PONTIFICIA UNIVERSIDAD CATOLICA DE CHILE
SCHOOL OF ENGINEERING

HIGH PERFORMANCE PRECONDITIONING AND PERTURBATION ANALYSIS APPLIED TO WAVE PROPAGATION PROBLEMS

PAUL ESCAPIL-INCHAUSPÉ

Thesis submitted to the Office of Graduate Studies
in partial fulfillment of the requirements for the degree of
Doctor in Engineering Sciences

Advisors:
CARLOS JEREZ-HANCKES
LEONARDO VANZI

Santiago de Chile, September 2021

PONTIFICIA UNIVERSIDAD CATOLICA DE CHILE
SCHOOL OF ENGINEERING

HIGH PERFORMANCE PRECONDITIONING AND PERTURBATION ANALYSIS APPLIED TO WAVE PROPAGATION PROBLEMS

PAUL LOUIS ESCAPIL-INCHAUSPÉ

Members of the Committee:

CARLOS JEREZ-HANCKES

LEONARDO VANZI

THOMAS FÜHRER

ELWIN VAN'T WOUT

IGNACIO MUGA

RALF HIPTMAIR

JUAN DE DIOS ORTÚZAR

Thesis submitted to the Office of Graduate Studies in partial fulfillment of the requirements for the Degree Doctor in Engineering Sciences

Santiago de Chile, September 2021

Thesis submitted to the Office of Research and Graduate Studies
in partial fulfillment of the requirements for the degree of
Doctor in Engineering Sciences

Santiago de Chile, September 2021

To my grandparents.

ACKNOWLEDGEMENTS

First of all, I want to express my gratitude to my advisor, Prof. Dr. Carlos Jerez-Hanckes. He was the first person who inspired and encouraged me to to start doing research because this is a field that I had never thought about exploring before. This was an important decision for me, and I would have never gone towards this direction without encountering him. He has been backing and trusting me for five years and is both a mentor and an example to me.

I am grateful to all the professors who made this project possible. First, I thank Prof. Dr. Christoph Schwab for receiving me at ETH Zürich in spring 2016, and Xavier Claeys for fruitful collaboration and several visits in Laboratoire Jacques-Louis Lion, Université Pierre et Marie Curie, Paris. I also want to highlight the work that I have done with Prof. Timo Betcke for his helpfulness and visits at University College London (UCL). I thank Prof. Dr. Juan Enrique Coeymans for his strong and unfailing friendship.

I am truly thankful to the Comittee members for their valuable feedback and their support.

I thank Dr. Simon Tournier for sharing his passion about programming, and for the stimulating discussions we always had. His philosophy of life was key to me. In addition, I want to mention that it was an enriching experience working with my fellow labmates Rubén Aylwin, José Pinto and Fernando Henriquez.

Furthermore, I want to mention the constant support from my family and my friends, especially those from Santiago and Bayonne. They gave me strength in both my Chilean and French lives, day after day. Finally, I am grateful to Rosario and her family for their support and affection.

This thesis research was supported by Facultad de Ingeniería of Pontificia Universidad Católica de Chile (PUC), Vicerrectoría de Investigation of PUC, Becas de Doctorado Nacional (Fondecyt) and Fondo Nacional de Desarrollo Científico y Tecnológico (Fondecyt) Regular 1171491.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ABSTRACT

Wave propagation is a fundamental physical phenomenon. Its simulation is key for a large number of applications in many science and engineering applications ranging from the design of antennas, sonars and aircraft to medical techniques in fluoroscopy and magnetic resonance imaging.

The behavior of electromagnetic waves in harmonic regime can be accurately represented by partial differential equations (Helmholtz and Maxwell). Taking into account that waves frequently propagate in unbounded domains, one can rely on integral equations set on the surface of the object, commonly solved via the so-called boundary element methods (BEM). BEM induces linear systems that are generally solved by iterative methods such as GMRES combined with preconditioning techniques. If so, the precision is controlled by the quality of the matrix induced by BEM, while the efficiency can be represented by the quality of the preconditioner.

In this thesis, we seek to fully understand the compromise between precision and efficiency in the resolution phase of discrete schemes, focusing on the complex case of BEM for Helmholtz and Maxwell equations. Among our main results, we introduce the novel bi-parametric operator preconditioning framework. Our findings are validated in three-dimensional BEM simulations for the Helmholtz and Maxwell equations, and high-performance applications are explored in uncertainty quantification.

