M, and Z are now denoted by $A^{(1)}$, $M^{(1)}$, and $Z^{(1,2)}$, respectively. In addition, $B^{(1)} = I$. We set a sequence of coarse-grid matrices

$$\hat{A}^{(j)} = A^{(j)} M^{(j)^{-1}} B^{(j)}, \quad j = 2, \ldots, m$$

where $A^{(j)} = Z^{(j-1,j)^T} A^{(j-1)} Z^{(j-1,j)}$, $M^{(j)} = Z^{(j-1,j)^T} M^{(j-1)} Z^{(j-1,j)}$, and $B^{(j)} = Z^{(j-1,j)^T} B^{(j-1)} Z^{(j-1,j)}$.

If the coarse-grid matrix on the coarsest level (i.e., $j = m$) is small, the associated Galerkin system

$$A^{(m)} M^{(m)^{-1}} B^{(m)} v_R^{(m)} = (v_R')^{(m)}$$

can be solved exactly. For $j = 2, \ldots, m-1$, the coarse-grid systems are solved approximately using a Krylov method. To accelerate the convergence, for each coarse-grid system, the following shift matrix

$$Q_{\hat{N}}^{(j)} = I - Z^{(j-1,j)} \hat{A}^{(j+1)^{-1}} Z^{(j-1,j)^T} \hat{A}^{(j)} + \omega^{(j)} \lambda_n^{(j)} Z^{(j-1,j)} \hat{A}^{(j)^{-1}} Z^{(j-1,j)^T},$$

is used as the (right) preconditioner, leading to the preconditioned system

$$A^{(j)} M^{(j)^{-1}} B^{(j)} Q_{\hat{N}}^{(j)} \overline{v}_R^{(j)} = (v_R')^{(j)}, \quad j = 2, \ldots, m-1.$$

At this stage, the only non-practical part of the implementation involves inversion of $M^{(j)}$. At $j = 1$, the inverse of $M^{(1)}$ is computed approximately by one multigrid iteration as is done for the standard shifted-Laplacian preconditioned Helmholtz system. Since $M^{(2)}$ is a coarse-grid approximation to $M^{(1)}$ and, in general, $M^{(j+1)}$

is a coarse-grid approximation to $M^{(j)}$, all $M^{(j)}, j = 2, \dots, m-1$ are also inverted approximately by one multigrid iteration; $M^{(m)}$ is inverted exactly due to its small size. Using $Z^{(j-1,j)}$ as the interpolation matrix in multigrid, clearly the multilevel and multigrid part now share the same coarse-grid matrices and interpolation matrices. They only need to be constructed once in the initialization phase.

Algorithm 2 summarizes the practical implementation of multilevel Krylov with multigrid iterations included, hence leading to the name "Multilevel Krylov-Multigrid" or MKMG. The Krylov method used for this implementation is

```
/* Initialization: */
Number of levels: $m$; Number of iterations on each level $it(j)$.
for $j = 1 : m$
   if $j = 1$
      set $A^{(1)} = A, M^{(1)} = M, B^{(1)} = I$
      set $\lambda_n^{(1)} = 1$ and choose $\omega^{(1)}$
   else
      construct $Z^{(j-1,j)}$
      compute $A^{(j)} = Z^{(j-1,j)^T} A^{(j-1)} Z^{(j-1,j)}$
      compute $M^{(j)} = Z^{(j-1,j)^T} M^{(j-1)} Z^{(j-1,j)}$
      compute $B^{(j)} = Z^{(j-1,j)^T} B^{(j-1)} Z^{(j-1,j)}$
      set $\lambda_n^{(j)} = 1$ and choose $\omega^{(j)}$
   end if
end for
/* Iteration phase: */
Set $j = 1$
function $x = $ MKMG($A^{(j)}, M^{(j)}, B^{(j)}, Z^{(j,j+1)}, b, it(j), \lambda_n^{(j)}, \omega^{(j)}, j$)
begin
   Set $x_0 = arbitrary$, $r_0 = b - A^{(j)}x_0$, $\beta = \|r_0\|$, $v_1 = r_0/\beta$
   for $\ell = 1, \dots, it(j)$ do
      /* Computation of $\hat{A}^{(j)} Q_{\hat{N}}^{(j)} v_\ell \equiv A^{(j)} M^{(j)^{-1}} B^{(j)} Q_{\hat{N}}^{(j)} v_\ell$ */
      /* Computation of $Q_{\hat{N}}^{(j)} v_\ell$: */
      $v_{M,\ell} = B^{(j)} v_\ell$
      /* Multigrid solve of shifted Laplacian $v_{M,\ell} = M^{(j)^{-1}} v_\ell$ */
      $v_{M,\ell} = $ MULTIGRID($M^{(j)}, v_\ell$, cycle, smooth)
      $s_\ell = A^{(j)} v_{M,\ell}$
      $t_\ell = s_\ell - \omega^{(j)} \lambda_n^{(j)} v_\ell$
      /* Restriction */
      $v'_{R,\ell} = Z^{(j,j+1)^T} t_\ell$
      $k = j + 1$
      if $k = m$
        /* Direct solve: */
        $v_{R,\ell} = B^{(m)^{-1}} M^{(m)} A^{(m)^{-1}} v'_{R,\ell}$
      else
```

```
            /* Recursive solve on coarse grid: */
            $v_{R,\ell} = \text{MKMG}(A^{(k)}, M^{(k)}, B^{(k)}, Z^{(k,k+1)}, v'_{R,\ell}, it(k), \lambda_n^{(k)}, \omega_n^{(k)}, k)$
        end if
        /* Interpolation: */
        $v_{I,\ell} = Z^{(j,j+1)} v_{R,\ell}$
        $q_\ell = v_\ell - v_{I,\ell}$
        /* End of computation of $Q_{\hat{N}}^{(j)} v^{(l)}$ */
        $p_\ell = B^{(j)} q_\ell$
        /* Multigrid solve of shifted Laplacian $g_\ell = M^{(j)^{-1}} p_\ell$ */
        $g_\ell = \text{MULTIGRID}(M^{(j)}, p_\ell, \text{cycle}, \text{smooth})$
        $w_\ell = A^{(j)} g_\ell$
        /* End of computation of $\hat{A}^{(j)} Q_{\hat{N}}^{(j)} v_\ell \equiv A^{(j)} M^{(j)^{-1}} B^{(j)} Q_{\hat{N}}^{(j)} v_\ell$ */
        /* FGMRES part */
        Compute $v_{\ell+1}$ and $H_\ell$ by orthogonalizing $w_\ell$ against $v_1, \ldots, v_\ell$
        Generate $G_\ell = [g_1, \ldots, g_\ell]$
        /* With $V_\ell = [v_1, \ldots, v_\ell]$ it holds $AG_\ell = H_\ell V_\ell$ */
        Compute $y^* = \text{argmin}_y \|\beta e_1 - H_\ell y\|_2$, and set $x_{\ell+1} = x_0 + G_\ell y^*$
        if ($j = 1$ and stopping criteria satisfied) then stop
    end for
end function
```

Algorithm 2: MKMG method

FGMRES [31] due to the non-constant preconditioners. Note that the action of $Q_{\hat{N}}^{(j)}$ on a vector v_ℓ is broken down as follows:

$$
\begin{aligned}
Q_{\hat{N}}^{(j)} v_\ell &= (I - Z^{(j-1,j)} \hat{A}^{(j+1)^{-1}} Z^{(j-1,j)^T} \hat{A}^{(j)} + \omega^{(j)} \lambda_n^{(j)} Z^{(j-1,j)} \hat{A}^{(j)^{-1}} Z^{(j-1,j)^T}) v_\ell \\
&= v_\ell - Z^{(j-1,j)} \hat{A}^{(j+1)^{-1}} Z^{(j-1,j)^T} (s_\ell - \omega^{(j)} \lambda_n^{(j)} v_\ell),
\end{aligned}
$$

where

$$
s_\ell = \hat{A}^{(j)} v_\ell = A^{(j)} M^{(j)^{-1}} B^{(j)} v_\ell.
$$

At the level j of the multilevel Krylov method, multigrid with $m - j$ levels is called to approximately invert $M^{(j)}$ with the corresponding coarse-grid matrices $M^{(j+1)}, \ldots, M^{(m)}$. Once the multilevel Krylov method reaches the level $j = m - 1$, the Galerkin problem at level $j = m$ is solved exactly.

Remark 4 In solving the Galerkin problems by a Krylov subspace method, a zero initial guess is always used. With this choice, the initial residual does not have to be computed explicitly because it is equal to the right-hand side vector of the Galerkin system. Hence, we can save one vector multiplication with $A^{(j)} M^{(j)^{-1}} B^{(j)}$.

Remark 5 At every level j, we require an estimate of $\lambda_n^{(j)}$. In our implementation, we set $\lambda_n^{(j)} = 1, j = 2, \ldots, m-1$.

5 Numerical Experiments

In this section we present numerical results for the 1D and 2D Helmholtz equation with Sommerfeld's boundary conditions. Otherwise stated, the numerical results are based on the practical MKMG. At each level $j > 1$ of MKMG, FGMRES [31] is applied to the preconditioned Galerkin system. In principle it is not necessary to use the same number of FGMRES iterations at each level. For the *practical* MKMG, we use the notation MKMG(6,2,2), for instance, to indicate that 6 FGMRES iterations are employed at level $j = 2$, 2 at level $j = 3$ and 2 at level $j = 4, \ldots, m-1$. At level $j = m$ the coarse-grid problem is solved exactly. Following [18], we employ one multigrid iteration to invert the shifted Laplacian, with an F-cycle, one pre- and postsmoothing, and Jacobi with underrelaxation ($\omega_R = 0.5$) as smoother. The coarsest level for both the multilevel Krylov and multigrid part consists of only one interior grid point. Convergence for the practical MKMG is declared if the initial relative residual at the finest level ($j = 1$) is reduced by FGMRES by six orders of magnitude.

5.1 1D Helmholtz

In this section, we use the same problem as the one in Sect. 3. Convergence results for the practical MKMG are shown in Tables 2 and 3. To obtain these convergence results, we have used the linear interpolation matrix Z in $Q_{\hat{N}}$. Results in the tables suggest convergence of MKMG which is only mildly dependent on the grid size h. Furthermore, the number of iterations to reach convergence increases only mildly with an increase in the wavenumber k. These results are worse than the ideal situation where the Galerkin system at the second level is solved exactly; cf. Table 1. The multilevel Krylov step, however, improves the convergence of the Krylov method with only shifted Laplacian preconditioner (shown in Table 2).

Table 2 Number of practical MKMG(6,2,2) iterations for 1D Helmholtz problems with constant wave number

g/w	$k = 20$	$k = 50$	$k = 100$	$k = 200$	$k = 500$
15	11 (19)	11 (29)	11 (43)	15 (66)	25 (138)
30	9 (18)	11 (28)	12 (42)	14 (68)	22 (136)
60	9 (18)	9 (28)	12 (43)	12 (68)	19 (141)

g/w stands for "# of grid points per wavelength". For comparison, the number of shifted-Laplacian preconditioned FGMRES iterations is shown in parentheses

Table 3 Number of practical MKMG(8,2,2) iterations and MKMG(8,2,1) (in parentheses) for 1D Helmholtz problems with constant wave number

g/w	$k = 20$	$k = 50$	$k = 100$	$k = 200$	$k = 500$
15	11 (11)	15 (16)	19 (18)	22 (21)	33 (33)
30	10 (10)	13 (13)	13 (13)	15 (15)	20 (20)
60	9 (9)	13 (13)	10 (12)	14 (14)	17 (18)

g/w stands for "# of grid points per wavelength"

The importance of the number of iterations at the second level in the practical MKMG can also be seen in Tables 2 and 3. While the convergence for MKMG(8,2,2) is slightly better than MKMG(6,2,2), this convergence is not, however, better than MKMG(8,2,1). In general, one FGMRES iteration at level $j \geq 4$ is sufficient for fast convergence, if enough iterations are spent on the second and third level. We note here that, for all cases, the ℓ_2 norms of the error at convergence fall below 10^{-5}.

5.2 2D Helmholtz

We consider 2D Helmholtz problems in a square domain with constant wavenumbers. At the boundaries, the first-order approximation to the Sommerfeld (non-reflecting) condition due to Engquist and Majda [11] is imposed. We consider problems where a source is generated in the middle of the domain.

Following the 1D case, the deflation subspace Z is chosen to be the same as the interpolation matrix in multigrid. For 2D cases, however, care should be taken in constructing the interpolation matrix Z. Consider a set of fine grid points defined by

$$\Omega_h := \{(x,y)\,|\,x = x_{i_x} = i_x h,\ y = y_{i_y} = i_y h,\ i_x = 1,\dots,N_{x,h},\ iy = 1,\dots,N_{y,h}\},$$

associated with the grid points on level $j = 1$. The set of grid points Ω_H corresponding to the coarse-grid level $j = 2$ is determined as follows. We assume that $(x_1, y_1) \in \Omega_H$ coincides with $(x_1, y_1) \in \Omega_h$, as illustrated in Fig. 7 (left). Starting from this point, the complete set of coarse-grid points is then selected according to the standard multigrid coarsening, i.e., by doubling the mesh size. This results in the coarse grid, with $H = 2h$,

$$\begin{aligned}\Omega_H := \ &\{(x,y)\,|\,x = x_{i_x} = (2i_x - 1)h,\ y = y_{i_y} = (2i_y - 1)h,\\ &i_x = 1,\dots,N_{x,H},\ i_y = 1,\dots,N_{y,H}\}.\end{aligned}$$

As shown in [18], this coarsening strategy leads to a good multigrid method for the shifted Laplacian preconditioner. Moreover, from a multilevel Krylov method point of view, this coarsening strategy results in larger projection subspaces than if, e.g., $(x_1, y_1) \in \Omega_H$ coincides with $(x_2, y_2) \in \Omega_h$.

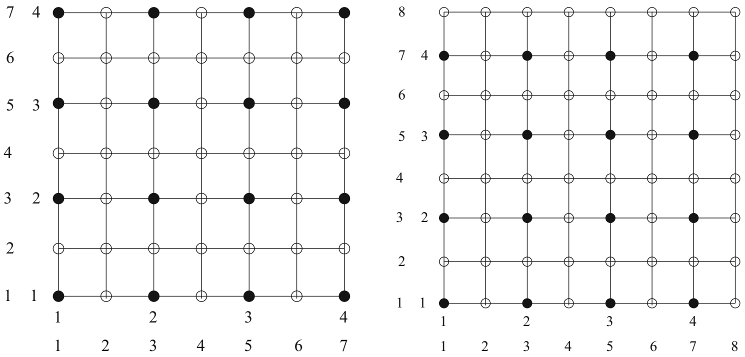


Fig. 7 Fine (*white circles*) and coarse (*black circles*) grid selections in 2D multigrid. *Black circles* also coincide with the fine grids. The *right figure* illustrates coarsening where the last indexed gridpoints do not coincide

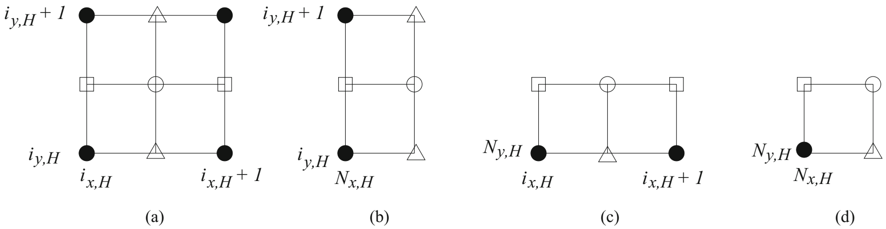


Fig. 8 Fine (*white circle*) and course (*black circle*) grid selection indicating the bilinear interpolation in 2D multigrid for (**a**) interior points, (**b**) points with fine grid points on the right boundary, (**c**) points with fine grid points on the upper boundary, and (**d**) points with fine grid points at the right top corner. *Black circles* coincide with the fine-grid points

Having defined the coarse-grid points according to Fig. 7 (left), the deflation vectors are determined by using the bilinear interpolation process of coarse-grid value into the fine grid as follows [38], for level 2 to level 1 (see Fig. 8a for the meaning of the symbols):

$$I_H^h v^{(1)}(x,y) = \begin{cases} v^{(2)}(x,y), & \text{for } \bullet, \\ \frac{1}{2}[v^{(2)}(x,y-h) + v^{(2)}(x,y+h)], & \text{for } \square, \\ \frac{1}{2}[v^{(2)}(x-h,y) + v^{(2)}(x+h,y)], & \text{for } \triangle, \\ \frac{1}{4}[v^{(2)}(x-h,y-h) + v^{(2)}(x-h,y+h & \\ \quad +v^{(2)}(x+h,y-h) + v^{(2)}(x+h,y+h)], & \text{for } \circ \end{cases}$$

In some cases, however, such a coarsening may result in the last-indexed coarse-grid points which do not coincide with the last-indexed fine-grid points. This is illustrated in Fig. 7 (right). There are three possible situations for such coarse-grid points, which are summarized in Fig. 8b–d. The interpolation associated with $(N_{x,h}h, jh)$, $(ih, N_{y,h}h)$, $(N_{x,h}h, N_{y,h}h) \in \Omega_h$ are given as follows:

- For fine-grid points $(x = N_{x,h}h, y = i_y h)$ (Fig. 8b)

$$I_H^h v^{(1)}(x,y) = \begin{cases} v^{(2)}(x,y), & \text{for } \bullet, \\ \frac{1}{2}[v^{(2)}(x,y-h) + v^{(2)}(x,y+h)], & \text{for } \square, \\ v^{(2)}(x-h,y), & \text{for } \triangle, \\ \frac{1}{2}[v^{(2)}(x-h,y-h) + v^{(2)}(x-h,y+h)], & \text{for } \circ\,. \end{cases}$$

- For fine-grid points $(x = i_x h, y = N_{y,h}h)$ (Fig. 8c)

$$I_H^h v^{(1)}(x,y) = \begin{cases} v^{(2)}(x,y), & \text{for } \bullet, \\ v^{(2)}(x,y-h), & \text{for } \square, \\ \frac{1}{2}[v^{(2)}(x-h,y) + v^{(2)}(x+h,y)], & \text{for } \triangle, \\ \frac{1}{2}[v^{(2)}(x-h,y-h) + v^{(2)}(x+h,y-h)], & \text{for } \circ\,. \end{cases}$$

- For fine-grid points $(x = N_{x,h}h, y = N_{y,h}h)$ (Fig. 8d)

$$I_H^h v^{(1)}(x,y) = \begin{cases} v^{(2)}(x,y), & \text{for } \bullet, \\ v^{(2)}(x,y-h), & \text{for } \square, \\ v^{(2)}(x-h,y), & \text{for } \triangle, \\ v^{(2)}(x-h,y-h), & \text{for } \circ\,. \end{cases}$$

Based on the interpolation matrix I_H^h, we set $Z^{(1,2)} = I_H^h$, and similarly for the coarser grid levels.

For benchmarking, we first show convergence results for the ideal MKMG, where the preconditioner $M^{(1)}$ is inverted exactly, and the coarse-grid matrix $A^{(2)} = Z^{(1,2)^T} A^{(1)} M^{(1)^{-1}} Z^{(1,2)}$ is computed explicitly and then inverted exactly (see Table 4). Note that due to excessive memory requirement for explicitly forming and storing $A^{(2)}$, we could not run the ideal MKMG for problems that require very fine grid (about 150^2) (indicated by "–" in Table 4). The results suggest, however, that the convergence is independent of h and k.

Possibly the closest practical MKMG to the ideal MKMG method is the one that uses two-level MK and an exact inversion of the preconditioner $M^{(1)}$. Convergence for this practical two-level MKMG is compared with the ideal ones in Table 4 (in parentheses). Even though slower than the ideal one, we can still attain a practically (h,k)-independent convergence, provided that the h is sufficiently small (which is often times needed for an accurate solution).

Table 4 Number of ideal MKMG iterations for 2D Helmholtz problems with constant wave number

g/w	$k = 20$	$k = 30$	$k = 40$	$k = 50$
15	4 (7)	4 (8)	4 (9)	4 (11)
20	4 (6)	4 (7)	4 (7)	– (8)
30	4 (6)	3 (6)	– (6)	– (7)

g/w stands for "# of grid points per wavelength", and "–" means "not computable". In parentheses are the number of iterations of two-level MK with exact inversion of the shifted Laplacian preconditioner

Table 5 Number of two-level MKMG iterations (two-level MK and two-level MG)

	k						
g/w	20	40	60	80	100	120	200
15	13 (9)	17 (12)	19 (15)	23 (14)	28 (21)	33 (26)	61 (48)
20	14 (10)	16 (11)	17 (12)	19 (13)	21 (15)	23 (17)	33 (25)
30	17 (12)	18 (13)	19 (14)	19 (14)	20 (14)	20 (14)	23 (16)

g/w stands for "# of grid points per wavelength". Performance of Jacobi smoother is compared with Gauss-Seidel (in parentheses)

Table 6 Number of practical MKMG(5,2,1) iterations for 2D Helmholtz problems with constant wave number

	k							
g/w	20	40	60	80	100	120	200	300
15	11	14	15	18	19	21	31	52
20	12	13	15	15	16	18	25	37
30	11	12	12	13	13	14	18	28

g/w stands for "# of grid points per wavelength"

Table 5 shows convergence results for two-level MKMG, with two-level MG steps. In this case, the method has two sources of inaccuracy: from the coarse-grid approximation and from an approximate inversion of the shifted-Laplacian preconditioner, expecting a slower convergence as compared to, e.g., the ideal MKMG is Table 4. The convergence is however only mildly dependent on k if h is taken sufficiently small, except for the low wavenumber, where the number of iterations to converge seems to behave irregularly in k. This convergence, based on Jacobi smoother in the MG steps, can be improved by using, for instance, one pre- and post-Gauss-Seidel smoothing; see in the same table between parentheses.

Finally, convergence results of the practical MKMG(5,2,1), MKMG(6,2,1), and MKMG(8,2,1) are shown in Tables 6, 7 and 8 for various wavenumbers. From these tables, for low grid resolutions (e.g., 15 grid points per wavelength) we observe convergence of MKMG, which tends to increase rapidly in terms of numbers of iterations, with increasing k. As observed in the previous tables, the convergence becomes less dependent on k if the grid size h is taken sufficiently small (e.g., 30 gridpoints per wavelength). Out of these three practical MKMG settings, MKMG(8,2,1) performs best, indicating the importance of an accurate solve of the second-level problem.

Table 7 Number of practical MKMG(6,2,1) iterations for 2D Helmholtz problems with constant wave number

	k							
g/w	20	40	60	80	100	120	200	300
15	11	14	14	18	18	20	28	47
20	12	13	15	15	16	17	25	36
30	11	12	12	13	13	14	16	25

g/w stands for "# of grid points per wavelength"

Table 8 Number of practical MKMG(8,2,1) iterations for 2D Helmholtz problems with constant wave number

	k							
g/w	20	40	60	80	100	120	200	300
15	11	14	14	17	18	21	27	39
20	12	13	15	14	15	16	20	28
30	11	12	12	12	13	14	15	19

g/w stands for "# grid points per wavelength"

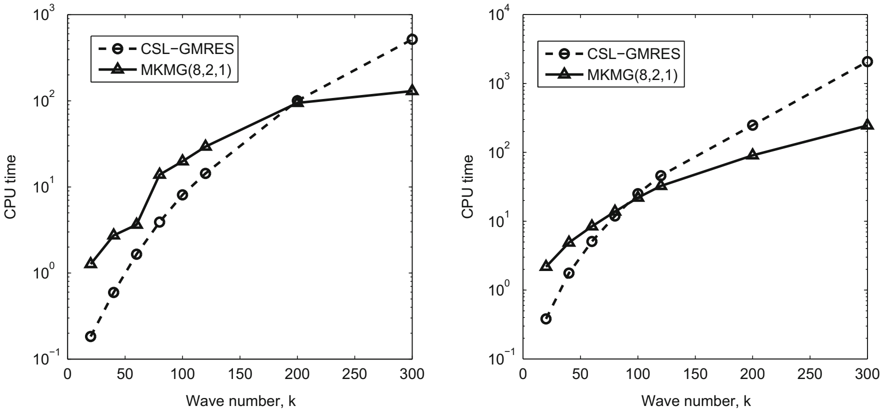


Fig. 9 CPU time for the iteration phase of MKMG(8,2,1) and CSL-preconditioned GMRES, with 15 (*left*) and 30 (*right*) gridpoints per wavelength

To gain insight into the total arithmetic operations needed by MKMG, in Fig. 9, we compare CPU time spent in the iteration phase of MKMG(8,2,1) and GMRES preconditioned by complex shifted Laplacian (CSL-GMRES). The elapsed time was measured on a LINUX maxchine with an Intel Xeon Processor for the iteration phase with the MATLAB command `tic ... toc`. We note that since the `for` loop is used heavily in the construction of the matrix, the measured overall time is too pessimistic, and thus is not presented. From Fig. 9 we observe that, for low wavenumbers, CSL-preconditioned GMRES is faster in CPU time than MKMG(8,2,1), but is outperformed by MKMG(8,2,1) when the wavenumber becomes sufficiently large.

6 Conclusions and Outlook

We have discussed a new multilevel method for solving the Helmholtz equation. The method is based on the multilevel Krylov framework. This framework consists of several ingredients which are a flexible Krylov method, a preconditioner, a shift operator which includes a restriction and prolongation and a subspace system. In contrast to multigrid methods the subspace system is solved by a few steps of a flexible Krylov method preconditioned recursively by the MK method.

The MK method is applied to the Helmholtz system preconditioned by the shifted-Laplacian preconditioner. With the latter inverted approximately by one multigrid iteration, this new method, called MKMG, is purely iterative and requires only matrix-vector multiplications. With the two methods combined, it is expected that issues related to indefiniteness and small eigenvalues can be resolved simultaneously.

Numerical results show that the method converges to the solution at a rate, which is mildly dependent on the grid size h and wavenumber k. Based on the convergence results of the ideal MKMG, we argue that it is possible to attain an almost (h,k)-independent convergence, provided that the preconditioner is inverted accurately and the coarse-grid matrix is approximated well. The construction of the coarse-grid approximation used in this chapter was mainly driven by practicality rather than accuracy. Numerical experiments nevertheless suggest that, so long as the preconditioner is inverted very accurately, an (h,k)-independent convergence can still be achieved. Accurate approximation of inverse of the preconditioner and the coarse-grid approximation are open problems and subject to future's research.

Some other open problems include spectral analysis within the ideal MKMG framework, for Helmholtz problems with PML boundary conditions. Spectral analysis and computations have suggested that better clustering is attained with radiation conditions than the Dirichlet conditions (considered as the worst case), and depending on the choice of the deflation matrix Z the eigenvalues are enclosed by a circle whose radius is smaller than 0.5.

Since the good convergence rate in MKMG is determined by an efficient interplay between MK an MG, ideally convergence analysis and spectral analysis should take into account the two components of MKMG. Given the complicated nature of the MK part, which in this case uses GMRES, with available convergence bounds not immediately useful, a quantitative analysis similar to LFA for multigrid that can guide into a proper choice of MKMG components seems to be still far away.

References

1. S. Abarbanel and D. Gottlieb. A mathematical analysis of the PML method. *J. Comput. Phys.*, 134:357–363, 1997.
2. S. Abarbanel and D. Gottlieb. On the construction and analysis of absorbing layers in CEM. *Appl. Numer. Math.*, 27:331–340, 1998.

3. I. Babuska, F. Ihlenburg, T. S. Trouboulis, and S. K. Gangaraj. A posteriori error estimation for finite element solutions of Helmholtz's equation. Part II: Estimation of the pollution error. *Internat. J. Numer. Methods Engrg.*, 40:3883–3900, 1997.
4. A. Bayliss, C. I. Goldstein, and E. Turkel. An iterative method for Helmholtz equation. *J. Comput. Phys.*, 49:443–457, 1983.
5. A. Bayliss, C. I. Goldstein, and E. Turkel. On accuracy conditions for the numerical computation of waves. *J. Comput. Phys.*, 59:396–404, 1985.
6. J. P. Berenger. A perfectly matched layer for the absorption of electromagnetic waves. *J. Comput. Phys.*, 114:185–200, 1994.
7. J. P. Berenger. Three-dimensional perfectly matched layer for the absorption of electromagnetic waves. *J. Comput. Phys.*, 127:363–379, 1996.
8. A. Brandt and I. Livshits. Wave-ray multigrid method for standing wave equations. *Electronic Transactions on Numerical Analysis*, 6:161–181, 1997.
9. M. Eiermann, O. G. Ernst, and O. Schneider. Analysis of acceleration strategies for restarted minimal residual methods. *J. Comput. Appl. Math.*, 123:261–292, 2000.
10. H. C. Elman, O. G. Ernst, and D. P. O'Leary. A multigrid method enhanced by Krylov subspace iteration for discrete Helmholtz equations. *SIAM J. Sci. Comput.*, 22:1291–1315, 2001.
11. B. Engquist and A. Majda. Absorbing boundary conditions for the numerical simulation of waves. *Math. Comp.*, 31:629–651, 1977.
12. B. Engquist and L. Ying. Sweeping preconditioner for the Helmholtz equation: moving perfectly matched layers. *Multiscale Model. Simul.*, 9:686–710, 2011.
13. Y. A. Erlangga and R. Nabben. Deflation and balancing preconditioners for Krylov subspace methods applied to nonsymmetric matrices. *SIAM J. Matrix Anal. Appl.*, 30:684–699, 2008.
14. Y. A. Erlangga and R. Nabben. Multilevel projection-based nested Krylov iteration for boundary value problems. *SIAM J. Sci. Comput.*, 30:1572–1595, 2008.
15. Y. A. Erlangga and R. Nabben. On a multilevel Krylov method for the Helmholtz equation preconditioned by shifted Laplacian. *E. Trans. Numer. Anal.*, 31:403–424, 2008.
16. Y. A. Erlangga and R. Nabben. Algebraic multilevel Krylov methods. *SIAM J. Sci. Comput.*, 31:3417–3437, 2009.
17. Y. A. Erlangga and R. Nabben. On the convergence of two-level Krylov methods for singular symmetric systems. 2015.
18. Y. A. Erlangga, C. W. Oosterlee, and C. Vuik. A novel multigrid-based preconditioner for the heterogeneous Helmholtz equation. *SIAM J. Sci. Comput.*, 27:1471–1492, 2006.
19. Y. A. Erlangga, C. Vuik, and C. W. Oosterlee. On a class of preconditioners for solving the Helmholtz equation. *Appl. Numer. Math.*, 50:409–425, 2004.
20. Y. A. Erlangga, C. Vuik, and C. W. Oosterlee. Comparison of multigrid and incomplete LU shifted-Laplace preconditioners for the inhomogeneous Helmholtz equation. *Appl. Numer. Math.*, 56:648–666, 2006.
21. J. Frank and C. Vuik. On the construction of deflation-based preconditioners. *SIAM J. Sci. Comput.*, 23:442–462, 2001.
22. M. J. Gander, I. G. Graham, and E. A. Spence. Applying GMRES to the Helmholtz equation with shifted Laplacian preconditioning: what is the largest shift for which wavenumber-independent convergence is guaranteed? *Numerische Mathematik*, 131 (3):567–614, 2015.
23. L. García Ramos and R. Nabben. On the spectrum of deflated matrices with applications to the deflated shifted Laplace preconditioner for the Helmholtz equation. 2016.
24. A. L. Laird and M. B. Giles. Preconditioned iterative solution of the 2D Helmholtz equation. Technical Report NA 02-12, Comp. Lab., Oxford Univ., 2002.
25. I. Livshits. A scalable multigrid method for solving indefinite Helmholtz equations with constant wave numbers. *Numerical Linear Algebra with Applications*, 21:177–193, 2014.
26. R. B. Morgan. A restarted GMRES method augmented with eigenvectors. *SIAM J. Matrix Anal. Appl.*, 16:1154–1171, 1995.
27. R. Nabben and C. Vuik. A comparison of deflation and the balancing preconditioner. *SIAM J. Sci. Comput.*, 27:1742–1759, 2006.

28. R. A. Nicolaides. Deflation of conjugate gradients with applications to boundary value problems. *SIAM J. Numer. Anal.*, 24:355–365, 1987.
29. C. W. Oosterlee. A GMRES-based plane smoother in multigrid to solve 3D anisotropic fluid flow problems. *J. Comput. Phys.*, 130:41–53, 1997.
30. J. Poulson, B. Engquist, S. Li, and L. Ying. A parallel sweeping preconditioner for heterogeneous 3D Helmholtz equations. *SIAM J. Sci. Comput.*, 35(3):C194–C212, 2013.
31. Y. Saad. A flexible inner-outer preconditioned GMRES algorithm. *SIAM J. Sci. Comput.*, 14:461–469, 1993.
32. Y. Saad. *Iterative Methods for Sparse Linear Systems.* SIAM, Philadelphia, 2003.
33. Y. Saad and M. H. Schultz. GMRES: a generalized minimal residual algorithm for solving nonsymmetric linear systems. *SIAM J. Sci. Stat. Comput.*, 7:856–869, 1986.
34. A. H. Sheikh, D. Lahaye, L. García Ramos, R. Nabben, and C. Vuik. Accelerating the Shifted Laplace Preconditioner for the Helmholtz Equation by Multilevel Deflation. *J. Comput. Phys.*, 322:473–490, 2016.
35. A. H. Sheikh, D. Lahaye, and C. Vuik. On the convergence of shifted Laplace preconditioner combined with multilevel deflation. *Numerical Linear Algebra with Applications*, 20(4):645–662, Apr 2013.
36. C. C. Stolk. A rapidly converging domain decomposition method for the Helmholtz equation. *J. Comput. Phys.*, 241:240–252, 2013.
37. J. Tang, R. Nabben, C. Vuik, and Y. A. Erlangga. Theoretical and numerical comparison of projection methods derived from deflation. *J. Sci. Comput.*, 39:340–370, 2009.
38. U. Trottenberg, C. Oosterlee, and A. Schüller. *Multigrid.* Academic Press, New York, 2001.
39. S. Tsynkov and E. Turkel. A cartesian perfectly matched layer for the Helmholtz equation. In L. Tourette and L. Harpern, editors, *Absorbing Boundaries and Layers, Domain Decomposition Methods Applications to Large Scale Computation*, pages 279–309. Springer, Berlin, 2001.
40. M. B. van Gijzen, Y. A. Erlangga, and C. Vuik. Spectral analysis of the discrete Helmholtz operator preconditioned with a shifted Laplacian. *SIAM J. Sci. Comput.*, 29:1942–1952, 2006.

A Geometric Multigrid Preconditioner for the Solution of the Helmholtz Equation in Three-Dimensional Heterogeneous Media on Massively Parallel Computers

H. Calandra, S. Gratton, and X. Vasseur

Abstract We consider the numerical simulation of acoustic wave propagation in three-dimensional heterogeneous media as occurring in seismic exploration. We focus on forward Helmholtz problems written in the frequency domain, since this setting is known to be particularly challenging for modern iterative methods. The geometric multigrid preconditioner proposed by Calandra et al. (Numer Linear Algebra Appl 20:663–688, 2013) is considered for the approximate solution of the Helmholtz equation at high frequencies in combination with dispersion minimizing finite difference methods. We present both a strong scalability study and a complexity analysis performed on a massively parallel distributed memory computer. Numerical results demonstrate the usefulness of the algorithm on a realistic three-dimensional application at high frequency.

1 Introduction

The efficient computation of wave propagation phenomena in three-dimensional heterogeneous media is of significant research interest in many environmental inverse problems [50, 58]. The core of these large scale nonlinear optimization

H. Calandra
TOTAL E&P Research and Technology USA, llc 1201 Louisiana, Suite 1800, Houston, TX 77002-4397, USA
e-mail: henri.calandra@total.com

S. Gratton
INPT-IRIT, University of Toulouse and ENSEEIHT, 2 Rue Camichel, BP 7122, F-31071 Toulouse Cedex 7, France
e-mail: serge.gratton@enseeiht.fr

X. Vasseur (✉)
ISAE-SUPAERO, 10 Avenue Edouard Belin, BP 54032, F-31055 Toulouse Cedex 4, France
e-mail: xavier.vasseur@isae.fr

Part of this work was performed while this author was a permanent researcher at CERFACS, Toulouse, France.

D. Lahaye et al. (eds.), *Modern Solvers for Helmholtz Problems*, Geosystems Mathematics, DOI 10.1007/978-3-319-28832-1_6

problems usually consists of the approximate solution of a linear system issued from the discretization of a Helmholtz scalar wave equation, typically written in the frequency domain. Hence, as discussed in this book, the design of efficient direct or iterative solvers for the resulting large scale linear systems is of paramount importance. In particular, efficient domain decomposition methods [17, 18, 25, 35, 47, 48] or multigrid methods [6, 7, 15, 16, 19–23, 26, 33, 37, 38, 42, 54–56] have been proposed in the past few years in this context; we also refer the reader to, e.g., [24, 52, Sect. 11.5] and references therein for comprehensive surveys.

In this chapter, we focus on the parallel performance of a geometric multigrid preconditioner for the solution of wave propagation problems related to acoustic imaging in seismic exploration. For such a purpose, we consider the geometric two-grid preconditioner proposed in [11] for the numerical solution of Helmholtz problems in three-dimensional heterogeneous media. This two-grid cycle is directly applied to the original Helmholtz operator and relies on an approximate coarse grid solution. A second multigrid method applied to a complex shifted Laplace operator is then used as a preconditioner when solving the coarse grid system to obtain an approximate coarse solution. In this chapter, we consider this preconditioner in relation with high order dispersion minimizing finite difference schemes to tackle propagation problems at relatively high frequencies. In particular, as main contributions, we investigate the strong scaling properties of the numerical method in a massively parallel setting and provide a complexity analysis related to a realistic test case in geophysics.

The chapter is organized as follows. We introduce both the continuous and discrete Helmholtz problems in Sect. 2. In Sect. 3, we describe the geometric multigrid preconditioner that is considered throughout the chapter. Numerical experiments performed in a massively parallel environment are reported in Sect. 4. Finally, conclusions are drawn in Sect. 5.

2 Problem Setting

We specify the continuous and discrete versions of the heterogeneous Helmholtz problem that we consider throughout this chapter.

2.1 Mathematical Formulation at the Continuous Level

Given a three-dimensional physical domain Ω_p of parallelepiped shape, the propagation of a wavefield in a heterogeneous medium can be modeled by the Helmholtz equation written in the frequency domain [50]

$$-\sum_{i=1}^{3} \frac{\partial^2 u}{\partial x_i^2} - \frac{(2\pi f)^2}{c^2} u = \delta(\mathbf{x} - \mathbf{s}), \quad \mathbf{x} = (x_1, x_2, x_3) \in \Omega_p. \tag{1}$$

In Eq. (1), the unknown u represents the pressure wavefield in the frequency domain, c the acoustic-wave velocity in $\mathrm{m\,s^{-1}}$, which varies with position, and f the frequency in Hertz. The source term $\delta(\mathbf{x}-\mathbf{s})$ represents a harmonic point source located at $\mathbf{s}=(s_1,s_2,s_3)\in\Omega_p$. The wavelength λ is defined as $\lambda=c/f$ and the wavenumber as $2\pi f/c$. A popular approach—the Perfectly Matched Layer formulation (PML) [4, 5]—has been used in order to obtain a satisfactory near boundary solution, without many artificial reflections. Artificial boundary layers are then added around the physical domain to absorb outgoing waves at any incidence angle as shown in [4]. We denote by Ω_{PML} the surrounding domain created by these artificial layers. This formulation leads to the following set of coupled partial differential equations with homogeneous Dirichlet boundary conditions imposed on Γ, the boundary of the domain

$$-\sum_{i=1}^{3}\frac{\partial^2 u}{\partial x_i^2}-\frac{(2\pi f)^2}{c^2}u=\delta(\mathbf{x}-\mathbf{s}) \quad \text{in} \quad \Omega_p, \tag{2}$$

$$-\sum_{i=1}^{3}\frac{1}{\xi_{x_i}(x_i)}\frac{\partial}{\partial x_i}\left(\frac{1}{\xi_{x_i}(x_i)}\frac{\partial u}{\partial x_i}\right)-\frac{(2\pi f)^2}{c^2}u=0 \quad \text{in} \quad \Omega_{PML}\backslash\Gamma, \tag{3}$$

$$u=0 \quad \text{on} \quad \Gamma, \tag{4}$$

where the one-dimensional ξ_{x_i} function represents the complex-valued damping function of the PML formulation in the i-th direction, selected as in [34]. The set of equations ((2)–(4)) defines the forward problem related to acoustic imaging in geophysics that will be considered in this chapter. We note that the proposed numerical method can be applied to other application fields, where wave propagation phenomena appear as well.

We also introduce the complex shifted Laplace operator defined as

$$-\sum_{i=1}^{3}\frac{\partial^2 u}{\partial x_i^2}-(1+i\beta)\frac{(2\pi f)^2}{c^2}u=\delta(\mathbf{x}-\mathbf{s}) \quad \text{in} \quad \Omega_p, \tag{5}$$

$$-\sum_{i=1}^{3}\frac{1}{\xi_{x_i}(x_i)}\frac{\partial}{\partial x_i}\left(\frac{1}{\xi_{x_i}(x_i)}\frac{\partial u}{\partial x_i}\right)-(1+i\beta)\frac{(2\pi f)^2}{c^2}u=0 \quad \text{in} \quad \Omega_{PML}\backslash\Gamma, \tag{6}$$

$$u=0 \quad \text{on} \quad \Gamma, \tag{7}$$

where the parameter $1+i\beta\in\mathbb{C}$ is called the complex shift.[1] This operator will play a significant role later in the definition of our multigrid preconditioner in Sect. 3.

[1] In [23] the authors have introduced the complex shifted Laplace with a negative imaginary part for the shift in the case of first- or second-order radiation boundary conditions. Due to the PML formulation considered in this paper, we have used a shift with positive imaginary part to derive an efficient preconditioner as explained in [36, Sect. 3.3.2].

2.2 Mathematical Formulation at the Discrete Level

2.2.1 Dispersion Minimizing Finite Difference Scheme

As frequently used in the geophysics community, we have considered a finite difference discretization of the Helmholtz problem ((2)–(4)) on a uniform equidistant Cartesian grid of size $n_x \times n_y \times n_z$. We denote later by h the corresponding mesh grid size, Ω_h the discrete computational domain and n_{PML} the number of points in each PML layer.

Since the standard second-order finite difference scheme is often found to be too dispersive [3, 13, 29, 47], we have considered dispersion minimizing finite difference schemes. These schemes are especially recommended when targeting the solution of heterogeneous Helmholtz problems at high frequency, since they provide a pollution-free solution [12, 34, 47, 51]. In the context of multilevel algorithms, these schemes are also relevant for the discretization of the coarse grid operator in order to provide the same dispersion level on both the coarse and fine scales [47]. This feature has also been found beneficial by several authors, see, e.g., [12, 47, 54]. Hereafter, we have considered the compact finite difference scheme proposed by Harari and Turkel [29] based on Padé approximations, which leads to a finite difference discretization with a 27 points stencil. This scheme is formally third-order accurate on general Cartesian grids and fourth-order accurate on uniform grids. Following [3], given reference values for both the frequency f_{ref} and the step size h_{ref} and denoting by q the discretization order of the finite difference scheme, we have used the following relation to determine the step size h, given a certain frequency f,

$$h^q f^{q+1} = h_{ref}^q f_{ref}^{q+1}. \tag{8}$$

2.2.2 Properties of the Discrete Linear System

The discretization of the forward problem ((2)–(4)) with the dispersion minimizing finite difference scheme leads to the following linear system $A_h x_h = b_h$, where $A_h \in \mathbb{C}^{n_h \times n_h}$ is a sparse complex matrix which is non Hermitian and non symmetric due to the PML formulation [5, 36, 45] and where $x_h, b_h \in \mathbb{C}^{n_h}$ represent the discrete frequency-domain pressure field and source, respectively. In addition, the right-hand side is usually very sparse. The condition (8) imposes to solve large systems of equations at the (usually high) frequencies of interest for the geophysicists, a task that may be too memory expensive for standard [45, 46] or advanced sparse direct methods exploiting hierarchically semi-separable structure [59, 60] on a reasonable number of cores of a parallel computer. Consequently, preconditioned Krylov subspace methods are most often considered and efficient preconditioners must be developed for such indefinite problems. We describe next in detail a multigrid preconditioner that has been proposed in [11] for the solution of the forward problem related to acoustic imaging.

3 A Geometric Multigrid Preconditioner

We describe the geometric two-grid preconditioner proposed in [11] and detail its salient properties. We first introduce notation related to multigrid methods to make easier the description of the multilevel algorithm. We conclude this section by briefly commenting the related parallel implementation.

3.1 Notation

The fine and coarse levels denoted by h and H are associated with discrete grids Ω_h and Ω_H, respectively. Due to the application in seismic exploration where structured grids are routinely used, a geometric construction of the coarse grid Ω_H is used. The discrete coarse grid domain Ω_H is then deduced from the discrete fine grid domain Ω_h by doubling the mesh size in each direction as classically done in vertex-centered geometric multigrid [49]. In the following, we assume that A_H represents a suitable approximation of the fine grid operator A_h on Ω_H. We also introduce $I_h^H : \mathscr{G}(\Omega_h) \to \mathscr{G}(\Omega_H)$ a restriction operator, where $\mathscr{G}(\Omega_k)$ denotes the set of grid functions defined on Ω_k. Similarly $I_H^h : \mathscr{G}(\Omega_H) \to \mathscr{G}(\Omega_h)$ will represent a given prolongation operator. More precisely, we select as a prolongation operator trilinear interpolation and as a restriction its adjoint which is often called the full weighting operator [28, 49]. We refer the reader to [53, Sect. 2.9] for a complete description of these operators in three dimensions.

3.2 Algorithm Overview

A two-grid preconditioner for the numerical solution of Helmholtz problems in three-dimensional heterogeneous media has been proposed in [11] in relation with second order finite difference discretization schemes. This two-grid cycle is directly applied to the original Helmholtz operator and relies on an approximate coarse grid solution. As shown in [36], the main difficulty is to find efficient approximate solution methods for the coarse level system $A_H z_H = v_H$. In this chapter, as in [11], we consider a preconditioning operator (the complex shifted Laplace operator) based on a different partial differential equation for which an efficient multilevel solution is possible. A second multigrid method applied to a complex shifted Laplace operator is then used as a preconditioner for the approximate solution of this coarse problem.

This combination of two cycles defined on two different hierarchies is detailed next. First, a two-grid cycle using Ω_h and Ω_H only (as fine and coarse levels, respectively) is applied to the original Helmholtz operator ((2)–(4)). A second sequence of grids $\Omega_k (k = 1, \cdots, l)$ with the finest grid Ω_l defined as $\Omega_l := \Omega_H$

Algorithm 1: Cycle applied to $A_h z_h = v_h$. $z_h = \mathscr{T}_{l,C}(v_h)$

```
Polynomial pre-smoothing: Apply $\vartheta$ cycles of GMRES($m_s$) to $A_h z_h = v_h$ with $\nu$ iterations of $\omega_h$-Jacobi as a right preconditioner to obtain the approximation $z_h^{\vartheta}$.
Restrict the fine level residual: $v_H = I_h^H(v_h - A_h z_h^{\vartheta})$.
Solve approximately the coarse problem $A_H z_H = v_H$ with initial approximation $z_H^0 = 0_H$:
    Apply $\vartheta_c$ cycles of FGMRES($m_c$) to $A_H z_H = v_H$ preconditioned by a cycle of multigrid applied to $S_l^{(\beta)} y_l = w_l$ on $\Omega_l \equiv \Omega_H$ to obtain the approximation $z_H$.
Perform the coarse level correction: $\widetilde{z_h} = z_h^{\vartheta} + I_H^h z_H$.
Polynomial post-smoothing: Apply $\vartheta$ cycles of GMRES($m_s$) to $A_h z_h = v_h$ with initial approximation $\widetilde{z_h}$ and $\nu$ iterations of $\omega_h$-Jacobi as a right preconditioner to obtain the final approximation $z_h$.
```

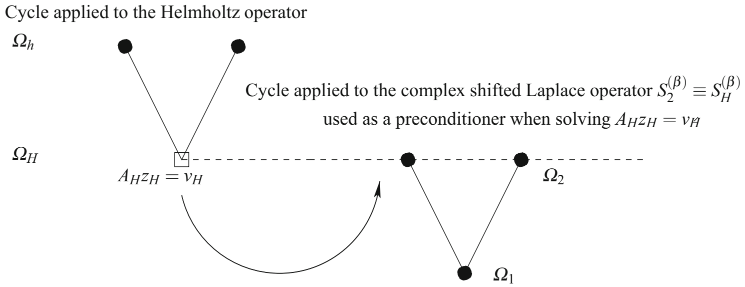

Cycle applied to the Helmholtz operator
Ω_h
Ω_H
$A_H z_H = v_H$
Cycle applied to the complex shifted Laplace operator $S_2^{(\beta)} \equiv S_H^{(\beta)}$
used as a preconditioner when solving $A_H z_H = v_H$
Ω_2
Ω_1

Fig. 1 Multigrid cycle applied to $A_h z_h = v_h$ sketched in Algorithm 1 (case of $\mathscr{T}_{2,V}$). The two-grid cycle is applied to the Helmholtz operator (*left part*), whereas the second two-grid cycle to be used as a preconditioner when solving the coarse grid problem $A_H z_H = v_H$ is shown on the *right part*. This second multigrid cycle acts on the complex shifted Laplace operator $S_H^{(\beta)}$ with β as a shift parameter

is introduced. On this second hierarchy, a multigrid cycle applied to a complex shifted Laplace operator $S_H^{(\beta)} := S_l^{(\beta)}$ is then used as a preconditioner when solving approximately the coarse level system $A_H z_H = v_H$ of the two-grid cycle. We note that the complex shifted Laplace operator $S_H^{(\beta)}$ is simply obtained by direct coarse grid discretization of Eqs. (5)–(7) on Ω_H.

The resulting cycle is sketched in Algorithm 1. The notation $\mathscr{T}_{l,C}$ of Algorithm 1 uses subscripts related to the cycle applied to the complex shifted Laplace operator with l denoting the number of grids of the second hierarchy and C referring to the cycling strategy which can be of V, F or W type.

As an illustration, Fig. 1 depicts the simplest configuration ($\mathscr{T}_{2,V}$) based on a two-grid cycle applied to the complex shifted Laplace operator. This cycle will be considered later in Sect. 4.

As explained in [11], this cycle leads to a variable nonlinear preconditioner which must be combined with an outer *flexible* Krylov subspace method [43, 44] and [57, Chap. 10]. In [11], the efficiency of the preconditioner in combination with FGMRES(5) [39] has been highlighted on both academic and realistic three-dimensional test problems. We investigate in the next section its performance when used in combination with a dispersion minimizing discretization scheme.

3.3 Parallel Implementation

The parallel implementation of the cycle proposed in Algorithm 1 is based on standard MPI (Message Passing Interface) [27]. We refer the reader to [53, Chap. 6] for details on the parallelization of geometric multigrid methods based on domain partitioning. In particular, the operations related to matrix-vector products, restriction and interpolation require local communications between neighbouring processes. As in [16, 56], the polynomial smoothing procedure is based on GMRES [41], which requires both local and global communications. Local and global communications also occur when solving the coarse grid system with the preconditioned FGMRES Krylov subspace method [39]. We refer the reader to [40, Chap. 11] for a discussion on parallel implementations of Krylov subspace methods. To take advantage of the current multicore based computer architectures, we note that the use of MPI and OpenMP would be relevant to consider. This is left to a future line of development.

We investigate in the next section the performance of the proposed geometric preconditioner, when a dispersion minimizing finite difference scheme is considered for the discretization of the Helmholtz problem.

4 Numerical Results on the SEG/EAGE Salt Dome Model

In this section, we illustrate the performance of the multigrid preconditioner used in combination with FGMRES(m) for the solution of the acoustic Helmholtz problem ((2)–(4)) on a realistic heterogeneous benchmark velocity model. The SEG/EAGE Salt dome model [2] is a velocity field containing a salt dome in a sedimentary embankment. It is defined in a parallelepiped domain of size $13.5 \times 13.5 \times 4.2\,\text{km}^3$. The minimum value of the velocity is $1500\,\text{m}\,\text{s}^{-1}$ and its maximum value is $4481\,\text{m}\,\text{s}^{-1}$, respectively. This test case is considered as challenging due to both the occurrence of a geometrically complex structure (salt dome) and to the truly large dimensions of the computational domain. We first analyse the strong scalability properties of the numerical method on this realistic application. Then we investigate numerically the complexity of the numerical method, i.e., the evolution of the memory requirements and computational times with respect to the number of unknowns. We first define the settings and parameters of the multigrid preconditioner used in this study.

4.1 Settings and Parameters

In the two-grid cycle of Algorithm 1, we consider as a smoother the case of one cycle of GMRES(2) preconditioned by two iterations of damped Jacobi ($\vartheta = 1$,

$m_s = 2$ and $\nu = 2$), a restarting parameter equal to $m_c = 10$ for the preconditioned GMRES method used on the coarse level and a maximal number of coarse cycles equal to $\vartheta_c = 10$. In the complex shifted multigrid cycle used as an approximate coarse solver, we use a shift parameter equal to $\beta = 0.5$ and two iterations of damped Jacobi as a smoother ($\nu_\beta = 2$). On the coarsest level we consider as an approximate solver one cycle of GMRES(10) preconditioned by two iterations of damped Jacobi ($\vartheta_\beta = 1, m_\beta = 10$ and $\nu_\beta = 2$). Finally, the relaxation coefficients considered in the Jacobi method are given by the following relation

$$(\omega_h, \omega_{2h}, \omega_{4h}) = (0.8, 0.8, 0.2).$$

We consider a value of the restarting parameter of the outer Krylov subspace method equal to $m = 5$ as in [10, 36]. The unit source is located at

$$(s_1, s_2, s_3) = (h\, n_{x_1}/2, h\, n_{x_2}/2, h\, (n_{PML} + 1))$$

where, e.g., n_{x_1} denotes the number of points in the first direction and n_{PML} is set to 10. A zero initial guess x_h^0 is selected and the iterative method is stopped when the Euclidean norm of the residual normalized by the Euclidean norm of the right-hand side satisfies the following relation

$$\frac{||b_h - A_h x_h||_2}{||b_h||_2} \le 10^{-5}.$$

This numerical study has been performed on Turing,[2] a IBM BG/Q computer located at IDRIS (each node of Turing is equipped with 16 PowerPC A2-64 bit cores at 1.6 GHz) using a Fortran 90 implementation with MPI in complex single precision arithmetic. Physical memory on a given node (16 cores) of Turing is limited to 16 GB.

4.2 *Strong Scalability Analysis*

We are interested in the strong scalability properties of the numerical method. Hence, we consider the acoustic wave propagation problem ((2)–(4)) at a fixed frequency (20 Hz) on a growing number of cores.The step size h is determined by relation (8) with $f_{ref} = 10$ Hz, $h_{ref} = 15$ and $q_{ref} = 4$. Table 1 collects the number of preconditioner applications (Prec) and computational time (Time) versus the number of cores. We note that the number of preconditioner applications (which corresponds to the number of outer Krylov subspace iterations) is found to be independent of the number of cores, which is a nice property. We also define a

[2] http://www.idris.fr/turing/.

Table 1 Strong scalability analysis

f (Hz)	Grid	# cores	Prec	T (s)	τ_s
20	2303 × 2303 × 767	16,384	29	586	1.00
20	2303 × 2303 × 767	32,768	29	302	0.97
20	2303 × 2303 × 767	65,536	29	164	0.89
20	2303 × 2303 × 767	131,072	29	87	0.84

Case of $\mathscr{T}_{2,V}$ applied as a preconditioner of FGMRES(5) for the heterogeneous velocity field EAGE/SEG Salt dome. Prec denotes the number of preconditioner applications, T the total computational time in seconds and τ_s the scaled parallel efficiency defined in relation (9)

scaled parallel efficiency as

$$\tau_s = \frac{T_{ref}}{T} / \frac{Cores}{Cores_{ref}}, \tag{9}$$

where T_{ref} and $Cores_{ref}$ denote reference values related to computational time and number of cores, respectively. A perfect scaling corresponds to the value of 1. In practice, we note that τ_s is close to this value. Only the last numerical experiment performed on 131,072 cores leads to a moderate decrease in terms of scaled parallel efficiency. This is partly due to the increased number of communications, which leads to a significant decrease of the ratio computation/communication.

4.3 Complexity Analysis

We next analyse the complexity of the numerical method with respect to the frequency or to the problem size, equivalently. In this numerical experiment, the number of cores is kept fixed to 131,072, while the frequency is growing from 15 Hz to 40 Hz, respectively. The case of $f = 40$ Hz leads to a linear system with approximately 56.7 billion of unknowns. The number of preconditioner applications (Prec), computational times (T) and memory requirements (M) are reported in Table 2. The number of preconditioner applications is rather moderate and is found to grow almost linearly with respect to the frequency. This linear dependency has been also observed for the complex shifted Laplace preconditioner in relation with other dispersion minimizing finite difference schemes [12], on problems of smaller size though. This behaviour is quite satisfactory, since huge linear systems can be solved in a reasonable amount of computational time on a parallel distributed memory machine.

Figure 2 shows the evolution of the computational time (T) versus the problem size. If N denotes the total number of unknowns, the computational time T is found to behave asymptotically as $N^{1.32}$. This is quite competitive with advanced sparse direct solution methods based on block low rank [1] or hierarchical matrix

Table 2 Complexity analysis

f (Hz)	Grid	# cores	Prec	T (s)	M (TB)
15	1586 × 1586 × 492	131,072	19	30	0.56
20	2303 × 2303 × 767	131,072	29	87	1.67
25	3071 × 3071 × 1023	131,072	37	236	3.79
30	3839 × 3839 × 1279	131,072	45	552	7.20
35	4607 × 4607 × 1535	131,072	57	1158	12.2
40	5631 × 5631 × 1791	131,072	69	2458	20.9

Case of $\mathscr{T}_{2,V}$ applied as a preconditioner of FGMRES(5) for the heterogeneous velocity field EAGE/SEG Salt dome. Prec denotes the number of preconditioner applications, T the total computational time in seconds and M the requested memory in TB

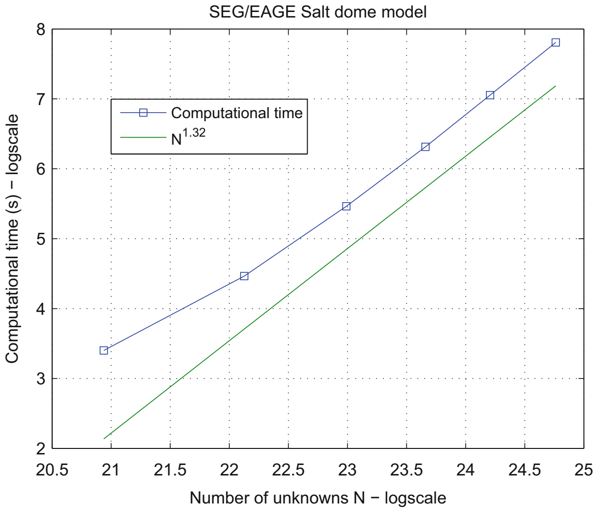


Fig. 2 Complexity analysis of the improved two-grid preconditioned Krylov subspace method. Evolution of computational time versus problem size. EAGE/SEG Salt dome. Results of Table 2

compression techniques [35, 59, 60]. To complement this study, it would be interesting to perform the same complexity analysis, now when addressing linear systems with multiple right-hand sides. Efficient block Krylov subspace methods based on block size reduction at each restart [9, 10] or at each iteration [8] have been proposed in this setting. This is left to a future line of research.

The evolution of the requested memory (M) versus the problem size is shown in Fig. 3. As expected, the memory requirements grow linearly with the number of unknowns, since no sparse factorization is involved neither at the global nor at local levels in the preconditioner. We remark that the benefit of the proposed method has to be viewed in the light of future parallel architectures with the most scalable architectures having limited memory per core.

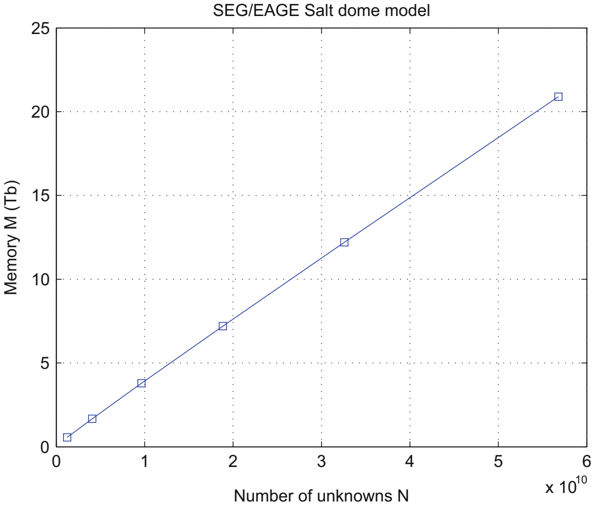


Fig. 3 Complexity analysis of the improved two-grid preconditioned Krylov subspace method. Evolution of memory requirements versus problem size. EAGE/SEG Salt dome. Results of Table 2

5 Summary and Outlook

In this chapter, we have focused on the performance of a geometric multigrid preconditioner for the solution of wave propagation problems related to acoustic imaging. We have proposed a two-grid preconditioner for the numerical solution of Helmholtz problems in three-dimensional heterogeneous media. This two-grid cycle is directly applied to the original Helmholtz operator and relies on an approximate coarse grid solution. A second multigrid method applied to a complex shifted Laplace operator is then used as a preconditioner to obtain the approximate coarse solution. We have highlighted the efficiency of the multigrid preconditioner on a concrete application in geophysics requiring the solution of problems of huge dimension (namely, billion of unknowns) in combination with dispersion minimizing finite difference schemes. Numerical results have demonstrated the usefulness of the combined algorithm on a realistic three-dimensional application at high frequency. Finally, a detailed complexity analysis has been provided to close this chapter.

We would like to mention three recent contributions for the solution of heterogeneous Helmholtz problems exhibiting attractive complexities and almost frequency independent rate of convergence. Zepeda-Núñez and Demanet [61] have proposed an algorithm based on the combination of domain decomposition techniques and integral equations with application to two-dimensional acoustic problems. Liu and Ying [31, 32] have proposed enhancements of the sweeping preconditioner leading to a $O(N)$ complexity for both the setup phase and the preconditioner application. A detailed performance comparison with the proposed numerical method on the same benchmark problem would be interesting to perform.

In the context of inverse problems in seismic, e.g., acoustic full waveform inversion, the solution of forward Helmholtz problems represents a major computational kernel, as outlined above. For that purpose, the geometric multigrid preconditioner used in combination with block Krylov subspace methods will play a key role to address the solution of linear systems with multiple right-hand sides efficiently; see [14] for a first attempt with a basic two-grid preconditioner developed in [36].

Advanced discretization methods based on Discontinuous Galerkin or high order finite element methods on unstructured grids are nowadays frequently used in geophysics for the solution of acoustic and/or elastic problems. Algebraic multigrid methods [53, Appendix A] could be used as well to extend the proposed geometric multigrid preconditioner and define an efficient numerical method in this setting; see, e.g., [6, 33] for related contributions.

Finally, we will have to reconsider the global algorithm to fully exploit the extreme core count of forthcoming parallel computers. Communication-avoiding or minimizing Krylov subspace methods [30] with asynchronous variants of the multigrid preconditioner should be developed in a near future to tackle this exciting new challenge.

Acknowledgements The authors would like to thank TOTAL for the financial support over the past years. They also would like to acknowledge GENCI (Grand Equipement National de Calcul Intensif) for the dotation of computing hours on the IBM BG/Q computer at IDRIS, France. This work was granted access to the HPC resources of IDRIS under allocation 2015065068 and 2016065068 made by GENCI.

References

1. P. Amestoy, C. Ashcraft, O. Boiteau, A. Buttari, J.-Y. L'Excellent, and C. Weisbecker. Improving multifrontal methods by means of block low-rank representations. *SIAM J. Sci. Comput.*, 37(3):A1451–A1474, 2015.
2. F. Aminzadeh, J. Brac, and T. Kunz. 3D Salt and Overthrust models. SEG/EAGE modeling series I, Society of Exploration Geophysicists, 1997.
3. A. Bayliss, C. I. Goldstein, and E. Turkel. An iterative method for the Helmholtz equation. *J. Comp. Phys.*, 49:443–457, 1983.
4. J.-P. Berenger. A perfectly matched layer for absorption of electromagnetic waves. *J. Comp. Phys.*, 114:185–200, 1994.
5. J.-P. Berenger. Three-dimensional perfectly matched layer for absorption of electromagnetic waves. *J. Comp. Phys.*, 127:363–379, 1996.
6. M. Bollhöfer, M. J. Grote, and O. Schenk. Algebraic multilevel preconditioner for the solution of the Helmholtz equation in heterogeneous media. *SIAM J. Sci. Comput.*, 31:3781–3805, 2009.
7. A. Brandt and I. Livshits. Wave-ray multigrid method for standing wave equations. *Electron. Trans. Numer. Anal.*, 6:162–181, 1997.
8. H. Calandra, S. Gratton, R. Lago, X. Vasseur, and L. M. Carvalho. A modified block flexible GMRES method with deflation at each iteration for the solution of non-hermitian linear systems with multiple right-hand sides. *SIAM J. Sci. Comput.*, 35(5):S345–S367, 2013.

9. H. Calandra, S. Gratton, R. Lago, X. Pinel, and X. Vasseur. Two-level preconditioned Krylov subspace methods for the solution of three-dimensional heterogeneous Helmholtz problems in seismics. *Numerical Analysis and Applications*, 5:175–181, 2012.
10. H. Calandra, S. Gratton, J. Langou, X. Pinel, and X. Vasseur. Flexible variants of block restarted GMRES methods with application to geophysics. *SIAM J. Sci. Comput.*, 34(2):A714–A736, 2012.
11. H. Calandra, S. Gratton, X. Pinel, and X. Vasseur. An improved two-grid preconditioner for the solution of three-dimensional Helmholtz problems in heterogeneous media. *Numer. Linear Algebra Appl.*, 20, pp. 663–688, 2013.
12. Z. Chen, D. Cheng, and T. Wu. A dispersion minimizing finite difference scheme and preconditioned solver for the 3D Helmholtz equation. *J. Comp. Phys.*, 231:8152–8175, 2012.
13. G. Cohen. *Higher-order numerical methods for transient wave equations*. Springer, 2002.
14. Y. Diouane, S. Gratton, X. Vasseur, L. N. Vicente, and H. Calandra. A parallel evolution strategy for an Earth imaging problem in geophysics. *Optimization and Engineering*, 17(1):3–26, 2016.
15. H. Elman, O. Ernst, D. O'Leary, and M. Stewart. Efficient iterative algorithms for the stochastic finite element method with application to acoustic scattering. *Comput. Methods Appl. Mech. Engrg.*, 194(1):1037–1055, 2005.
16. H. C. Elman, O. G. Ernst, and D. P. O'Leary. A multigrid method enhanced by Krylov subspace iteration for discrete Helmholtz equations. *SIAM J. Sci. Comput.*, 23:1291–1315, 2001.
17. B. Engquist and L. Ying. Sweeping preconditioner for the Helmholtz equation: moving perfectly matched layers. *Multiscale Modeling and Simulation*, 9:686–710, 2011.
18. B. Engquist and L. Ying. Sweeping preconditioner for the Helmholtz equation: Hierarchical matrix representation. *Comm. Pure Appl. Math.*, 64:697–735, 2011.
19. Y. A. Erlangga. *A robust and efficient iterative method for the numerical solution of the Helmholtz equation*. PhD thesis, TU Delft, 2005.
20. Y. A. Erlangga. Advances in iterative methods and preconditioners for the Helmholtz equation. *Archives of Computational Methods in Engineering*, 15:37–66, 2008.
21. Y. A. Erlangga and R. Nabben. On a multilevel Krylov method for the Helmholtz equation preconditioned by shifted Laplacian. *Electron. Trans. Numer. Anal.*, 31:403–424, 2008.
22. Y. A. Erlangga, C. Oosterlee, and C. Vuik. A novel multigrid based preconditioner for heterogeneous Helmholtz problems. *SIAM J. Sci. Comput.*, 27:1471–1492, 2006.
23. Y. A. Erlangga, C. Vuik, and C. Oosterlee. On a class of preconditioners for solving the Helmholtz equation. *Appl. Num. Math.*, 50:409–425, 2004.
24. O. Ernst and M. J. Gander. Why it is difficult to solve Helmholtz problems with classical iterative methods. In O. Lakkis I. Graham, T. Hou and R. Scheichl, editors, *Numerical Analysis of Multiscale Problems*. Springer, 2011.
25. C. Farhat, A. Macedo, and M. Lesoinne. A two-level domain decomposition method for the iterative solution of high frequency exterior Helmholtz problems. *Numer. Math.*, 85:283–308, 2000.
26. M. Gander, I. G. Graham, and E. A. Spence. Applying GMRES to the Helmholtz equation with shifted Laplacian preconditioning: What is the largest shift for which wavenumber-independent convergence is guaranteed? *Numer. Math.*, 131:567–614, 2015.
27. W. Gropp, E. Lusk, and A. Skjellum. *Using MPI: Portable Parallel Programming with the Message-Passing Interface*. MIT Press, 1999.
28. W. Hackbusch and U. Trottenberg. *Multigrid methods*. Springer, 1982. Lecture Notes in Mathematics, vol. 960, Proceedings of the conference held at Köln-Porz, November 23–27 1981.
29. I. Harari and E. Turkel. Accurate finite difference methods for time-harmonic wave propagation. *J. Comp. Phys.*, 119:252–270, 1995.
30. M. Hoemmen. *Communication-avoiding Krylov subspace methods*. PhD thesis, University of California, Berkeley, Department of Computer Science, 2010.
31. F. Liu and L. Ying Additive sweeping preconditioner for the Helmholtz equation ArXiv e-prints, 2015. http://arxiv.org/abs/1504.04058.

32. F. Liu and L. Ying Recursive sweeping preconditioner for the 3D Helmholtz equation ArXiv e-prints, 2015. http://arxiv.org/abs/1502.07266.
33. S. P. MacLachlan and C. W. Oosterlee. Algebraic multigrid solvers for complex-valued matrices. *SIAM J. Sci. Comput.*, 30:1548–1571, 2008.
34. S. Operto, J. Virieux, P. R. Amestoy, J.-Y. L'Excellent, L. Giraud, and H. Ben Hadj Ali. 3D finite-difference frequency-domain modeling of visco-acoustic wave propagation using a massively parallel direct solver: A feasibility study. *Geophysics*, 72–5:195–211, 2007.
35. J. Poulson, B. Engquist, S. Li, and L. Ying. A parallel sweeping preconditioner for heterogeneous 3d Helmholtz equations. *SIAM J. Sci. Comput.*, 35:C194–C212, 2013.
36. X. Pinel. *A perturbed two-level preconditioner for the solution of three-dimensional heterogeneous Helmholtz problems with applications to geophysics*. PhD thesis, CERFACS, 2010. TH/PA/10/55.
37. B. Reps, W. Vanroose, and H. bin Zubair. On the indefinite Helmholtz equation: complex stretched absorbing boundary layers, iterative analysis, and preconditioning. *J. Comp. Phys.*, 229:8384–8405, 2010.
38. C. D. Riyanti, A. Kononov, Y. A. Erlangga, R.-E. Plessix, W. A. Mulder, C. Vuik, and C. Oosterlee. A parallel multigrid-based preconditioner for the 3D heterogeneous high-frequency Helmholtz equation. *J. Comp. Phys.*, 224:431–448, 2007.
39. Y. Saad. A flexible inner-outer preconditioned GMRES algorithm. *SIAM J. Scientific and Statistical Computing*, 14:461–469, 1993.
40. Y. Saad. *Iterative Methods for Sparse Linear Systems.* SIAM, Philadelphia, 2003. Second edition.
41. Y. Saad and M. H. Schultz. GMRES: A generalized minimal residual algorithm for solving nonsymmetric linear systems. *SIAM J. Scientific and Statistical Computing*, 7:856–869, 1986.
42. A.H. Sheikh, D. Lahaye, and C. Vuik. On the convergence of shifted Laplace preconditioner combined with multilevel deflation. *Numer. Linear Algebra Appl.*, 20:645–662, 2013.
43. V. Simoncini and D. B. Szyld. Flexible inner-outer Krylov subspace methods. *SIAM J. Numer. Anal.*, 40:2219–2239, 2003.
44. V. Simoncini and D. B. Szyld. Recent computational developments in Krylov subspace methods for linear systems. *Numer. Linear Algebra Appl.*, 14:1–59, 2007.
45. F. Sourbier, S. Operto, J. Virieux, P. Amestoy, and J. Y. L' Excellent. FWT2D : a massively parallel program for frequency-domain full-waveform tomography of wide-aperture seismic data - part 1: algorithm. *Computer & Geosciences*, 35:487–495, 2009.
46. F. Sourbier, S. Operto, J. Virieux, P. Amestoy, and J. Y. L' Excellent. FWT2D : a massively parallel program for frequency-domain full-waveform tomography of wide-aperture seismic data - part 2: numerical examples and scalability analysis. *Computer & Geosciences*, 35:496–514, 2009.
47. C. Stolk, M. Ahmed, and S. K. Bhowmik. A multigrid method for the Helmholtz equation with optimized coarse grid correction. *SIAM J. Sci. Comput.*, 36:A2819–A2841, 2014.
48. C. Stolk. A rapidly converging domain decomposition method for the Helmholtz equation. *J. Comp. Phys.*, 241:240–252, 2013.
49. K. Stüben and U. Trottenberg. Multigrid methods: fundamental algorithms, model problem analysis and applications. In W. Hackbusch and U. Trottenberg, editors, *Multigrid methods, Koeln-Porz, 1981, Lecture Notes in Mathematics, volume 960*. Springer, 1982.
50. A. Tarantola. *Inverse problem theory and methods for model parameter estimation*. SIAM, 2005.
51. E. Turkel, D. Gordon, R. Gordon, and S. Tsynkov. Compact 2D and 3D sixth order schemes for the Helmholtz equation with variable wavenumber. *J. Comp. Phys.*, 232:272–287, 2013.
52. A. Toselli and O. Widlund. *Domain Decomposition methods - Algorithms and Theory*. Springer Series on Computational Mathematics, Springer, 34, 2005.
53. U. Trottenberg, C. W. Oosterlee, and A. Schüller. *Multigrid.* Academic Press Inc., 2001.
54. N. Umetani, S. P. MacLachlan, and C. W. Oosterlee. A multigrid-based shifted Laplacian preconditioner for fourth-order Helmholtz discretization. *Numer. Linear Algebra Appl.*, 16:603–626, 2009.

55. M. B. van Gijzen, Y. A. Erlangga, and C. Vuik. Spectral analysis of the discrete Helmholtz operator preconditioned with a shifted Laplacian. *SIAM J. Sci. Comput.*, 29:1942–1958, 2007.
56. W. Vanroose, B. Reps, and H. bin Zubair. A polynomial multigrid smoother for the iterative solution of the heterogeneous Helmholtz problem. Technical Report, University of Antwerp, Belgium, 2010. http://arxiv.org/abs/1012.5379.
57. P. S. Vassilevski. *Multilevel Block Factorization Preconditioners, Matrix-based Analysis and Algorithms for Solving Finite Element Equations*. Springer, New York, 2008.
58. J. Virieux and S. Operto. An overview of full waveform inversion in exploration geophysics. *Geophysics*, 74(6):WCC127–WCC152, 2009.
59. S. Wang, M. V. de Hoop, and J. Xia. Acoustic inverse scattering via Helmholtz operator factorization and optimization. *J. Comp. Phys.*, 229:8445–8462, 2010.
60. S. Wang, M. V. de Hoop, and J. Xia. On 3D modeling of seismic wave propagation via a structured parallel multifrontal direct Helmholtz solver. *Geophysical Prospecting*, 59:857–873, 2011.
61. L. Zepeda-Núñez and L. Demanet. The method of polarized traces for the 2D Helmholtz equation. *J. Comp. Phys.*, 308:347–388, 2016.

Part III
Implementations and Industrial Applications

In this part real applications based on Helmholtz solvers and their challenges are described and solved.

Some Computational Aspects of the Time and Frequency Domain Formulations of Seismic Waveform Inversion

René-Édouard Plessix

Abstract Seismic waveform inversion relies on efficient solutions of the elasto-dynamic wave equations. The associated inverse problem can be formulated either in the time domain or in the frequency domain. The choice between these two approaches mainly depends on their numerical efficiency. Here, I discuss some of the computational aspects of the frequency-domain solution of the visco-acoustic vertical transverse isotropic wave equations based on a Krylov subspace iterative solver and a complex shifted Laplace preconditioner. In the context of least-square migration or non-linear impedance inversion, the frequency domain approaches are currently not attractive because a complete frequency band response is required. However, in the context of waveform tomography when a small number of frequency responses are inverted, the frequency-domain approaches become relevant, especially when viscous effects are modeled, depending on the geological context.

1 Introduction

Characterizing and understanding the structure of the Earth interior and its evolution represent a major goal in global seismology and exploration geophysics. Very few, and only sparsely sampled, direct observations exist, for instance outcrops and well logs. Geophysicists rely on measurements of the responses to physical phenomena that interact with the Earth. They infer information on the Earth structure through an inversion, also called imaging, approach. Surface seismic data, that measure the earth responses to a mechanical excitation, provide indirect information on the elastic properties of the Earth. In the Earth, the elasto-dynamic wave equations govern the propagation of the seismic waves [1, 2] and the seismic waves diffract on the Earth elastic discontinuities. Figure 1 displays a typical shot gather from an active marine seismic acquisition. One distinguishes different types of events notably the early transmitted arrivals at long offsets (distances between source and

R.-É. Plessix (✉)
Shell Global Solutions International, Kesslerpark 1, 2288 GS Rijswijk, The Netherlands
e-mail: reneedouard.plessix@shell.com

D. Lahaye et al. (eds.), *Modern Solvers for Helmholtz Problems*, Geosystems Mathematics, DOI 10.1007/978-3-319-28832-1_7

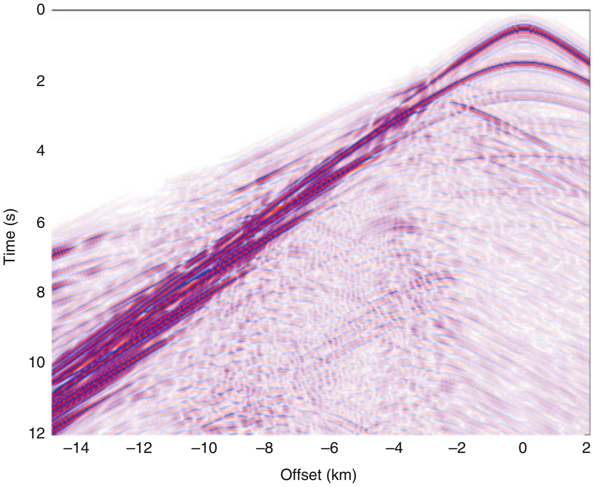


Fig. 1 A typical shot gather from a marine (synthetic) seismic acquisition corresponding to the pressure response to an explosive source. The source and receivers are a few meters below the sea surface

receivers), and the reflected waves at shot offsets. The goal of seismic imaging is to reconstruct an Earth model from the seismic data set. Due to the propagative nature of the seismic waves in the Earth, in a high frequency representation, an event can be characterized by its traveltime and its amplitude once deconvolved from the source signature [2, 52]. One imaging approach consists of first picking the traveltimes and the amplitudes of dedicated events, then carrying out a traveltime inversion to retrieve the propagation velocities, and an amplitude versus offset (AVO) inversion to deduce impedance variations at the discontinuities. Another imaging approach consists of directly minimizing the data misfit between observed and modeled seismic data [7, 27, 46], which is known as seismic full waveform inversion (FWI).

Seismic full waveform inversion relies on our ability to synthesize a seismic experiment, that is to solve efficiently the elasto-dynamic wave equations since a seismic data set consists of thousands to hundreds of thousand shot gathers in real applications. This raises the question of the modeling domain [28]. In the time domain, the elasto-dynamic wave equations are hyperbolic equations that can be efficiently solved with a time marching and an explicit scheme. In the frequency domain, the elasto-dynamic wave equations are parabolic equations and their discretization gives a linear system. This linear system is challenging to invert because it is indefinite, has negative and positive eigenvalues, and is very large for 3D real-sized applications. A review of the different modeling approaches can be found in [50]. The choice of the modeling domain, that is frequency or time, depends on the behavior of the inverse problem, principally the possibility to invert or not the frequencies independently. Due to the lack of low frequencies in seismic data, the least-squares data misfit is a very oscillatory function [46] because of the cycle skipping between observed and modeled data. Unfortunately, the computation cost of solving the elasto-dynamic wave equations forces us to use local optimization techniques to invert this misfit function. With a local optimization, it was recognized soon after the introduction of the waveform inversion [27, 45] that the behavior is different for reflection and transmission data [24]. With transmission data, the smooth part, that is the low or intermediate wavenumbers, of the velocities could be retrieved. This leads to waveform tomography [24, 41]. With reflection data, the rough part, that is the high wavenumbers of the impedance, could be retrieved. This leads to reverse time migration and non-linear impedance inversion [27, 45]. The physical reasons of this behavior are explained for instance in [24, 31, 41, 49]. Figures 2 and 3 illustrate the two different behaviors in two very simple cases. In the transmission case, the gradient of the least-squares data misfit function obtained with the narrow band data set is a good estimation of the velocity perturbation. A frequency continuation approach can be used in waveform tomography, and in fact is recommended, to retrieve the low-to-intermediate wavenumbers of the velocities. One can carry out the inversion per frequency (or per small groups of frequencies) [41, 49]. In the reflection case, the resolution of the image/gradient is

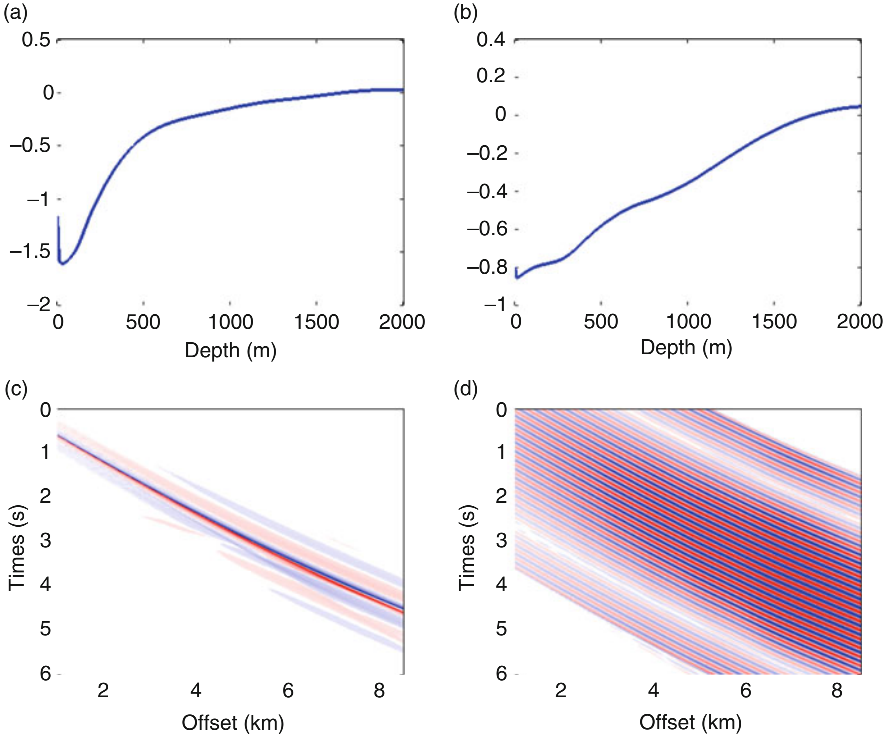


Fig. 2 Gradients and data residuals with long offset transmission data. The observed data are computed with a linear velocity ($v(z) = 1500 + 0.4z$), with the depth z in meter and the velocity v in m/s. The gradient and data residuals are computed with a perturbed linear velocity ($v(z) = 1500 + 0.37z$). The gradient should then be negative. The full band data set corresponds to a flat source spectrum between 2 and 30 Hz. The narrow band data set corresponds to a source spectrum between 4.9 and 5.1 Hz. (**a**) Gradient (full band). (**b**) Gradient (narrow band). (**c**) Data residual (full band). (**d**) Data residual (narrow band)

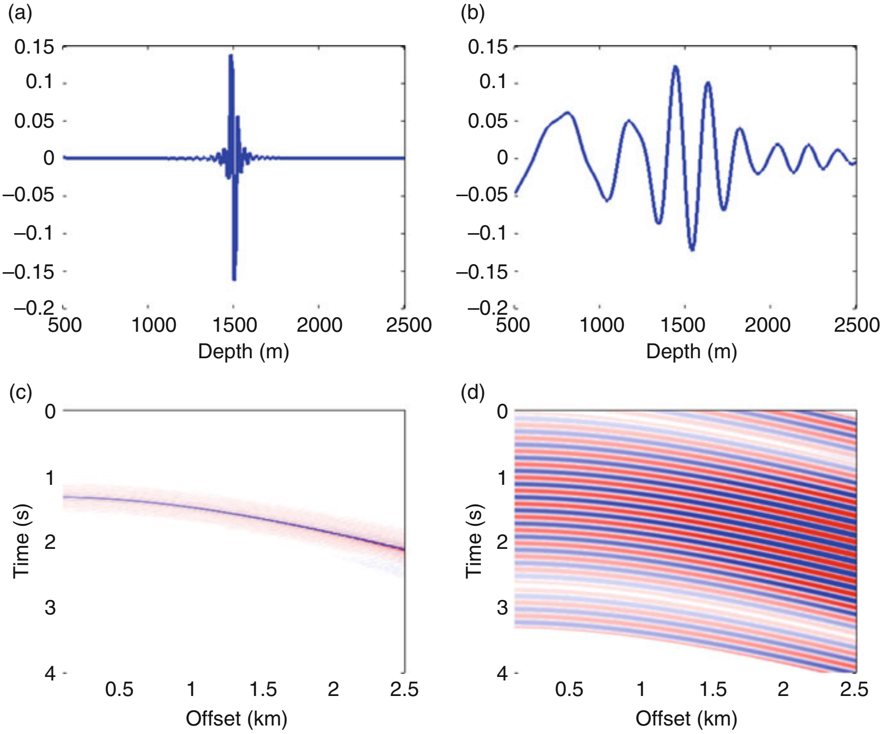


Fig. 3 Gradients and data residuals with short offset reflection data. The observed data are computed in a two-layer velocity model (the velocity of the first layer is 1500 m/s and the one of the second layer 2000 m/s with the interface at 1.5 km depth). The gradient and data residual are computed with a 1500 m/s homogeneous layer. The gradient should indicate the discontinuity convolved with the source wavelet square. The full band data set corresponds to a flat source spectrum between 2 and 30 Hz. The narrow band data set corresponds to a source spectrum between 4.9 and 5.1 Hz. (**a**) Gradient (full band). (**b**) Gradient (narrow band). (**c**) Data residual (full band). (**d**) Data residual (narrow band)

determined by the bandwidth. For migration and impedance inversion, a sufficiently large frequency band needs to be inverted.