Keywords: Galerkin methods, preconditioning, iterative linear solvers, perturbation analysis, wave propagation, uncertainty quantification.

RESUMEN

La propagación de ondas es uno de los fenómenos físicos fundamentales cuya simulación es clave para un gran número de aplicaciones en ingeniería eléctrica que van desde el diseño de antenas, sonares y aviones a aplicaciones médicas en radioscopia o imágenes por resonancia magnética.

El comportamiento de las ondas electromagnéticas en régimen armónico puede ser representado mediante ecuaciones diferenciales parciales (Helmholtz y Maxwell). Teniendo en cuenta que las ondas se propagan frecuentemente en dominios no acotados, se suele recurrir a ecuaciones integrales sobre en la superficie del objeto, resueltas comúnmente por medio de método de elementos de frontera (BEM). Los métodos BEM inducen sistemas lineales que se resuelven en general mediante métodos iterativos tales como GMRES, combinados con precondicionamiento. En aquellos casos, la precisión se controla con la calidad de la matriz inducida por BEM, mientras la eficiencia puede ser representada por la calidad del precondicionador.

El presente proyecto busca abordar la problemática del compromiso entre precisión y eficiencia en la fase de resolución de los esquemas discretos, enfocándose en el caso sumamente complejo de BEM para las ecuaciones de Helmholtz y Maxwell. A lo largo del proyecto, se introduce el nuevo método de precondicionamiento por operadores biparamétrico. Se validan los resultados teóricos propuestos con aplicaciones de BEM para las ecuaciones de Helmholtz y Maxwell y se exploran soluciones de alto rendimiento en cuantificación de incertidumbre.

Palabras Claves: Métodos de Galerkin, precondicionamiento, métodos iterativos, análisis perturbativo, propagación de ondas, cuantificación de incertidumbre.

1. INTRODUCTION

1.1. Motivation

Electromagnetic waves propagation is described by Maxwell's equations, introduced in the 1860s. In harmonic regime and in homogeneous media, they can be recasted as partial differential equations (PDEs), referred to as "Maxwell" in this chapter, and give rise to another scalar PDE, "Helmholtz" under more restrictive requirements (Monk et al., 2003; Nédélec, 2001). Furthermore, the latter describes the propagation of acoustic waves in homogeneous media. Cases of interest include the analysis of perfect electric conductors (PECs) (Andriulli et al., 2008), perfect magnetic conductors (PMCs), impedance or transmission problems (TP) (Colton & Kress, 2012). These configurations are taken into account by enforcing boundary conditions (BCs). They allow to recast such problems as boundary value problems (BVPs) (Ern & Guermond, 2013).

Resolution of BVPs is key in a number of fields in engineering for design and optimization purposes (Aylwin, Jerez-Hanckes, Schwab, & Zech, 2020; Allaire & Schoenauer, 2007). As explained before, they generally are presented in the form "PDE in a domain + BC on the boundary". Still, they appeal to complex notions of functional analysis, and thus it is paramount to enclose them into a proper mathematical framework in order to ensure key properties such as existence, uniqueness and stability of the solution provided by the model (Ern & Guermond, 2013). Besides, their resolution is complex as no simple analytical solution is available except for simple configurations—e.g., balls, squares or cubes.

Consequently, discretization methods are approaches of choice, as they allow to generate sequences converging asymptotically toward the exact continuous solution, under a moderate number of operations (linear o quasi-linear) (Steinbach, 2007). On the other hand, the dramatic increase in computing capacities over the last decades has allowed these methods to become ubiquitous, ensuing a large number of new algorithms and resurgence of applied mathematics.

The numerical resolution of the aforementioned problems is commonly carried out by:

(i) An analysis of continuous differential operators in reflexive Banach spaces (Steinbach, 2007; Ern & Guermond, 2013);

(ii) The discretization of these operators by means of spectral bases (Haldenwang, Labrosse, Abboudi, & Deville, 1984; Shen, Tang, & Wang, 2011), or by discretizing the domain considered by some low-order method such as the finite element method (FEM) (Brenner & Scott, 2007; Ern & Guermond, 2013);

(iii) The resolution of a linear system from the impedance matrix (or precision matrix) assembled.