In waveform tomography, a frequency-domain solver of the elasto-dynamic wave equations becomes attractive when it is faster than the time-domain scheme to compute one frequency response. In non-linear migration or impedance inversion, it becomes attractive only when it is faster than the time-domain scheme to compute the time response after Fourier transform. The linear system associated to the discretization of the frequency-domain approach can be solved either with a direct solver or with an iterative solver [34, 38]. The advantage of a direct solver resides in its efficient computation of the frequency responses due to many source points, that is many right-hand sides, after having performed the LU decomposition [28, 33, 51]. However, the LU decomposition remains challenging for large 3D problems. The difficulties with the iterative solver come from the indefiniteness of the system and the definition of an efficient preconditioner. Several different approaches have been recently proposed [18, 19, 21, 25, 35]. A Krylov subspace iterative solver with a complex shifted Laplace preconditioner [21, 22] gives a robust approach at seismic frequencies and has been used to successfully invert real-sized 3D data sets [39, 40].

In this work, I discuss the computational aspects of the time-domain formulation and the frequency-domain formulation based on the shifted Laplace preconditioner in the context of seismic waveform inversion. I do not include the developments based on deflation and multilevel Krylov method that improve the convergence of the iterative solver [23]. A priori, these new developments would not change the main conclusions of this paper. They could potentially increase the relevance of the iterative solver when the frequency-domain formulation competes with the time-domain formulation. More works in these directions are required to evaluate them in real-sized applications and when high-order finite-difference schemes are used to reduce the problem size. For this discussion, I consider real-sized seismic imaging problems, and anisotropy and viscous effects. I only focus on exploration geophysics applications that invert body compression waves. This means that only acoustic wave equations are discussed. Solving the elastic frequency-domain wave equations with large 3D problems remains a challenge for both the direct and iterative solvers [9].

I first review the general modeling aspects of the time and frequency domain formulations and their numerical complexity. Then I present some numerical results of the iterative solver with a complex Laplace shifted preconditioner. I present a land example and a marine example because the iterative solver behaves differently. Finally, I compare the computational aspects of the time and frequency domain formulations of the waveform inversion.

2 The Modeling Aspect

Given a source excitation and earth medium properties, a seismic shot gather can be synthesized by solving the time-domain visco-elasto-dynamic equations [1, 15, 47]:

$$\begin{cases} \rho(\mathrm{x})\partial_{tt}u_i(\mathrm{x}_s,\mathrm{x},t) = \partial_j\sigma_{ij}(\mathrm{x}_s,\mathrm{x},t) + f_i(\mathrm{x}_s,\mathrm{x},t); \\ \sigma_{ij}(\mathrm{x}_s,\mathrm{x},t) = \int dt'\,\partial_t\psi_{ijkl}(\mathrm{x},t-t')\partial_k u_l(\mathrm{x}_s,\mathrm{x},t') + m_{ij}(\mathrm{x}_s,\mathrm{x},t); \\ \mathrm{d}^{mod}(\mathrm{x}_s,\mathrm{x}_r,t) = \mathrm{S}_u(\mathrm{x}_s,\mathrm{x}_r)\mathrm{u}(\mathrm{x}_s,\mathrm{x}_r,t) + \mathrm{S}_\sigma(\mathrm{x}_s,\mathrm{x}_r)\sigma(\mathrm{x}_s,\mathrm{x}_r,t); \end{cases} \tag{1}$$

where t is the time, x the coordinate vector of the sub-surface point, x_s of the source point and x_r of the receiver point. $\mathrm{u} = (u_i)$ is the particle displacement vector, $\sigma = (\sigma_{ij})$ the stress tensor, $\mathrm{f} = (f_i)$ the source force vector, $\mathrm{m} = (m_{ij})$ the source moment tensor, ρ the density and $\psi = (c_{ijkl})$ the stress-strain relaxation tensor that depends on time to account for the visco-elastic effects, that is the history of the strains in Hooke's law. d^{mod} denotes the modeled seismic traces, it may correspond to particle displacement, velocity or acceleration or pressure and S_u and S_σ are sampling operators.

The frequency-domain visco-elasto dynamic equation is simply obtained by Fourier transform:

$$\begin{cases} -\omega^2\rho(\mathrm{x})u_i(\mathrm{x}_s\mathrm{x},\omega) = \partial_j\sigma_{ij}(\mathrm{x}_s,\mathrm{x},\omega) + f_i(\mathrm{x}_s,\mathrm{x},\omega); \\ \sigma_{ij}(\mathrm{x}_s,\mathrm{x},\omega) = c_{ijkl}(\mathrm{x},\omega)\partial_k u_l(\mathrm{x}_s,\mathrm{x},\omega) + m_{ij}(\mathrm{x}_s,\mathrm{x},\omega); \\ \mathrm{d}^{mod}(\mathrm{x}_s,\mathrm{x}_r,\omega) = \mathrm{S}_u(\mathrm{x}_s,\mathrm{x}_r)\mathrm{u}(\mathrm{x}_s,\mathrm{x}_r,\omega) + \mathrm{S}_\sigma(\mathrm{x}_s,\mathrm{x}_r)\sigma(\mathrm{x}_s,\mathrm{x}_r,\omega); \end{cases} \tag{2}$$

with ω the angular frequency. I use the same symbols in the time and frequency domains. The time-convolution is transformed into a multiplication in the frequency domain with $\mathrm{c} = (c_{ijkl})$ the frequency-dependent complex stress-strain stiffness tensor. This makes the frequency formulation attractive when modeling visco-elastic effects. Indeed, we can consider any frequency dependency.

Many discretizations of these two systems have been proposed after having added some boundary conditions (see for instance the references in [50]). For the discussion here, I consider the same discretization for the spatial derivatives of the time and frequency formulations.

The discretization of the frequency formulation, system (2), gives an implicit linear system of the type:

$$\begin{cases} \mathrm{A}(\mathrm{x},\omega)\mathrm{u}(\mathrm{x}_s,\omega) = \mathrm{f}(\mathrm{x}_s,\omega); \\ \mathrm{d}^{mod}(\mathrm{x}_s,\mathrm{x}_r,\omega) = \mathrm{S}(\mathrm{x}_s,\mathrm{x}_r)\mathrm{u}(\mathrm{x}_s,\omega); \end{cases} \tag{3}$$

where A is a complex matrix depending on the earth parameters, u the vector of the discretized fields containing particle displacements or stresses, f the source vector and S the sampling operator. Each frequency response is computed independently.

Evaluating a general time convolution at each time step of the discretization of the time formulation, system (1), would lead to an unaffordable numerical scheme. To obtain an efficient time-domain formulation, the structure of the time (hence frequency) dependency of the strain-stress tensor is restricted to a series of P standard linear solid models [10, 13]:

$$\psi_{ijkl}(\mathrm{x},t) = c^0_{ijkl}(\mathrm{x})\left(1 + \sum_{p=1}^{P}(Q^p_{ijkl}(\mathrm{x}))^{-1}\exp(-\omega_p t)\right)H(t) \tag{4}$$

with $\mathrm{c}^0 = (c^0_{ijkl})$ the stiffness tensor at 0 Hz frequency, ω_p the angular relaxation frequency of the pth standard linear solid model, $(Q^p_{ijkl})^{-1}$ the strength of the pth standard linear solid model proportional to the inverse of the quality factor and H the Heaviside function.

After introducing the memory variables r^p_{ij}, the time-domain visco elasto-dynamic equations read [10, 13]:

$$\begin{cases} \rho(\mathrm{x})\partial_{tt}u_i(\mathrm{x}_s\mathrm{x},t) = \sum_j \partial_j\sigma_{ij}(\mathrm{x}_s,\mathrm{x},t) + f_i(\mathrm{x}_s,\mathrm{x},t); \\ \sigma_{ij}(\mathrm{x}_s,\mathrm{x},t) = \sum_{kl} c^0_{ijkl}(\mathrm{x})(1 + \sum_p (Q^p_{ijkl}(\mathrm{x}))^{-1})\partial_k u_l(\mathrm{x}_s,\mathrm{x},t) + \sum_p r^p_{ij} m_{ij}(\mathrm{x}_s,\mathrm{x},t); \\ \partial_t r^p_{ij}(\mathrm{x}_s,\mathrm{x},t) = -\omega_p r^p_{ij}(\mathrm{x}_s,\mathrm{x},t) - \sum_{kl}\omega_p(Q^p_{ijkl}(\mathrm{x}))^{-1}\partial_k u_l(\mathrm{x}_s,\mathrm{x},t). \end{cases} \tag{5}$$

The evolution equations for the P memory variables are obtained after time derivation of the memory variables defined by:

$$r_{ij}^{p}(t) = -\sum_{kl} c_{ijkl}^{0} \int dt' \; \omega_p (Q_{ijkl}^{p})^{-1} \exp(-\omega_p(t-t'))H(t-t')\partial_k u_l(t). \tag{6}$$

These equations are stiff. To avoid a too small time stepping, an implicit Crank-Nicholson scheme is generally used. This still leads to an efficient explicit overall scheme when we stagger the time discretization of the stresses and memory variables in a first-order velocity-stress formulation or when we stagger the time discretization of the memory variables and the displacements or stresses in a second-order formulation (although the stability condition is reduced compared to the pure elastic case). With the initial conditions $\mathrm{u}(\mathrm{x}_s, t_{-2}) = \mathrm{u}(\mathrm{x}_s, t_{-1}) = 0$, a general form of the discrete scheme is:

$$\begin{cases} \mathrm{u}(\mathrm{x}_s, t_n) = \mathrm{B}_1(\mathrm{x})\mathrm{u}(\mathrm{x}_s, t_{n-1}) + \mathrm{B}_2(\mathrm{x})\mathrm{u}(\mathrm{x}_s, t_{n-2}) + \mathrm{f}(\mathrm{x}_s, t_n); \\ \mathrm{d}^{mod}(\mathrm{x}_s, \mathrm{x}_r, t_n) = \mathrm{S}(\mathrm{x}_s, \mathrm{x}_r)\mathrm{u}(\mathrm{x}_s, t_n); \end{cases} \tag{7}$$

where B_1 and B_2 are matrices depending on the earth parameters, t_n the time discretization, u the vector field containing particle displacements or stresses and memory variables, f the source vector and S the sampling operator. I abuse the notations since for instance I use the same notations u and f in Eqs. (3) and (7) whereas they have a different meaning.

The discrete Fourier transform relates the frequency and time representations of the model data. This assumes that d is periodic which is not the case, although the signal vanishes at long time making the Fourier transform meaningful.

With a finite-difference, finite-volume or finite-element spatial discretization scheme the size of u is in $O(M)$ with M the number of grid points or cells ($O(M)$ means of the order of M). The matrices A, B_1 and B_2 are sparse matrices with $O(M)$ nonzero elements. In the rest of the text, I assume that $M = N^3$, which corresponds to a 3D regular discretization of a cubic problem with N the number of points in each direction. The frequency-domain linear system can be solved either with a direct method [33, 51] or with an iterative method [19, 22, 38]. The computational time complexities are with N_s the number of source points, N_t the number of time steps, N_ω the number of frequencies, N_{it} the number of iterations of the iterative scheme [34, 38]:

1. For the time-domain approach: $N_s N_t N^3$. In the time domain the spatial discretization is governed by the maximum frequency, f_{max}, one needs to model. Hence N is proportional to f_{max}. The stability condition tells us that N_t is also proportional to f_{max} when the grid spacing is adapted to frequency. This leads to the classic time domain complexity in $N_s f_{max}^4$.
2. For the frequency-domain approach with an iterative solver: $N_s N_\omega N_{it} N^3$. In the frequency domain, the spatial discretization can be adapted to the frequency, f, one models. If we want to model only one frequency the complexity is $N_s N_{it} f^3$.

If we want to model a frequency band regularly discretized up to f_{max}, the complexity is in $N_s N_{it} f_{max}^4$ because $\int_0^{f_{max}} f^3 df = \frac{f_{max}^4}{4}$. When N_{it} is proportional to f, this gives a complexity in $N_s f_{max}^5$.

3. For the frequency-domain approach with a direct solver: $N_\omega N^6$ (since the bandwidth of the matrix is N^2 with a finite-difference scheme). With a direct solver, the LU decomposition of the matrix A is the most expensive part. The approach therefore becomes almost independent of the number of source terms. At least the complexity does not increase as long as N_s is not larger than $O(N^2)$ which is the case for realistic seismic applications. This gives a complexity in f^6 when modeling only one frequency and in f_{max}^7 when modeling a full frequency band.

In this complexity analysis, I just discuss the order of magnitude in terms of number of points. The actual computational time of the simulations is given by C times the complexity order. This multiplicative constant, C, depends on the earth model and does play a role in the efficiency of the different approaches. I shall attempt to address this point later. This discussion on the complexity of the different approaches shows that the time-domain approach appears the most efficient one in 3D when modeling a complete frequency band or a time response. Indeed, the Shannon-Nyquist theorem tells us that N_ω is in $O(N_t)$, hence $O(N)$. Even with $N_s = O(N^2)$, the time-domain formulation in $O(N^6)$ has a smaller complexity than the frequency-domain formulation, in $O(N^7)$ with the direct solver or in $O(N_{it}N^6)$ with the iterative solver, that is $O(N^7)$ when N_{it} is proportional to f.

In real applications, the computational domain may be adapted to the acquisition geometry in order to reduce the spatial discretization. With the time-domain formulation the shots are processed separately and we can define a computational domain per shot. Similarly with the frequency-domain formulation and an iterative solver, although there are studies to process multiple right-hand sides. The number of right-hand sides would a-priori stay small because of the memory constraints of current computer architectures and the communication costs. So, for simplicity, we consider that with the iterative solver we process one shot at a time. With the frequency domain formulation and a direct solver, the situation is different because the LU decomposition of the matrix A is the most expensive part. Consequently, we a-priori want a single matrix for all the right-hand sides. With a streamer acquisition, when a boat tows the sources and receivers, the domain covered by all the shots is much larger than the one covered by only one shot. This makes the direct solver approach much more expensive a-priori, since the complexity is in N^6. In an ocean bottom node (OBN) acquisition, most of the nodes are active when a source is triggered. Therefore, all the shots more or less illuminate the same part of the earth and we can see the acquisition as a fixed spread acquisition. We can then consider a common computational domain for all the shots. With this OBN acquisition, when modeling only one (or a few) frequency response, the approach with the direct solver may be competitive as long as the number of shots is in $O(N^2)$. This is generally the case in modern OBN 3D acquisition. The application of the frequency-domain formulation with a direct solver is therefore limited to this fixed-spread acquisition. Moreover most of the current applications consider only an acoustic wave equation and it remains unclear whether an elastic approach would soon be feasible in a realistic situation. Indeed, because the shear velocities are smaller than the compressional velocities, let us say between 2 and 5 times smaller in typical cases, N is larger by a factor 2 to 5 in typical elastic cases. The computational time of the LU decomposition increases by a factor 64 to 15,625 in the elastic case, and the memory requirement, that is in $O(N^4)$, by a factor between 16 and 625. To reduce this very high computational cost, approximate LU decompositions are currently proposed [34].

The recent developments on the Helmholtz iterative solver leads to a number of iterations in $O(N^\alpha)$, that is in $O(f^\alpha)$ with α generally smaller or equal to 1. With a deflation approach and multi-level Krylov solvers, α may be smaller than 1 [23]. This means that the iterative solver may compete with the direct solver and is more flexible. However, we should say that currently with an OBN survey, the published results seems to indicate that the direct solver is probably more efficient, although the two approaches have the same complexity. Most of the published works concentrate on the isotropic acoustic case. Accounting for anisotropy, that is crucial in real applications, can be challenging with the iterative solver that relies on a preconditioner built through a multi-grid approach. On the other hand, the iterative solver with a complex shifted Laplace preconditioner performs better in presence of viscous effects. Before discussing the seismic imaging with the time and frequency domain approaches, I then present some results obtained with the complex shifted Laplace preconditioner and a BI-CGSTAB Krylov solver. The above complexity analysis indicates that the frequency-domain approach could be relevant with waveform tomography. For non-linear impedance waveform inversion, that requires the modeling of a large frequency band, the frequency-domain iterative solver could become competitive only when the number of iterations, N_{it}, is in $O(1)$, that is frequency independent. The direct solver is a-priori not competitive for non-linear impedance inversion.

3 An Iterative Solver for the Vertical Transverse Isotropy Visco-Acoustic Wave Equation

Solving iteratively the linear system (3) is challenging because the matrix A is indefinite with positive and negative eigenvalues. Over the years, different subspace methods have been tested with different preconditioners [18, 21, 35]. The complex shifted Laplace preconditioner shows an interesting behavior [21, 22] and has been successfully implemented to invert seismic data [38]. Here I describe its behavior with a BI-CGSTAB iterative solver [48] and a multi-grid approximation of the inverse of the preconditioner when solving the visco-acoustic vertical transversely isotropic (VTI) wave equations. Improvements to the approach, based notably on deflation and projection to tackle the issue of the eigenvalues close to zero [23] are not considered here.

Researchers in exploration geophysics have developed anisotropic wave equations in order to better approximate the kinematics of the P-waves without accounting for the shear velocities [3, 17]. In this way, the spatial discretization can be relatively coarse since the slow shear velocity are not considered. This leads to efficient schemes in seismic imaging from active data when we focus on the compressional waves. However, these acoustic anisotropic wave equations, obtained by zeroing the shear velocities, are not physical. In an elastic anisotropic medium, the kinematics of the P-waves depends on the shear velocities. The relevance of this acoustic approximation imposes weak anisotropy. Because of the natural layering in the earth crust, a common anisotropy is the vertical transverse isotropy (VTI). Under the VTI assumption, the second-order visco-acoustic wave equations read, for instance:

$$\begin{cases} -\frac{\omega^2}{\rho v_n^2(1-\imath/Q)}p_n - \partial_x\frac{1}{\rho}\partial_x p_h - \partial_y\frac{1}{\rho}\partial_y p_h - \frac{1}{\sqrt{1+2\delta}}\partial_x\frac{1}{\rho}\partial_x\frac{p_n}{\sqrt{1+2\delta}} = s_n; \\ -\frac{\omega^2}{\rho v_n^2(1-\imath/Q)}p_h - (1+2\eta)\partial_x\frac{1}{\rho}\partial_x p_h - (1+2\eta)\partial_y\frac{1}{\rho}\partial_y p_h - \frac{1}{\sqrt{1+2\delta}}\partial_x\frac{1}{\rho}\partial_x\frac{p_n}{\sqrt{1+2\delta}} = s_h; \end{cases} \tag{8}$$

with v_n the normal moveout (NMO) velocity, η the anelliptic parameter, δ the stretched parameter, ρ the density, $p_n = -\sqrt{1+2\delta}\,\sigma_{xx} = -\sqrt{1+2\delta}\,\sigma_{yy}$ and $p_h = -\sigma_{zz}$ the 'NMO' and 'horizontal' pressures and s_n and s_h the source terms. From the stiffness coefficients, we have $1+2\varepsilon = c_{11}/c_{33}$, $\sqrt{1+2\delta} = c_{13}/c_{33}$, $\eta = (\varepsilon-\delta)/(1+2\delta)$ and $v_n = \sqrt{1+2\delta}\sqrt{c_{33}/\rho}$. The parameter Q is the quality factor. The pure isotropic case corresponds to $Q = \infty$. Here, we consider only one Q factor, which means that the viscous effect is assumed similar for the NMO and horizontal velocities. These equations are completed with standard boundary conditions, for instance absorbing conditions with the perfectly matched layer (pml) conditions [8] and free-surface conditions on the top of the earth model.

The spatial discretization of this system gives a non-symmetric discrete system A, Eq. (3), with the vector u containing the discretized elements of p_h and p_v and the vector f the discretized elements of s_h and s_v. Compared to the isotropic case, we lost the symmetry (which is a consequence of working with a non-physical wave equation).

The linear system associated to the discretization of Eq. (8) can be solved with a Krylov subspace iterative method [42] after preconditioning to speed up the convergence. With isotropic systems, the complex shifted preconditioner significantly improves the convergence [21]. With such a preconditioner, the number of iterations of the Krylov subspace solvers becomes roughly proportional to frequency [22, 38]. From a physical point of view this preconditioner corresponds to the discretization of a heavily damped wave equation. One obtains the equations for the preconditioner by replacing $1/(1-\imath/Q)$ by $\beta_r + \imath\beta_i$ in system (8) where β_r and β_i are the real and imaginary parts of the complex shift added to the equations.

After discretization, we obtain the preconditioning system:

$$\mathrm{Mu} = \mathrm{f}, \tag{9}$$

with u the unknown vector and f the source vector.

Now we aim to solve the preconditioned system, that is equivalent to system (3), with v an intermediate unknown vector:

$$\mathrm{AM}^{-1}\mathrm{v} = \mathrm{f}; \quad \text{with } \mathrm{Mu} = \mathrm{v}. \tag{10}$$

Solving system (9) exactly could be costly. With a sufficiently large β_i the system (9) becomes close to a diffusive system and one cycle of a multi-grid solver [12] leads to a fair approximation of M^{-1} [22, 38]. However, a too large β_i results in a poor preconditioner. In practice $\beta_r = 1$ and $\beta_i = 0.5$ gives a satisfactory compromise for the isotropic acoustic applications [22, 38]. This more or less corresponds to a quality factor of 2. I use the same choice in the VTI examples and solve the preconditioned system AM^{-1} with a BI-CGSTAB algorithm [48].

I discretize the spatial derivatives with standard high-order finite differences with only a few (3 to 4) points per wavelength and code up a matrix-free implementation. This is an advantage of this iterative solver approach over the direct solver where more dedicated compact stencils are developed to avoid large bandwidth and hence fill-in during the LU decomposition [34, 51]. In the examples, I use a 8th-order finite-difference scheme.

For the multi-grid solver, I follow the approach described in [38]. I approximately solve $\mathrm{Mu} = \mathrm{v}$ with one multigrid V-cycle. The smoother is one iteration of a Gauss-Seidel algorithm and the prolongation operator a trilinear interpolation. With the isotropic wave equation, I used the standard full-coarsening approach [12, 38]. I also test a line relaxation in the depth direction and a semi-coarsening that splits the coarsening in the horizontal plane and in the depth axis to evaluate whether this improves the convergence in the anisotropic case. However, we should remember that this approach requires the inversion of a tri-diagonal matrix because of the line relaxation and the coarsening is performed in two steps. I implemented the VTI case via a block matrix approach. At each discretization point of the grid, I lump the two unknowns, p_h and p_v. The tri-diagonal matrix for the line relaxation becomes in fact a 2-by-2 block tri-diagonal matrix.

To analyze the convergence of this iterative solver I generate the pressure responses in a land and a marine environment for different frequencies. In both examples, an explosive source is positioned 6 m below the free surface. On the other edges of the model absorbing boundary conditions are implemented. The models, Fig. 4, are derived from the SEG/EAGE overthrust model [4]. They are 20 km wide in both lateral direction and 4.5 km deep for the land example and 5 km deep for the marine one. For the marine example, I have added a 500 m water column. I define a η-field based on the velocity field. In the simulation, I took $\delta = 0$. For the viscous model, a constant Q value in the earth is used of either 50 and 100. In the marine case, the value of Q in the water layer is always very large (10^6) since the water layer hardly attenuates the seismic frequencies. During the computations, the grid sampling is adapted to frequency with 4 points per minimum wavelength in the marine example and 3.2 in the land one. For the land case, the minimum wavelength corresponds to a velocity of 2000 m/s and for the marine one to a velocity of 1500 m/s.

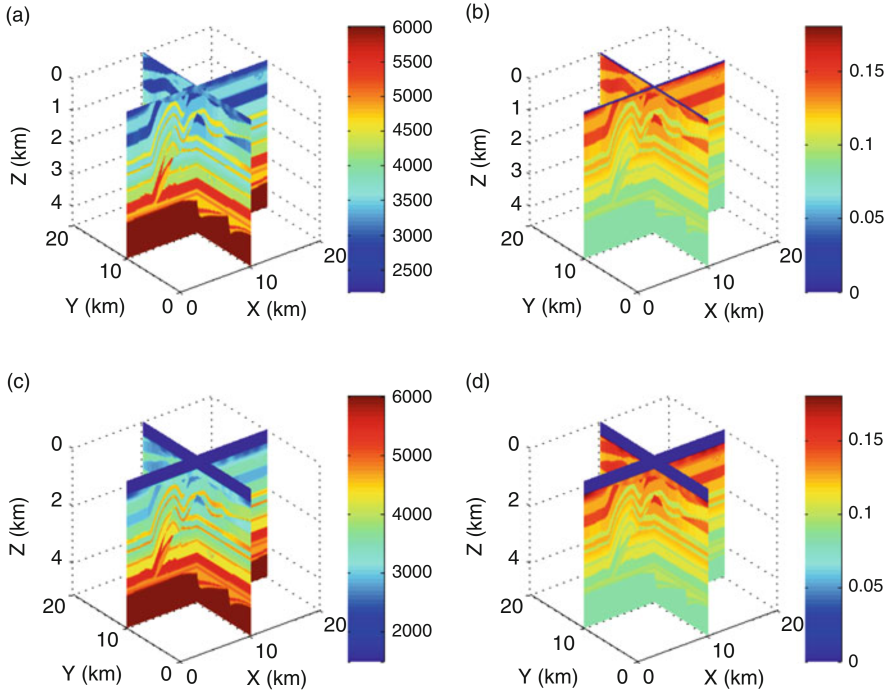


Fig. 4 Velocity and η-parameter for the land and marine examples. The two earth models are similar except for the 500 m water column in the marine case. (**a**) NMO velocity (land case). (**b**) η-Parameter (land case). (**c**) NMO velocity (marine case). (**d**) η-Parameter (water case)

In Figs. 5 and 6, the 'NMO' pressure fields at 2 and 8 Hz are displayed. In the marine case, a wave is channeled in the water, this corresponds to the multiple reflections between the free surface and the water bottom. These responses are obtained with a 10^{-3} stopping criterion on the normalized norm of the residual. The convergence history of the BI-CGSTAB algorithm is plotted in Fig. 7. For the isotropic and anisotropic computations in the land case and for the isotropic computations in the marine one a full-coarsening approach was used. However, for the anisotropic computations in the marine case a line-relaxation and semi-coarsening approach was used because it requires fewer iterations as discussed later. As expected, the convergence of the BI-CGSTAB is not monotonic. This could explain some of the oscillations in the number of iterations per frequency plots. Other Krylov subspace solvers, such that GMRES and IDR, may give different behaviors [9, 22, 42, 44].

To further analyze the convergence of the BI-CGSTAB, I model the responses for different frequencies between 2 and 8 Hz. The computational domain sizes are given in Table 1 for the land example and in Table 2 for the marine example. The number of unknowns is twice the number of grid points for the anisotropic computations because at each grid point we have two pressures, p_n and p_h.

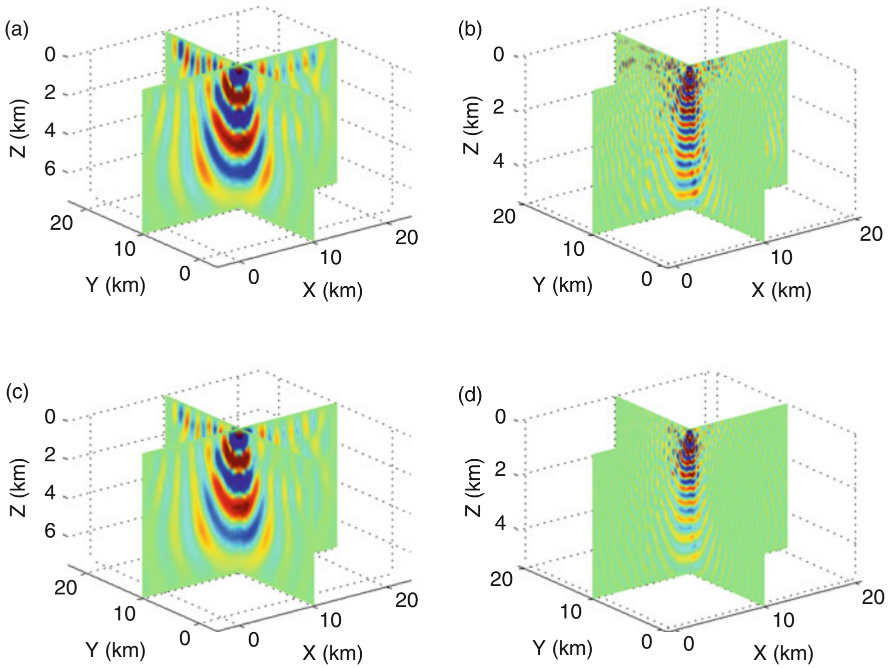


Fig. 5 Pressure fields at 2 and 8 Hz with an anisotropic ($Q = \infty$) and visco-anisotropic ($Q = 50$) wave equations in the land case. (**a**) Anisotropic, 2 Hz. (**b**) Anisotropic, 8 Hz. (**c**) Visco-anisotropic, 2 Hz. (**d**) Visco-anisotropic, 8 Hz

Figure 8 shows the number of BI-CGSTAB iterations required for convergence with a full-coarsening for the isotropic and anisotropic computations and with a semi-coarsening and a line relaxation for the anisotropic computations. The number of iterations more or less linearly increases with frequency when the computation grid is adapted to keep the number of points per wavelength fixed. The anisotropy or the change of coarsening do not change this behavior that was earlier reported for the isotropic case [22, 38]. One notices that the multiplicative constant, C, in the complexity analysis, that is the slope of the linear regression line, differs from one simulation to another. Moreover the behavior of the convergence of the iterative solver varies. Whereas in the land example, the use of a semi-coarsening and line-relaxation approach increases the number of iterations, in the marine case the number of iterations decreases. This result illustrates one of the challenges of this iterative solver in practical geophysical applications. Given an earth model, the number of required iterations to obtain the seismic responses is difficult to predict; hence the computational cost of the approach is not easily predictable. We should however mention that the shifted Laplace preconditioner gives a robust solver even in a VTI medium. In the marine example, the anisotropic runs require more iterations than the isotropic case. This is not the case in the land example. This behavioral difference may be a consequence of the channeled wave (that is the

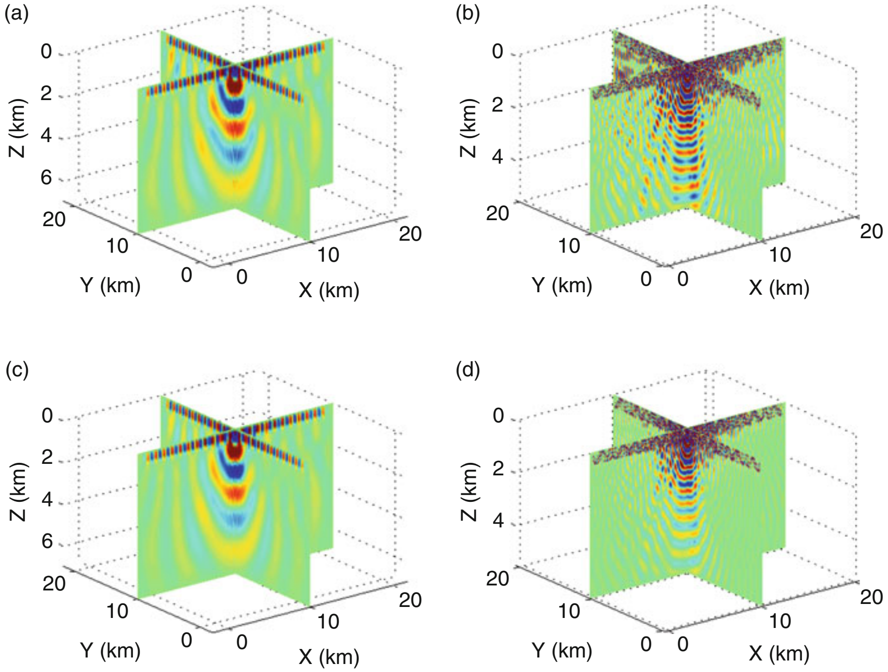


Fig. 6 'NMO' pressure fields at 2 and 8 Hz with an anisotropic ($Q = \infty$) and visco-anisotropic ($Q = 50$ except in the water) wave equations in the marine case. (**a**) Anisotropic, 2 Hz. (**b**) Anisotropic, 8 Hz. (**c**) Visco-anisotropic, 2 Hz. (**d**) Visco-anisotropic, 8 Hz

presence of the so-called water-bottom multiples) in the marine example. Indeed, in this example, there is a large earth model contrast at the water bottom. This means that most of the energy reflects on the water bottom and stays in the water column. The convergence rate of the Krylov subspace iterative solver decreases in this case. Although the wave in the water is not a standing wave because of the absorbing conditions on the lateral edges of the model, this behavior is somewhat similar. The shifted Laplace preconditioner is generally less efficient with Dirichlet boundary conditions (that are reflecting conditions) than with absorbing boundary conditions. The frequency dependency of the behavior of these partially channel waves may explain the non-monotonic increase of the number of BI-CGSTAB iterations versus frequency.

Computationally, the semi-coarsening and line-relaxation approach is not beneficial even in the marine example. Despite the reduction of the number of Bi-CGSTAB iterations, the cost per grid point remains higher than with the full-coarsening approach because of the significant increase of the computational cost per iteration, as illustrated in Fig. 9.

The number of BI-CGSTAB iterations versus frequency for the viscous applications are plotted in Figs. 10 and 11. The case $Q = 100$ corresponds to a relatively

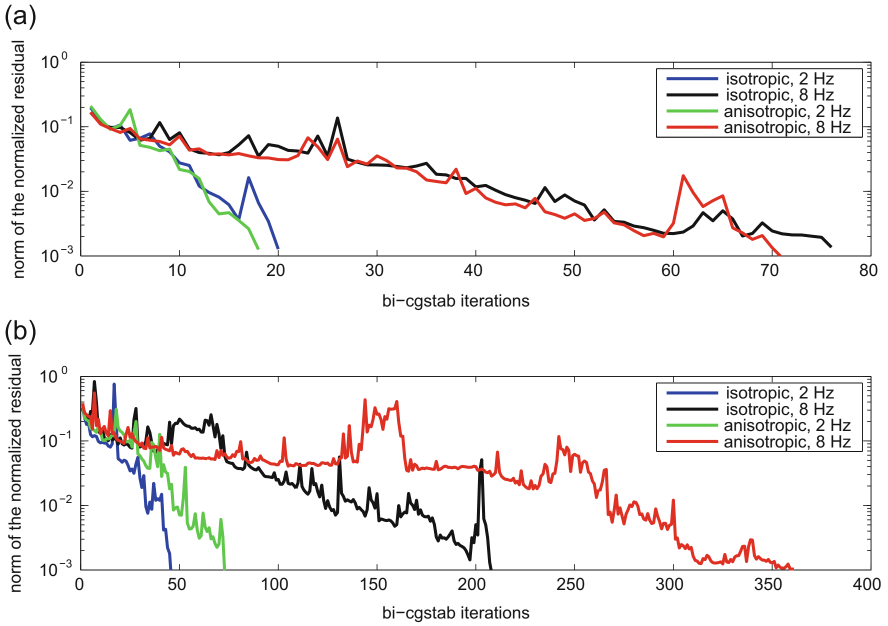


Fig. 7 Convergence history of the BI-CGSTAB algorithm. For the isotropic and anisotropic computations in the land case and for the isotropic computations in the marine case a full-coarsening approach was used. For the anisotropic computations in the marine case, a line relaxation and semi-coarsening approach was used. (**a**) Land case. (**b**) Marine case

Table 1 Number of grid points for the computations in the land case

Frequency (Hz)	2	3	4	5	6	7	8
Number of points in *x* or *y*	85	117	149	181	213	245	277
Number of points in *z*	25	33	40	47	55	62	69
Total number of points (in millions)	0.18	0.45	0.89	1.54	2.50	3.72	5.30

Table 2 Number of grid points for the computations in the marine case

Frequency (Hz)	2	3	4	5	6	7	8
Number of points in *x* or *y*	127	181	234	287	341	394	447
Number of points in *z*	38	51	65	79	92	106	119
Total number of points (in millions)	0.61	1.67	3.56	6.51	10.70	16.46	23.78

small attenuation and the case $Q = 50$ corresponds to a significant attenuation in crustal applications (although locally Q could be smaller, for instance due to gas accumulation). Again the two examples display different behaviors. In the land examples, the number of iterations significantly decreases with decreasing Q values as expected. In the marine example, because no viscous effects are modeled in the water, the number of iterations decreases less between the acoustic and visco-

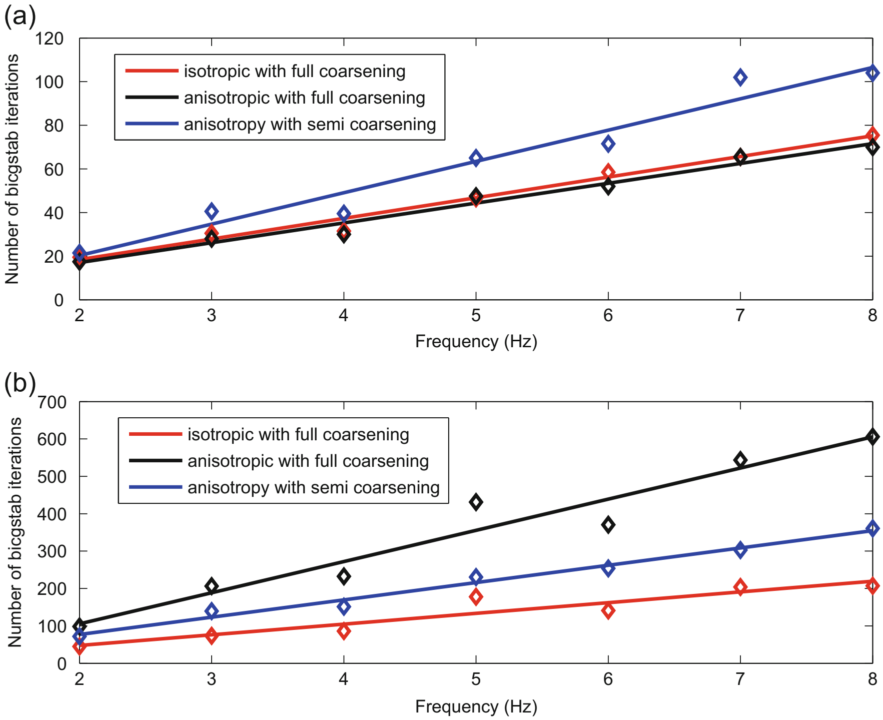


Fig. 8 Comparison between the full-coarsening approach and the line-relaxation and semi-coarsening approach. The *diamonds* correspond to the actual numbers and the *lines* correspond to the best linear fits through the points of a given type of modeling versus frequencies. (**a**) Land case. (**b**) Marine case

acoustic simulations. Again, the modeling of the waves in the water column seems to govern the convergence rate of the BI-CGSTAB iterative solver. In Tables 3 and 4 I give the slope coefficient of the linear regression that fits the number of iterations per frequency to quantify this observation.