Provided that waves frequently propagate in unbounded domains, their resolution by FEM can be problematic. For this reason, Green's function-based methods are a method of choice. They give birth to boundary integral equations (BIEs)—defined on the surface of the studied object, and are solved using boundary element methods (BEM) (Sauter & Schwab, 2010; Steinbach, 2007; Nédélec, 2001).

However, the originated matrices are dense, and when many degrees of freedom are used, the resulting linear system is not suitable for direct resolution, calling for iterative methods (Saad, 2003; Nevanlinna, 1993). Moreover, in the case of first-kind Fredholm integral equations (Sauter & Schwab, 2010), the poor conditioning of the matrices can generate serious convergence problems (Steinbach, 2007). For this reason, the BEM community devotes many efforts to find efficient preconditioners in order to consequently reduce the number of iterations (Steinbach, 2007; Sauter & Schwab, 2010; Thierry, 2014).

1.2. Wave Scattering Problem

As described before, this thesis seeks to provide mathematical and high performance computational tools to solve wave diffraction problems (Helmholtz, Maxwell) in harmonic regime and homogeneous media by BEM. General wave scattering problems can be represented as BVPs as follows in Problem 1:

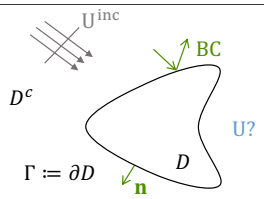
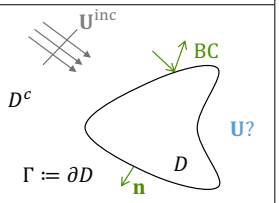
	Acoustics	Electromagnetics
		
Parameters	$\mu, c > 0, \kappa := \omega/c$	$\mu, \epsilon > 0, \kappa := \omega\sqrt{\epsilon\mu}$
Traces	$\gamma_0 \mathrm{U} := \mathrm{U}\|_\Gamma$ $\gamma_1 \mathrm{U} := \mathbf{n} \cdot \nabla \mathrm{U}\|_\Gamma$	$\gamma_0 \mathbf{U} := \mathbf{U}\|_\Gamma \times \mathbf{n}$ $\gamma_1 \mathbf{U} := \frac{1}{\kappa}(\mathbf{curl}\, \mathbf{U}\|_\Gamma \times \mathbf{n})$
Operator	$-\Delta \mathrm{U} - \kappa^2 \mathrm{U}$	$\mathbf{curl}\, \mathbf{curl}\, \mathbf{U} - \kappa^2 \mathbf{U}$

TABLE 1.1. Overview of wave propagation problems.

PROBLEM 1. *Let $D \subset \mathbb{R}^d$, $d = 2, 3$, be a bounded Lipschitz domain. Given $\kappa > 0$ and an incident wave $\mathrm{U}^{\mathrm{inc}}$ such that $\mathsf{L}_\kappa \mathrm{U}^{\mathrm{inc}} = 0$, seek $\mathrm{U} := \mathrm{U}^{\mathrm{scat}} + \mathrm{U}^{\mathrm{inc}}$ in $D^c := \mathbb{R}^d \backslash \overline{D}$ such that*

$$\begin{cases} \mathsf{L}_\kappa \mathrm{U} = 0 & \text{in } D^c, \\ \mathrm{BC}(\mathrm{U}) = 0 & \text{on } \Gamma, \\ \mathrm{RC}(\mathrm{U}^{\mathrm{scat}}, \kappa) & \text{for } r \to \infty. \end{cases}$$

For each case, the PDE in the domain is given by a partial differential operator L_κ while BCs are given on the boundary. The domain D^c being unbounded, radiation (boundary) conditions (RC) are provided at infinity (Nédélec, 2001). Also, the incident field acts as a source term. Notations of Problem 1 are summed up in Table 1.1.

1.3. Analysis in reflexive Banach spaces

Questions of interest such as well-posedness of Problem 1 require a precise analysis. A problem is said to be well posed—according to Hadamard nomenclature—if the following properties hold (Ern & Guermond, 2013):

(i) It admits a solution;

(ii) The solution is unique;

(iii) The solution is endowed with a stability property, namely it is controlled by the data.

The *variational approach* is simple and well suited for a whole class of approximation methods. The BVPs are transformed into an entirely different kind of problem, allowing for an analysis in reflexive Banach spaces (Megginson, 2012; Steinbach, 2007), on which one can lay well-posedness results (refer to Chapter 2 in (Ern & Guermond, 2013) for an exhaustive summary of variational problems in abstract form).

Therefore, one has to approximate the continuous solution, the latter issuing approximation methods such as the Galerkin methods (see e.,g., (Ern & Guermond, 2013, Chapter 2)). Provided adapted assumptions on the approximation space, these allow to transfer well-posedness from continuous to discrete setting. A discrete approximation is obtained at the cost of solving a linear system of the form:

$$\mathbf{A}\mathbf{u} = \mathbf{b}. \tag{1.1}$$

1.4. Boundary Integral Methods

Galerkin methods rely on the use of certain approximation spaces, such as splines, finite elements or spectral bases. A method of choice is FEM (Ern & Guermond, 2013). It consists in meshing the object, and using piecewise defined basis functions—i.e. with local support—to obtain the approximation function. Refer to (Brenner & Scott, 2007) and the references therein for a comprehensive introduction to FEM.

This thesis is particularly devoted to study wave propagation problems, which are commonly defined on unbounded domains. This causes the classical FEM to be impractical, providing the mesh generation for an unbounded volume which is problematic. To amend this, one can approximate the RC—e.g. with an approximation of the Steklov-Poincaré operator (Sauter & Schwab, 2010) to make the domain bounded. As opposed to the later, when the exterior domain is homogeneous, one can resort to Second Green's formula

(Sauter & Schwab, 2010) and arrive at an equivalent problem, which consists in a BIE—inducing a boundary integral operator (BIO), posed on the boundary $\Gamma := \partial D$ with D introduced in Problem 1. Thus, one meshes the boundary and applies classical FEM: these methods are referred to as BEM. This proceeding allows to lower the dimension of the problem, but induces further costs:

(i) The originated matrices are dense and poorly conditioned, preventing the use of direct linear solvers;

(ii) BEM is not straightforward to implement;

(iii) Trace spaces—spaces defined on the boundary, are rather technical.

Taking into account these issues, BEM can be a method of choice due to its precision. We refer readers to Chapter 1 in (Sauter & Schwab, 2010) for a complete introduction to BEM. Without going much more into detail, we emphasize that the theoretical results in this thesis go further than BEM. Still, we decided to use the application to BEM for wave scattering as: (i) a way to improve existing results for the BEM community; and (ii) as a proof of concept for the newly created paradigms and techniques in this thesis.

1.5. Preconditioning and Iterative Solvers

Previously mentioned challenges justify the use for preconditioners to solve the induced linear systems. To begin with, a "good" preconditioner is a matrix $\mathbf{C}$ that can be assembled relatively simply, and such that:

$$\mathbf{CA} \approx \mathbf{I}, \tag{1.2}$$

with $\mathbf{I}$ the identity matrix.

Additionally, we report two important remarks:

(i) One has to define in what sense $\mathbf{CA}$ approximates well the identity. It is worthy to be able to quantify the size of the perturbation of identity. For instance, one

can arrive on continuous level at:

$$\mathsf{CA} = \mathsf{I} + \mathsf{K}, \tag{1.3}$$

with K a compact operator. The latter will ensure that the discrete spectrum will cluster at 1, benefiting to iterative solvers.

(ii) Dependence with respect to parameters is of importance. For example, one will value preconditioners that generates a sequence of linear systems such that the spectral condition numbers:

$$\kappa_S(\mathbf{C}(h)\mathbf{A}(h)) \leq \mathrm{K}, \ h \to 0^+, \tag{1.4}$$

i.e. bounded condition numbers with h, also referred to as *h-independent* condition numbers.

1.6. Fast BEM

BEM induces dense matrices, requiring a memory cost and number of operations per matrix-vector product growing as $\mathcal{O}(N^2)$ with N the number of degrees of freedom of the linear system. However, integral operators have their own characteristics of interest: their Green function $g(x, y)$ generally admits a decay of the form $\mathcal{O}\|x - y\|^{-\alpha}$ with a parameter $\alpha > 0$, as well as its derivatives, with other parameters. That is, they have a singular behavior for $x = y$ and that the value of the interactions decreases when further interactions are considered. Then, the intuition is that high resolution methods are obtained by solving singularities carefully and by approximating distant interactions with less precision.