These computational tests illustrate that the convergence of the BI-CGSTAB iterative solver with a shifted Laplace preconditioner is more or less linear in frequency when the number of points per wavelength is kept fixed under the viscous-acoustic and visco-anisotropic (VTI) assumption. The convergence rate depends not only on the type of equations but also on geology. The presence of channeled waves or strong reverberations can decrease the convergence rate as illustrated by the marine example.

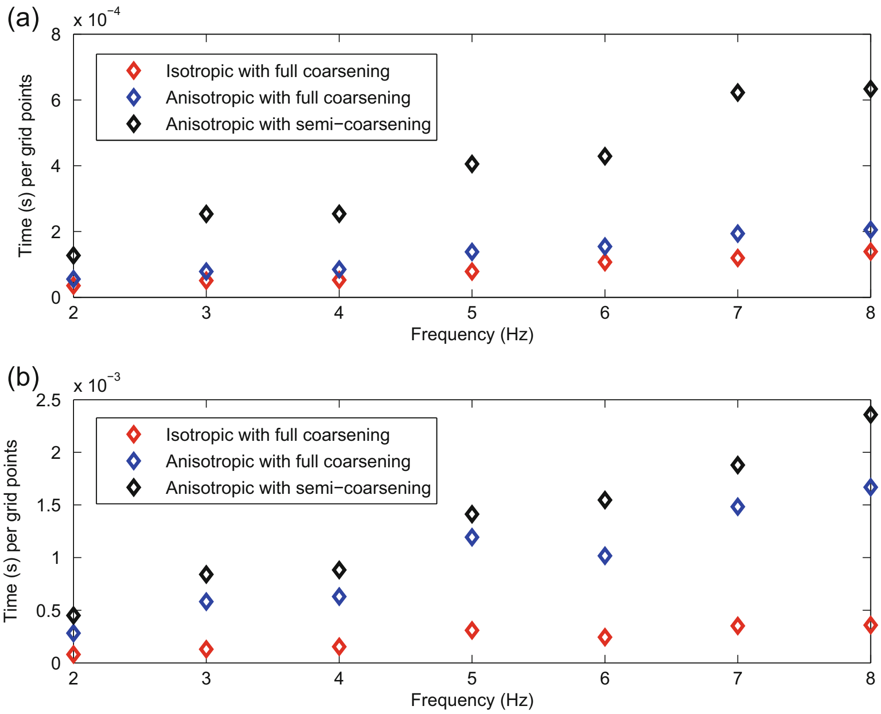


Fig. 9 Computation time per grid point comparison between the full-coarsening approach and the line-relaxation and semi-coarsening approach. The computational time is just indicative because it does depend on the system. (**a**) Land case. (**b**) Marine case

4 The Least-Square Seismic Imaging Problem

The goal of seismic imaging is to retrieve the earth medium properties, ρ, the density and c, the stiffness coefficients or the velocities. This problem can be formulated as an inverse problem which consists of minimizing a least-square data misfit. Because the source wavelet is often not perfectly known a match filter, $\alpha = (\alpha(\mathrm{x}_s, t))$, is added to the Earth parameters. This leads to the following data misfit function:

$$\begin{aligned} J(\rho, \mathrm{c}, \alpha) &= \tfrac{1}{2} \textstyle\int d\mathrm{x}_s d\mathrm{x}_r dt \\ &\times \left|\left|\mathrm{W}(\mathrm{x}_s, \mathrm{x}_r, t) \left(\textstyle\int dt' \, \alpha(\mathrm{x}_s, t-t')\mathrm{d}^{mod}(\mathrm{x}_s, \mathrm{x}_r, t') - \mathrm{d}^{obs}(\mathrm{x}_s, \mathrm{x}_r, t)\right)\right|\right|^2 \end{aligned} \tag{11}$$

with W a data weight matrix that selects certain events in the seismic traces. We shall call e the data residuals:

$$e(\mathrm{x}_s, \mathrm{x}_r, t) = \int dt' \, \alpha(\mathrm{x}_s, t-t')\mathrm{d}^{mod}(\mathrm{x}_s, \mathrm{x}_r, t') - \mathrm{d}^{obs}(\mathrm{x}_s, \mathrm{x}_r, t). \tag{12}$$

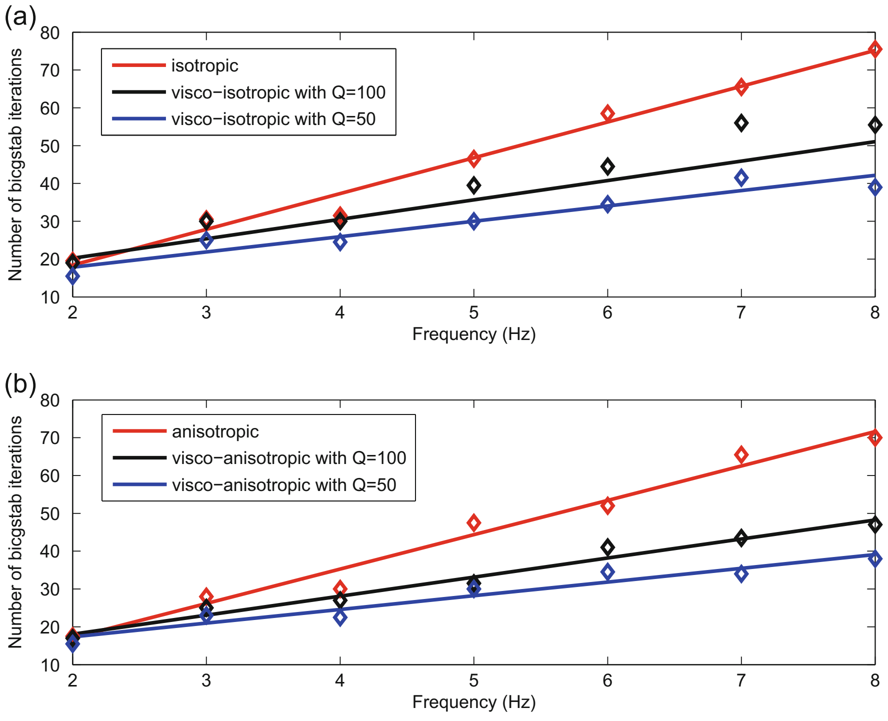


Fig. 10 Number of BI-CGSTAB iterations to compute the pressure responses for different frequencies with the land case. The *diamonds* correspond to the actual numbers and the *lines* correspond to the best linear fits through the points of a given type of modeling versus frequencies. For the isotropic and anisotropic runs a full-coarsening approach is used. (**a**) Isotropic land case. (**b**) Anisotropic land case

This formulation corresponds to the time-domain formulation. Regularization terms could be added and different norms could be used. However, in this paper, I focus only on the least-square misfit term.

An active seismic survey can contain thousands to hundred of thousands shots. Therefore, the minimization of the misfit function, J, requires a very large number of solutions of the elasto-dynamic equations. We may then ask ourselves what the most efficient way is to solve the elastic-dynamic system. Thanks to Parceval's equality, the misfit function, Eq. (11), reads in the frequency domain:

$$\begin{aligned} J(\rho, \mathrm{c}, \alpha) &= \frac{1}{2} \int d\mathrm{x}_s d\mathrm{x}_r d\omega \\ &\times \left|\left| \int d\omega' \, \mathrm{W}(\mathrm{x}_s, \mathrm{x}_r, \omega - \omega') \left(\alpha(\mathrm{x}_s, \omega') \mathrm{d}^{mod}(\mathrm{x}_s, \mathrm{x}_r, \omega') - \mathrm{d}^{obs}(\mathrm{x}_s, \mathrm{x}_r, \omega') \right) \right|\right|^2 . \end{aligned} \tag{13}$$

In discrete form, the equivalence between the time and frequency formulation is obtained when the time and frequency discretizations satisfy the Shannon-Nyquist theorem.

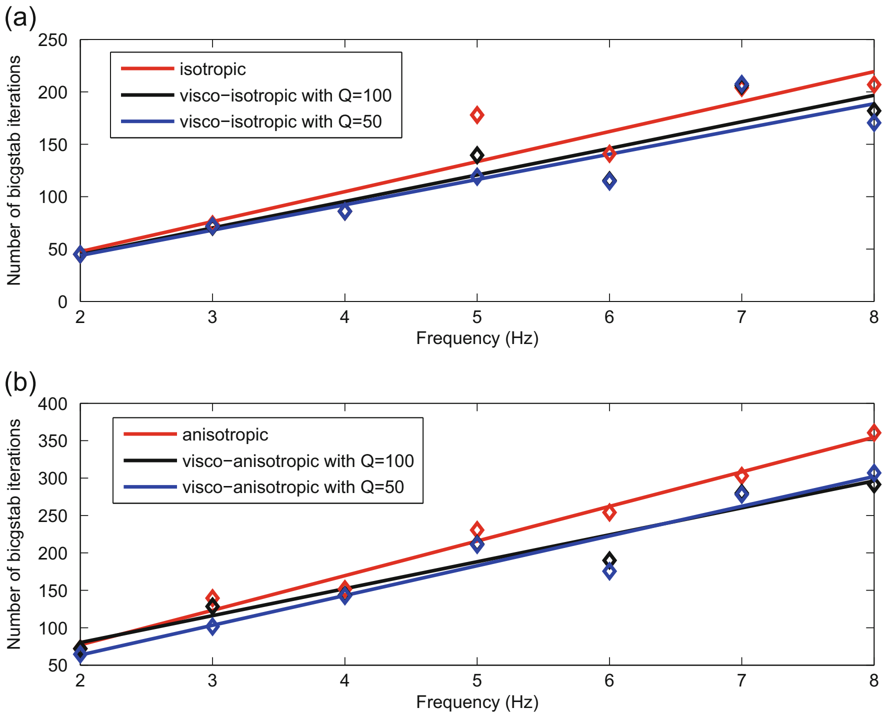


Fig. 11 Number of BI-CGSTAB iterations to compute the pressure responses for different frequencies with the marine case. The *diamonds* correspond to the actual numbers and the *lines* correspond to the best linear fits through the points of a given type of modeling versus frequencies. For the isotropic runs a full-coarsening approach is used and for the anisotropic runs a line-relaxation and semi-coarsening approach. (**a**) Isotropic marine case. (**b**) Anisotropic marine case

Table 3 Slope coefficient of the best linear fits with the land example

Modeling type	iso full	viso100 full	viso50 full	ani full	vani100 full	vani50 full	ani semi
Slope coefficient	9.46	5.14	4.05	9.09	5.04	3.62	14.39

In the modeling type denomination, iso stand for isotropic, ani for anisotropic, viso for visco-isotropic, vani for visco-anisotropic, full for full-coarsening, semi for semi-coarsening, 100 for Q=100 and 50 for Q=50

Table 4 Slope coefficient of the best linear fits with the marine example

Modeling type	iso full	viso100 full	viso50 full	ani semi	vani100 semi	vani50 semi	ani full
Slope coefficient	28.64	25.33	24.15	46.25	36	39.72	83.38

In the modeling type denomination, iso stand for isotropic, ani for anisotropic, viso for visco-isotropic, vani for visco-anisotropic, full for full-coarsening, semi for semi-coarsening, 100 for Q=100 and 50 for Q=50

The presence of a general time-dependent data weight function in Eq. (11) causes a mixing of frequencies in Eq. (13). This considerably reduces the attractiveness of the frequency formulation. The opposite would happen if we would consider a frequency-dependent data weight per receiver. In this case, a time convolution between the weights and the data residuals would be required in the time-domain misfit function, Eq. (11). This convolution would reduce to a multiplication in the frequency-domain formulation. However, there is a major difference. When solving the time-domain wave equations, we automatically compute the response from time 0 to the maximum recording time and we can perform the convolution without extra solutions of the wave equations. This explains why the estimation of the match filter in the time-domain formulation does not significantly increase the computational time, especially when the match filter length remains small with respect to the seismic trace length. In the frequency domain, we solve each frequency separately and the complexity analysis of the frequency-domain solvers shows that they can be advantageous only if we can formulate the inverse problem per frequency too. This implies that the data weights should be taken independently of time, except in the particular case of $W(x_s, x_r, t) = \exp(-s(t - t_0(\mathrm{x}_s, x_r)))$ with s a real positive number and t_0 a time depending on the source and receiver positions. In this case we have:

$$\begin{aligned} W(\mathrm{x}_s, \mathrm{x}_r, t)e(\mathrm{x}_s, \mathrm{x}_r, t) &= \frac{1}{\sqrt{2\pi}} \int d\omega \, \exp(-s(t - t_0(\mathrm{x}_s, x_r))e(\mathrm{x}_s, \mathrm{x}_r, t) \exp(\imath\omega t) \\ &= \frac{1}{\sqrt{2\pi}} \exp(t_0(\mathrm{x}_s, x_r)) \int d\omega \, e(\mathrm{x}_s, \mathrm{x}_r, t) \exp(\imath(\omega + \imath s)t). \end{aligned} \tag{14}$$

This corresponds to a Laplace-Fourier transform of the residuals [11, 43]. One directly obtains the Laplace-Fourier transform response by replacing the real frequency ω by a complex one, $\omega + \imath s$, in the forward system (8). Using a complex frequency adds an overall damping factor to the wave equations. The effect is similar to the complex shift of the shifted Laplace preconditioner. Therefore, the iterative solver should be quite efficient with large s values as illustrated in Table 5. With a large s value, the choice $\beta_i = 0.5$ is not optimal. For instance, with $s = 10$, we could take $\beta_i = 0$ and obtain a faster convergence (notably when we change the stopping

Table 5 Number of BI-CGSTAB iterations to compute the Laplace-Fourier response of the marine example with the full-coarsening approach and $\beta_r = 1$ and $\beta_i = 0.5$

Frequency (Hz)	2	3	4	5	6	7	8
Isotropic with s=10	1	1	2	2	2	2	2
Isotropic with s=1	3	4	5	6	6	7	8
Isotropic wit s=0.1	16	22	31	39	43	50	61
Isotropic with s=0	45	74	86	178	141	204	207
Anisotropic with s = 10	1	1	2	2	2	2	2
Anisotropic with s=1	3	4	5	6	6	7	8
Anisotropic with s=0.1	23	34	55	61	73	96	104
Anisotropic with s=0	99	207	233	431	371	554	606

Table 6 Number of grid points for the time-domain computations in the land case

Frequency (Hz)	2	3	4	5	6	7	8
Number of points in x or y	120	161	203	244	286	328	369
Number of points in z	41	50	60	69	79	88	98
Total number of points (in millions)	0.59	1.30	2.47	4.11	6.47	9.47	13.34

Table 7 Number of grid points for the time-domain computations in the marine case

Frequency (Hz)	2	3	4	5	6	7	8
Number of points in x or y	175	244	314	383	452	522	591
Number of points in z	57	75	92	110	128	145	163
Total number of points (in millions)	1.75	4.47	9.08	16.14	26.16	39.51	56.93

criteria to 10^{-6} to better modeling the large dynamic range of the Laplace-Fourier solution). Indeed, with a large s value, the Laplace-Fourier wave equation behaves like a diffusive equation that can be solved with a multi-grid solver similarly to the diffusive electromagnetic equations [6, 30, 37].

The frequency-domain responses can also be computed by Fourier transform of the time responses. In the context of seismic imaging, we can either model the responses with a sinusoidal source wavelet or we can use a large frequency band source wavelet [32]. The latter provides in fact all the frequencies up to a maximum frequency which could be advantageous in seismic imaging. To compare the efficiency of the iterative solver with the time-domain formulation I simulate the time-domain responses with a flat spectrum source with a high frequency cut taper. The size of the taper is 30 % of the nominal frequency value. This for instance means that at 2 Hz the maximum frequency supported by the grid is 2.6 Hz and at 8 Hz, 10.4 Hz. With this choice, the discretization grid for the time-domain simulations, Tables 6 and 7, are larger than the one for the frequency-domain ones, Tables 1 and 2. I also choose a maximum recording time of 12 s. Different tapers and maximum recording times would change the computational time of the time-domain approach. Here, I have tried to choose realistic values without optimizing them. This should give an indication of how the computational costs of the time and frequency domain formulations compare.

Figures 12 and 13 display the computation times versus frequency of the different simulations on one core. These computation times are just indicative because they depend on the computer architecture, the load of the machine and the code optimization. Accounting for the viscous effects in the time domain formulation increases the computation time since the number of operations per time step increases due to the additional memory variable equations. Moreover the stability condition decreases with decreasing Q values, hence the number of time steps increases. This explains why the computation time increases with decreasing Q values. The situation is reverse with the frequency-domain formulation since accounting for the viscous effects decreases the number of BI-CGSTAB iterations. With the land example, Fig. 12, the frequency-domain viscous simulations become

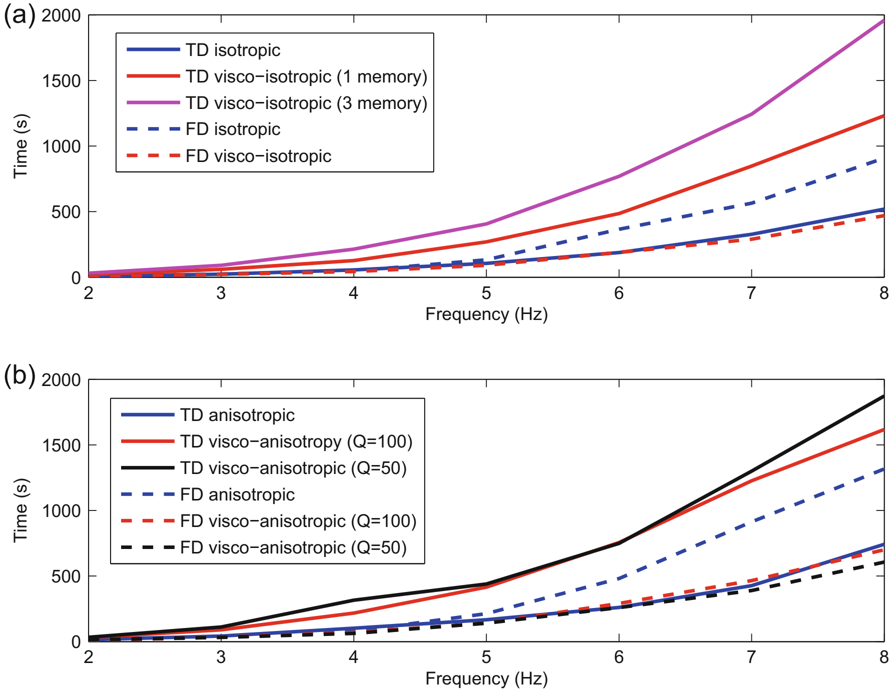


Fig. 12 Computation time versus frequency to simulate the seismic responses of the land example with the time and frequency domain formulations. In the viscous simulation the quality factor is equal to 100 for the isotropic case. For the time-domain visco-isotropic simulations, I used one or three memory variables and for the time-domain visco-anisotropic simulations, I used one memory variable. (**a**) Isotropic land example. (**b**) Anisotropic land example

significantly faster than the time-domain ones while it was not the case for the pure isotropic simulations. With the marine example, Fig. 13, the situation is different. On one hand, the computational times of frequency-domain visco-isotropic simulations are similar to the ones of the time-domain simulations with one memory variable and shorter than the one of the time-domain simulations with three memory variables. On another hand, the frequency-domain visco-anisotropic simulations remain more expensive than the time-domain ones. The presence of the non-viscous water layer and the multiple reflections at the water bottom reduce the efficiency of the iterative solver as already discussed. The ratio of the frequency-domain computation time divided by the time-domain computation time for different simulation types are plotted in Fig. 14. For most of the cases, the ratio is between 0.5 and 2 meaning that the multiplicative constant, C, of the complexity analysis of the two approaches is roughly similar. However with the marine case, the anisotropic simulation is about one order of magnitude more expensive with the frequency-domain iterative solver than with the time-domain approach. In this case, the multiplication constant, C, of the complexity analysis is almost 10 times higher. With the iterative solver, the

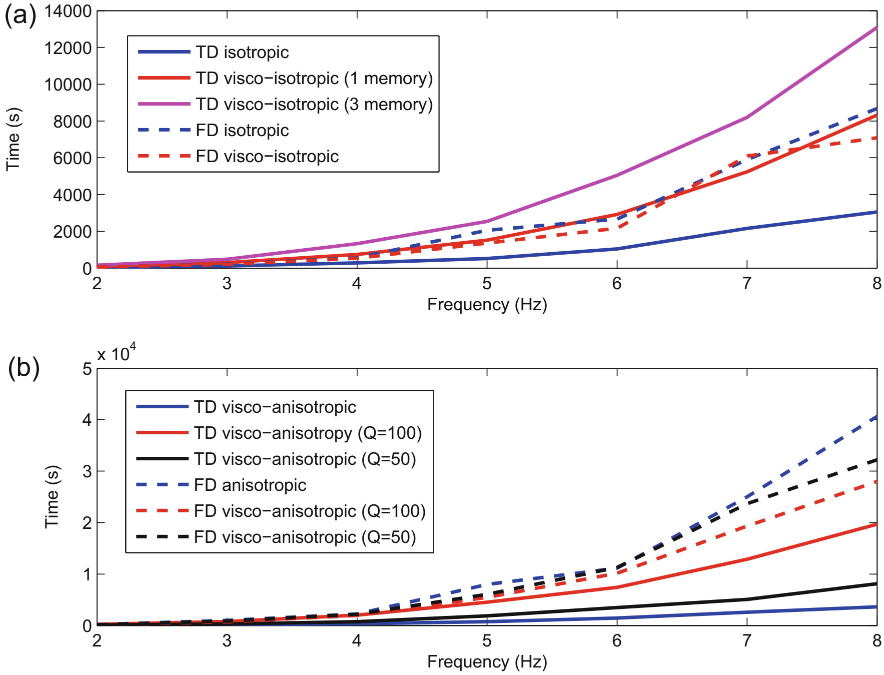


Fig. 13 Computation time versus frequency to simulate the seismic responses of the land example with the time and frequency domain formulations. In the viscous simulation the quality factor is equal to 50 for the isotropic case. For the time-domain visco-isotropic simulations, I used one or three memory variables and for the time-domain visco-anisotropic simulations, I used one memory variable. (**a**) Isotropic marine example. (**b**) Anisotropic marine example

multiplicative constant of the complexity analysis significantly depends on the earth model. When we consider the Laplace-Fourier approach with s values between 0.1 and 10, the iterative solver is significantly faster than the time-domain approach when s is large, Fig. 15.

From a simulation point of view, these results illustrate that the frequency-domain iterative solver could be an alternative to the time-domain approach only when a few frequency responses are required that is with waveform tomography [38, 41, 49], especially when viscous effects are modeled. Nonetheless, the number of frequencies required to properly carry out a multi-parameter inversion remains a question, which can challenge the use of the iterative solver, except when a Laplace-Fourier approach is chosen [43]. With migration and non-linear impedance waveform inversion, the time-domain approach remains the preferred option in 3D since the full band frequency is required to obtain a high resolution image from the reflected waves.

Although the simulation governs most of the complexity aspect of an imaging problem, other aspects, related to the minimization of the misfit function J, Eq. (11) or (13), may influence our choice of modeling algorithms. One of them is the

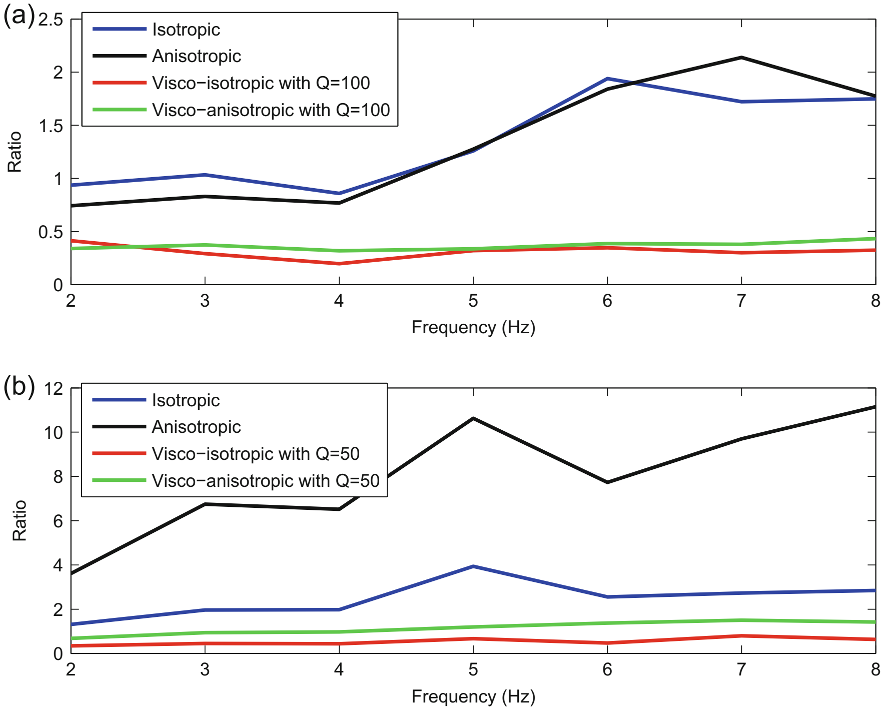


Fig. 14 Ratio of the frequency-domain computation times with the time-domain computation time for different simulations. (**a**) Land example. (**b**) Marine example

memory requirement. Because of the numerical cost of solving the wave equations, the minimization of J is carried out with a local optimization technique. The gradient is efficiently obtained through the adjoint state technique [14, 36].

In the time-domain formulation, we first solve the adjoint equations with the initial conditions $\mathrm{v}(\mathrm{x}_s, \mathrm{x}_r, t_{-2}) = \mathrm{v}(\mathrm{x}_s, \mathrm{x}_r, t_{-1}) = 0$:

$$\mathrm{v}(\mathrm{x}_s, t_n) = \mathrm{B}_1^T \mathrm{v}(\mathrm{x}_s, t_{n-1}) + \mathrm{B}_2^T \mathrm{v}(\mathrm{x}_s, t_{n-2}) - \sum_{x_r} \mathrm{S}^T(\mathrm{x}_s, \mathrm{x}_r)\mu(\mathrm{x}_s, \mathrm{x}_r, t_{N-n}), \qquad (15)$$

with v the adjoint state variables, $T = t_N$ the maximum time, and $\mu(\mathrm{x}_s, \mathrm{x}_r, t_n) = \frac{\partial J}{\partial d^{mod}(\mathrm{x}_s, \mathrm{x}_r, t_n)}$ and ^{T} the transpose operator. I have written an equation with initial boundary conditions because I have reversed the time of the adjoint source μ. The gradient of the misfit function with the time-domain formulation reads:

$$\frac{\partial J}{\partial m_k} = -\sum_{x_s} \sum_{n=0}^{N} \mathrm{v}^T(\mathrm{x}_s, t_n) \left(\frac{\partial \mathrm{B}_1}{\partial m_k} \mathrm{u}(\mathrm{x}_s, t_{n-1}) + \frac{\partial \mathrm{B}_2}{\partial m_k} \mathrm{u}(\mathrm{x}_s, t_{n-2}) \right), \qquad (16)$$

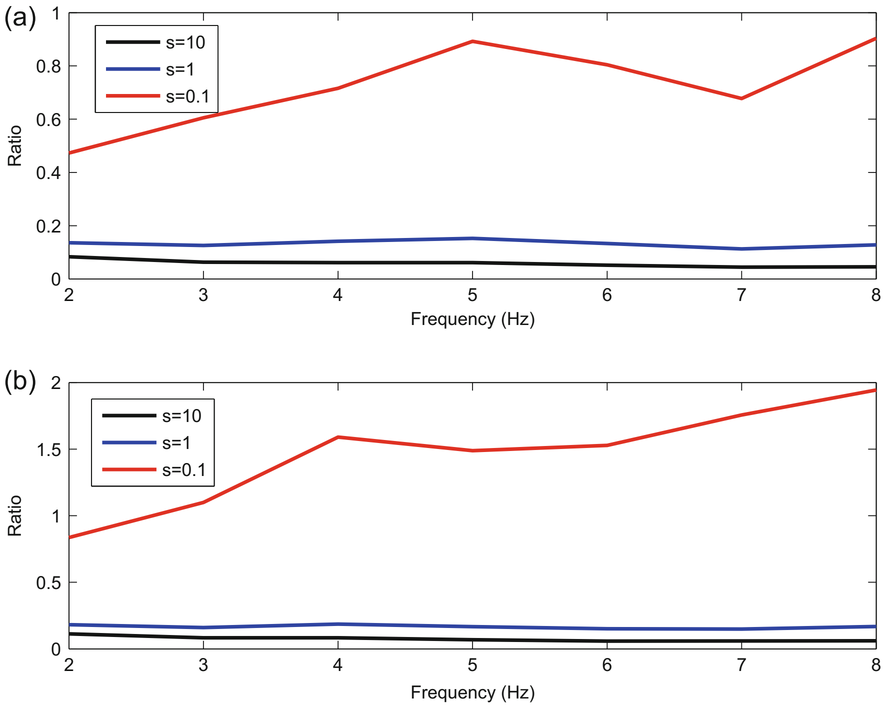


Fig. 15 Ratio of the frequency-domain computation times with the time-domain computation time for different values of the s parameter in the Laplace-Fourier transform. (**a**) Isotropic marine example. (**b**) Anisotropic marine example

with m_k one of the model parameters describing ρ or c.

To compute the adjoint state variables, $\mathrm{v}(\mathrm{x}_s, t_n)$, all the state variables, $\mathrm{u}(\mathrm{x}_s, t_n)$, have to be computed for the time of the adjoint source has been reversed. This leads to an implementation challenge for the time-domain formulation because in 3D one cannot store all the state variables in core memory. In practice, the states are either saved on disk or a check-pointing approach is implemented that requires additional computation [5, 26]. With the non-viscous simulation, the states can be recomputed backwards after having saved the values of the state variables on the boundaries during the forward modeling [16]. A crude estimation of the computation time of the gradient in the time-domain formulation is then about three times the one of the forward system, since solving of adjoint/backward equations is about as expensive as solving the state/forward equations.

In the frequency domain, the adjoint equations read:

$$\mathrm{A}^*(\omega)\mathrm{v}(\mathrm{x}_s,\omega) = \sum_{x_r} \mathrm{S}^T(\mathrm{x}_s,\mathrm{x}_r)\mu(\mathrm{x}_s,\mathrm{x}_r,\omega), \tag{17}$$

with v the adjoint state variables (I again abuse the notation and reuse the same symbols), $\mu(\mathrm{x}_s, \mathrm{x}_r, \omega) = \frac{\partial J}{\partial d^{mod}(\mathrm{x}_s, \mathrm{x}_r, \omega)}$ and * the complex conjugate.

The gradient of the misfit function with the frequency-domain formulation reads:

$$\frac{\partial J}{\partial m_k} = -\sum_{x_s} \sum_{\omega} \mathrm{v}^*(\mathrm{x}_s, \omega) \frac{\partial \mathrm{A}(\omega)}{\partial m_k} \mathrm{u}(\mathrm{x}_s, \omega). \tag{18}$$

The frequencies can be processed sequentially with frequency independent data weights. The state variables, $\mathrm{u}(\mathrm{x}_s, \omega)$, can be stored in memory. Solving the system with A^* is roughly equivalent to solving the system with A. A crude estimation of the computation time of the gradient in the frequency domain formulation is twice the one of the forward system.

The factor 2/3 in favor of the frequency-domain modeling does not significantly change the conclusions on the choice of the modeling domain. Nonetheless it increases the attractiveness of the frequency domain approach when viscous effects are considered. The memory requirement may even be more in favor of a frequency-domain formulation when one uses truncated Newton optimization [20, 29].

5 Conclusions

A Krylov subspace iterative solver with a complex shifted Laplace preconditioner gives a robust frequency domain solution of the vertical transverse isotropic visco-acoustic wave equations. Based on two numerical examples, I have shown that in the context of waveform tomography, this iterative solver could be an alternative to standard time-domain schemes, notably when viscous effects are taken into account or when using data weights that exponentially damp the seismic traces with time allowing us to apply a Laplace-Fourier transform of the data residuals. The convergence rate of the iterative solver however suffers in presence of wave guides although in the examples it stays proportional to frequency. This means that the number of iterations to reach converge does depend on the geological structures. Though the number of the iterations increases more or less linearly with frequency independently of the earth model, the multiplicative constant in the complexity analysis significantly depends on the structural features of the earth model. The time-domain approach hence remains the most flexible and predictable approach. It can serve for both waveform tomography and non-linear impedance waveform inversion because a complete frequency band response is computed at once. Nevertheless, when considering the memory requirements, together with the development on the preconditioner to limit the influence of the near-zero eigenvalues, the iterative solver could potentially supersede the time-domain approach if the frequency dependency of the number of iterations is reduced and the solver is efficient with anisotropic wave equations. This still needs to be proven with real multi-parameter waveform inversions when one needs to simultaneously invert several, let us say five to ten, frequencies at a time to obtain a reliable earth model.

References

1. Achenbach, J.: Wave Propagation in Elastic Solids, North-Holland (1973).
2. Aki, K, Richards, P.: Quantitative Seismology, Vol. I, Freeman & Co (1980).
3. Alkhalifah, T.: An acoustic wave equation for anisotropic media: Geophysics, **65**, 1239–1250 (2000).
4. Aminzadeh, F., Brac, J., Kunz, T.: 3-D salt and overthrust models, SEG/EAGE 3-D Modeling Series no.1, SEG (1997).
5. Anderson, J.E., Tan, L., Wang, D.: Time-reversal checkpointing methods for RTM and FWI, Geophysics, **77**, S93–S103 (2012).
6. Aruliah, D. A., Ascher, U. A.: Multigrid preconditioning for Krylov methods for time-harmonic Maxwells equations in 3D, SIAM J. Sci. Comput., **24**, 702–18 (2003).
7. Bamberger, A., Chavent, G., Hemon, C., Lailly, P.: Inversion of normal incidence seismograms, Geophysics, 47, 757–770 (1982).
8. Bérenger, J.-P.: A perfectly matched layer for absorption of electromagnetic waves, J. Comput. Phys., **114**, 185–200 (1994).
9. Baumann, M., van Gijzen, M.B.: Nested Krylov Methods for shifted linear systems, SIAM Journal on Scientific Computing, **37**, S90–S112 (2015).
10. Blanch, J., Robertson, J.O.A., Symes, W.W.: Modeling of a constant Q: methodology and algorithm for an efficient and optimally inexpensive viscoelastic technique, Geophysics, **60**, 176–184 (1995).
11. Brenders A.J., Pratt, R.G.: Full waveform tomography for lithospheric imaging: results from a blind test in a realistic crustal model, Geophysical Journal International, **168**, 133–151 (2007).
12. Briggs, W.L., Henson, V.E., McCormick, S.F.: A multigrid tutorial, 2nd ed., SIAM (2000).
13. Carcione, J.M.: Wave fields in real media:Wave propagation in anisotropic, anelastic and porous media: Pergamon Press (2011).
14. Chavent G.: Nonlinear least squares for inverse problems: theoretical foundations and step-by-step guide for applications, Springer (2009).
15. Christensen, R.M.: Theory of viscoelasticity - An introduction, Academic Press Inc (1982).
16. Clapp, R. G.: Reverse time migration with random boundaries, 79th Annual International Meeting, SEG, Expanded Abstracts, 2809–2813 (2009).
17. Duveneck, E., Milcik, P., Bakker, P.M., Perkins, C.: Acoustic VTI wave equations and their applications for anisotropic reverse-time migration: 78th Annual International Meeting, SEG, Expanded Abstract, 2186–2189 (2008).
18. Elman, H., Ernst, O., O' Leary, D.: A multigrid based preconditioner for heterogeneous Helmholtz equation, SIAM Journal on Scientific Computing, **23**, 1291–1315 (2001).
19. Engquist, B., Ying, L.: Sweeping preconditioner for the Helmholtz equation; Hierarchical matrix representation, Communications on Pure and Applied Mathematics, LXIV, 0697–0735 (2011).
20. Epanomeritakis, I., Akçelik V., Ghatta,s O., Bielak, J.: A Newton-CG method for large-scale three- dimensional elastic full waveform seismic inversion, Inverse Problems, **24**, 1–26 (2008).
21. Erlangga, Y.A., Vuik, C., Oosterlee, C.: On a class of preconditioners for the Helmholtz equation, Applied Numerical Mathematics, **50**, 409–425 (2004).
22. Erlangga, Y.A., Vuik, C., Oosterlee, C.: A novel multigrid based preconditioner for heterogeneous Helmholtz problems, SIAM Journal on Scientific Computing, **27**, 1471–1492 (2006).
23. Erlangga, Y.A., Nabben, R.: On a multilevel Krylov method for the Helmholtz equation preconditioned by shifted laplacian, Electronic transactions on Numerical Analysis, **31**, 408–424 (2008).

24. Gauthier O., Virieux J., Tarantola A.: Two-dimensional nonlinear inversion of seismic waveform: numerical results. Geophysics 51, 1387–1403 (1986).
25. Gordon, D., Gordon, R.: Robust and highly scalable parallel solution of the Helmholtz equation with large wave numbers, Journal of computation and applied mathematics, **237**, 182–196 (2013).
26. Griewank, A., Walther, A.: Algorithm 799: An implementation of checkpointing for the reverse or adjoint mode of computational differentiation: ACM Transactions on Mathematical Software, **26**(1), 19–45 (2000).
27. Lailly, P.: The seismic inverse problem as a sequence of before stack migrations: Conference on Inverse Scattering, Theory and Application, Society of Industrial and Applied Mathematics, Expanded Abstracts, 206–220 (1983).
28. Marfurt, K.J.: Accuracy of finite-difference and finite-element modeling of the scalar and elastic wave equations. Geophysics, **49**, 533–549 (1984).
29. Métivier L., Brossier, R., Virieux, J., Operto, S.: Full Waveform Inversion and the Truncated Newton Method, SIAM J. Sci. Comput., **35**, B401–437 (2013).
30. Mulder, W. A.: A multigrid solver for 3D electromagnetic diffusion Geophys. Prosp., **54**, 633–649 (2006).
31. Mulder, W.A., Plessix, R.-É.: Exploring some issues in acoustic full waveform inversion, Geophysical Prospecting, **56**, 827–841 (2008).
32. Nihei, K.T., Li, X.: Frequency response modelling of seismic waves using finite difference time domain with phase sensitive detection (TDPSD), Geophysical Journal International, **169**, 1069–1078 (2006).
33. Operto, S., Virieux, J., Amestoy, P., L'Excellent, J.Y., Giraud, L.: 3D finite-difference frequency-domain modeling of viscoacoustic wave propagation using a massively parallel direct solver: a feasibility study, Geophysics, **72**, SM195–SM211 (2007).
34. Operto, S., Brossier, R., Combe, L., Métivier, L.,Ribodetti, A., Virieux,J., 2014. Computationally-efficient three-dimensional visco-acoustic finite difference frequency-domain seismic modeling in vertical transversely isotropic media with sparse direct solver, Geophysics, **79**,T257–T275 (2014).
35. Plessix, R.-É, Mulder, W.A.: Separation of variables as a preconditioner for an iterative Helmholtz solver, Applied Numerical Mathematics, **44**, 385–400 (2003).
36. Plessix, R.-É.: A review of the adjoint-state method for computing the gradient of a functional with geophysical applications, Geophys. J. Int., **167**, 495–503 (2006).
37. Plessix, R.-É., Darnet, M,Mulder, W. A.: An approach for 3D multi-source, multi-frequency CSEM modeling Geophysics **72** SM177–84 (2007).
38. Plessix, R.-É.: A Helmholtz iterative solver for 3D seismic-imaging problems, Geophysics, **72**, SM185–SM194 (2007).
39. Plessix, R.-É.: Three-dimensional frequency-domain full-waveform inversion with an iterative solver, Geophysics, **74**,WCC149–WCC157 (2009).
40. Plessix, R.-É., Perkins, C.: Full waveform inversion of a deep water ocean bottom dataset. First Break **28**, 71–78 (2010).
41. Pratt, R. G., Song, Z.M., Williamson, P.R., Warner, M.: Two-dimensional velocity model from wide-angle seismic data by wavefield inversion: Geophysical Journal International, **124**, 323–340 (1996).
42. Saad, Y.: Iterative methods for linear systems, Second edition, SIAM (2003).
43. Shin, C., Cha, Y.H.: Waveform inversion in the Laplace-Fourier domain, Geophysical Journal International, **171**, 1067–1079 (2009).
44. Sonneveld, P., van Gijzen, M.B.: IDR(s): A family of simple and fast algorithms for solving large nonsymmetric systems of linear equations, SIAM Journal on Scientific Computing, **31**, 1035–1062 (2008).
45. Tarantola, A.: Inversion of seismic reflection data in the acoustic approximation, Geophysics, **49**, 1259–1266 (1984).
46. Tarantola A.: Inverse Problem Theory, Elsevier (1987).

47. Tarantola, A.: Theoretical background for the inversion of seismic waveforms, including elasticity and attenuation, Pure Appl. Geophys., **128**, 365–399 (1988).
48. van der Vorst, H.A.: Bi-CGSTAB: A fast and smoothly converging variant of Bi-CG for solution of nonsymmetric linear systems, SIAM J. Sci. Stat. Comput., **13**, 631–644 (1992).
49. Virieux, J., Operto, S.: An overview of full waveform inversion in exploration geophysics, Geophysics, **74**, WCC1–WCC2 (2009).
50. Virieux, J, Calandra A., Plessix, R.-É.: A review of the spectral, pseudo-spectral, finite-difference and finite-element modelling techniques for geophysical imaging, Geophysical Prospecting, **59**, 794–813 (2011).
51. Wang, S., de Hoop, M.V., Xia, J.: On 3D modeling of seismic wave propagation via a structured parallel multifrontal direct Helmholtz solver, Geophys. Prospect., **59**, 857–873 (2011).
52. Yilmaz, O., 2001. Seismic Data Analysis, SEG.