This idea generated the FMM algorithm (Darve, 2000), introduced by Greengard and Rokhlin, and considered one of the ten best algorithms of the twentieth century. Another more algebraic approach is that of hierarchical matrices (H-mat) (Bebendorf, 2008; Bebendorf, Bollhöfer, & Bratsch, 2013), which use low-rank approximations of the far interactions of the impedance matrix. These apply to a general class of integral operators, called

pseudo-differential operators (Bebendorf, 2008) to which the Helmholtz and Maxwell kernels belong, among others. These methods usually allow to reduce the memory requirements and the cost of the matrix-vector product to $\mathcal{O}(N \log N)$.

However, in the case of oscillatory operators such as those considered throughout, the quasi-linear behavior is not conserved for high frequencies, despite being quite stable numerically, according to (Betcke, van't Wout, & Gélat, 2017). These problems can be solved, using for example directional matrices (Engquist, Ying, et al., 2009), requiring a more complex implementation work. Finally, the discretization of the integral operators involves calculation of integrals on discrete meshes. This is carried out by means of quadrature rules that induce approximation errors.

1.7. Summary of Contributions and Outline

The main contributions of this thesis are the following:

- Extension of the Operator Preconditioning (OP) framework to Petrov-Galerkin methods;
- Introduction of the Bi-Parametric OP framework with application to iterative solvers;
- Application of the Bi-Parametric OP framework to the Multiplicative Calderón Preconditioning for the Electric Field Integral Equation;
- Efficient application of the First-Order Sparse Boundary Element approximation in the context of Uncertainty Quantification for Helmholtz Scattering Problems by random shapes.

This thesis is structured as follows:

(i) In Chapter 2, we extend the operator preconditioning framework (Hiptmair, 2006) to Petrov-Galerkin methods while accounting for parameter-dependent perturbations of both variational forms and their preconditioners, as occurs when performing numerical approximations. By considering different perturbation parameters for the original form and its preconditioner, our bi-parametric abstract

setting leads to robust and controlled schemes. For Hilbert spaces, we derive exhaustive linear and super-linear convergence estimates for iterative solvers delivering h-independent convergent schemes, when preconditioning with low-accuracy or, equivalently, high compression approximations.

(ii) In Chapter 3, we consider the standard Calderón preconditioning for the EFIE. We apply the Bi-Parametric OP framework based on hierarchical matrices. We split solution and preconditioner accuracies, significantly reducing computation times and memory requirements while retaining the good properties of the original Calderón preconditioner. Numerical experiments validate our claims for increasingly complex settings, yielding results comparable to those given by algebraic techniques such as near-field preconditioners and providing insights into further research avenues;

(iii) In Chapter 4, we consider the numerical solution of time-harmonic acoustic scattering by obstacles with uncertain geometries for Dirichlet, Neumann, impedance and transmission boundary conditions. In particular, we aim to quantify diffracted fields originated by small stochastic perturbations of a given relatively smooth nominal shape. Using first-order shape Taylor expansions, we derive tensor deterministic first-kind boundary integral equations for the statistical moments of the scattering problems considered. These are then approximated by sparse tensor Galerkin discretizations via the combination technique (Griebel et al. (Griebel, Schneider, & Zenger, 1990; Griebel & Harbrecht, 2014)). We supply extensive numerical experiments confirming the predicted error convergence rates with poly-logarithmic growth in the number of degrees of freedom and accuracy in approximation of the moments.

Finally, we represent in Figure 1.1 the methodology of this thesis and the main contributions in green.

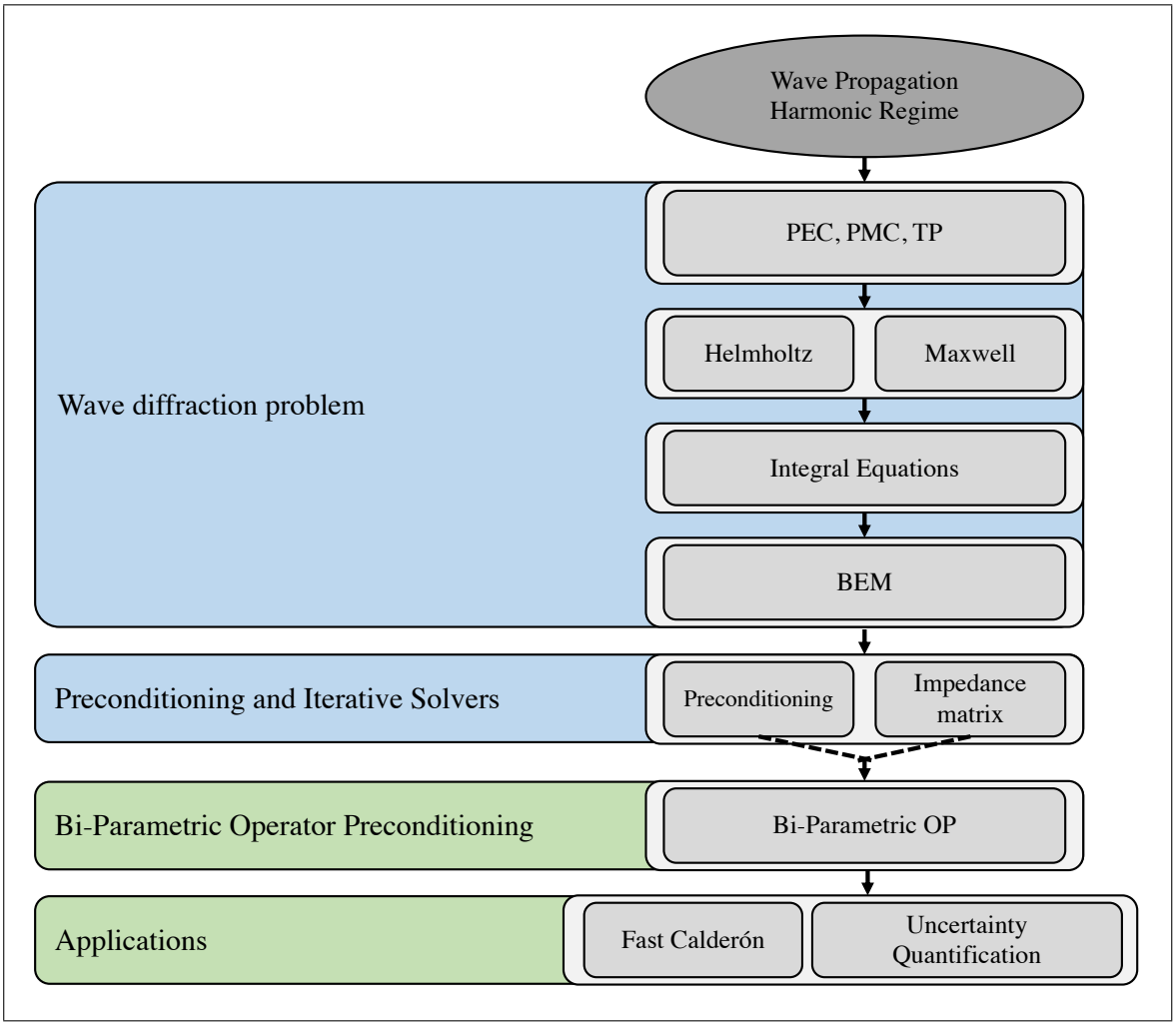


FIGURE 1.1. Global overview of the thesis environment with main contributions in green.

2. BI-PARAMETRIC OPERATOR PRECONDITIONING

This chapter was submitted to Computers & Mathematics with Applications in November, 2020 (under review).

2.1. Introduction

Variational equations—*continuous weak forms* (Betcke, Scroggs, & Śmigaj, 2020, Section 3.1)—in suitably defined reflexive Banach spaces X, Y, or equivalently (Ern & Guermond, 2013, Proposition A.21) as operator equations—*continuous strong forms* (Betcke et al., 2020, Section 3.1)—have successfully been employed to model a plethora of phenomena, particularly in the form of integro-differential equations. In general, one can only approximate solutions by solving linear systems or *matrix equations* arising from the continuous infinite-dimensional counterparts. Galerkin methods are a widely accepted choice to derive such linear systems due to their solid theoretical and practical understanding. Specifically, *Petrov-Galerkin (PG) methods* provide a generic framework for operator equations with operators of the form $\mathsf{A} : X \to Y'$, allowing to choose different trial and test spaces. Within PG methods, one finds *Bubnov-Galerkin (BG) methods*, namely, the case when $\mathsf{A} : X \to X'$ as well as PG for endomorphisms (PGE), i.e., $\mathsf{A} : X \to X$, as in second-kind Fredholm integral equations, wherein A is a compact perturbation of the identity in X.