Optimized Schwarz Domain Decomposition Methods for Scalar and Vector Helmholtz Equations

X. Antoine and C. Geuzaine

Abstract In this chapter we review Schwarz domain decomposition methods for scalar and vector Helmholtz equations, with a focus on the choice of the associated transmission conditions between the subdomains. The methods are analyzed in both acoustic and electromagnetic settings, and generic weak formulations directly amenable to finite element discretization are presented. An open source solver along with ready-to-use examples is freely available online for further testing.

1 Introduction

Solving high-frequency time-harmonic wave problems is a very challenging problem, encountered in many physical applications, from acoustic noise propagation to seismology and geophysical exploration to electromagnetic radiation. Among the various approaches for numerical simulation, the Finite Element Method (FEM) with an Absorbing Boundary Condition (ABC) or a Perfectly Matched Layer (PML) is well suited for tackling complex geometrical configurations and heterogeneous media. The brute-force application of the FEM in the high-frequency regime however requires the solution of extremely large, complex-valued and possibly indefinite linear systems [39]. Direct sparse solvers do not scale well for such large-size problems, and Krylov subspace iterative solvers exhibit slow convergence or diverge, while efficiently preconditioning proves difficult [24]. Domain decomposition methods provide an alternative, iterating between subproblems of smaller sizes, amenable to sparse direct solvers [49].

X. Antoine
Institut Elie Cartan de Lorraine, Université de Lorraine, Inria Nancy-Grand Est EPI SPHINX, F-54506 Vandoeuvre-lès-Nancy Cedex, France
e-mail: xavier.antoine@univ-lorraine.fr

C. Geuzaine (✉)
Institut Montefiore B28, Université de Liège, B-4000 Liège, Belgium
e-mail: cgeuzaine@ulg.ac.be

D. Lahaye et al. (eds.), *Modern Solvers for Helmholtz Problems*, Geosystems Mathematics, DOI 10.1007/978-3-319-28832-1_8

In [36], Lions introduced a converging Schwarz domain decomposition method without overlap for the Laplace equation by using Fourier-Robin boundary conditions on the interfaces instead of the standard Dirichlet or Neumann continuity conditions. For scalar or vector Helmholtz equations, these methods need to be adapted to lead to converging iterative algorithms. The first developments in this direction were introduced by Després [13, 14], who used simple impedance boundary conditions on the interfaces. A great variety of more general impedance conditions has been proposed since these early works, leading to so-called optimized Schwarz domain decomposition methods for time-harmonic wave problems [1, 9–11, 14–16, 19, 20, 25, 27, 43–45]. These methods can be used with or without overlap between the subdomains, and their convergence rate strongly depends on the transmission condition. Optimal convergence is obtained by using as transmission condition on each interface the non-local Dirichlet-to-Neumann (DtN) map [42] related to the complementary of the subdomain of interest [40, 41]. For acoustic waves, this DtN map links the normal derivative and the trace of the acoustic pressure on the interface. For electromagnetic waves, it links the magnetic and the electric surface currents (and is referred to in this case as the Magnetic-to-Electric, or MtE, map) [19]. However, using the DtN leads to a very expensive numerical procedure in practice, as this operator is non-local. Practical algorithms are thus based on local approximations of these operators, both for the acoustic case [9–11, 13, 27] and the electromagnetic one [1, 14–16, 20, 21, 43–45]. Recently, PMLs have also been used for this same purpose [23, 47, 51, 52].

In this chapter we provide a concise review of the most common transmission operators for optimized Schwarz methods applied to time-harmonic acoustic and electromagnetic wave problems, with the corresponding mathematical background. We analyze the behavior of these transmission operators on a model problem and derive generic weak formulations in view of their implementation in finite element codes. All the formulations are readily available for testing on several acoustic and electromagnetic cases using the open source GetDDM environment (http://onelab.info/wiki/GetDDM) [33, 48], based on the finite element solver GetDP (http://getdp.info) [17, 18, 28] and the mesh generator Gmsh (http://gmsh.info) [31, 32].

2 Scalar Helmholtz Equation: Acoustic Waves

Let Ω^- be an open subset of $\mathbb{R}^d (d = 1, 2, 3)$ with boundary $\Gamma := \partial\Omega^-$. The exterior domain of propagation is the complementary connected set defined by $\Omega^+ = \mathbb{R}^d \setminus \overline{\Omega^-}$. When considering a time-harmonic incident wave u^{inc}, the obstacle Ω^- creates a complex-valued scattered field u which is solution of the following problem

$$\begin{cases} (\Delta + k^2)u = 0 & \text{in } \Omega^+, \\ u = -u^{\text{inc}} & \text{on } \Gamma, \\ u \text{ outgoing}, & \end{cases} \tag{1}$$

fixing the time dependence under the form $e^{-i\omega t}$. The Laplacian operator is $\Delta = \sum_{i=1}^{d} \partial_{x_i}^2$ and the real-valued strictly positive wavenumber is given by $k = \omega/c$ (where $c = c(\mathbf{x})$ is the local speed of sound in the propagation medium). We denote by $\mathbf{a} \cdot \overline{\mathbf{b}}$ the inner product between two complex-valued vectors $\mathbf{a}$ and $\mathbf{b}$ in $\mathbb{C}^3$. We designate by $\overline{z}$ the complex conjugate of $z \in \mathbb{C}$ and the associated norm is $||\mathbf{a}|| := \sqrt{\mathbf{a} \cdot \overline{\mathbf{a}}}$. In this chapter, we fix a Dirichlet boundary condition on Γ corresponding to the sound-soft obstacle case. Nevertheless, any other condition can be studied like e.g. for a Neumann, Fourier or even for a penetrable obstacle. The outgoing condition at infinity, better known as Sommerfeld radiation condition ($\imath$ being the square root of -1), is added

$$\lim_{\|\mathbf{x}\| \to \infty} \|\mathbf{x}\|^{\frac{d-1}{2}} \left(\nabla u \cdot \frac{\mathbf{x}}{\|\mathbf{x}\|} - \imath k u \right) = 0.$$

This allows to prove that the solution to (1) is unique. In addition, this translates the property that the scattered field u is directed from Ω^- to infinity.

To numerically compute the solution to problem (1) by using e.g. the finite element method, Ω^+ has to be truncated. This can be realized for example by introducing a Perfectly Matched Layer (PML) [7, 12] or a fictitious boundary Γ^∞ with an Absorbing Boundary Condition (ABC) [6, 22] (see e.g. [4] for a review). If we consider an ABC on a fictitious boundary, we have to compute a field $\hat{u}$ approximating u on the finite domain Ω with boundary $\Gamma^\infty \bigcup \Gamma$. After merging the notations $\hat{u}$ and u for simplicity, the problem to be solved is

$$\begin{cases} (\Delta + k^2)u = 0 & \text{in } \Omega, \\ u = -u^{\text{inc}} & \text{on } \Gamma, \\ \partial_{\mathbf{n}} u + \mathcal{B} u = 0 & \text{on } \Gamma^\infty, \end{cases} \tag{2}$$

where the unit normal vector $\mathbf{n}$ is directed outside Ω (and thus inside Ω^- on Γ). The simplest local ABC, i.e., the Sommerfeld radiation condition at finite distance (zeroth-order condition), is obtained by setting

$$\mathcal{B} u = -\imath k u. \tag{3}$$

The extension to more accurate ABCs or PMLs is standard [34].

2.1 Domain Decomposition and Transmission Operators

Let us consider now that Ω is decomposed into N_{dom} disjoint subdomains Ω_i (the *substructures*) without overlap. For every $i = 0, \ldots, N_{\text{dom}}-1$, we set $\Gamma_i = \Gamma \bigcap \partial\Omega_i$, $\Gamma_i^\infty = \Gamma^\infty \bigcap \partial\Omega_i$, and, for $j = 0, \ldots, N_{\text{dom}}-1, j \neq i$, we introduce the transmission boundary $\Sigma_{ij} = \Sigma_{ji} = \overline{\partial\Omega_i \bigcap \partial\Omega_j}$. To simplify, let $D := \{0, \ldots, N_{\text{dom}} - 1\}$ be the set of indices of the subdomains, and for $i \in D$, let $D_i := \{j \in D$ such that $j \neq i$ and $\Sigma_{ij} \neq \emptyset\}$ be the set of indices of the subdomains sharing at least a point with Ω_i (such a domain is said to be *connected* to Ω_i). Finally, for all $i \in D$, the unit normal $\mathbf{n}_i$ is directed into the exterior of Ω_i and thus inside the obstacle Ω^- (if $\Gamma_i \neq \emptyset$).

Then the additive Schwarz domain decomposition method follows the steps at iteration $n+1$

1. For all $i \in D$, compute u_i^{n+1} solution to the boundary-value problem

$$\begin{cases} (\Delta + k^2)u_i^{n+1} = 0 & \text{in } \Omega_i, \\ u_i^{n+1} = -u^{\text{inc}} & \text{on } \Gamma_i, \\ \partial_{\mathbf{n}_i} u_i^{n+1} + \mathcal{B} u_i^{n+1} = 0 & \text{on } \Gamma_i^{\infty}, \\ \partial_{\mathbf{n}_i} u_i^{n+1} + \mathcal{S} u_i^{n+1} = g_{ij}^n & \text{on } \Sigma_{ij}, \quad \forall j \in D_i. \end{cases} \tag{4}$$

2. For all $i \in D$ and $j \in D_i$, update the interface unknowns with respect to the relation

$$g_{ji}^{n+1} = -\partial_{\mathbf{n}_i} u_i^{n+1} + \mathcal{S} u_i^{n+1} = -g_{ij}^n + 2\mathcal{S} u_i^{n+1}, \qquad \text{on } \Sigma_{ij}. \tag{5}$$

The operator $\mathcal{S}$ is a transmission operator that will be described later. A more compact writing of the $(n+1)$th iteration is

1. For all $i \in D$, compute the volume solution u_i^{n+1} of problem (4), which is written here as $u_i^{n+1} = \mathscr{V}_i(u^{\text{inc}}, g^n)$, where $g^n = (g_{ji}^n)_{i \in D, j \in D_i}$ is the vector that collects all the contributions related to the interface unknowns.
2. For all $i \in D$ and $j \in D_i$, update the surface fields g_{ji}^{n+1} following relation (5). This is written as $g_{ji}^{n+1} = \mathscr{T}_{ji}(g_{ij}^n, u_i^{n+1})$ in what follows.

In the boundary-value problem (4), only the case of Dirichlet sources is considered; however, any kinds such as volume sources could be handled similarly in the algorithm. These sources are called *physical* sources in contrast with the *artificial* sources g_{ij}^n related to the transmission boundaries.

The algorithm described by (4) and (5) can be understood as a Jacobi iteration for a linear operator equation. For every $n \in \mathbb{N}$, the field u_i^{n+1} can be decomposed by linearity as $u_i^{n+1} = v_i^{n+1} + \tilde{u}_i^{n+1}$, with

$$v_i^{n+1} = \mathscr{V}_i(u^{\text{inc}}, 0) \qquad \text{and} \qquad \tilde{u}_i^{n+1} = \mathscr{V}_i(0, g^n). \tag{6}$$

The function v_i^{n+1} does not depend on the iteration n and can be written as $v_i := v_i^n,\ \forall n \in \mathbb{N}, \forall i \in D$. Therefore, Eq. (5) can be written

$$g_{ji}^{n+1} = \mathscr{T}_{ji}(g_{ij}^n, u_i^{n+1}) = \mathscr{T}_{ji}(g_{ij}^n, \tilde{u}_i^{n+1}) + 2\mathcal{S} v_i, \qquad \text{on } \Sigma_{ij}. \tag{7}$$

Let us define the vector $b = (b_{ji})_{i\in D, j\in D_i}$, with $b_{ji} = 2(\mathcal{S}v_i)|_{\Sigma_{ij}}$, and $\mathcal{A} : g^n \mapsto \mathcal{A}g^n$ as the operator such that

$$\forall i \in D \quad \begin{cases} \tilde{u}_i^{n+1} = \mathscr{V}_i(0, g^n), \\ (\mathcal{A}g^n)_{ji} = \mathscr{T}_{ji}(g_{ij}^n, \tilde{u}_i^{n+1}), \ \forall j \in D_i. \end{cases} \tag{8}$$

One iteration of the domain decomposition method writes

$$g^{n+1} = \mathcal{A}g^n + b. \tag{9}$$

This can be interpreted as an iteration of the Jacobi method for solving the system

$$(\mathcal{I} - \mathcal{A})g = b, \tag{10}$$

where the identity operator is $\mathcal{I}$. An interesting consequence of (10) is that any iterative linear solver can be used for solving the equation. For example, Krylov subspace methods can be applied such as GMRES [46]. When a Krylov subspace solver is used, the resulting method is called a substructured preconditioner [26].

An important remark is that the iteration unknowns in (9), (10) are the surface quantities g and not the volume unknowns u. To get the volume quantities from the surface unknowns, $u_i = \mathscr{V}_i(u^{\text{inc}}, g)$ needs to be solved on every subdomain Ω_i. Algorithm 1 summarizes the Schwarz method with Krylov solver.

The convergence rate of the iterative solver is strongly related to the choice of the transmission operator $\mathcal{S}$ [10]. The so-called Dirichlet-to-Neumann (DtN) map for the complement of each subdomain [40, 41] appears as being optimal. Unfortunately, this operator is nonlocal and consequently costly to use in an iterative solver. An alternative approach consists in using local approximations based on polynomial or rational approximations of the total symbol of the surface DtN operator in the free-space, or a volume representation through PMLs. We detail

Algorithm 1: Schwarz algorithm with Krylov solver

1. Compute the right-hand side b

$$\begin{cases} \forall i \in D, \quad v_i = \mathscr{V}_i(u^{\text{inc}}, 0), \\ \forall i \in D, \forall j \in D_i, \quad b_{ji} = \mathscr{T}_{ji}(0, v_i). \end{cases}$$

2. Solve the following system $(\mathcal{I} - \mathcal{A})g = b$ iteratively by using a Krylov subspace solver, where the operator $\mathcal{A}$ is given by (8).
3. At convergence, compute the solution: $\forall i \in D, \quad u_i = \mathscr{V}_i(u^{\text{inc}}, g)$.

below four specific examples which are also implemented in GetDDM for a generic transmission boundary Σ

- *Evanescent Modes Damping Algorithm* [9, 11]:

$$\mathcal{S}_{\mathrm{IBC}}(\chi)u = (-\imath k + \chi)u,$$

where χ is a real-valued constant. This zeroth-order polynomial approximation is a generalization of the well-known Després condition [14], which corresponds to $\chi = 0$. We will denote this family of impedance transmission conditions as IBC(χ) in what follows.
- *Optimized second-order transmission condition* [27]:

$$\mathcal{S}_{\mathrm{GIBC}}(a, b)u = au + b\Delta_\Sigma u, \tag{11}$$

where Δ_Σ designates the Laplace-Beltrami operator on Σ, and a and b are two complex-valued numbers computed by solving a min-max optimization problem involving the rate of convergence (spectral radius) of the iteration operator. At the symbol level, this condition yields a second-order polynomial approximation of the DtN symbol. In the following, this family of generalized impedance transmission conditions is denoted by GIBC(a, b). A zeroth-order optimized condition can be built similarly.
- *Padé-localized square-root transmission condition* [10]:

$$\mathcal{S}_{\mathrm{GIBC}}(N_p, \alpha, \varepsilon)u = -\imath k C_0 u - \imath k \sum_{\ell=1}^{N_p} A_\ell \mathrm{div}_\Sigma\left(\frac{1}{k_\varepsilon^2}\nabla_\Sigma\right)\left(\mathcal{I} + B_\ell \mathrm{div}_\Sigma\left(\frac{1}{k_\varepsilon^2}\nabla_\Sigma\right)\right)^{-1} u, \tag{12}$$

setting

$$k_\varepsilon = k + i\varepsilon. \tag{13}$$

The complex-valued coefficients C_0, A_ℓ and B_ℓ are

$$C_0 = e^{\imath\alpha/2} R_{N_p}\left(e^{-\imath\alpha} - 1\right), \quad A_\ell = \frac{e^{-\frac{\imath\alpha}{2}} a_\ell}{(1 + b_\ell(e^{-\imath\alpha} - 1))^2}, \quad B_\ell = \frac{e^{-\imath\alpha} b_\ell}{1 + b_\ell(e^{-\imath\alpha} - 1)}. \tag{14}$$

The parameter α is a rotation angle in the complex plane (usually taken as $\pi/4$) and R_{N_p} are the standard real-valued Padé approximations of order N_p of $\sqrt{1+z}$

$$R_{N_p}(z) = 1 + \sum_{\ell=1}^{N_p} \frac{a_\ell z}{1 + b_\ell z},$$

with

$$a_\ell = \frac{2}{2N_p+1}\sin^2\left(\frac{\ell\pi}{2N_p+1}\right) \qquad \text{and} \qquad b_\ell = \cos^2\left(\frac{\ell\pi}{2N_p+1}\right). \tag{15}$$

This transmission condition is a complex-valued rational approximation [37] of the nonlocal pseudodifferential operator

$$\mathcal{S}_{\mathrm{GIBC(sq},\varepsilon)}u = -\imath k\sqrt{1+\mathrm{div}_\Sigma\left(\frac{1}{k_\varepsilon^2}\nabla_\Sigma\right)}u.$$

Fixing $\varepsilon = 0$ leads to the principal symbol of the exact DtN operator for the half-space. The introduction of the parameter ε regularizes this operator to model glancing rays at the surface of a curved interface. An optimal choice of ε is explained below in Sect. 2.2. In what follows, we denote this family of generalized impedance transmission conditions as GIBC(N_p, α, ε) and GIBC(sq, ε), respectively.

- *PML transmission condition* [23, 47, 51, 52]: The operator $\mathcal{S}_{\mathrm{PML}}(\sigma)$ is constructed by appending a layer Ω^{PML} to the transmission interface, in which a PML transformation with absorption profile σ is applied. For example, in cartesian coordinates, the singular profile

$$\sigma(x_{\mathrm{PML}}) = \frac{1}{k(x_{\mathrm{PML}}-\delta)}$$

can be used, where δ corresponds to the thickness of the PML layer and x_{PML} is the local coordinate inside the PML [8, 38].

All these methods are referred to as optimized Schwarz domain decomposition methods. Note that GIBC(N_p, α, ε) and PML(σ) have in common that they introduce additional surface/volume unknowns, whereas the other two transmission conditions do not. Also, the first three transmission conditions can be formulated explicitly through sparse surface equations (see e.g. the weak formulations (20)–(25) below), while a sparse formulation of the PML transmission condition requires a volume representation (see e.g. (26)–(27)), a surface representation being dense [50].

2.2 Convergence Analysis on a Model Problem

To study the impact of the various transmission conditions on the convergence of DDM, we analyze the model problem depicted in Fig. 1 which couples two subdomains: a disk-shaped *bounded* subdomain Ω_1 of radius R_0 and an *unbounded*

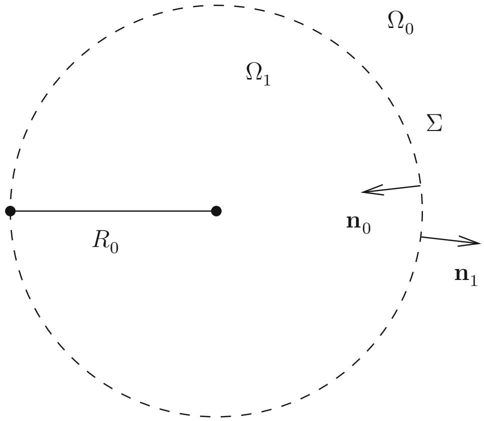


Fig. 1 Model problem with two subdomains and a circular interface

domain $\Omega_0 = \mathbb{R}^2 \setminus \Omega_1$:

$$\Omega_0 := \{\mathbf{x} \in \mathbb{R}^2, \ \ |\mathbf{x}| > R_0\}, \quad \Omega_1 := \{\mathbf{x} \in \mathbb{R}^2, \ \ |\mathbf{x}| < R_0\}, \tag{16}$$

with $\partial\Omega_0 = \partial\Omega_1 = \Sigma$. We analyze the spectral properties of the iteration operator $\mathcal{A}$ obtained from the domain decomposition algorithm coupling these two subdomains. Understanding the coupling of a curved bounded and unbounded subdomains allows us to clarify the main properties that one could not be analyzed by considering two bounded (e.g. a square domain divided in two) or two unbounded (e.g. two half-planes) subdomains. The considered model problem essentially contains the main features arising when solving exterior scattering problems in homogeneous media. It is thus not directly applicable to the PML-based transmission conditions, which introduce a fictitious heterogeneous medium, even for a radial profile.

For this problem, the iteration operator $\mathcal{A}$ can be expanded as $\mathcal{A} = \sum_{m=-\infty}^{+\infty} \mathcal{A}_m e^{\imath m\theta}$. We report in Fig. 2 the modal spectral radius $\rho(\mathcal{A}_m)$ with respect to the Fourier mode m for the transmitting boundary conditions IBC(0), IBC($k/2$), GIBC(a, b) and GIBC(sq, 0). We fix $k = 6\pi$, $R_0 = 1$ and the maximal number of modes is set to $m^{\max} = [10kR_0]$ (where $[10kR_0]$ denotes the integer part of $10kR_0$). Clearly, IBC(0) leads to a spectral radius equal to 1 for the evanescent modes, which is improved by IBC($k/2$)—for which the radius of convergence is always strictly less than one. Using GIBC(a, b) further improves over IBC($k/2$), particularly for large spatial modes m. We recall here that the GIBC(a, b) method is based on optimizing the coefficients a and b in relation (11) according to a min-max problem posed in the Fourier space [10, 27]. For the square-root transmission condition with $\varepsilon = 0$ (GIBC(sq,0)), we clearly observe an optimal convergence rate in the evanescent part of the spectrum. We also see a significant improvement over the IBC(0), IBC(χ) and GIBC(a, b) algorithms on the propagating modes. The damping parameter ε can be optimized to further improve the spectrum of

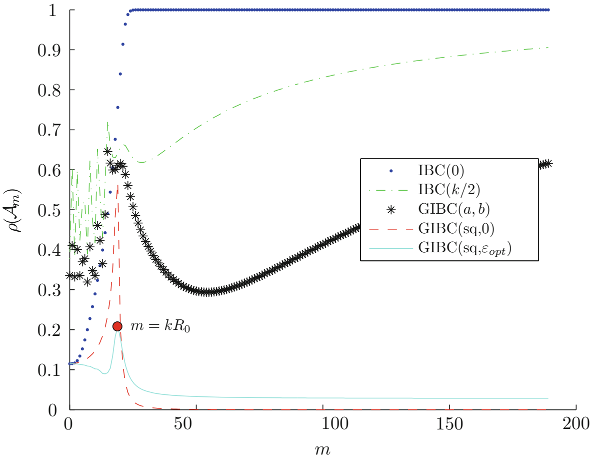


Fig. 2 Spectral radius of the modal iteration operator $\mathcal{A}_m$ vs. the Fourier mode m

the iteration operator corresponding to the modes in the transition zone. The optimization problem can be formulated as a min-max problem: find $\varepsilon_{\text{opt}} > 0$ such that it minimizes the spectral radius $\rho(\mathcal{A}_m)$ of the iteration operator (associated with GIBC(sq,ε)) for the mode $m \in \mathbb{Z}$ where it is maximal. Mathematically, this leads to solving the problem

$$\rho^{\text{sq},\varepsilon_{\text{opt}}} = \min_{\varepsilon \in \mathbb{R}^+} \left(\max_{m \in \mathbb{Z}} |\rho(\mathcal{A}_m)| \right), \tag{17}$$

resulting in the estimate $\varepsilon_{\text{opt}} = 0.4k^{1/3}\mathcal{H}^{2/3}$ [10] of the optimal value of the damping parameter, where $\mathcal{H}$ is the mean curvature on Σ. We see in Fig. 2 that the spectral radius of the iteration operator is indeed locally minimized for ε_{opt}.

Fast convergence of the GMRES solver is known to be strongly linked to the existence of eigenvalues clustering of the operator to solve, i.e. $(I - \mathcal{A})$ in our case. We report in Fig. 3 (left) the spectrum of the iteration operator for IBC(0), IBC($k/2$), GIBC(a, b) and GIBC(sq,ε_{opt}) (again for $kR_0 = 6\pi$ and $m^{\max} = [10kR_0]$). For all transmission operators, the spectrum lies in the right half-plane, which makes the GMRES converging. Nevertheless, many eigenvalues spread out in the complex plane for IBC(0). A slightly better clustering occurs for IBC($k/2$) and GIBC(a, b), while there is an excellent clustering of the eigenvalues for GIBC(sq, ε_{opt}). Most particularly, only a few eigenvalues associated with the propagating modes do not cluster but are very close to $(1, 0)$. In addition, the eigenvalues linked to the evanescent modes seem to cluster at $(1, 0)$. The eigenvalues clustering for the evanescent modes can be shown in numerical experiments to lead to a quasi-optimal GMRES convergence rate that is independent of the density of discretization points per wavelength n_λ [10].

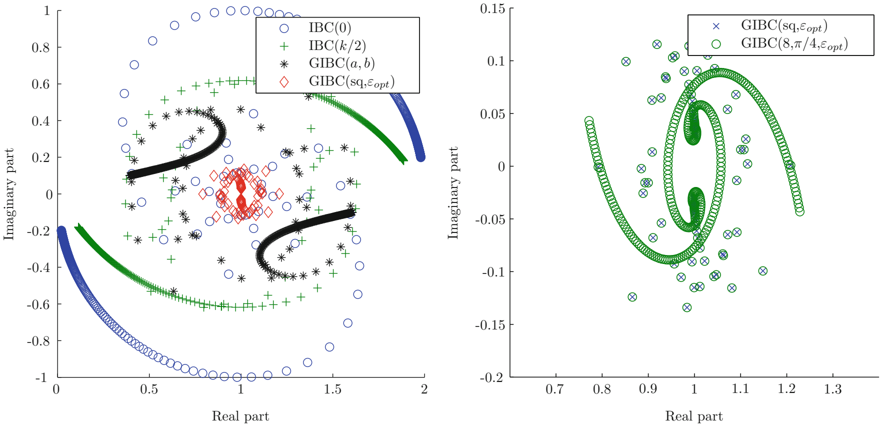


Fig. 3 *Left*: Eigenvalues distribution in the complex plane for $(I - \mathcal{A})$ and different transmission operators. *Right*: Eigenvalues distribution in the complex plane for the exact and Padé-localized square-root transmission operator of order 4

As said before, the square-root operator (2.1) is a first-order nonlocal pseudodifferential operator. Therefore, it is impractical in a finite element setting since it would lead to consider full complex-valued matrices at the transmission interfaces. Fortunately, a localization process of this operator can be efficiently realized and based on partial differential (local) operators to have a sparse matrix representation. In [3, 35, 37], this is done by using a rotating branch-cut approximation of the square-root and next applying complex Padé approximants of order N_p, leading to the transmission operator (12). We report in Fig. 3 (right) the spectrum of the modal iteration operators GIBC(sq, ε_{opt}) and GIBC($4, \pi/4, \varepsilon_{\text{opt}}$). As already noticed, there is an almost perfect clustering of the eigenvalues for GIBC(sq, ε_{opt}). As expected, the larger N_p, the better the approximation of the spectrum of the square-root. Moreover, N_p allows to adjust the spectrum accuracy for large modes m (evanescent modes which numerically correspond to mesh refinement in a finite element context). Numerical simulations show that in practice relatively small values of N_p ($N_p = 2, 4, 8$) give optimal convergence results.

2.3 *Weak Formulations*

For the finite element approximation, we consider some variational formulations. Two kinds of PDEs are involved when using optimized Schwarz methods: firstly, a volume system (in the present case, the scalar Helmholtz equation) given by $\mathscr{V}_i$, and, secondly, a surface system on the transmission interfaces, fixed by $\mathscr{T}_{ji}$. The variational formulations are first provided for a general transmission operator $\mathcal{S}$. To simplify the presentation, we consider the situation where no contribution comes from $\partial\Sigma_{ij}$ through an integration by parts. However, in some cases (e.g. when $\Sigma_{ij} \bigcap \Gamma^{\infty} \neq \emptyset$), a special attention must be directed towards the inclusion of these terms into the variational formulations.

Without loss of generality, we only detail the case of a particular subdomain Ω_i, for $i \in D$, without incident wave contribution (i.e. homogeneous Dirichlet boundary condition). We consider the general setting where PML layers $\Omega_i^{\mathrm{PML}} = \cup_{j\in D_i}\Omega_{ij}^{\mathrm{PML}}$ are potentially appended to the artificial interfaces Σ_{ij}, and define $\Omega_i^* := \Omega_i \cup \Omega_i^{\mathrm{PML}}$. In what follows, the space $H^1(\Omega_i^*) := \{\tilde{u}_i \in L^2(\Omega_i^*) \text{ such that } \nabla\tilde{u}_i \in (L^2(\Omega_i^*))^3\}$ is the classical Sobolev space and $H_0^1(\Omega_i^*)$ is the space of functions $\tilde{u}_i \in H^1(\Omega_i^*)$ such that $\tilde{u}_i|_{\Gamma_i} = 0$, which slightly differs from its usual definition (the Dirichlet condition is here set only on part of $\partial\Omega_i^*$). Then,

- the volume PDE $\tilde{u}_i^{n+1} = \mathscr{V}_i(0, g^n)$ has the following weak formulation

$$\begin{cases} \text{Find } \tilde{u}_i^{n+1} \text{ in } H_0^1(\Omega_i^*) \text{ such that, for every } \tilde{u}_i' \in H_0^1(\Omega_i^*) : \\ \displaystyle\int_{\Omega_i} \nabla\tilde{u}_i^{n+1}\cdot\nabla\tilde{u}_i' \,\mathrm{d}\Omega_i - \int_{\Omega_i} k^2\tilde{u}_i^{n+1}\tilde{u}_i' \,\mathrm{d}\Omega_i + \int_{\Gamma_i^{\infty}} \mathcal{B}\tilde{u}_i^{n+1}\tilde{u}_i' \,\mathrm{d}\Gamma_i^{\infty} \\ \qquad\qquad\displaystyle + \sum_{j\in D_i}\int_{\Sigma_{ij}} \mathcal{S}\tilde{u}_i^{n+1}\tilde{u}_i' \,\mathrm{d}\Sigma_{ij} = \sum_{j\in D_i}\int_{\Sigma_{ij}} g_{ij}^n\tilde{u}_i' \,\mathrm{d}\Sigma_{ij}, \end{cases} \tag{18}$$

- and the surface PDE $g_{ji}^{n+1} = \mathscr{T}_{ji}(g_{ij}^n, \tilde{u}_i^{n+1})$ has the following one:

$$\begin{cases} \text{Find } g_{ji}^{n+1} \text{ in } H^1(\Sigma_{ij}) \text{ such that, for every } g_{ji}' \in H^1(\Sigma_{ij}) : \\ \displaystyle\int_{\Sigma_{ij}} g_{ji}^{n+1}g_{ji}' \,\mathrm{d}\Sigma_{ij} = -\int_{\Sigma_{ij}} g_{ij}^n g_{ji}' \,\mathrm{d}\Sigma_{ij} + 2\int_{\Sigma_{ij}} \mathcal{S}\tilde{u}_i^{n+1}g_{ji}' \,\mathrm{d}\Sigma_{ij}. \end{cases} \tag{19}$$

Depending on the choice of the transmission operator $\mathcal{S}$, the quantities $\int_{\Sigma_{ij}} \mathcal{S}\tilde{u}_i^{n+1}\tilde{u}_i' \,\mathrm{d}\Sigma_{ij}$ and $\int_{\Sigma_{ij}} \mathcal{S}\tilde{u}_i^{n+1}g_{ji}' \,\mathrm{d}\Sigma_{ij}$ write as follows:

- IBC(χ):

$$\int_{\Sigma_{ij}} \mathcal{S}\tilde{u}_i^{n+1}\tilde{u}_i' \,\mathrm{d}\Sigma_{ij} := \int_{\Sigma_{ij}} (-\imath k + \chi)\tilde{u}_i^{n+1}\tilde{u}_i' \,\mathrm{d}\Sigma_{ij}; \tag{20}$$

$$\int_{\Sigma_{ij}} \mathcal{S}\tilde{u}_i^{n+1}g_{ji}' \,\mathrm{d}\Sigma_{ij} := \int_{\Sigma_{ij}} (-\imath k + \chi)\tilde{u}_i^{n+1}g_{ji}' \,\mathrm{d}\Sigma_{ij}. \tag{21}$$

- GIBC(a, b):

$$\int_{\Sigma_{ij}} \mathcal{S}\tilde{u}_i^{n+1}\tilde{u}_i' \,\mathrm{d}\Sigma_{ij} := \int_{\Sigma_{ij}} a\tilde{u}_i^{n+1}\tilde{u}_i' \,\mathrm{d}\Sigma_{ij} - \int_{\Sigma_{ij}} b\nabla\tilde{u}_i^{n+1}\cdot\nabla\tilde{u}_i' \,\mathrm{d}\Sigma_{ij}; \tag{22}$$

$$\int_{\Sigma_{ij}} \mathcal{S}\tilde{u}_i^{n+1}g_{ji}' \,\mathrm{d}\Sigma_{ij} := \int_{\Sigma_{ij}} a\tilde{u}_i^{n+1}g_{ji}' \,\mathrm{d}\Sigma_{ij} - \int_{\Sigma_{ij}} b\nabla\tilde{u}_i^{n+1}\cdot\nabla g_{ji}' \,\mathrm{d}\Sigma_{ij}. \tag{23}$$

- GIBC(N_p, α, ε):

$$\int_{\Sigma_{ij}} \mathcal{S}\tilde{u}_i^{n+1}\tilde{u}_i' \,\mathrm{d}\Sigma_{ij} := -\imath k C_0 \int_{\Sigma_{ij}} \tilde{u}_i^{n+1}\tilde{u}_i' \,\mathrm{d}\Sigma_{ij} + \imath k \sum_{\ell=1}^{N_p} A_\ell \int_{\Sigma_{ij}} \frac{1}{k_\varepsilon^2} \nabla_{\Sigma_{ij}}\varphi_\ell \cdot \nabla_{\Sigma_{ij}}\tilde{u}_i' \,\mathrm{d}\Sigma_{ij}, \tag{24}$$

where, for every $\ell = 1, \dots, N_p$, the function φ_ℓ is obtained through the resolution of

$$\begin{cases} \text{Find } \varphi_\ell \text{ in } H^1(\Sigma_{ij}) \text{ such that, for every } \varphi_\ell' \in H^1(\Sigma_{ij}) : \\ -\displaystyle\int_{\Sigma_{ij}} \tilde{u}_i^{n+1}\varphi_\ell' \,\mathrm{d}\Sigma_{ij} - B_\ell \int_{\Sigma_{ij}} \frac{1}{k_\varepsilon^2} \nabla_{\Sigma_{ij}}\varphi_\ell \cdot \nabla_{\Sigma_{ij}}\varphi_\ell' \,\mathrm{d}\Sigma_{ij} + \int_{\Sigma_{ij}} \varphi_\ell \cdot \varphi_\ell' \,\mathrm{d}\Sigma_{ij} = 0; \end{cases}$$

$$\int_{\Sigma_{ij}} \mathcal{S}\tilde{u}_i^{n+1} g_{ji}' \,\mathrm{d}\Sigma_{ij} := -\imath k C_0 \int_{\Sigma_{ij}} \tilde{u}_i^{n+1} g_{ji}' \,\mathrm{d}\Sigma_{ij} - \imath k \sum_{\ell=1}^{N_p} \frac{A_\ell}{B_\ell} \int_{\Sigma_{ij}} (\tilde{u}_i^{n+1} - \varphi_\ell) g_{ji}' \,\mathrm{d}\Sigma_{ij}. \tag{25}$$

- PML(σ):

$$\int_{\Sigma_{ij}} \mathcal{S}\tilde{u}_i^{n+1}\tilde{u}_i' \,\mathrm{d}\Sigma_{ij} := \int_{\Omega_{ij}^{\mathrm{PML}}} D\nabla\tilde{u}_i^{n+1} \cdot \nabla\tilde{u}_i' \,\mathrm{d}\Omega_{ij}^{\mathrm{PML}} - \int_{\Omega_{ij}^{\mathrm{PML}}} k^2 E\, \tilde{u}_i^{n+1}\tilde{u}_i' \,\mathrm{d}\Omega_{ij}^{\mathrm{PML}}; \tag{26}$$

$$\int_{\Sigma_{ij}} \mathcal{S}\tilde{u}_i^{n+1} g_{ji}' \,\mathrm{d}\Sigma_{ij} := \int_{\Omega_{ij}^{\mathrm{PML}}} D\nabla\tilde{u}_i^{n+1} \cdot \nabla g_{ji}' \,\mathrm{d}\Omega_{ij}^{\mathrm{PML}} - \int_{\Omega_{ij}^{\mathrm{PML}}} k^2 E\, \tilde{u}_i^{n+1} g_{ji}' \,\mathrm{d}\Omega_{ij}^{\mathrm{PML}}, \tag{27}$$

where $D = \mathrm{diag}(\frac{1}{\gamma_x}, \gamma_x, \gamma_x)$ and $E = \gamma_x$, with $\gamma_x(x_{\mathrm{PML}}) = 1 + \frac{\imath}{\omega}\sigma_x(x_{\mathrm{PML}})$, that is, we consider a 1D PML with an absorption function that grows only in the direction normal to the interface. In (27) the domain of definition of the test functions g_{ji}' on Σ_{ij} is extended to the neighboring PML layer $\Omega_{ij}^{\mathrm{PML}}$, effectively resulting at the discrete level in the integration of the functions associated with the nodes of the interface in the layer of volume elements connected to the interface.

3 Vector Helmholtz Equation: Electromagnetic Waves

We now consider the case of an incident electromagnetic wave $\mathbf{E}^{\mathrm{inc}}$ illuminating a perfectly conducting obstacle Ω^- with boundary Γ, in a three dimensional medium. The scattered electric field $\mathbf{E}$ is solution to the following exterior electromagnetic scattering problem:

$$\left\{\begin{aligned} \textbf{curl curl E} - k^2\mathbf{E} &= 0, && \text{in } \Omega^+,\\ \gamma^T(\mathbf{E}) &= -\gamma^T(\mathbf{E}), && \text{on } \Gamma,\\ \lim_{\|\mathbf{x}\|\to\infty} \|\mathbf{x}\| \left(\frac{\mathbf{x}}{\|\mathbf{x}\|} \times \textbf{curl E} + \imath k\mathbf{E}\right) &= 0, \end{aligned}\right. \tag{28}$$

where $k := 2\pi/\lambda$ is again the wavenumber and λ the wavelength, $\mathbf{n}$ is the outward unit normal to Ω^+ (thus, inward to the obstacle) and γ^T is the tangential component trace operator

$$\gamma^T : \mathbf{v} \longmapsto \mathbf{n} \times (\mathbf{v} \times \mathbf{n}).$$

The **curl** operator is defined by $\textbf{curl a} := \nabla \times \mathbf{a}$, for a complex-valued vector field $\mathbf{a} \in \mathbb{C}^3$, and the notation $\mathbf{a} \times \mathbf{b}$ designates the cross product between two complex-valued vectors $\mathbf{a}$ and $\mathbf{b}$. The last equation of system (28) is the so-called Silver-Müller radiation condition at infinity, which provides the uniqueness of the solution to the scattering boundary-value problem (28).

As in the acoustic case, solving (28) numerically with a volume discretization method requires the truncation of the exterior propagation domain with a PML or with an ABC on a fictitious boundary Γ^∞ surrounding Ω^-. For an ABC the problem to be solved is then defined on the bounded domain Ω, with boundaries Γ and Γ^∞:

$$\left\{\begin{aligned} \textbf{curl curl E} - k^2\mathbf{E} &= 0, && \text{in } \Omega,\\ \gamma^T(\mathbf{E}) &= -\gamma^T(\mathbf{E}), && \text{on } \Gamma,\\ \gamma^t(\textbf{curl E}) + \mathcal{B}(\gamma^T(\mathbf{E})) &= 0, && \text{on } \Gamma^\infty, \end{aligned}\right. \tag{29}$$

with γ^t the tangential trace operator:

$$\gamma^t : \mathbf{v} \longmapsto \mathbf{n} \times \mathbf{v}.$$

As above, the unit normal $\mathbf{n}$ is outwardly directed to Ω and, to simplify, the solution of the above problem is still designated by $\mathbf{E}$. The operator $\mathcal{B}$ is an approximation of the Magnetic-to-Electric (MtE) operator. The well-known Silver-Müller ABC at finite distance is obtained with $\mathcal{B} = \imath k$, similar to (3) for acoustics modulo the sign (due to the trace operator definitions). The extension to more accurate ABCs or PMLs is standard.

3.1 Domain Decomposition and Transmission Operators

The optimized Schwarz domain decomposition without overlap for the Maxwell problem (29) can be set up in exactly the same way as for the scalar Helmholtz equation. The domain Ω is decomposed as described in Sect. 2.1, and the same notations are used. The iterative Jacobi algorithm for the computation of the electric fields $(\mathbf{E}_i^{n+1})_{i\in D}$ at iteration $n+1$ involves, first, the solution of the N_{dom} following problems

$$\begin{cases} \mathbf{curl}\,\mathbf{curl}\,\mathbf{E}_i^{n+1} - k^2\,\mathbf{E}_i^{n+1} = \mathbf{0}, & \text{in } \Omega_i, \\ \gamma_i^T(\mathbf{E}_i^{n+1}) = -\gamma_i^T(\mathbf{E}^{\mathrm{inc}}), & \text{on } \Gamma_i, \\ \gamma_i^t(\mathbf{curl}\,\mathbf{E}_i^{n+1}) + \mathcal{B}(\gamma_i^T(\mathbf{E}_i^{n+1})) = \mathbf{0}, & \text{on } \Gamma_i^\infty, \\ \gamma_i^t(\mathbf{curl}\,\mathbf{E}_i^{n+1}) + \mathcal{S}(\gamma_i^T(\mathbf{E}_i^{n+1})) = \mathbf{g}_{ij}^n, & \text{on } \Sigma_{ij}, \forall j \in D_i, \end{cases} \tag{30}$$

and then forming the quantities $\mathbf{g}_{ji}^{n+1}$ through

$$\mathbf{g}_{ji}^{n+1} = \gamma_i^t(\mathbf{curl}\,\mathbf{E}_i^{n+1}) + \mathcal{S}(\gamma_i^T(\mathbf{E}_i^{n+1})) = -\mathbf{g}_{ij}^n + 2\mathcal{S}(\gamma_i^T(\mathbf{E}_i^{n+1})), \quad \text{on } \Sigma_{ij}, \tag{31}$$

where, for $i \in D$, $\mathbf{E}_i = \mathbf{E}_{|\Omega_i}$, $\mathcal{S}$ is a transmission operator through the interfaces Σ_{ij} and γ_i^t and γ_i^T are the local tangential trace and tangential component trace operators:

$$\gamma_i^t : \mathbf{v}_i \longmapsto \mathbf{n}_i \times \mathbf{v}_{i|\partial\Omega_i} \quad \text{and} \quad \gamma_i^T : \mathbf{v}_i \longmapsto \mathbf{n}_i \times (\mathbf{v}_{i|\partial\Omega_i} \times \mathbf{n}_i),$$

with $\mathbf{n}_i$ the outward-pointing unit normal to Ω_i.

Following the same procedure as in Sect. 2.1, we introduce the two families of operators $(\mathscr{V}_i)_{i\in D}$ and $(\mathscr{T}_{ji})_{i\in D, j\in D_i}$ as:

1. $\mathbf{E}_i^{n+1} = \mathscr{V}_i(\mathbf{E}^{\mathrm{inc}}, \mathbf{g}^n) \iff \mathbf{E}_i^{n+1}$ is solution of problem (30), where $\mathbf{g}^n = (\mathbf{g}_{ji}^n)_{i\in D, j\in D_i}$ collects all the unknowns at iteration n;
2. $\mathbf{g}_{ji}^{n+1} = \mathscr{T}_{ji}(\mathbf{g}_{ij}^n, \mathbf{E}_i^{n+1}) \iff \mathbf{g}_{ji}^{n+1}$ is solution of problem (31).

By linearity, we decompose the field $\mathbf{E}_i^{n+1}$ as $\mathbf{E}_i^{n+1} = \mathbf{F}_i^{n+1} + \widetilde{\mathbf{E}}_i^{n+1}$, where

$$\mathbf{F}_i^{n+1} = \mathscr{V}_i(\mathbf{E}^{\mathrm{inc}}, 0) \qquad \text{and} \qquad \widetilde{\mathbf{E}}_i^{n+1} = \mathscr{V}_i(0, \mathbf{g}^n). \tag{32}$$

The quantity $\mathbf{F}_i^{n+1}$ is independent of the iteration number n and can hence be written as $\mathbf{F}_i := \mathbf{F}_i^n$, $\forall n \in \mathbb{N}, \forall i \in D$. The whole algorithm can then be recast into a linear system:

$$(\mathcal{I} - \mathcal{A})\,\mathbf{g} = \mathbf{b}, \tag{33}$$

that can be solved by a Krylov subspace solver.

As in the scalar case, for a vector $\mathbf{g}^n$, the quantity $\mathcal{A}\mathbf{g}^n$ is given by, for $i \in D$ and $j \in D_i$, $(\mathcal{A}\mathbf{g}^n)_{ji} = \mathscr{T}_{ji}\left(\mathbf{g}_{ij}^n, \widetilde{\mathbf{E}}_i^{n+1}\right)$. The information about the incident wave is contained in the right-hand side: $\mathbf{b}_{ji} = \mathscr{T}_{ji}\,(0, \mathbf{F}_i)$. The domain decomposition algorithm for the Maxwell system is then exactly the same as the one described in

Algorithm 1 for the scalar Helmholtz equation, by formally replacing v_i, u^{inc}, g and u_i by $\mathbf{F}_i, \mathbf{E}^{\text{inc}}, \mathbf{g}$ and $\mathbf{E}_i$, respectively.

Similarly to the acoustic case, optimal convergence of the domain decomposition algorithm would be achieved by using the (nonlocal) MtE operator as transmission condition. Local approximations based on polynomial or rational approximations of the total symbol of the surface free-space MtE have been proposed, as well as volume representations through Perfectly Matched Layers. We detail four of those approximations below, for a generic transmission boundary Σ:

- *Zeroth-order transmission condition* [14]:

$$\mathcal{S}_{\text{IBC}(0)}(\gamma^T(\mathbf{E})) = \imath k \gamma^T(\mathbf{E}). \tag{34}$$

- *Optimized second-order transmission condition* [45]:

$$\mathcal{S}_{\text{GIBC}}(a,b)(\gamma^T(\mathbf{E})) = \imath k \left(\mathcal{I} + \frac{a}{k^2}\nabla_\Sigma \text{div}_\Sigma\right)^{-1} \left(\mathcal{I} - \frac{b}{k^2}\mathbf{curl}_\Sigma \text{curl}_\Sigma\right)(\gamma^T(\mathbf{E})), \tag{35}$$

where the curl operator is the dual operator of **curl** and where a and b are chosen so that an optimal convergence rate is obtained for the (TE) and (TM) modes; see [45] for the expression of a and b in the half-plane case. An optimized transmission condition using a single second-order operator was proposed in [1]:

$$\mathcal{S}_{\text{GIBC}}(a)(\gamma^T(\mathbf{E})) = \imath k a \left(\mathcal{I} - \frac{1}{k^2}\mathbf{curl}_\Sigma \text{curl}_\Sigma\right)(\gamma^T(\mathbf{E})). \tag{36}$$

- *Padé-localized square-root transmission condition* [19, 21]:

$$\mathcal{S}_{\text{GIBC}}(N_p,\alpha,\varepsilon)(\gamma^T(\mathbf{E})) = \imath k \left(C_0 + \sum_{\ell=1}^{N_p} A_\ell X(\mathcal{I} + B_\ell X)^{-1}\right)^{-1} \left(\mathcal{I} - \mathbf{curl}_\Sigma \frac{1}{k_\varepsilon^2}\text{curl}_\Sigma\right)(\gamma^T(\mathbf{E})), \tag{37}$$

with $X := \nabla_\Sigma \frac{1}{k_\varepsilon^2}\text{div}_\Sigma - \mathbf{curl}_\Sigma \frac{1}{k_\varepsilon^2}\text{curl}_\Sigma$, and where k_ε, C_0, A_ℓ and B_ℓ are defined by (13) and (14). This transmission condition corresponds to a rational approximation of the nonlocal operator

$$\mathcal{S}_{\text{GIBC(sq},\varepsilon)}(\gamma^T(\mathbf{E})) = \imath k (\mathcal{I} + X)^{-1/2} \left(\mathcal{I} - \mathbf{curl}_\Sigma \frac{1}{k_\varepsilon^2}\text{curl}_\Sigma\right)(\gamma^T(\mathbf{E})),$$

which for $\varepsilon = 0$ is the principal symbol of the exact MtE operator for the half-space. As in the scalar Helmholtz case, the parameter ε is introduced to regularize this operator for grazing rays on curved interfaces, and the rational approximation generalizes the polynomial approximations underlying (34), (36) and (35).

- *PML transmission condition* [51, 52]: The operator $\mathcal{S}_{\text{PML}}(\sigma)$ is constructed by appending a layer Ω^{PML} to the transmission interface, into which a PML transformation with absorption profile σ is applied in the same way as for the acoustic case.

3.2 Convergence Analysis on a Model Problem

In order to study the convergence rate and spectral properties of the DDM algorithm we consider a similar setting as for the scalar case, but in three dimensions: the whole domain $\Omega = \mathbb{R}^3$ is separated in two curved subdomains Ω_1 and Ω_2 by a spherical boundary of radius R_0

$$\Omega_0 := \{\mathbf{x} \in \mathbb{R}^3, |\mathbf{x}| > R_0\}, \quad \Omega_1 := \{\mathbf{x} \in \mathbb{R}^3, |\mathbf{x}| < R_0\}, \tag{38}$$

with $\partial\Omega_0 = \partial\Omega_1 := \Sigma$. Again, in this homogeneous medium setting we only consider the transmission operators that lead to a sparse surface representation. Using the same strategy as in Sect. 2.2, we fix $R_0 = 1$ and $k = 6\pi$, and consider a maximal number of modes $m^{\max} = [10kR]$. We report on Fig. 4 the modal spectral radius $\rho(\mathcal{A}_m)$ for the transmission conditions IBC(0), GIBC(a), GIBC(a, b) and GIBC(sq, ε). For GIBC(a) and GIBC(a, b), the optimal parameters a and b are numerically computed by solving the min-max problem

$$\min_{(a,b)\in\mathbb{C}^2} \max_{m\geq 1} \rho(\mathcal{A}_m) \tag{39}$$

with the Matlab function `fminsearch`. Analytical solutions of (39) for the half-space case are provided in [1] for GIBC(a) and in [45] for GIBC(a, b). Contrary

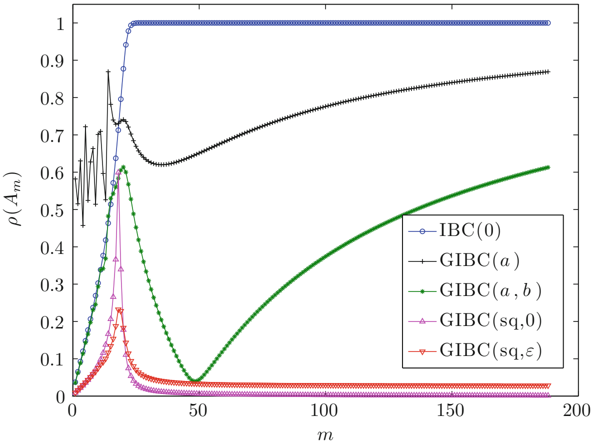


Fig. 4 Spectral radius of the modal iteration operator $\mathcal{A}_m$ vs. the Fourier mode m

to the scalar Helmholtz case and to the half-space case [16], where IBC(0) leads to a convergence factor that is exactly 1 for the evanescent modes, in this model problem IBC(0) leads to $\rho(\mathcal{A}_m) < 1$ in the whole spectrum, although $\rho(\mathcal{A}_m)$ is very close to 1 for the evanescent modes, which results in a globally slowly converging DDM. For GIBC(a), we see that $\rho(\mathcal{A}_m) < 1$, for all m, which is improved further for GIBC(a, b). GIBC(sq, 0) leads to a better convergence rate still, which can furthermore be optimized in the transition zone by using GIBC(sq, ε) (with the value parameter $\varepsilon = 0.4k^{1/3}R^{-2/3}$). Finally, a numerical study using the exact series solution shows that GIBC(a) can lead to a spectral radius larger than one if the parameter a is chosen as in the half-plane case, which highlights the need for careful geometry-dependent optimization of the parameters.

The history of the GMRES residual with respect to the number of iterations #iter is displayed on Fig. 5 (left) for the various transmission conditions. As we can observe, there is a hierarchy in the convergence curves that is directly connected to the increasing order of the GIBCs, the best convergence being obtained for GIBC(sq, ε). Note that when using GIBC(a, b) with the optimal parameters for the half-plane, the number of iterations is about the same as for GIBC(a). Also, numerical tests show that using the Jacobi method instead of GMRES can lead to a convergence failure for IBC(0), GIBC(a) and GIBC(a, b). The eigenvalues distribution of the operator $(\mathcal{I} - \mathcal{A})$ is displayed on Fig. 5 (right). As in the scalar Helmholtz case, the improvement in the clustering of the eigenvalues around $(1, 0)$ is again observed when improving the approximation of the MtE. Finally, it is shown in [21] that the localization of GIBC(sq, ε) using Padé approximants behaves very similarly to the scalar Helmholtz case.

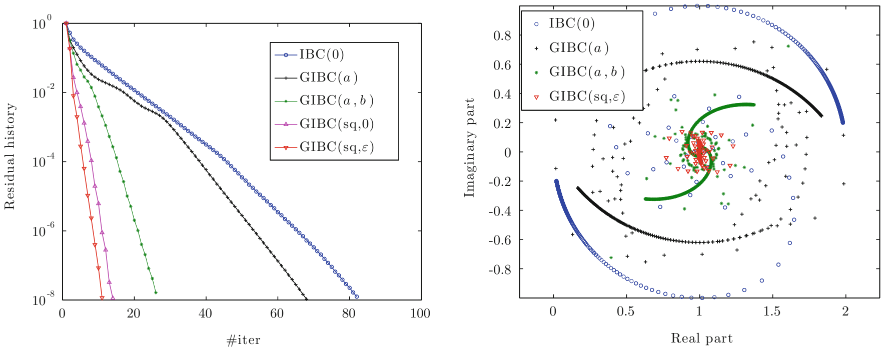


Fig. 5 *Left*: Residual history of GMRES vs. #iter for the various transmission conditions. *Right*: Eigenvalues distribution of the operator $(\mathcal{I} - \mathcal{A})$ for the different transmission conditions

3.3 Weak Formulations

Without loss of generality, only the case of a particular subdomain Ω_i, for $i \in D$, with no incident wave (homogeneous Dirichlet boundary condition) is detailed. We consider the same general setting as in the scalar Helmholtz case, i.e., where the PML layers $\Omega_i^{\mathrm{PML}} = \cup_{j \in D_i} \Omega_{ij}^{\mathrm{PML}}$ are potentially appended to the artificial interfaces Σ_{ij}, and define $\Omega_i^* := \Omega_i \cup \Omega_i^{\mathrm{PML}}$. The space of complex-valued curl-conforming vector fields on Ω_i^* is denoted by $\mathbf{H}(\mathbf{curl}, \Omega_i^*) := \{\mathbf{W} \in (L^2(\Omega_i^*))^3$ such that $\mathbf{curl}(\mathbf{W}) \in (L^2(\Omega_i^*))^3\}$. The functional space $\mathbf{H}_0(\mathbf{curl}, \Omega_i^*)$ is the space of functions $\mathbf{W}_i$ in $\mathbf{H}(\mathbf{curl}, \Omega_i^*)$ such that $\gamma_i^T(\mathbf{W}_i) = 0$ on $\Gamma_i = 0$ (the boundary condition is only imposed on a part $\partial\Omega_i^*$).

- The volume PDE $\widetilde{\mathbf{E}}_i^{n+1} = \mathscr{V}_i(0, \mathbf{g}^n)$ has the following weak formulation:

$$\begin{cases} \text{Find } \widetilde{\mathbf{E}}_i^{n+1} \in \mathbf{H}_0(\mathbf{curl}, \Omega_i) \text{ such that, for every } \widetilde{\mathbf{E}}_i' \in \mathbf{H}_0(\mathbf{curl}, \Omega_i) : \\ \displaystyle\int_{\Omega_i} \mathbf{curl}\, \widetilde{\mathbf{E}}_i^{n+1} \cdot \mathbf{curl}\, \widetilde{\mathbf{E}}_i' \, \mathrm{d}\Omega_i - \int_{\Omega_i} k^2 \widetilde{\mathbf{E}}_i^{n+1} \cdot \widetilde{\mathbf{E}}_i' \, \mathrm{d}\Omega_i - \int_{\Gamma_i^\infty} \mathcal{B}(\gamma_i^T(\widetilde{\mathbf{E}}_i^{n+1})) \cdot \widetilde{\mathbf{E}}_i' \, \mathrm{d}\Gamma_i^\infty \\ \qquad\qquad \displaystyle -\sum_{j \in D_i} \int_{\Sigma_{ij}} \mathcal{S}(\gamma_i^T(\widetilde{\mathbf{E}}_i^{n+1})) \cdot \widetilde{\mathbf{E}}_i' \, \mathrm{d}\Sigma_{ij} = -\sum_{j \in D_i} \int_{\Sigma_{ij}} \mathbf{g}_{ij}^n \cdot \widetilde{\mathbf{E}}_i' \, \mathrm{d}\Sigma_{ij}. \end{cases} \tag{40}$$

- The surface PDE $\mathbf{g}_{ji}^{n+1} = \mathscr{T}_{ji}(\mathbf{g}_{ij}^n, \widetilde{\mathbf{E}}_i^{n+1})$ has the following one:

$$\begin{cases} \text{Find } \mathbf{g}_{ji}^{n+1} \text{ in } \mathbf{H}(\mathbf{curl}, \Sigma_{ij}) \text{ such that, for every } \mathbf{g}_{ji}' \in \mathbf{H}(\mathbf{curl}, \Sigma_{ij}) : \\ \displaystyle\int_{\Sigma_{ij}} \mathbf{g}_{ji}^{n+1} \cdot \mathbf{g}_{ji}' \, \mathrm{d}\Sigma_{ij} = -\int_{\Sigma_{ij}} \mathbf{g}_{ij}^n \cdot \mathbf{g}_{ji}' \, \mathrm{d}\Sigma_{ij} + 2\int_{\Sigma_{ij}} \mathcal{S}(\gamma_i^T(\widetilde{\mathbf{E}}_i^{n+1})) \cdot \mathbf{g}_{ji}' \, \mathrm{d}\Sigma_{ij}. \end{cases}$$

On the transmission boundaries, we have:

- IBC(0):

$$\int_{\Sigma_{ij}} \mathcal{S}(\gamma_i^T(\widetilde{\mathbf{E}}_i^{n+1})) \cdot \widetilde{\mathbf{E}}_i' \, \mathrm{d}\Sigma_{ij} := \int_{\Sigma_{ij}} \imath k(\gamma_i^T(\widetilde{\mathbf{E}}_i^{n+1})) \cdot \widetilde{\mathbf{E}}_i' \, \mathrm{d}\Sigma_{ij}; \tag{41}$$

$$\int_{\Sigma_{ij}} \mathcal{S}(\gamma_i^T(\widetilde{\mathbf{E}}_i^{n+1})) \cdot \mathbf{g}_{ji}' \, \mathrm{d}\Sigma_{ij} := \int_{\Sigma_{ij}} \imath k(\gamma_i^T(\widetilde{\mathbf{E}}_i^{n+1})) \cdot \mathbf{g}_{ji}' \, \mathrm{d}\Sigma_{ij}. \tag{42}$$

- GIBC(a, b):

$$\int_{\Sigma_{ij}} \mathcal{S}(\gamma_i^T(\widetilde{\mathbf{E}}_i^{n+1})) \cdot \widetilde{\mathbf{E}}_i' \, \mathrm{d}\Sigma_{ij} := \int_{\Sigma_{ij}} \imath k \mathbf{r} \cdot \widetilde{\mathbf{E}}_i' \, \mathrm{d}\Sigma_{ij}, \tag{43}$$

where the function $\mathbf{r} \in \mathbf{H}(\mathbf{curl}, \Sigma_{ij})$ is obtained through the solution of

$$\begin{cases} \text{Find } \mathbf{r} \text{ in } \mathbf{H}(\mathbf{curl}, \Sigma_{ij}) \text{ and } \rho \text{ in } H^1(\Sigma_{ij}) \text{ such that } \forall\ \mathbf{r}' \in \mathbf{H}(\mathbf{curl}, \Sigma_{ij}) \\ \text{and } \forall \rho' \in H^1(\Sigma_{ij}) : \\ -\displaystyle\int_{\Sigma_{ij}} \frac{a}{k^2} \nabla_{\Sigma_{ij}} \rho \cdot \mathbf{r}' \, \mathrm{d}\Sigma_{ij} - \int_{\Sigma_{ij}} \mathbf{r} \cdot \mathbf{r}' \, \mathrm{d}\Sigma_{ij} + \int_{\Sigma_{ij}} \gamma_i^T (\widetilde{\mathbf{E}}_i^{n+1}) \cdot \mathbf{r}' \, \mathrm{d}\Sigma_{ij} \\ \qquad\qquad - \displaystyle\int_{\Sigma_{ij}} \frac{b}{k^2} \operatorname{curl}_{\Sigma_{ij}} (\gamma_i^T (\widetilde{\mathbf{E}}_i^{n+1})) \operatorname{curl}_{\Sigma_{ij}} \mathbf{r}' \, \mathrm{d}\Sigma_{ij} = 0, \\ \displaystyle\int_{\Sigma_{ij}} \rho \rho' \, \mathrm{d}\Sigma_{ij} + \int_{\Sigma_{ij}} \mathbf{r} \cdot \nabla_{\Sigma_{ij}} \rho' \, \mathrm{d}\Sigma_{ij} = 0; \end{cases} \tag{44}$$

$$\int_{\Sigma_{ij}} \mathcal{S}(\gamma_i^T (\widetilde{\mathbf{E}}_i^{n+1})) \cdot \mathbf{g}_{ji}' \, \mathrm{d}\Sigma_{ij} := \int_{\Sigma_{ij}} \imath k \mathbf{r} \cdot \mathbf{g}_{ji}' \, \mathrm{d}\Sigma_{ij}. \tag{45}$$

- GIBC(N_p, α, ε):

$$\int_{\Sigma_{ij}} \mathcal{S}(\gamma_i^T (\widetilde{\mathbf{E}}_i^{n+1})) \cdot \widetilde{\mathbf{E}}_i' \, \mathrm{d}\Sigma_{ij} := \int_{\Sigma_{ij}} \imath k \mathbf{r} \cdot \widetilde{\mathbf{E}}_i' \, \mathrm{d}\Sigma_{ij}, \tag{46}$$

where the function $\mathbf{r} \in \mathbf{H}(\mathbf{curl}, \Sigma_{ij})$ is obtained through the solution of

$$\begin{cases} \text{Find } \mathbf{r} \text{ in } \mathbf{H}(\mathbf{curl}, \Sigma_{ij}), \text{ and for } \ell = 1, \ldots, N_p, \boldsymbol{\varphi}_\ell \text{ in } \mathbf{H}(\mathbf{curl}, \Sigma_{ij}) \text{ and } \rho_\ell \text{ in } H^1(\Sigma_{ij}) \\ \text{such that } \forall \mathbf{r}' \in \mathbf{H}(\mathrm{curl}, \Sigma_{ij}), \forall \boldsymbol{\varphi}_\ell' \in \mathbf{H}(\mathrm{curl}, \Sigma_{ij}) \text{ and } \forall \rho_\ell' \in H^1(\Sigma_{ij}) : \\ \displaystyle\int_{\Sigma_{ij}} C_0 \mathbf{r} \cdot \mathbf{r}' \, \mathrm{d}\Sigma_{ij} - \int_{\Sigma_{ij}} \gamma_i^T (\widetilde{\mathbf{E}}_i^{n+1}) \cdot \mathbf{r}' \, \mathrm{d}\Sigma_{ij} + \int_{\Sigma_{ij}} \frac{1}{k_\varepsilon^2} \operatorname{curl}_{\Sigma_{ij}} (\gamma_i^T (\widetilde{\mathbf{E}}_i^{n+1})) \operatorname{curl}_{\Sigma_{ij}} \mathbf{r}' \, \mathrm{d}\Sigma_{ij} \\ \qquad + \displaystyle\sum_{\ell=1}^{N_p} A_\ell \left[\int_{\Sigma_{ij}} \nabla_{\Sigma_{ij}} \rho_\ell \cdot \mathbf{r}' \, \mathrm{d}\Sigma_{ij} - \int_{\Sigma_{ij}} \frac{1}{k_\varepsilon^2} \operatorname{curl}_{\Sigma_{ij}} \boldsymbol{\varphi}_\ell \operatorname{curl}_{\Sigma_{ij}} \mathbf{r}' \, \mathrm{d}\Sigma_{ij} \right] = 0, \\ \displaystyle\int_{\Sigma_{ij}} \boldsymbol{\varphi}_\ell \cdot \boldsymbol{\varphi}_\ell' \, \mathrm{d}\Sigma_{ij} + B_\ell \left[\int_{\Sigma_{ij}} \nabla_{\Sigma_{ij}} \rho_\ell \cdot \boldsymbol{\varphi}_\ell' \, \mathrm{d}\Sigma_{ij} - \int_{\Sigma_{ij}} \frac{1}{k_\epsilon^2} \operatorname{curl}_{\Sigma_{ij}} \boldsymbol{\varphi}_\ell \operatorname{curl}_{\Sigma_{ij}} \boldsymbol{\varphi}_\ell' \, \mathrm{d}\Sigma_{ij} \right] \\ \qquad - \displaystyle\int_{\Sigma_{ij}} \mathbf{r} \cdot \boldsymbol{\varphi}_\ell' \, \mathrm{d}\Sigma_{ij} = 0, \quad \ell = 1, \ldots, N_p, \\ \displaystyle\int_{\Sigma_{ij}} \rho_\ell \rho_\ell' \, \mathrm{d}\Sigma_{ij} + \int_{\Sigma_{ij}} \frac{1}{k_\varepsilon^2} \boldsymbol{\varphi}_\ell \cdot \nabla_{\Sigma_{ij}} \rho_\ell' \, \mathrm{d}\Sigma_{ij} = 0, \quad \ell = 1, \ldots, N_p; \end{cases} \tag{47}$$

$$\int_{\Sigma_{ij}} \mathcal{S}(\gamma_i^T (\widetilde{\mathbf{E}}_i^{n+1})) \cdot \mathbf{g}_{ji}' \, \mathrm{d}\Sigma_{ij} := \int_{\Sigma_{ij}} \imath k \mathbf{r} \cdot \mathbf{g}_{ji}' \, \mathrm{d}\Sigma_{ij}. \tag{48}$$

- PML(σ):

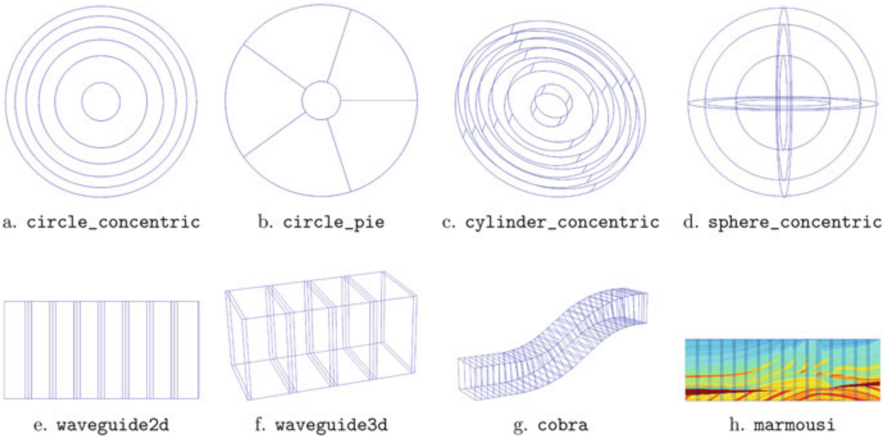
$$\begin{aligned} \int_{\Sigma_{ij}} \mathcal{S}(\gamma_i^T (\widetilde{\mathbf{E}}_i^{n+1})) \cdot \widetilde{\mathbf{E}}_i' \, \mathrm{d}\Sigma_{ij} := \int_{\Omega_{ij}^{\mathrm{PML}}} D^{-1} \, \mathbf{curl}\, \widetilde{\mathbf{E}}_i^{n+1} \cdot \mathbf{curl}\, \widetilde{\mathbf{E}}_i' \, \mathrm{d}\Omega_{ij}^{\mathrm{PML}} \\ - \int_{\Omega_{ij}^{\mathrm{PML}}} D \, k^2 \widetilde{\mathbf{E}}_i^{n+1} \cdot \widetilde{\mathbf{E}}_i' \, \mathrm{d}\Omega_{ij}^{\mathrm{PML}}; \end{aligned} \tag{49}$$

$$\int_{\Sigma_{ij}} \mathcal{S}(\gamma_i^T(\widetilde{\mathbf{E}}_i^{n+1}))\cdot \mathbf{g}'_{ji}\,\mathrm{d}\Sigma_{ij} := \int_{\Omega_{ij}^{\mathrm{PML}}} D^{-1}\,\mathbf{curl}\,\widetilde{\mathbf{E}}_i^{n+1}\cdot\mathbf{curl}\,\mathbf{g}'_{ji}\,\mathrm{d}\Omega_{ij}^{\mathrm{PML}} - \int_{\Omega_{ij}^{\mathrm{PML}}} D\,k^2\widetilde{\mathbf{E}}_i^{n+1}\cdot \mathbf{g}'_{ji}\,\mathrm{d}\Omega_{ij}^{\mathrm{PML}}, \tag{50}$$

where the tensor D is defined as for the acoustic case and the test functions $\mathbf{g}'_{ji}$ are again extended to the volume of the PML layers.

4 Numerical Implementation

The domain decomposition methods analyzed above are all readily available for testing using finite element methods in the open source GetDDM software environment [33, 48], available online on the web site of the ONELAB projet [29, 30]: http://onelab.info/wiki/GetDDM. GetDDM is based on the open source finite element solver GetDP (http://getdp.info) [17, 18, 28] and the open source mesh generator Gmsh (http://gmsh.info) [31, 32]. Various 2D and 3D test-cases are provided online (see Fig. 6) for both acoustic and electromagnetic wave problems, as well as detailed instructions on how to build the software for parallel computer architectures. Pre-compiled, serial versions of the software for Windows, MacOS and Linux are also available for development and testing.



a. circle_concentric
b. circle_pie
c. cylinder_concentric
d. sphere_concentric
e. waveguide2d
f. waveguide3d
g. cobra
h. marmousi

Fig. 6 Sample models available online at http://onelab.info/wiki/GetDDM. (**a**)–(**d**) Acoustic or electromagnetic (**c** and **d** only) scattering by cylindrical or spherical obstacles, with concentric or radial subdomains [10, 21]. (**e**) and (**f**) Guided acoustic or electromagnetic waves in rectangular waveguides [51]. (**g**) Guided acoustic or electromagnetic waves in the COBRA benchmark defined by the JINA98 workgroup [52]. (**h**) Acoustic waves in the underground Marmousi model [47]

While GetDDM is written in C++, all the problem-specific data (geometry description, finite element formulation with appropriate transmission condition, domain decomposition algorithm) are directly written in input ASCII text files, using the code's built-in language. This general implementation allows to solve a wide variety of problems with the same software, without recompilation, and hides all the complexities of the finite element implementation from the end-user (in particular the MPI-based parallelization). Moreover, the software is designed to work both on small- and medium-scale problems (on a workstation, a laptop, a tablet or even a mobile phone) and on large-scale problems on high-performance computing clusters, without changing the input files.

One of the main features of the environment is the closeness between the input data files and the symbolic mathematical expressions of the problems. In particular, the weak formulations presented in Sects. 2.3 and 3.3 are directly transcribed symbolically in the input files. For example, the relevant terms of the finite element formulation for the Maxwell problem using IBC(0) as transmission condition are directly written as follows in the input file:

```
Galerkin { [ Dof{Curl E~{i}}, {Curl E~{i}} ];
           In Omega~{i}; Integration I; Jacobian V; }
Galerkin { [ -k[]^2 * Dof{E~{i}}, {E~{i}} ];
           In Omega~{i}; Integration I; Jacobian V; }
Galerkin { [ -I[] * k[] * N[] /\ (Dof{E~{i}} /\ N[]), {E~{i}}];
           In GammaInf~{i}; Integration I; Jacobian S; }
Galerkin { [ g~{i}[], {E~{i}} ];
           In Sigma~{i}; Integration I; Jacobian S; }
Galerkin { [ -I[] * k[] * N[] /\ (Dof{E~{i}} /\ N[]), {E~{i}}];
           In Sigma~{i}; Integration I; Jacobian S; }
```

where `Dof{E~{i}}` corresponds to the discrete unknown in the ith subdomain `Omega~{i}` and `[.,.]` denotes the inner product. Other transmission conditions are implemented in a similar way, as is the update relation. The parallel implementation of the iterative algorithm uses the built-in function `IterativeLinearSolver`, which takes as argument the operations that implement the matrix-vector product required by Krylov subspace solvers, and is based on PETSc [5] and MUMPS [2] for the parallel (MPI-based) implementation of the linear algebra routines.

For illustration purposes, Fig. 7 presents some other cases that have been solved using GetDDM. Published references are provided, which contain further information about the specific test cases, mathematical models and numerical results.

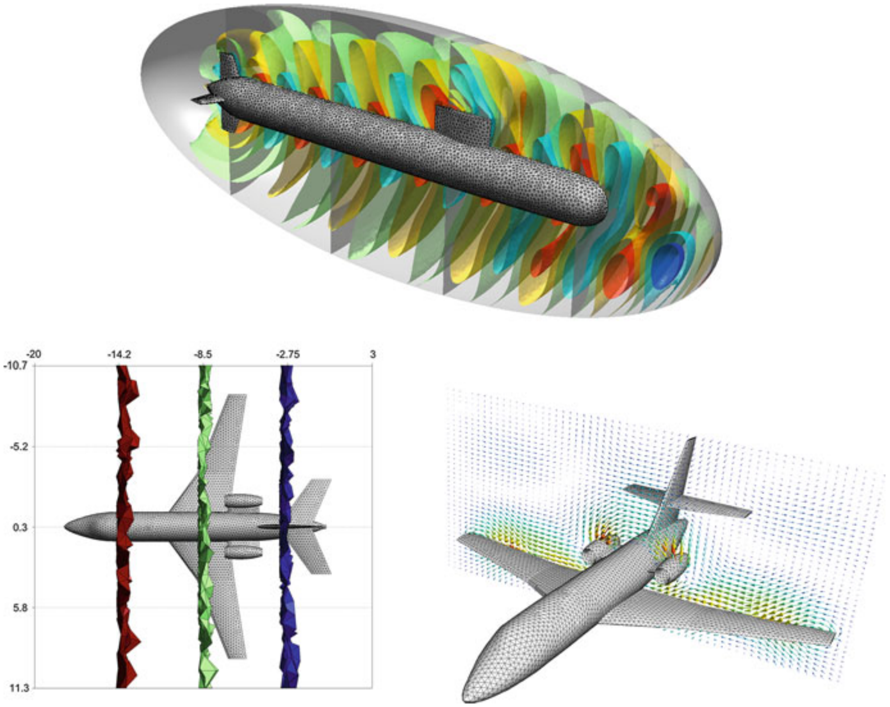


Fig. 7 Sample models solved with GetDDM. *Top*: acoustic waves around a submarine (image reproduced from [10]). *Bottom*: electromagnetic waves around a Falcon aircraft (images reproduced from [21])

Acknowledgements This work was supported in part by the Wallonia-Brussels Federation (ARC grant for Concerted Research Actions ARC WAVES 15/19-03), the Belgian Science Policy (PAI grant P7/02), the Walloon Region (WIST3 grants ONELAB and ALIZEES), the French ANR (grant MicroWave NT09 460489 “Programme Blanc”) and the “EADS Foundation” (High-BRID project, grant 089-1009-1006). Computational resources have been provided by CÉCI, funded by F.R.S.-FNRS (Fonds de la Recherche Scientifique) under grant n°2.5020.11, and the Tier-1 supercomputer of the Fédération Wallonie-Bruxelles, funded by the Walloon Region under grant n°1117545.

References

1. A. Alonso-Rodriguez and L. Gerardo-Giorda. New nonoverlapping domain decomposition methods for the harmonic Maxwell system. *SIAM J. Sci. Comput.*, 28(1):102–122, 2006.
2. P. R. Amestoy, I. S. Duff, J. Koster, and J.-Y. L’Excellent. A fully asynchronous multifrontal solver using distributed dynamic scheduling. *SIAM Journal on Matrix Analysis and Applications*, 23(1):15–41, 2001.

3. X. Antoine, M. Darbas, and Y.Y. Lu. An improved surface radiation condition for high-frequency acoustic scattering problems. *Comput. Methods Appl. Mech. Engrg.*, 195(33–36):4060–4074, 2006.
4. X. Antoine, C. Geuzaine, and K. Ramdani. *Wave Propagation in Periodic Media - Analysis, Numerical Techniques and Practical Applications*, volume 1, chapter Computational Methods for Multiple Scattering at High Frequency with Applications to Periodic Structures Calculations, pages 73–107. Progress in Computational Physics, 2010.
5. S. Balay, M. F. Adams, J. Brown, P. Brune, K. Buschelman, V. Eijkhout, W. D. Gropp, D. Kaushik, M. G. Knepley, L. Curfman McInnes, K. Rupp, B. F. Smith, and H. Zhang. PETSc Web page. http://www.mcs.anl.gov/petsc, 2015.
6. A. Bayliss, M. Gunzburger, and E. Turkel. Boundary conditions for the numerical solution of elliptic equations in exterior regions. *SIAM J. Appl. Math.*, 42(2):430–451, 1982.
7. J.-P. Berenger. A perfectly matched layer for the absorption of electromagnetic waves. *J. Comput. Phys.*, 114(2):185–200, 1994.
8. A. Bermúdez, L. Hervella-Nieto, A. Prieto, and R. Rodríguez. An optimal perfectly matched layer with unbounded absorbing function for time-harmonic acoustic scattering problems. *J. Comput. Phys.*, 223(2):469–488, 2007.
9. Y. Boubendir. An analysis of the BEM-FEM non-overlapping domain decomposition method for a scattering problem. *J. Comput. Appl. Math.*, 204(2):282–291, 2007.
10. Y. Boubendir, X. Antoine, and C. Geuzaine. A quasi-optimal non-overlapping domain decomposition algorithm for the Helmholtz equation. *Journal of Computational Physics*, 231(2):262–280, 2012.
11. Y. Boubendir, A. Bendali, and M. B. Fares. Coupling of a non-overlapping domain decomposition method for a nodal finite element method with a boundary element method. *Internat. J. Numer. Methods Engrg.*, 73(11):1624–1650, 2008.
12. F. Collino and P. Monk. The perfectly matched layer in curvilinear coordinates. *SIAM J. Sci. Comput.*, 19(6):2061–2090 (electronic), 1998.
13. B. Després. *Méthodes de décomposition de domaine pour les problèmes de propagation d'ondes en régime harmonique. Le théorème de Borg pour l'équation de Hill vectorielle.* PhD thesis, Rocquencourt, 1991. Thèse, Université de Paris IX (Dauphine), Paris, 1991.
14. B. Després, P. Joly, and J. E. Roberts. A domain decomposition method for the harmonic Maxwell equations. In *Iterative methods in linear algebra (Brussels, 1991)*, pages 475–484, Amsterdam, 1992. North-Holland.
15. V. Dolean, J. M. Gander, S. Lanteri, J.-F. Lee, and Z. Peng. Optimized Schwarz methods for curl-curl time-harmonic Maxwell's equations. 2013.
16. V. Dolean, M. J. Gander, and L. Gerardo-Giorda. Optimized Schwarz methods for Maxwell's equations. *SIAM J. Sci. Comput.*, 31(3):2193–2213, 2009.
17. P. Dular and C. Geuzaine. GetDP Web page, http:://getdp.info, 2015. [online]. available: http::// getdp.info.
18. P. Dular, C. Geuzaine, F. Henrotte, and W. Legros. A general environment for the treatment of discrete problems and its application to the finite element method. *IEEE Transactions on Magnetics*, 34(5):3395–3398, September 1998.
19. M. El Bouajaji, X Antoine, and C. Geuzaine. Approximate local magnetic-to-electric surface operators for time-harmonic Maxwell's equations. *Journal of Computational Physics*, 279(15):241–260, 2014.
20. M. El Bouajaji, V. Dolean, M. Gander, and S. Lanteri. Optimized Schwarz methods for the time-harmonic Maxwell equations with damping. *SIAM Journal on Scientific Computing*, 34(4):A2048–A2071, 2012.
21. M. El Bouajaji, B. Thierry, X. Antoine, and C. Geuzaine. A quasi-optimal domain decomposition algorithm for the time-harmonic Maxwell's equations. *Journal of Computational Physics*, 294(1):38–57, 2015.
22. B. Engquist and A. Majda. Absorbing boundary conditions for the numerical simulation of waves. *Math. Comp.*, 31(139):629–651, 1977.

23. B. Engquist and L. Ying. Sweeping preconditioner for the Helmholtz equation: moving perfectly matched layers. *Multiscale Model. Simul.*, 9(2):686–710, 2011.
24. O.G. Ernst and M.J. Gander. Why it is difficult to solve Helmholtz problems with classical iterative methods. In Ivan G. Graham, Thomas Y. Hou, Omar Lakkis, and Robert Scheichl, editors, *Numerical Analysis of Multiscale Problems*, volume 83 of *Lecture Notes in Computational Science and Engineering*, pages 325–363. Springer Berlin Heidelberg, 2012.
25. M. Gander. Optimized Schwarz methods. *SIAM Journal on Numerical Analysis*, 44(2):699–731, 2006.
26. M. Gander and L. Halpern. Méthode de décomposition de domaine. Encyclopédie électronique pour les ingénieurs, 2012.
27. M. J. Gander, F. Magoulès, and F. Nataf. Optimized Schwarz methods without overlap for the Helmholtz equation. *SIAM J. Sci. Comput.*, 24(1):38–60 (electronic), 2002.
28. C. Geuzaine. GetDP: a general finite-element solver for the de Rham complex. In *PAMM Volume 7 Issue 1. Special Issue: Sixth International Congress on Industrial Applied Mathematics (ICIAM07) and GAMM Annual Meeting, Zürich 2007*, volume 7, pages 1010603–1010604. Wiley, 2008.
29. C. Geuzaine, F. Henrotte, E. Marchandise, J.-F. Remacle, P. Dular, and R. Vazquez Sabariego. ONELAB: Open Numerical Engineering LABoratory. *Proceedings of the 7th European Conference on Numerical Methods in Electromagnetism (NUMELEC2012)*, 2012.
30. C. Geuzaine, F. Henrotte, E. Marchandise, J.-F. Remacle, and R. Vazquez Sabariego. ONELAB Web page, http:://onelab.info, 2015. [online]. available: http:://onelab.info.
31. C. Geuzaine and J.-F. Remacle. Gmsh Web page, http:://gmsh.info, 2015. [online]. available: http:://gmsh.info.
32. C. Geuzaine and J.-F. Remacle. Gmsh: A 3-D finite element mesh generator with built-in pre- and post-processing facilities. *Internat. J. Numer. Methods Engrg.*, 79(11):1309–1331, 2009.
33. C. Geuzaine, B. Thierry, N. Marsic, D. Colignon, A. Vion, S. Tournier, Y. Boubendir, M. El Bouajaji, and X. Antoine. An open source domain decomposition solver for time-harmonic electromagnetic wave problems. In *Proceedings of the 2014 IEEE Conference on Antenna and Measurements and Applications, CAMA 2014*, November 2014.
34. D. Givoli. Computational absorbing boundaries. In Steffen Marburg and Bodo Nolte, editors, *Computational Acoustics of Noise Propagation in Fluids - Finite and Boundary Element Methods*, pages 145–166. Springer Berlin Heidelberg, 2008.
35. R. Kerchroud, X. Antoine, and A. Soulaimani. Numerical accuracy of a Padé-type non-reflecting boundary condition for the finite element solution of acoustic scattering problems at high-frequency. *International Journal for Numerical Methods in Engineering*, 64(10):1275–1302, 2005.
36. P.-L. Lions. On the Schwarz alternating method. III. A variant for nonoverlapping subdomains. In *Third International Symposium on Domain Decomposition Methods for Partial Differential Equations (Houston, TX, 1989)*, pages 202–223. SIAM, Philadelphia, PA, 1990.
37. FA Milinazzo, CA Zala, and GH Brooke. Rational square-root approximations for parabolic equation algorithms. *Journal of the Acoustical Society of America*, 101(2):760–766, FEB 1997.
38. A. Modave, E. Delhez, and C. Geuzaine. Optimizing perfectly matched layers in discrete contexts. *International Journal for Numerical Methods in Engineering*, 99(6):410–437, 2014.
39. A. Moiola and E. A. Spence. Is the Helmholtz equation really sign-indefinite? *SIAM Rev.*, 56(2):274–312, 2014.
40. F. Nataf. Interface connections in domain decomposition methods. *NATO Science Series II*, 75, 2001.
41. F. Nataf and F. Nier. Convergence rate of some domain decomposition methods for overlapping and nonoverlapping subdomains. *Numer. Math.*, 75:357–377, 1997.
42. J.-C. Nédélec. *Acoustic and electromagnetic equations*, volume 144 of *Applied Mathematical Sciences*. Springer-Verlag, New York, 2001. Integral representations for harmonic problems.
43. Z. Peng and J. Lee. A scalable nonoverlapping and nonconformal domain decomposition method for solving time-harmonic Maxwell equations in $\mathbb{R}^3$. *SIAM Journal on Scientific Computing*, 34(3):A1266–A1295, 2012.

44. Z. Peng, V. Rawat, and J.-F. Lee. One way domain decomposition method with second order transmission conditions for solving electromagnetic wave problems. *Journal of Computational Physics*, 229(4):1181–1197, 2010.
45. V. Rawat and J.-F. Lee. Nonoverlapping domain decomposition with second order transmission condition for the time-harmonic Maxwell's equations. *SIAM J. Scientific Computing*, 32(6):3584–3603, 2010.
46. Y. Saad and M. H. Schultz. GMRES: a generalized minimal residual algorithm for solving nonsymmetric linear systems. *SIAM J. Sci. Statist. Comput.*, 7(3):856–869, 1986.
47. C. Stolk. A rapidly converging domain decomposition method for the Helmholtz equation. *Journal of Computational Physics*, 241(0):240–252, 2013.
48. B. Thierry, A.Vion, S. Tournier, M. El Bouajaji, D. Colignon, N. Marsic, X. Antoine, and C. Geuzaine. GetDDM: an open framework for testing optimized Schwarz methods for time-harmonic wave problems. *Submitted to Computer Physics Communications*, 2015.
49. A. Toselli and O. Widlund. *Domain decomposition methods—algorithms and theory*, volume 34 of *Springer Series in Computational Mathematics*. Springer-Verlag, Berlin, 2005.
50. A. Vion, R. Bélanger-Rioux, L. Demanet, and C. Geuzaine. A DDM double sweep preconditioner for the Helmholtz equation with matrix probing of the DtN map. In *Mathematical and Numerical Aspects of Wave Propagation WAVES 2013*, June 2013.
51. A. Vion and C. Geuzaine. Double sweep preconditioner for optimized Schwarz methods applied to the Helmholtz problem. *Journal of Computational Physics*, 266(0):171–190, 2014.
52. A. Vion and C. Geuzaine. Parallel double sweep preconditioner for the optimized Schwarz algorithm applied to high frequency Helmholtz and Maxwell equations. In *LNCSE, Proc. of DD22*, 2014.

Computationally Efficient Boundary Element Methods for High-Frequency Helmholtz Problems in Unbounded Domains

Timo Betcke, Elwin van ’t Wout, and Pierre Gélat

Abstract This chapter presents the application of the boundary element method to high-frequency Helmholtz problems in unbounded domains. Based on a standard combined integral equation approach for sound-hard scattering problems we discuss the discretization, preconditioning and fast evaluation of the involved operators. As engineering problem, the propagation of high-intensity focused ultrasound fields into the human rib cage will be considered. Throughout this chapter we present code snippets using the open-source Python boundary element software BEM++ to demonstrate the implementation.

1 Introduction

The boundary element method (BEM) is an efficient and competitive tool to solve large-scale high-frequency Helmholtz problems in bounded or unbounded domains. Recent developments in fast matrix compression and preconditioning for boundary integral operators have pushed the computational limit of high-frequency boundary element computations such that problems in three dimensions with over a hundred wavelengths across the domain can be solved on a single workstation [47]. Furthermore, the availability of high-level software libraries allows for a convenient implementation of different boundary integral formulations [41]. This combination makes it possible to solve large-scale problems of engineering interest effectively with the BEM.

T. Betcke
Department of Mathematics, University College London, London, UK
e-mail: t.betcke@ucl.ac.uk

E. van ’t Wout (✉)
School of Engineering and Faculty of Mathematics, Pontificia Universidad Católica de Chile, Santiago, Chile
e-mail: e.wout@uc.cl

P. Gélat
Department of Mechanical Engineering, University College London, London, UK
e-mail: p.gelat@ucl.ac.uk

D. Lahaye et al. (eds.), *Modern Solvers for Helmholtz Problems*, Geosystems Mathematics, DOI 10.1007/978-3-319-28832-1_9

This chapter will deal with exterior scattering of sound waves. In this case, a bounded domain $\Omega^- \subset \mathbb{R}^3$ is immersed in a homogeneous unbounded region $\Omega^+ := \mathbb{R}^3 \setminus \overline{\Omega^-}$ and excited by a harmonic wave with a fixed wavenumber k. Notice that the object has to be bounded but not necessarily connected. The main objective is the computation of the total wave field u^{tot} obtained from the scattering of an incident wave field u^{inc} at the object. For rigid objects, we have a sound-hard condition at the boundary Γ, which is assumed to be Lipschitz continuous with unit normal direction $\hat{\mathbf{n}}$ outward pointing. This scattering problem is modeled by the Helmholtz system

$$-\Delta u^{\text{tot}} - k^2 u^{\text{tot}} = 0 \text{ in } \Omega^+, \tag{1a}$$

$$\frac{\partial u^{\text{tot}}}{\partial \hat{\mathbf{n}}} = 0 \text{ on } \Gamma, \tag{1b}$$

$$\lim_{|\mathbf{x}| \to \infty} |\mathbf{x}| \left(\frac{\partial u^{\text{sca}}}{\partial |x|} - iku^{\text{sca}} \right) = 0 \tag{1c}$$

where the last equation is the Sommerfeld radiation condition at infinity. Here, u^{sca} denotes the scattered field, such that $u^{\text{tot}} = u^{\text{inc}} + u^{\text{sca}}$. The scatterer object is assumed to be impenetrable, hence $u^{\text{tot}} = 0$ in Ω^-.

Helmholtz problems are often solved with computational methods such as finite-difference, finite-element and spectral techniques. As opposed to these volume-based algorithms, we will use the surface-based BEM [39, 40, 43]. The basic idea behind the BEM is to reformulate the Helmholtz system into a boundary integral formulation and solve the scattering problem on the surface itself. In this chapter we will review the design of boundary integral equations with an emphasis on large-scale scattering problems at high frequencies. For this case, it is necessary to use modern matrix compression and preconditioning techniques. We will apply these state-of-the-art techniques to a challenging problem arising from medical high-intensity focused ultrasound simulations [25]. In [47] we have published an earlier version of some of the techniques presented in this chapter. There, more details about the engineering application can be found. Here, we give a more detailed analysis of the boundary integral formulations, include other formulations as well and explain the compression technique. Furthermore, this chapter uses a newer version of BEM++ which allows us to perform experiments on a larger scale.

The explicit use of the acoustic Green's function gives the BEM some major advantages compared to standard computational methods. First of all, the Sommerfeld radiation condition (1c) is exactly satisfied by boundary integral representations. There is thus no need for absorbing boundary conditions to artificially truncate the exterior region, as is required for volume-based discretization techniques [28]. This makes the BEM a natural choice for solving scattering problems in unbounded domains. Another positive effect from the Green's function is that well-chosen discretizations are essentially free of pollution and dispersion, even for low order discretizations using piecewise constant basis functions [29]. Furthermore, since the model equations live on the boundary only, surface meshes are being used. These are often easier to generate for complex geometries compared to volume meshes.

On the other hand, the BEM is not free of problems. For instance, it is crucial to carefully choose the correct type of boundary integral equation formulation. In particular for high-frequency problems it is necessary to choose a formulation that does not suffer from breakdown at certain resonant frequencies [1, 2]. This will be the topic of Sect. 2.

In the case of large-scale simulations, the discrete system of equations is typically being solved with iterative linear solvers, which are asymptotically more efficient than direct solvers [3]. Furthermore, these methods mainly rely on matrix-vector multiplications, which are relatively easy to parallelize and for which acceleration algorithms are available. However, the required number of iterations can easily become prohibitively large for high-frequency problems, especially for the classical boundary integral formulations. In Sect. 3 we therefore review various operator preconditioning techniques for high-frequency applications and numerically assess their performance in Sect. 5.2.

A naive discretization of the boundary integral operators would lead to dense matrix problems and a complexity of $\mathscr{O}\left(N^2\right)$ for the assembly and the matrix-vector product, where N is the number of elements. For a fixed number of surface elements per wavelength, i.e., $N \sim k^2$, the complexity will therefore scale as $\mathscr{O}\left(k^4\right)$. This is only feasible for small-scale problems. For large-scale applications it is vital to use acceleration schemes that reduce the computation time and memory footprint to realistic measures for present-day computer architectures. The most prominent of such methods are Fast Multiple Methods (FMM) [16, 17, 23] and hierarchical matrix techniques ($\mathscr{H}$-matrices and their $\mathscr{H}^2$ and HSS variants) [6, 8, 32, 34, 48]. They achieve a complexity of $\mathscr{O}(N)$ or $\mathscr{O}(N\log(N))$ for the matrix-vector multiplication, depending on the frequency regime and the specific implementation. In Sect. 4 we will discuss the behavior of classical $\mathscr{H}$-matrix techniques for exterior scattering problems in more detail. While their complexity with respect to a growing wavenumber k is asymptotically not as good as high-frequency FMM, they are kernel-independent, relatively easy to implement and offer good performance for a wide range of application relevant frequencies.

The numerical implementation of a high-frequency BEM is challenging, mainly because of the necessity of specialized acceleration techniques and quadrature rules for singular integral operators. In Sect. 5 we will introduce the open-source software library BEM++ [41] which has been used to perform all computational experiments in this chapter. This library was originally developed at University College London and provides a comprehensive Python toolbox to setup and solve Laplace, Helmholtz and Maxwell problems via the BEM. Matrix compression is integrated and various preconditioners are available for the efficient solution of large-scale problems. Fast computations are achieved because the core discretization and compression routines are written in C++. All these routines are accessible via a high-level Python interface, which provides a user-friendly programming environment. We will present code examples to demonstrate how, with only a limited amount of high-level instructions, an entire BEM simulation can be performed with

BEM++. Tutorials in the form of IPython notebooks can be downloaded from the website of the BEM++ project (www.bempp.org).

Finally, in Sect. 6 we present the application of the fast BEM to a realistic problem arising from medical treatment planning in high-frequency focused ultrasound. The described problem will lead to a system with around half a million unknowns and simulates over one hundred wavelengths across the computational domain. This has been solved with BEM++ on a single workstation, thus confirming the capabilities of the efficient BEM presented in this chapter.

2 Boundary Integral Formulations of High-Frequency Scattering

In this section we review the standard combined field equations for boundary integral formulations of high-frequency scattering. Details and proofs of the statements given here can be found in standard textbooks such as [39, 40, 43]. A recent overview article of novel mathematical developments for high-frequency scattering formulations based on hybrid numerical-asymptotic methods is also given in [15]. While these hybrid numerical-asymptotic methods have the potential to solve scattering problems on certain geometries with an almost wavenumber independent convergence, they are not yet suitable for larger industrial applications with realistic meshes.

2.1 Surface Representation of the Scattering Model

The reformulation of the exterior model into a surface model necessitates operators that map between the volume $\Omega^- \cup \Omega^+$ and the boundary Γ. The map from the volume to the boundary is provided by the *trace operators*, which are denoted by γ. More specifically, the Dirichlet trace operators γ_0^- and γ_0^+ are defined as the limit values of a field towards the interface from the interior and exterior domain, respectively, and the Neumann trace operators γ_1^- and γ_1^+ are the corresponding normal derivatives. On the other hand, the *potential operators* map from the surface to the volume. They are defined as

$$(\mathscr{V}\psi)(\mathbf{x}) := \int_\Gamma G(\mathbf{x},\mathbf{y})\psi(\mathbf{y})\,\mathrm{d}\Gamma(\mathbf{y}) \qquad \text{for } \mathbf{x} \in \Omega^- \cup \Omega^+, \tag{2}$$

$$(\mathscr{K}\phi)(\mathbf{x}) := \int_\Gamma \partial_{n(\mathbf{y})} G(\mathbf{x},\mathbf{y})\phi(\mathbf{y})\,\mathrm{d}\Gamma(\mathbf{y}) \qquad \text{for } \mathbf{x} \in \Omega^- \cup \Omega^+ \tag{3}$$

and are called the single-layer and double-layer potential operators, respectively. Here, ψ and ϕ denote surface potentials that live on the boundary only. The function

$G(\mathbf{x}, \mathbf{y})$ is the acoustic Green's function defined by

$$G(\mathbf{x}, \mathbf{y}) := \frac{e^{ik|\mathbf{x}-\mathbf{y}|}}{4\pi |\mathbf{x}-\mathbf{y}|} \qquad \text{for } \mathbf{x} \neq \mathbf{y} \tag{4}$$

and $\partial_{n(\mathbf{y})} G(\mathbf{x}, \mathbf{y})$ is its normal derivative along $\hat{\mathbf{n}}$ with respect to $\mathbf{y}$.

Using the single-layer and double-layer potential operator one can derive a *representation formula* for any radiating solution u of the Helmholtz equation as

$$u(\mathbf{x}) = (\mathscr{V}\psi)(\mathbf{x}) - (\mathscr{K}\phi)(\mathbf{x}) \quad \text{for } \mathbf{x} \in \Omega^- \cup \Omega^+ \tag{5}$$

with

$$\psi = \gamma_1^- u - \gamma_1^+ u, \tag{6a}$$

$$\phi = \gamma_0^- u - \gamma_0^+ u \tag{6b}$$

being the jumps of the solution across the interface.

Taking the trace or normal derivative of both sides of the equality in Eq. (5) will result in an equation that is fully defined on the boundary. This necessitates the analysis of the traces and normal derivatives of potential operators. One can show that the following boundary operators are well defined almost everywhere if Γ is piecewise smooth:

$$(V\psi)(\mathbf{x}) := \int_\Gamma G(\mathbf{x}, \mathbf{y})\psi(\mathbf{y})\, \mathrm{d}\Gamma(\mathbf{y}) \qquad \text{for } \mathbf{x} \in \Gamma, \tag{7}$$

$$(K\phi)(\mathbf{x}) := \int_\Gamma \partial_{n(\mathbf{y})} G(\mathbf{x}, \mathbf{y})\phi(\mathbf{y})\, \mathrm{d}\Gamma(\mathbf{y}) \qquad \text{for } \mathbf{x} \in \Gamma, \tag{8}$$

$$(T\psi)(\mathbf{x}) := \int_\Gamma \partial_{n(\mathbf{x})} G(\mathbf{x}, \mathbf{y})\psi(\mathbf{y})\, \mathrm{d}\Gamma(\mathbf{y}) \qquad \text{for } \mathbf{x} \in \Gamma, \tag{9}$$

$$(D\phi)(\mathbf{x}) := -\partial_{n(\mathbf{x})} \int_\Gamma \partial_{n(\mathbf{y})} G(\mathbf{x}, \mathbf{y})\phi(\mathbf{y})\, \mathrm{d}\Gamma(\mathbf{y}) \qquad \text{for } \mathbf{x} \in \Gamma. \tag{10}$$

Moreover, for piecewise smooth Γ the following *jump relations* are defined almost everywhere:

$$V\psi = \gamma_0^-(\mathscr{V}\psi) = \gamma_0^+(\mathscr{V}\psi), \tag{11}$$

$$K\phi = \gamma_0^-(\mathscr{K}\phi) + \frac{1}{2}\phi = \gamma_0^+(\mathscr{K}\phi) - \frac{1}{2}\phi, \tag{12}$$

$$T\psi = \gamma_1^-(\mathscr{V}\psi) - \frac{1}{2}\psi = \gamma_1^+(\mathscr{V}\psi) + \frac{1}{2}\psi, \tag{13}$$

$$D\phi = -\gamma_1^-(\mathscr{K}\phi) = -\gamma_1^+(\mathscr{K}\phi). \tag{14}$$

For the precise definition in the general Lipschitz case see e.g. [43, Chap. 6].

The operators V, K, T, and D are called the single-layer, double-layer, adjoint double-layer and hypersingular boundary integral operator, respectively, and satisfy the mapping properties

$$\begin{aligned} &V : \mathscr{H}^{-\frac{1}{2}}(\Gamma) \to \mathscr{H}^{\frac{1}{2}}(\Gamma), && K : \mathscr{H}^{\frac{1}{2}}(\Gamma) \to \mathscr{H}^{\frac{1}{2}}(\Gamma), \\ &T : \mathscr{H}^{-\frac{1}{2}}(\Gamma) \to \mathscr{H}^{-\frac{1}{2}}(\Gamma), && D : \mathscr{H}^{\frac{1}{2}}(\Gamma) \to \mathscr{H}^{-\frac{1}{2}}(\Gamma) \end{aligned}$$

for fractional Sobolev spaces $\mathscr{H}^{\frac{1}{2}}(\Gamma)$ and $\mathscr{H}^{-\frac{1}{2}}(\Gamma)$. In addition, the identity boundary operator is denoted by I. Boundary integral equations can now readily be derived by taking traces of representation formulas. The simplest forms are based on the normal derivative of the single-layer or double-layer potential operator only. Drawback of these operators is their nontrivial nullspace at resonant frequencies. An effective approach to mitigate the breakdown at resonances is to consider combined field integral equations that are uniquely solvable for all real wavenumbers.

2.2 The Burton-Miller Combined Boundary Integral Equation

A classical combined field integral equation for the scattering problem (1) is the Burton-Miler formulation [13]. This formulation is free of spurious resonances and the unique solution has a direct interpretation as the trace of the exterior total field on the boundary Γ. We start with the direct representation (5) of the scattered field, i.e., $u^{\text{sca}} = \mathscr{V}\psi - \mathscr{K}\phi$ where the surface potentials ψ and ϕ are given by the jumps of the scattered field across the boundary and can be simplified as

$$\begin{aligned} \psi &= \gamma_1^- u^{\text{sca}} - \gamma_1^+ u^{\text{sca}} = \gamma_1^-(u^{\text{tot}} - u^{\text{inc}}) + \gamma_1^+ u^{\text{inc}} = 0, \\ \phi &= \gamma_0^- u^{\text{sca}} - \gamma_0^+ u^{\text{sca}} = \gamma_0^-(u^{\text{tot}} - u^{\text{inc}}) - \gamma_0^+(u^{\text{tot}} - u^{\text{inc}}) = -\gamma_0^+ u^{\text{tot}} \end{aligned}$$

because the total field is zero in the interior and the incident wave field smooth across the boundary. This reduces the representation formula to

$$u^{\text{sca}} = \mathscr{K}(\varphi), \qquad \varphi = \gamma_0^+ u^{\text{tot}}. \tag{15}$$

Taking the exterior Neumann trace γ_1^+ of this representation formula yields

$$-\gamma_1^+ u^{\text{inc}} = -D\varphi \tag{16}$$

where the boundary condition and jump relation (14) have been used. The interior Dirichlet trace γ_0^- of the representation formula results in

$$-\gamma_0^+ u^{\text{inc}} = K\varphi - \frac{1}{2}\varphi \tag{17}$$

where the zero interior field, jump relation (12) and smoothness of the incident wave field have been used.

Both boundary integral equations (16) and (17) solve the scattering problem for the same surface potential. Any linear combination will therefore solve the scattering problem as well. That is, for a coupling parameter $\eta \in \mathbb{C}$, the Burton-Miller formulation

$$A_\eta \varphi = u^{\text{inc}} + \eta \partial_n u^{\text{inc}} \tag{18}$$

with

$$A_\eta := \left(\tfrac{1}{2}I - K\right)\varphi + \eta D\varphi$$

solves the scattering problem with the representation formula (15). The Burton-Miller formulation is uniquely solvable for $\Im(\eta) \neq 0$ and $\eta = i/k$ is a good choice of coupling parameter [36].

2.3 Regularizing the Burton-Miller Formulation

We notice that the Burton-Miller formulation (18) is not without problems. The operator $\left(\frac{1}{2}I - K\right)$ is minus the interior trace of the double layer potential operator $\mathscr{K}$ and maps from $\mathscr{H}^{\frac{1}{2}}(\Gamma)$ into $\mathscr{H}^{\frac{1}{2}}(\Gamma)$, whereas the hypersingular operator D maps from $\mathscr{H}^{\frac{1}{2}}(\Gamma)$ into $\mathscr{H}^{-\frac{1}{2}}(\Gamma)$. A solution to this mismatch in mapping characteristics is to consider regularized combined field operators [12]. For a regularization operator

$$\mathscr{R} : \mathscr{H}^{-\frac{1}{2}}(\Gamma) \to \mathscr{H}^{\frac{1}{2}}(\Gamma),$$

the regularized Burton-Miller formulation reads

$$\left(\tfrac{1}{2}I - K\right)\varphi + \mathscr{R}D\varphi = u^{\text{inc}} + \mathscr{R}\partial_n u^{\text{inc}}, \tag{19}$$

where now the operator $A_{\mathscr{R}} := \left(\frac{1}{2}I - K\right) + \mathscr{R}D$ is well defined on $\mathscr{H}^{\frac{1}{2}}(\Gamma)$. The design of sophisticated regularization techniques forms the basis of the efficient preconditioning strategies discussed in Sect. 3.

2.4 Indirect Formulations

An alternative approach to obtaining a combined field integral equation for the scattering problem (1) is to use an indirect representation of the scattered field as the linear combination

$$u^{\mathrm{sca}} = -i\mu\mathscr{V}\phi + \mathscr{K}(\mathscr{R}\phi) \tag{20}$$

where regularization with $\mathscr{R}$ has been applied. Taking the exterior Neumann trace γ_1^+ on both sides and using $\partial_n u^{\mathrm{inc}} = -\partial_n u^{\mathrm{sca}}$ on boundary Γ results in

$$-\partial_n u^{\mathrm{inc}} = i\mu\left(\tfrac{1}{2}I - T\right)\phi - D(\mathscr{R}\phi). \tag{21}$$

Traditionally, Eq. (21) without the regularization is called the Brakhage-Werner formulation [9]. In [11] it is suggested to use $\mu = 1$ for high-frequency scattering problems.

2.5 Boundary Element Methods

For the discretization of boundary integral operators typically either collocation or Galerkin methods are used. While collocation methods are easier to implement, the Galerkin method has advantages with respect to coupling with finite element methods, symmetry of the resulting operators, and assembly on non-smooth domains. Here, we focus on Galerkin methods for the Burton-Miller formulation (18).

Let Γ_h be a triangulation of Γ with n nodes $\hat{\mathbf{x}}_j$, $j = 1, \ldots, n$. Let ϕ_j be a continuous piecewise linear function defined on Γ_h such that $\phi_j(\hat{\mathbf{x}}_i) = \begin{cases} 1, & i = j \\ 0, & i \neq j \end{cases}$. Let us denote by $V_h := \left\{\sum_{j=1}^n v_j\phi_j,\ v_j \in \mathbb{C}\right\}$ the space spanned by the nodal basis functions ϕ_j. Define the standard real dual pairing

$$\langle\varphi, \vartheta\rangle := \int_\Gamma \varphi(\mathbf{y}) \cdot \vartheta(\mathbf{y})\,\mathrm{d}\Gamma(\mathbf{y}). \tag{22}$$

The Galerkin discretization of the Burton-Miller formulation is now given as the discrete matrix problem

$$\overline{A}_\eta\mathbf{v} = \mathbf{b}$$

with $\left[\overline{A}_\eta\right]_{ij} = \langle A_\eta\phi_j, \phi_i\rangle$ and $b_i = \langle u^{\mathrm{inc}}, \phi_i\rangle + \langle\eta\partial_n u^{\mathrm{inc}}, \phi_i\rangle$.

The matrix $\overline{A}_\eta$ is given as $\overline{A}_\eta = \frac{1}{2}\overline{I} - \overline{K} + \eta\overline{D}$, where the individual matrix entries are computed as

$$\begin{aligned}
\left[\overline{I}\right]_{ij} &= \int_\Gamma \phi_i(\mathbf{x})\phi_j(\mathbf{x})\ \mathrm{d}\Gamma(\mathbf{x}),\\
\left[\overline{K}\right]_{ij} &= \int_\Gamma \phi_i(\mathbf{x}) \int_\Gamma \partial_{n(\mathbf{y})} G(\mathbf{x},\mathbf{y})\phi_j(\mathbf{y})\, \mathrm{d}\Gamma(\mathbf{y})\, \mathrm{d}\Gamma(\mathbf{x}),\\
\left[\overline{D}\right]_{ij} &= -\int_\Gamma \phi_i(\mathbf{x})\partial_{n(\mathbf{x})} \int_\Gamma \partial_{n(\mathbf{y})}\partial_{n(\mathbf{y})} G(\mathbf{x},\mathbf{y})\phi(\mathbf{y})\, \mathrm{d}\Gamma(\mathbf{y})\, \mathrm{d}\Gamma(\mathbf{x})\\
&= \int_\Gamma \int_\Gamma G(\mathbf{x},\mathbf{y}) \left(\mathrm{curl}_\Gamma \phi_i(\mathbf{x}) \cdot \mathrm{curl}_\Gamma \phi_j(\mathbf{y})\right)\ \mathrm{d}\Gamma(\mathbf{y})\, \mathrm{d}\Gamma(\mathbf{x})\\
&\quad - k^2 \int_\Gamma \int_\Gamma G(\mathbf{x},\mathbf{y})\phi_i(\mathbf{x})\phi_j(\mathbf{y})\ (\hat{\mathbf{n}}(\mathbf{x})\cdot\hat{\mathbf{n}}(\mathbf{y}))\ \mathrm{d}\Gamma(\mathbf{y})\, \mathrm{d}\Gamma(\mathbf{x}).
\end{aligned}$$

For the hypersingular operator D the last formula follows from integration by parts and leads to a weakly singular integral. We also note that $\overline{D}^T = \overline{D}$ and $\overline{K}^T = \overline{T}$, where $\overline{T}$ is the discretization of the adjoint double-layer boundary operator.

Evaluating these integrals requires singularity-adapted quadrature rules. A general fully numerical quadrature scheme based on regularizing coordinate transformations is described in [40]. However, this scheme can still lead to large errors in situations such as sharp edges, two parallel triangles that are close to each other, and almost degenerate triangles. Alternative quadrature schemes that can deal with some of these issues are described for example in [37].

If instead of a scalar η we use a regularizing operator $\mathscr{R}$, then the operator $A_\mathscr{R}$ is well defined on $\mathscr{H}^{\frac{1}{2}}(\Gamma)$ and we can formulate a variational problem to find $\phi \in \mathscr{H}^{\frac{1}{2}}(\Gamma)$ such that

$$\langle A_\mathscr{R}\phi, \vartheta\rangle = \langle u^{\mathrm{inc}}, \vartheta\rangle + \langle \mathscr{R}\partial_n u^{\mathrm{inc}}, \vartheta\rangle, \quad \forall\vartheta \in \mathscr{H}^{-\frac{1}{2}}(\Gamma),$$

where we now interpret the dual pairing $\langle\cdot,\cdot\rangle$ as a dual pairing on $\mathscr{H}^{\frac{1}{2}}(\Gamma) \times \mathscr{H}^{-\frac{1}{2}}(\Gamma)$. The corresponding discrete left-hand-side matrix is then given as

$$\overline{A_\mathscr{R}} := \frac{1}{2}\overline{I} - \overline{K} + \overline{\mathscr{R}}\,\overline{I}^{-1}\,\overline{D},$$

where $[\mathscr{R}]_{ij} = \langle \mathscr{R}\phi_j, \phi_i\rangle$. To analyze the Galerkin variational formulation, techniques as discussed in [12] can now be used.

The discretization above uses the same space V_h of continuous piecewise linear nodal basis functions to discretize $\mathscr{H}^{\frac{1}{2}}(\Gamma)$ and $\mathscr{H}^{-\frac{1}{2}}(\Gamma)$. However, we use the space $\mathscr{H}^{-\frac{1}{2}}(\Gamma)$ to represent Neumann data. Hence, this approximation is only suitable if the boundary Γ is sufficiently smooth to support continuous Neumann data. For more general Lipschitz domains we can expect discontinuities and a more natural basis of $\mathscr{H}^{-\frac{1}{2}}(\Gamma)$ is a space of discontinuous piecewise constant functions. A stable dual pairing between continuous nodal basis functions and a space of piecewise constant discontinuous functions can be achieved by defining the discontinuous functions on the dual grid [33].

3 Operator Preconditioners for High-Frequency Problems

The classical Burton-Miller formulation suffers from poor convergence for high-frequency problems on general domains. The main reason is that the hypersingular operator D acts like an unbounded differential operator from $\mathscr{H}^{\frac{1}{2}}(\Gamma)$ to $\mathscr{H}^{-\frac{1}{2}}(\Gamma)$. As explained in Sect. 2.3, including a regularization operator fixes the mismatch in function spaces. Being an operator preconditioner, this regularization should be carefully chosen such that it improves the conditioning of the discrete system [33, 35, 42]. In practice, the regularization is ideally designed such that the resulting boundary integral operator is a compact perturbation of the identity operator.

In this section we will focus on two types of regularization, based on a high-frequency approximation of the Neumann-to-Dirichlet (NtD) map and the single-layer boundary operator. These operator preconditioners do not depend on the discretization method and can readily be combined with acceleration schemes such as $\mathscr{H}$-matrix compression.

3.1 OSRC Preconditioning

The On-Surface Radiation Condition (OSRC) preconditioner is based on the idea of finding a local surface approximation of the NtD map [4, 5, 20]. For $\vartheta \in \mathscr{H}^{-\frac{1}{2}}(\Gamma)$ we define the exterior Neumann-to-Dirichlet map $N_{\text{ex}}^{+} : \mathscr{H}^{-\frac{1}{2}}(\Gamma) \to \mathscr{H}^{\frac{1}{2}}(\Gamma)$ as $N_{\text{ex}}^{+}(\vartheta) := \gamma_0^{+} u_\vartheta$, where u_ϑ is the solution of the exterior Helmholtz problem

$$\begin{aligned} -\Delta u_\vartheta - k^2 u_\vartheta &= 0 \text{ in } \Omega^{+}, \\ \frac{\partial u_\vartheta}{\partial \hat{\mathbf{n}}} &= \vartheta \text{ on } \Gamma, \\ \lim_{|\mathbf{x}|\to\infty} |\mathbf{x}| \left(\frac{\partial u_\vartheta}{\partial |x|} - iku_\vartheta \right) &= 0. \end{aligned}$$

Using the NtD map it follows from the exterior Calderón projector [43, Sect. 7.5] that

$$\left(\frac{1}{2} I - T - DN_{\text{ex}}^{+} \right) \vartheta = \vartheta \tag{23}$$

for $\vartheta \in \mathscr{H}^{-\frac{1}{2}}(\Gamma)$. Assume that an approximation $\widetilde{N_{\mathrm{ex}}^{+}}$ of the NtD map is given. Then, after discretization, we obtain

$$\left(\frac{1}{2}\overline{I} - \overline{T} - \overline{D}\,\overline{I}^{-1}\overline{\widetilde{N_{\mathrm{ex}}^{+}}}\right) v \approx \overline{I}v.$$

Notice that since $\overline{T}^{T} = \overline{K}$ and $\overline{D}^{T} = \overline{D}$ the transpose of the left-hand-side operator equals the regularized Burton-Miller operator with $\overline{\mathscr{R}}^{T} = -\overline{\widetilde{N_{\mathrm{ex}}^{+}}}$. This shows that a good approximation to the NtD map results in an excellent preconditioner.

Unfortunately, the NtD map is a non-local pseudo-differential operator whose computation itself involves the solution of an exterior Helmholtz problem which makes its direct use as preconditioner impractical. However, there are efficient approximations that can be used. We have already encountered the most basic approximation, namely $N_{\mathrm{ex}}^{+} \approx \frac{1}{ik}$ giving the classical Burton-Miller operator with $\eta = i/k$. Alternatively, a more accurate approximation of the NtD map can be derived as

$$N^{osrc} = \frac{1}{ik}\left(1 + \frac{\Delta_\Gamma}{k_\epsilon^2}\right)^{-1/2} \tag{24}$$

where Δ_Γ denotes the surface Laplace-Beltrami operator [4, 5]. The occurrence of singularities is prevented with the use of a damped wavenumber $k_\epsilon = k(1 + i\epsilon)$. Based on a spectral analysis on a sphere, a good choice of damping is $\epsilon = 0.4(kR)^{-2/3}$ with R the radius of the object [20]. Localization of this operator is achieved with a Padé approximation of size n and a nonzero branch cut, typically 4 and $\pi/3$, respectively. The application of the OSRC operator is now reduced to solving a set of $(n + 1)$ surface Helmholtz equations with complex-valued wavenumber. The solution procedure of these local operators can efficiently be performed with sparse LU-factorization.

The OSRC-preconditioned Burton-Miller formulation

$$\left(\tfrac{1}{2}I - K\right)\varphi - N^{osrc}D\varphi = u^{\mathrm{inc}} - N^{osrc}\partial_n u^{\mathrm{inc}} \tag{25}$$

is uniquely solvable in $\mathscr{H}^{\frac{1}{2}}(\Gamma)$ on a smooth surface, for any wavenumber and nonzero damping factor [20]. Moreover, the boundary integral operator reduces to

$$\left(\tfrac{1}{2}I - K\right)\varphi - N^{osrc}D\varphi = \left(\frac{1}{2} + \frac{k_\epsilon}{2k}\right)I + C \tag{26}$$

for a compact operator C if Γ is sufficiently smooth. This is a second kind Fredholm integral equation and has a clustering of eigenvalues, resulting in fast convergence of linear solvers.

3.2 Regularization by Single-Layer Boundary Operators

Another strategy to achieve regularization of the hypersingular operator is to consider the single-layer potential. With Calderón identities [43, Corollary 6.19], one can show that

$$\begin{aligned} DV &= \tfrac{1}{4}I - T^2, \\ VD &= \tfrac{1}{4}I - K^2. \end{aligned}$$

Hence, if Γ is sufficiently smooth, then the product of the single-layer and the hypersingular boundary operator is a compact perturbation of a scaled identity. However, the single-layer operator alone is not a good choice of a regularizer due to the existence of resonances. A solution was proposed in [11], where the single-layer boundary operator V_κ with wavenumber $\kappa = ik/2$ was investigated as regularizer for the Brakhage-Werner formulation (21). Specifically,

$$i\left(\tfrac{1}{2}I - T\right)\varphi - DV_\kappa\varphi = -\partial_n u^{\text{inc}}, \tag{27}$$

for a coupling parameter $\mu = 1$. Similarly, this regularization can also be applied to the Burton-Miller formulation (19). For sufficiently smooth Γ this formulation is again a perturbation of a scaled identity because $V_\kappa D = (V + C)D$, where C is a compact operator [12, Lemma 2.1] and V is the single-layer operator for the original wavenumber k. The imaginary-wavenumber single-layer operator can be evaluated relatively cheap as it allows a very efficient low-rank representation.

4 Fast $\mathscr{H}$-Matrix Assembly

Hierarchical ($\mathscr{H}$-)matrix compression based on adaptive cross approximation (ACA) is a widely used technique to assemble boundary integral operators in a compressed format. It has a complexity of $\mathscr{O}(N \log N)$ for compression and evaluation of matrix-vector products, where N denotes the number of global degrees of freedom. This approach is relatively easy to implement, easily parallelizable, and builds a direct algebraic representation of the compressed operator that allows very fast matrix-vector products, compared to FMM. Main disadvantages are the longer setup time and often significantly higher memory consumption than FMM. However, particularly for low-frequency or non-oscillatory problems the performance is often excellent. Moreover, even though standard $\mathscr{H}$-matrix compression does not scale well asymptotically as $k \to \infty$, its practical performance even for higher-frequency problems is often very good as we will see in this and the following sections.

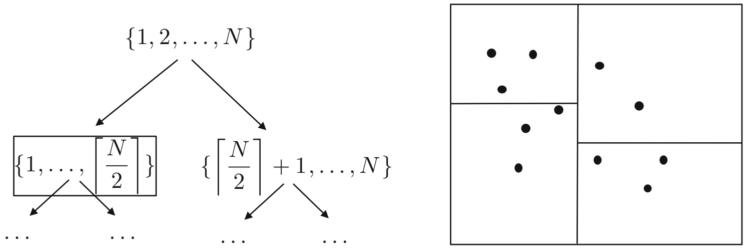

$\{1, 2, \ldots, N\}$
$\{1, \ldots, \lceil \frac{N}{2} \rceil\}$
$\{\lceil \frac{N}{2} \rceil + 1, \ldots, N\}$
. . .
. . .
. . .
. . .

Fig. 1 Division of degrees of freedom into a cluster tree

4.1 The Fundamentals of $\mathscr{H}$-Matrix Compression

In this section we give a brief overview of the main ideas of $\mathscr{H}$-matrix compression. More details can be found in [7, 32]. The $\mathscr{H}$-matrix compression is based on a geometric subdivision of the set of degrees of freedom (dofs) I in the boundary element mesh into a cluster tree $T(I)$. On each level the dofs are subdivided into two geometrically separated sets, as depicted in Fig. 1. The leafs of the cluster tree are reached when the number of dofs in each subdivision is below a specified tolerance. Given a set of dofs I for the test functions and a set of dofs J for the basis functions in the BEM discretization a block cluster tree $T(I \times J)$ is now constructed as follows.

1. The root of the block cluster tree is the product index set $b_0 = \tau \times \sigma$ with $\tau = I$ and $\sigma = J$.
2. Given a node $b' = \tau' \times \sigma'$ of the block cluster tree, where τ' and σ' are nodes of the corresponding cluster trees $T(I)$ and $T(J)$:
   - Stop the recursion if the current node satisfies an admissibility condition or if one of the cluster tree nodes σ' and τ' is a leaf node.
   - If the recursion is not stopped, define the sons of the block cluster tree node b' as the set $\{\tau_1' \times \sigma_1', \tau_1' \times \sigma_2', \tau_2' \times \sigma_1', \tau_2' \times \sigma_2'\}$ for the sons τ_i' and σ_j', $i, j = 1, 2$ of the cluster tree nodes τ' and σ'.

The admissibility condition is satisfied if the geometric bounding boxes X and Y associated with the cluster nodes τ' and σ' satisfy a separability condition. A frequently used condition is given as

$$\min\{\mathrm{diam}(X), \mathrm{diam}(Y)\} \le \alpha\, \mathrm{dist}(X, Y).$$

Here, diam denotes the diameter of a bounding box and dist the distance of two bounding boxes. The parameter α controls how strongly separated X and Y must be so that the admissibility condition is satisfied. By default, BEM++ uses a weaker condition given as

$$\mathrm{dist(X, Y)} > 0.$$

This works sufficiently well in practice and usually leads to a fewer number of blocks on the block cluster tree.

Once the generation of the block cluster tree has been completed, a compressed representation of the BEM matrix $\overline{A}$ can be assembled as follows. Let $b' = \tau' \times \sigma' \in \mathscr{L}(T(I \times J))$, the set of all leaf blocks of the block cluster tree $T(I \times J)$.

- If b' is not admissible, then evaluate all entries of $\overline{A}_{\tau' \times \sigma'}$, the restriction of $\overline{A}$ onto the index set $\tau' \times \sigma'$, directly and store the corresponding dense representation.
- If b' is admissible, then store a low rank representation $\overline{A}_{\tau' \times \sigma'} \approx U_{b'} \times V_{b'}^H$, where $U_{b'}$ is of dimension $|\tau'| \times t$ and $V_{b'}$ is of dimension $|\sigma'| \times t$ where t denotes the local rank.

To obtain a low-rank representation, a frequently used algorithm is Adaptive Cross Approximation (ACA). It is a heuristic algorithm that often works remarkably well and allows an approximate error control to determine the local rank t adaptively given a global error bound. However, most importantly, ACA only needs to compute a small fraction of the elements of the original matrix so that even very large BEM discretizations can be assembled on standard workstation systems.

Finally, often the above described compression procedure is intermixed with a recompression scheme in which after the compression of individual son blocks of a block cluster tree node b' a compression of b' itself is attempted using information from the sons. If this needs less memory than the original son representations, then the low-rank compression of b' itself is used instead and the sons deleted.

4.2 The $\mathscr{H}$-Matrix Compression at High Frequencies

The above described compression scheme is very efficient for low or non-oscillatory problems. However, for high-frequency problems the minimum rank required in each admissible block grows with the wavenumber. Let us consider the block cluster leaf node $b' = \tau' \times \sigma'$ and the corresponding bounding boxes X and Y. Given the Green's function $G(\mathbf{x}, \mathbf{y})$, the efficiency of the above described $\mathscr{H}$-matrix compression depends on the number t_ϵ, such that

$$\left\| G(\mathbf{x}, \mathbf{y}) - \sum_{j=1}^{t_\epsilon} g_j(\mathbf{x}) h_j(\mathbf{y}) \right\|_{X \times Y} < \epsilon$$

for given ϵ. The number t_ϵ is the minimum number of terms needed for a low-rank representation of the Green's function with accuracy ϵ. In [22] it is shown that

$$k^{2-\delta} \lesssim t_\epsilon \lesssim k^{2+\delta}, \quad \forall \delta > 0. \tag{28}$$

The overall computational cost of compression and evaluation is linear with respect to the rank estimate t in the admissible blocks, that is, the complexity scales like

Table 1 The performance of the $\mathscr{H}$-matrix compression of the single-layer boundary operator V on the unit sphere with varying wavenumber

k	N	Memory (Mb)	Compression (%)	Time (s)	Growth rate β
1	114	0.19	94.6	8.3E − 2	–
5	2136	39.6	56.9	0.53	1.83
10	7832	255	27.3	2.29	1.43
20	30,404	1.62E3	11.5	16.6	1.36
30	68,078	4.75E3	6.71	36.6	1.34
40	120,500	1.03E4	4.63	72.4	1.35
50	188,146	1.84E4	3.41	1.3E2	1.30
60	270,276	2.99E4	2.68	2.05E2	1.33
70	367,276	4.44E4	2.16	3.22E2	1.30
80	480,024	6.37E4	1.81	4.67E2	1.34

$\mathscr{O}(tN \log N)$. However, the rank t is dependent on N in high-frequency scattering. We typically choose a fixed number of dofs per wavelength, that is $N \sim k^2$. Together with (28) it therefore follows that $t \sim N$ giving an overall asymptotic complexity of $\mathscr{O}\left(N^2 \log N\right)$ for $\mathscr{H}$-matrix compression. This would make $\mathscr{H}$-matrices unfeasible for large-scale problems in the limit $k \to \infty$.

Fortunately, in practice the behavior seems much better for realistic wavenumbers. In Table 1 we show performance results for the compression of the standard single-layer boundary operator V with piecewise constant basis functions on the unit sphere for varying wavenumbers. We discretize the sphere with around 10 elements per wavelength, that is, $h = 2\pi/(10k)$. For the ACA we choose an error tolerance of 10^{-5}, which is sufficient for a wide range of applications. The timing results were done on a 20 cores, two processor Intel Xeon E5-2670 workstation with 2.5 GHz and 192 GB RAM. The compression rate measures how much memory the $\mathscr{H}$-matrix requires compared to a dense matrix of the same size. Recompression was not enabled. Also, BEM++ currently ignores the symmetry of the single-layer boundary operator, which could give another factor two saving. For the highest wavenumber $k = 80$ with 480 thousand elements the assembly time is roughly 7.8 min and the memory consumption is 62 GB.

It is interesting to measure the growth rate of the memory in dependence on N. We assume a memory growth of $\mathscr{O}(N^{\beta})$ for some $\beta > 0$. The last column in Table 1 shows estimates for β by comparing the memory growth from one wavenumber to the next. The effective exponent is around 1.3, which is significantly better than the asymptotic worst-case consideration given above and makes it possible to apply $\mathscr{H}$-matrices to many realistic high-frequency problems.

4.3 Modern Developments

The standard $\mathscr{H}$-matrix approximations are popular for many applications because of their ease of implementation and relatively good performance. However, recent FMM developments can significantly outperform classical $\mathscr{H}$-matrix techniques. While FMM uses hierarchical basis information to propagate information from the sources to the targets this is not the case for $\mathscr{H}$-matrices. A remedy for this is given by $\mathscr{H}^2$-matrices [8]. These are algebraically equivalent to FMM and refine the $\mathscr{H}$-matrix format by exploiting hierarchical information within the cluster bases. This reduces the complexity of compression and matrix-vector product for low-frequency problems to $\mathscr{O}(N)$ instead of $\mathscr{O}(N \log N)$. A novel development specifically for high-frequency problems are wideband $\mathscr{H}$-matrix techniques. They exploit that within a cone of opening angle $\theta \sim \frac{1}{k}$ the source and target clusters admit low-rank representations even for large wavenumber [23]. The difficulty is that these novel wideband $\mathscr{H}$-matrix approaches need to deal with a very large number of small block clusters. The implementation in [6] uses a mixture of $\mathscr{H}$-matrix approximations for the near-field and $\mathscr{H}^2$-matrix approximations for the far-field to efficiently deal with this large number of block clusters.

5 High-Frequency Boundary Element Simulations with BEM++

Boundary integral formulations can conveniently be implemented with the open-source library BEM++ [41]. As will be shown in this section, only high-level instructions are necessary to perform a BEM simulation. Apart from the code snippets in this section, an IPython notebook of the OSRC-preconditioned Burton-Miller formulation can be downloaded from the BEM++ website (www.bempp.org).

5.1 Creating and Solving an OSRC-Preconditioned Burton-Miller Formulation

In the following we will describe the implementation and solution of the OSRC-preconditioned Burton-Miller formulation for the scattering of a plane wave incident field

$$u^{\text{inc}}(x, y, z) = e^{ikx}$$

which travels in the x-direction.

The BEM++ framework can be used as a Python library, imported with the usual command.

```
import bempp.api
```

The first step for the implementation of a boundary element simulation is to specify the model data such as incident wave field and scatterer object. In this example we specify the incident field by defining a corresponding Python function. Other ways of specifying boundary data are also possible.

A Python function that specifies an incident field takes as input arguments the location `x`, normal direction `n`, and optionally the region `domain_index` of the object. The following two functions specify the incident field and its normal derivative. The NumPy array `result` stores the value of the function in each dimension.

```
k = 4.5
def dirichlet_fun(x, n, domain_index, result):
    result[0] = np.exp(1j*k * x[0])
def neumann_fun(x, n, domain_index, result):
    result[0] = 1j*k * n[0] * np.exp(1j*k * x[0])
```

Several canonical objects can readily be created with BEM++, such as a sphere, cube and ellipsoid. Optionally, the mesh size `h` can be passed, e.g. to guarantee an oversampling of ten elements per wavelength. The import of arbitrary triangular surface meshes in Gmsh format [27] is also possible. Alternatively, the node and connectivity information of a mesh can be specified. In the following we define the mesh of an ellipsoid with radius 3 in the *x*-direction and 1 in the other directions.

```
h = 2*np.pi / (10 * k)
grid = bempp.api.shapes.ellipsoid(3, 1, 1, h=h)
```

As finite element space, the BEM++ library provides continuous and discontinuous polynomial function spaces up to high-order and also function spaces defined on the barycentric mesh. Here, we only need the standard P1-elements.

```
space = bempp.api.function_space(grid, 'P', 1)
```

The native BEM++ object `GridFunction` provides functionality to store boundary data of the wave fields and also projections of the excitation field onto the boundary element space.

```
dirichlet_data = \
  bempp.api.GridFunction(space, fun=dirichlet_fun)
neumann_data = \
  bempp.api.GridFunction(space, fun=neumann_fun)
```

The creation of the boundary integral operators requires the specification of the mapping properties on the boundary element spaces, i.e., the domain, range and dual-to-range (test) space. For Galerkin discretization only the domain and the test space are required. The range space allows the implementation of an operator algebra that automatically creates the correct mass matrix transformations. This will be needed in the following. The OSRC-approximated NtD operator only requires one space object associated with a space of continuous functions to discretize the underlying Laplace-Beltrami operator, where it is always assumed that the domain, range and dual to range space are identical.

```
id = bempp.api.operators.boundary.sparse.\
     identity(space, space, space)
from bempp.api.operators.boundary.helmholtz import *
dlp = double_layer(space, space, space, k)
hyp = hypersingular(space, space, space, k)
ntd = osrc_ntd(space, k)
```

The created boundary integral operators are abstract objects, for which basic linear algebra operations such as addition and multiplication are available. The BEM++ library will take care of the correct mapping properties and uses mass-matrix transformations where necessary. Combined field boundary integral formulations can thus conveniently be created with the following high-level instructions.

```
bm_osrc_model = 0.5 * id - dlp - ntd * hyp
bm_osrc_data = dirichlet_data - ntd * neumann_data
```

Here, we have shown the creation of the OSRC-preconditioned Burton-Miller formulation (25). Other formulations can be implemented similarly.

So far, we have defined the boundary integral formulation with abstract objects. The actual discretization of the operators is not being performed until necessary or explicitly called. Instead of calling the weak formulation, we opt to compute the strong formulation which is the weak formulation with additional mass matrix preconditioning. By default, the matrix assembly is performed with $\mathscr{H}$-matrix compression enabled. The right-hand-side vector is given by the coefficients of the excitation data.

```
bm_osrc_matrix = bm_osrc_model.strong_form()
bm_osrc_rhs = bm_osrc_data.coefficients
```

The obtained matrix and right-hand-side vector can be interpreted by the SciPy library. This allows for solving the discrete system with its GMRES implementation.

```
from scipy.sparse.linalg import gmres
bm_osrc_sol,info = gmres(bm_osrc_matrix, bm_osrc_rhs)
```

The surface potential can readily be visualized with e.g. Gmsh but BEM++ also provides functionality to compute the scattered field outside the boundary. For this, an array of locations `points` have to be created on which the exterior field will be computed.

```
bm_osrc_pot = bempp.api.GridFunction(space, \
  coefficients=bm_osrc_sol)
from bempp.api.operators.potential.helmholtz import *
dlp_nearfield = double_layer(space, points, k)
bm_osrc_scattered = dlp_nearfield * bm_osrc_pot
```

The resulting field can then be exported for further processing or directly plotted using a Python plotting library.

5.2 Numerical Results

In this section we present some numerical results on canonical test shapes which demonstrate the performance of the formulations discussed in the previous sections. An application problem with realistic data from medical engineering will be presented in Sect. 6.

5.2.1 Stability in the Presence of Resonances

A prime advantage of the combined field integral equations over simpler formulations is stability at resonance frequencies. For example, the double-layer formulation (17) has a nontrivial nullspace at resonance frequencies, which are explicitly known for special geometries such as a cube. To this end, let us consider a unit-sized cube near the two resonances of $k = \pi\sqrt{1+1+3^2} = 10.42$ and $k = \pi\sqrt{1+2^2+3^2} = 11.75$. The mesh is created with an oversampling of ten elements per wavelength.

```
grid = bempp.api.shapes.cube(h=2*np.pi/(10*k))
```

The incident wave field is given by a plane wave field traveling in the positive x-direction and P1-elements are used for discretization. As a linear solver, the GMRES method available from the SciPy library has been used with a tolerance of 1.0E−5.

As can be seen in Fig. 2, the number of iterations used by the GMRES solver clearly depends on the choice of boundary integral formulation. The number of iterations for the Burton-Miller formulation and its preconditioned variant are constant for this small frequency range. The peaks at the resonance frequencies indicate the breakdown of the double-layer formulation. While at these low frequencies the convergence is still reasonable, this becomes problematic for high frequencies where the modal density increases.

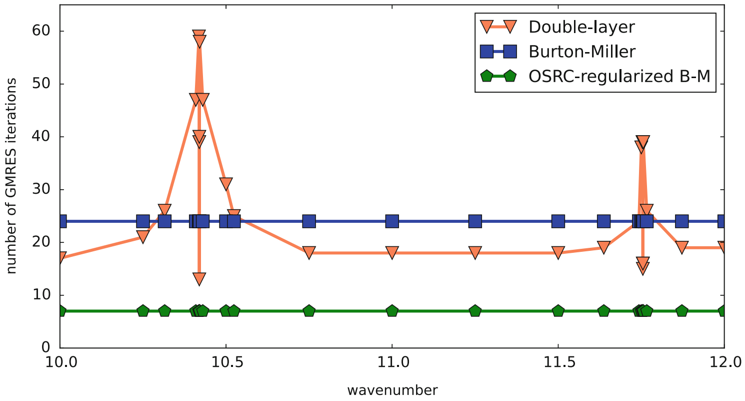


Fig. 2 The GMRES convergence for different model formulations near two resonance frequencies

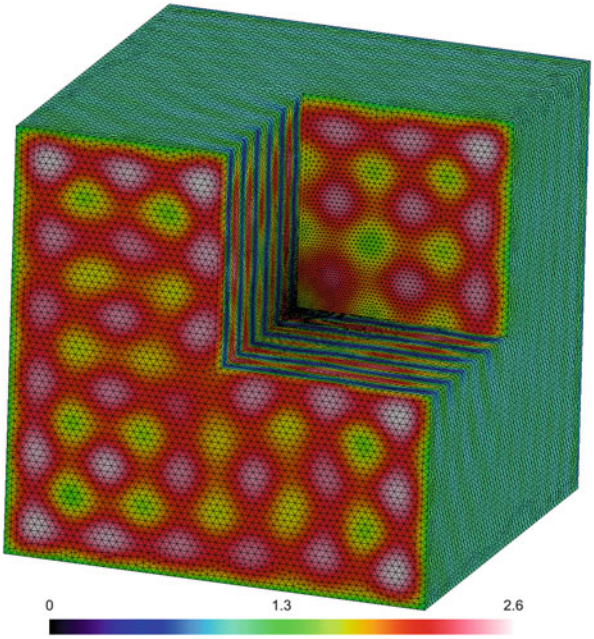


Fig. 3 The magnitude of the surface potential on the re-entrant cube for wavenumber $k = 37$

5.2.2 Performance with Frequency at a Re-entrant Cube

Although the combined field formulations are stable with respect to resonances, their convergence will deteriorate when increasing the frequency. The use of regularization is expected to improve the convergence, as explained in Sect. 2.3. Here, we will test this on a re-entrant cube of unit dimension, meshed with an oversampling of ten elements per wavelength.

```
grid=bempp.api.shapes.reentrant_cube(h=2*np.pi/(10*k))
```

The solution of the Burton-Miller formulation for $k = 37$ has been depicted in Fig. 3. For this wavenumber, the size of the object measures ten wavelengths across and 28,068 degrees of freedom are present.

The performance with respect to frequency of four different formulations will be assessed with this test case: the Burton-Miller formulation (18), its OSRC-preconditioned variant (25), the Brakhage-Werner formulation (21), and its complex-wavenumber single-layer regularized variant (27). For the standard Brakhage-Werner formulation we choose $\mathscr{R} = 1/k$ as a resemblance to the Burton-Miller formulation. As linear solver, the GMRES algorithm without restart is being used. Both the number of iterations and the wall-clock time of the linear solver are depicted in Fig. 4.

The experiment clearly shows that the use of regularization does have a big impact on the performance of the linear solver. The OSRC preconditioner and complex single-layer regularization both reduce the number of iterations consider-

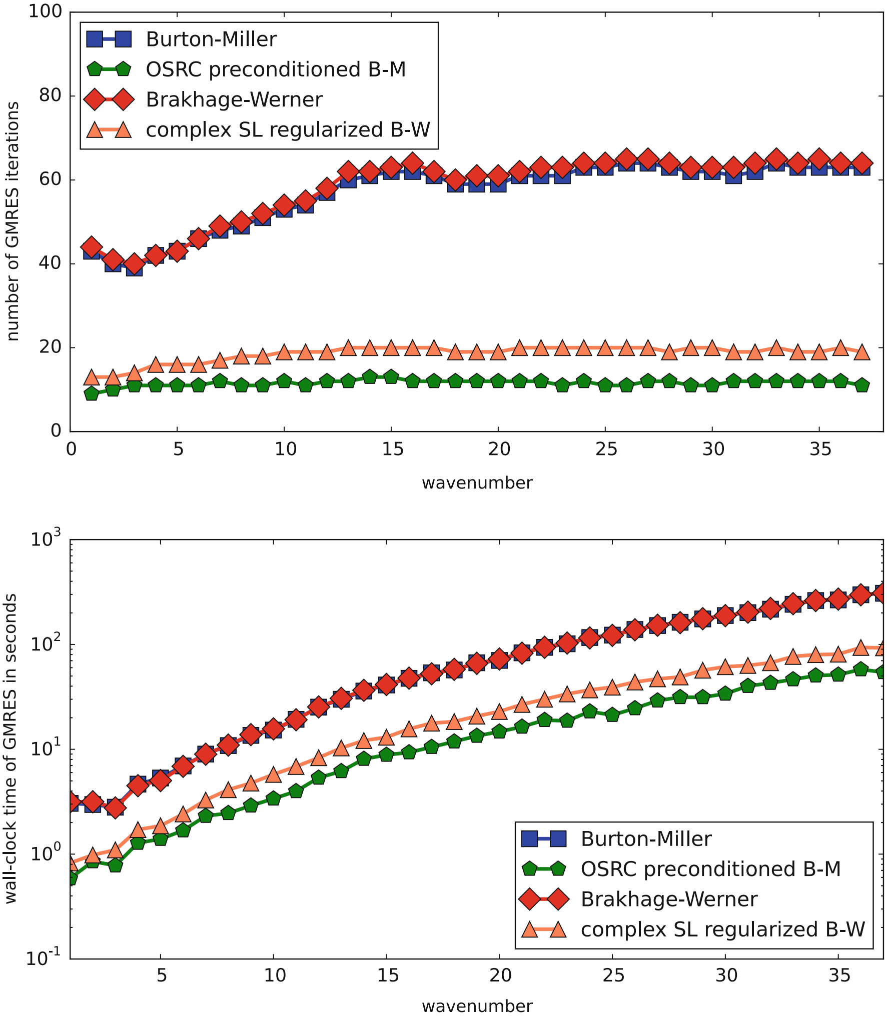


Fig. 4 The GMRES convergence for different model formulations on a re-entrant cube

ably compared with the classical Burton-Miller and Brakhage-Werner formulations. The reduction of number of iterations with the preconditioning strategies was not achieved at the price of much computational overhead. More precisely, compared to the classical formulations, the preconditioning results in an average overhead of 1.6 and 1.8 % per iteration for OSRC and complex single-layer regularization, respectively. However, both require additional initial setup time. For the OSRC this is the computation of sparse LU decompositions of the surface Helmholtz problems and for the complex-single layer regularization it is the $\mathscr{H}$-matrix assembly of the compressed single-layer operator. For the presented examples, both are small compared to the assembly times of the other operators involved in the Burton-Miller and Brakhage-Werner formulations.

6 HIFU Treatment

This section describes the application of the fast BEM techniques to a challenging problem of importance in medical engineering. To reduce the health risks of open surgery, clinicians are increasingly inclined to use modern non-invasive techniques, such as High-Intensity Focused Ultrasound (HIFU) treatment. Computational methods have the potential to improve the patient-specific treatment planning. Here, we will consider the case of transcostal HIFU, where the presence of the ribs has a significant influence on the sound propagation. Since the computational model is based on an exterior scattering problem, the BEM is perfectly suited as numerical solution technique.

6.1 Application to a Realistic High-Frequency Problem in HIFU Treatment

Surgery is the most effective local therapy for treating solid malignancies [18]. However, surgery to remove tumors in specific organs, such as the liver, still presents considerable challenges [14], with prognoses for the patients remaining poor [46]. The significant negative side effects associated with surgical interventions have led to an ongoing quest for safer, more efficient and better tolerated alternatives. In recent years, there has been a notable shift away from open surgery towards less invasive procedures, such as laparoscopic and robotic surgery, and also energy-based methods for in situ tumor destruction. The latter include embolization, radiofrequency, microwave and laser ablation, cryoablation and HIFU [18]. HIFU is a medical procedure which uses high-amplitude ultrasound to heat and ablate a localized region of tissue. Typically, the ultrasound is generated by a focused transducer located outside the human body. As the ultrasound propagates through tissue and at high acoustic intensities, absorption of the energy can induce local tissue necrosis targeted within a well-defined volume without damaging the overlying tissue [44]. Currently, HIFU is the only non-ionizing intervention capable of completely non-invasive ablation. The clinical acceptance of HIFU has grown in recent years, leading to its FDA approval for treating uterine fibroids, prostate cancer and for the palliative treatment of bone metastases.

Whilst the feasibility of HIFU for the treatment of cancer of the liver has been demonstrated [19], there remain a number of significant challenges which currently hinder its more widespread clinical application. The liver is located in the upper-right portion of the abdominal cavity under the diaphragm and to the right of the stomach. When administering a HIFU treatment in view of destroying tumors of the liver, the ultrasonic transducer is positioned outside the body and typically coupled to the abdomen via a region of water. Rib bone, which both absorbs and reflects ultrasound strongly, may therefore narrow the acoustic window between the transducer and the tumor. Hence, a common side effect of focusing ultrasound in regions located behind the rib cage is the overheating of bone and surrounding tissue, which can lead to skin burns at the ribs [38]. Furthermore, the presence of ribs can lead to aberrations at the focal region due to effects of diffraction [25].

One of the minimal technical specifications of a HIFU system for the treatment of liver tumors should be to transmit energy either in between, below, or through the ribs without damaging the ribs or causing a skin burn [45]. A means of addressing this requirement is via a patient-specific treatment planning protocol based on numerical simulations carried out using the patient's anatomical data. Such a protocol could provide a standardized framework by which HIFU may be optimized to treat tumors of the liver without adverse effects. The role of numerical models also extends to pre-clinical experiments on soft tissue and bone mimicking phantoms. As there remain substantial metrological challenges when carrying out such physical experiments, validated numerical models play a key role in planning this work and interpreting its outcome.

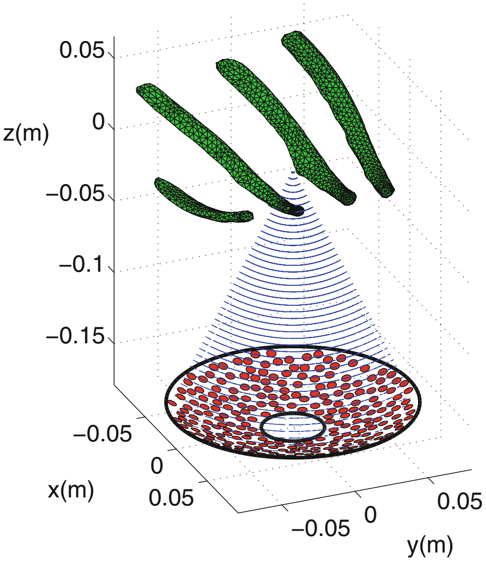
6.2 Methodology

As the ultrasonic waves propagate from the surface of the transducer to the focal region, they will encounter water and soft tissue, including skin and fat, and rib bone, before finally reaching the liver. Different soft tissue types tend to bear acoustic properties similar to those of water. The speed of propagation of longitudinal waves in these media is generally comparable, and is approximately $1500\,\mathrm{m\,s^{-1}}$ [21]. The same is true of the density [21], which is around $1000\,\mathrm{kg\,m^{-3}}$. Ribs however act as strong scatterers, owing to their higher acoustic impedance relative to that of soft tissue. A first step towards treating the problem of scattering of a HIFU field by the rib cage is therefore to consider the ribs as being immersed in an infinite homogeneous medium with acoustic properties representative of those of soft tissue. The modeling of the scattering of the field of a HIFU array by human ribs can then be considered as an exterior scattering problem. This can be efficiently treated using the BEM [26]. The optimal transducer excitation frequency for HIFU of the liver has been established to be around 1–1.5 MHz. At frequencies below 1 MHz, the cavitation threshold in tissue decreases, thus creating the risk of unwanted cavitation at pre-focal regions. At frequencies above 1.5 MHz, since attenuation in soft tissue is roughly proportional to frequency, the resulting focal intensities may be too low to achieve tissue necrosis, particularly in the case of deep-seated tumors. For transcostal HIFU, this implies that the wavelengths in soft tissue will be around 1.0–1.5 mm. The computational domain being approximately $20\,\mathrm{cm}\times 20\,\mathrm{cm}\times 20\,\mathrm{cm}$ reinforces the notion that it is advantageous to employ a computational method which does not rely on a volumetric mesh, which strengthens the case for using the BEM.

The advent of multi-element array transducers driven by multi-channel electronics offers significant advantages over concave single-element piezoelectric devices. Multi-element transducers have the ability to compensate for tissue and bone heterogeneities and to steer the beam electronically by adjusting the time delays in each channel to produce constructive interference at the required location, thus minimizing the requirement for mechanical repositioning of the transducer during treatment. A pseudo-random arrangement of the circular planar elements on the surface of the transducer is often opted for. This has been shown to minimize the formation of side lobes when design constraints place a limit on the amount of elements that can be used and on the spacing between these elements [24]. Figure 5 depicts a mesh of four ribs, together with a spherical section transducer array, with 256 pseudo-randomly distributed elements. The array is positioned so that its geometric focus is located at an intercostal space, approximately 3 cm deep into the rib cage.

In order to address the scattering problem, a suitable description of the incident acoustic field and its normal derivative on the surface of the ribs must be arrived at. In the case of multi-element transducers, the incident acoustic pressure field is commonly modeled as a superposition of plane circular piston sources [24]. The spatial component of the acoustic field of such a circular source may be represented by the Rayleigh integral, which can be solved using numerical quadrature techniques [47].

Fig. 5 Position of ribs relative to a HIFU array for an intercostal treatment, approximately 3 cm deep into the rib cage



6.3 Computational Results

In Sect. 6.2, it was proposed that, in first instance, a physical model for HIFU treatment planning of the liver could be formulated as an exterior scattering problem. The BEM is ideally suited to tackle such problems. The strict requirement of frequencies in the MHz range necessitates the use of fast solution techniques, such as operator preconditioning and matrix compression. Here, we will use the OSRC-preconditioned Burton-Miller formulation with $\mathscr{H}$-matrix compression since this has experimentally proved to be the most effective technique.

The scattering object is given by a human rib cage model [25], consisting of the four ribs closest to the liver. The ribs are rigid and immersed in an infinite domain where the speed of sound is $1500\,\mathrm{m\,s^{-1}}$, as is typical for water and soft tissue. The ultrasound excitation is generated by a multi-element transducer array of 256 piston sources. The field generated by each element is modeled with a numerical quadrature rule, resulting in a total of 38,144 point sources. The frequency of the ultrasound field is 1 MHz, which corresponds to a wave length of 1.5 mm. The diameter of the ribcage model is 20.3 cm, which makes it 135 times larger than the wave length.

The surface mesh at the ribs consists of triangles with a maximum width of 0.18 mm, thus representing each wavelength with at least 8 elements. The boundary element space of continuous piecewise linear elements contains 479,124 degrees of freedom. The experiment has been performed on a high-specification workstation of eight quad-cores with a clock rate of 2.8 GHz each. The shared memory is 264 GB.

Standard values for the parameters in the OSRC-preconditioner have been used, namely a size of 4 and a branch cut of $\pi/3$ for the Padé approximation. The GMRES solver of SciPy has been used with a default termination criterion of 10^{-5} and finished the solution in 19 iterations and 6:59 min only.

The assembly of the dense matrices has been performed with $\mathscr{H}$-matrix compression with an ϵ-value of 10^{-5}, a maximum rank of 1000 and a maximum block size of 100,000. The assembly of the boundary operators took 5 h and 16 min. Where the storage of dense matrices would have needed in excess of 7 TB memory, the compressed matrices required 194 GB only. The compression rates are 2.08 and 3.31 % for the single-layer and hypersingular boundary operator, respectively.

The total field exterior to the rib cage was computed on a vertical plane and is visualized in Fig. 6. The reflected waves are clearly visible, along with a shadow region behind the ribs. The influence of the scattering on the focal region is not significant in this configuration: the energy is still bundled in the desired region. The realistic wave field for this challenging object confirms the capability of our modern BEM implementation to simulate acoustic scattering at high frequencies.

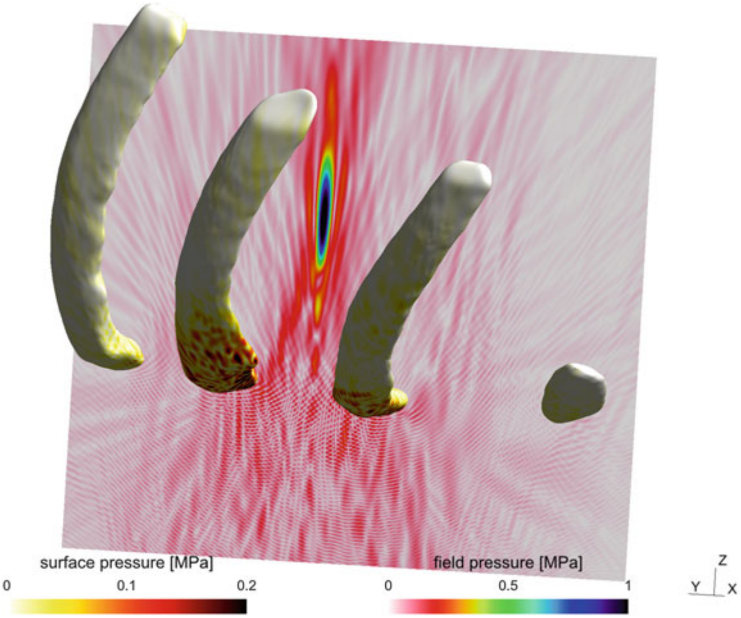


Fig. 6 The computational results of the HIFU model. At the surface the magnitude of the surface potential $\varphi = u^{\text{tot}}|_{\Gamma}$ and on the exterior plane with $x = 0$ the magnitude of the total wave field $u^{\text{inc}} + \mathscr{K}\varphi = u^{\text{tot}}$ have been visualized

7 Discussion

In this chapter we have demonstrated efficient BEM formulations for exterior acoustic problems, their fast implementation using the open-source BEM++ library, and performance results when applied to a realistic high-frequency problem. Modern preconditioning strategies for the Burton-Miller formulation based on OSRC or complex wavenumber single-layer boundary operators are highly effective and lead to a small number of GMRES iterations for each right-hand side. Even though the applicability of the BEM to large-scale simulations has been confirmed in this chapter, there is still a need for faster computations. A goal is to incorporate the BEM in an optimization routine for the configuration of HIFU transducer arrays. This necessitates the solution of the BEM formulation for multiple right-hand-side vectors. When such an implementation could be achieved effectively, this would bring the BEM a step closer to actual application in a clinical environment.

Significant speed improvements are still possible with respect to the discretization of the boundary operators. While the $\mathscr{H}$-matrix based discretization described in this chapter performs well for many Helmholtz problems, a direct improvement is possible by moving towards $\mathscr{H}^2$-matrix techniques. They allow for a considerable memory reduction [8], but similar to $\mathscr{H}$-matrices, they are not asymptotically optimal at high frequencies.

For problems with only few right-hand sides, high-frequency FMM methods [16, 30] are very efficient. Yet, they are less suited for problems with many right-hand sides due to their often slower matrix-vector product. Wideband hierarchical matrix techniques such as the one presented in [6] combine fast algebraic matrix-vector products with asymptotic optimal complexity as $k \to \infty$.

A potential improvement to the limitations at high-frequencies may be the development of fast approximate direct solvers. While there has been considerable progress for low-frequency problems (see e.g. [10]), the development of fast approximate direct solvers that scale well as $k \to \infty$ remains elusive. The most promising approach may be based on butterfly compression techniques. A butterfly recompression scheme for an approximate $\mathscr{H}$-matrix LU decomposition is described in [31]. The results in this paper are impressive but still require an initial compression using standard $\mathscr{H}$-matrices.

While there is a wealth of software available for finite element discretizations there are still few open-source packages for boundary element problems. The BEM++ library is continuously being developed and aims to integrate modern technologies as they become relevant for practical applications. We have given a demonstration of BEM++ in this chapter. Many more example applications including Maxwell problems are described at the website www.bempp.org.

References

1. S. Amini and P. J. Harris. A comparison between various boundary integral formulations of the exterior acoustic problem. *Computer methods in applied mechanics and engineering*, 84(1):59–75, 1990.
2. S. Amini and Stephen Martin Kirkup. Solution of Helmholtz equation in the exterior domain by elementary boundary integral methods. *Journal of Computational Physics*, 118(2):208–221, 1995.
3. S. Amini and N. D. Maines. Preconditioned Krylov subspace methods for boundary element solution of the Helmholtz equation. *International Journal for Numerical Methods in Engineering*, 41(5):875–898, 1998.
4. Xavier Antoine and Marion Darbas. Alternative integral equations for the iterative solution of acoustic scattering problems. *The Quarterly Journal of Mechanics and Applied Mathematics*, 58(1):107–128, 2005.
5. Xavier Antoine and Marion Darbas. Generalized combined field integral equations for the iterative solution of the three-dimensional Helmholtz equation. *ESAIM: Mathematical Modelling and Numerical Analysis*, 41(01):147–167, 2007.
6. M. Bebendorf, C. Kuske, and R. Venn. Wideband nested cross approximation for Helmholtz problems. *Numerische Mathematik*, 130(1):1–34, Jul 2014.
7. Mario Bebendorf. *Hierarchical matrices*. Springer, 2008.
8. Steffen Börm. *Efficient numerical methods for non-local operators*, volume 14 of *EMS Tracts in Mathematics*. European Mathematical Society (EMS), Zürich, 2010. $\mathscr{H}^2$-matrix compression, algorithms and analysis.
9. Helmut Brakhage and Peter Werner. Über das Dirichletsche Aussenraumproblem für die Helmholtzsche Schwingungsgleichung. *Archiv der Mathematik*, 16(1):325–329, 1965.
10. James Bremer, Adrianna Gillman, and Per-Gunnar Martinsson. A high-order accurate accelerated direct solver for acoustic scattering from surfaces. *BIT Numerical Mathematics*, 55(2):367–397, Jul 2014.

11. Oscar Bruno, Tim Elling, and Catalin Turc. Regularized integral equations and fast high-order solvers for sound-hard acoustic scattering problems. *International Journal for Numerical Methods in Engineering*, 91(10):1045–1072, 2012.
12. A. Buffa and R. Hiptmair. Regularized combined field integral equations. *Numerische Mathematik*, 100(1):1–19, 2005.
13. A. J. Burton and G. F. Miller. The application of integral equation methods to the numerical solution of some exterior boundary-value problems. *Proceedings of the Royal Society of London A: Mathematical, Physical and Engineering Sciences*, 323(1553):201–210, 1971.
14. Charles H. Cha, M. Wasif Saif, Brett H. Yamane, and Sharon M. Weber. Hepatocellular carcinoma: Current management. *Current Problems in Surgery*, 47(1):10–67, 2010.
15. Simon N. Chandler-Wilde, Ivan G. Graham, Stephen Langdon, and Euan A. Spence. Numerical-asymptotic boundary integral methods in high-frequency acoustic scattering. *Acta Numerica*, 21:89–305, Apr 2012.
16. H Cheng, W Y Crutchfield, Z Gimbutas, L F Greengard, J F Ethridge, J Huang, V Rokhlin, N Yarvin, and J Zhao. A wideband fast multipole method for the Helmholtz equation in three dimensions. *Journal of Computational Physics*, 216(1):300–325, 2006.
17. Weng Cho Chew, Eric Michielssen, J. M. Song, and Jian-Ming Jin. *Fast and efficient algorithms in computational electromagnetics*. Artech House, Inc., 2001.
18. David Cranston. A review of high intensity focused ultrasound in relation to the treatment of renal tumours and other malignancies. *Ultrasonics Sonochemistry*, 27:654–658, 2015.
19. Lawrence Crum, Michael Bailey, Joo Ha Hwang, Vera Khokhlova, and Oleg Sapozhnikov. Therapeutic ultrasound: Recent trends and future perspectives. *Physics Procedia*, 3(1):25–34, 2010. International Congress on Ultrasonics, Santiago de Chile, January 2009.
20. Marion Darbas, Eric Darrigrand, and Yvon Lafranche. Combining analytic preconditioner and fast multipole method for the 3-D Helmholtz equation. *Journal of Computational Physics*, 236:289–316, 2013.
21. Francis A. Duck. *Physical properties of tissues – A comprehensive reference book*. Academic, 1990.
22. B. Engquist and H. Zhao. Approximate Separability of Green's Function for High Frequency Helmholtz Equations. Technical report.
23. Björn Engquist and Lexing Ying. Fast directional multilevel algorithms for oscillatory kernels. *SIAM J. Sci. Comput.*, 29(4):1710–1737 (electronic), 2007.
24. Leonid R. Gavrilov and Jeffrey W. Hand. *High-Power Ultrasound Phased Arrays for Medical Applications*. Nova, 2014.
25. P. Gélat, G. ter Haar, and N. Saffari. A comparison of methods for focusing the field of a HIFU array transducer through human ribs. *Physics in Medicine and Biology*, 59(12):3139–3171, 2014.
26. Pierre Gélat, Gail ter Haar, and Nader Saffari. Modelling of the acoustic field of a multi-element HIFU array scattered by human ribs. *Physics in Medicine and Biology*, 56(17):5553–5581, 2011.
27. Christophe Geuzaine and Jean-François Remacle. Gmsh: A 3-D finite element mesh generator with built-in pre-and post-processing facilities. *International Journal for Numerical Methods in Engineering*, 79(11):1309–1331, 2009.
28. Dan Givoli. Computational absorbing boundaries. In *Computational Acoustics of Noise Propagation in Fluids-Finite and Boundary Element Methods*, pages 145–166. Springer, 2008.
29. I. G. Graham, M. Löhndorf, J. M. Melenk, and E. A. Spence. When is the error in the h-BEM for solving the Helmholtz equation bounded independently of k ? *BIT Numerical Mathematics*, 55(1):171–214, Sep 2014.
30. Nail A Gumerov and Ramani Duraiswami. A broadband fast multipole accelerated boundary element method for the three dimensional Helmholtz equation. *The Journal of the Acoustical Society of America*, 125(1):191–205, 2009.
31. Han Guo, Jun Hu, and Eric Michielssen. On MLMDA/Butterfly Compressibility of Inverse Integral Operators. *IEEE Antennas and Wireless Propagation Letters*, 12:31–34, 2013.
32. Wolfgang Hackbusch. *Hierarchical matrices: Algorithms and Analysis*. Springer, 2015.

33. R. Hiptmair. Operator Preconditioning. *Computers & Mathematics with Applications*, 52(5):699–706, Sep 2006.
34. Maryna Kachanovska. Hierarchical matrices and the high-frequency fast multipole method for the Helmholtz equation with decay. Technical Report 13, MPI Leipzig, 2014.
35. Robert C. Kirby. From functional analysis to iterative methods. *SIAM review*, 52(2):269–293, 2010.
36. Rainer Kress. Minimizing the condition number of boundary integral operators in acoustic and electromagnetic scattering. *The Quarterly Journal of Mechanics and Applied Mathematics*, 38(2):323–341, 1985.
37. Marc Lenoir and Nicolas Salles. Evaluation of 3-D Singular and Nearly Singular Integrals in Galerkin BEM for Thin Layers. *SIAM Journal on Scientific Computing*, 34(6):A3057–A3078, Jan 2012.
38. Jun-Lun Li, Xiao-Zhou Liu, Dong Zhang, and Xiu-Fen Gong. Influence of ribs on the nonlinear sound field of therapeutic ultrasound. *Ultrasound in Medicine and Biology*, 33(9):1413–1420, 2007.
39. Jean-Claude Nédélec. *Acoustic and electromagnetic equations: integral representations for harmonic problems*, volume 144. Springer Science & Business Media, 2001.
40. Stefan A. Sauter and Christoph Schwab. *Boundary element methods*, volume 39 of *Springer Series in Computational Mathematics*. Springer-Verlag, Berlin, 2011.
41. Wojciech Śmigaj, Timo Betcke, Simon Arridge, Joel Phillips, and Martin Schweiger. Solving boundary integral problems with BEM++. *ACM Transactions on Mathematical Software (TOMS)*, 41(2):6, 2015.
42. O. Steinbach and W.L. Wendland. The construction of some efficient preconditioners in the boundary element method. *Advances in Computational Mathematics*, 9(1–2):191–216, 1998.
43. Olaf Steinbach. *Numerical approximation methods for elliptic boundary value problems: finite and boundary elements*. Springer Science & Business Media, 2007.
44. G. ter Haar, D. Sinnett, and I. Rivens. High intensity focused ultrasound-a surgical technique for the treatment of discrete liver tumours. *Physics in Medicine and Biology*, 34(11):1743–1750, 1989.
45. J.-F. Aubry *et al.* Transcostal high-intensity-focused ultrasound: ex vivo adaptive focusing feasibility study. *Journal of Therapeutic Ultrasound*, 1:1–13, 2013.
46. James S. Tomlinson *et al.* Actual 10-year survival after resection of colorectal liver metastases defines cure. *Journal of Clinical Oncology*, 25(29):4575–4580, 2007.
47. Elwin van 't Wout, Pierre Gélat, Timo Betcke, and Simon Arridge. A fast boundary element method for the scattering analysis of high-intensity focused ultrasound. *The Journal of the Acoustical Society of America*, 138(5):2726–2737, 2015.
48. Jianlin Xia, Shivkumar Chandrasekaran, Ming Gu, and Xiaoye S. Li. Fast algorithms for hierarchically semiseparable matrices. *Numerical Linear Algebra with Applications*, 17(6):953–976, Dec 2010.