Most relevant applications lead to large linear systems solved by iterative methods (Saad, 2003) such as Krylov (subspace) methods (Saad & Schultz, 1986, Chapters 6 and 7) as direct inversion quickly becomes computationally impractical. For real symmetric (resp. complex Hermitian) positive definite matrices, the standard choice is the conjugate gradient method (CG) (Hestenes & Stiefel, 1952), whereas the general minimal residual method (GMRES) and its m-restarted variant GMRES(m) (Saad & Schultz, 1986) are common alternatives for nonsingular indefinite complex matrices. For these methods, convergence of the residual strongly depends on matrix properties inherited from the continuous (resp. discrete) operator. Features such as the field-of-values (FoV) or singular values

distributions are key to obtain residual convergence bounds (Sarkis & Szyld, 2007; Beckermann, Goreinov, & Tyrtyshnikov, 2005; Steinbach, 2007; Nevanlinna, 1993). Yet, convergence for these methods can be slow, with performance commonly deteriorating as the linear system dimension increases. Thus, the need for robust preconditioning techniques.

For a linear system $\mathbf{Au} = \mathbf{b}$, in our case spawned by any PG method, preconditioning consists in the application of a (left) preconditioner $\mathbf{P}$ such that

$$\mathbf{PAu} = \mathbf{Pb}.$$

We say that the preconditioner $\mathbf{P}$ is good if: (i) it is relatively cheap to compute; and (ii) the product $\mathbf{PA}$ approximates the identity matrix or iterative solvers perform better than on the original linear system. In this note, we focus on the framework of operator preconditioning (OP). Successfully applied to BG methods (Hiptmair, 2006; S. H. Christiansen & Nédélec, 2000)—denoted OP-BG—, we aim at extending OP to general PG methods (OP-PG) as well as understanding the effects of numerical perturbations in iterative solvers.

Fundamentally, OP relies on finding suitable endomorphic operator equations, i.e. mappings onto the same function spaces, leading to bounded *spectral condition numbers*. In the BG setting, one has reflexive Banach spaces X, V and an operator $\mathsf{A} : X \to X'$, for which one considers another operator $\mathsf{C} : V \to V'$, such that

$$\text{(OP-BG)} \qquad \begin{array}{ccc} X & \xrightarrow{\mathsf{A}} & X' \\ {\scriptstyle \mathsf{M}^{-1}}\uparrow & & \downarrow{\scriptstyle \mathsf{M}^{-*}} \\ V' & \xleftarrow[\mathsf{C}]{} & V \end{array}, \tag{2.1}$$

with $\mathsf{M} : X \to V'$ linking the domain of A and the range of C. The preconditioning operator is then $\mathsf{P} := \mathsf{M}^{-1}\mathsf{C}\mathsf{M}^{-*}$. Similarly, opposite-order OP has been considered for PG methods (Andreev, 2013), particularly in the context of pseudo-differential operators (Steinbach & Wendland, 1998; McLean & Steinbach, 1999; Hsiao & Wendland, 2008),

i.e. for $\mathsf{A} : X \to Y'$ and $\mathsf{C} : Y' \to X$, with[1] $\mathsf{M} = \mathsf{N} =: \mathsf{I}$, leading to

$$\text{(opposite-order OP)} \qquad X \underset{\mathsf{C}}{\overset{\mathsf{A}}{\rightleftarrows}} Y' \ . \tag{2.2}$$

However, there is no known result for general OP-PG, which would encompass both OP-BG and opposite-order PG. This entails considering the following more general framework. For reflexive Banach spaces X, Y, V and W, and a preconditioner $\mathsf{C} : V \to W'$ to $\mathsf{A} : X \to Y'$, we need to build the commuting diagram:

$$\text{(OP-PG)} \qquad \begin{array}{ccc} X & \xrightarrow{\ \mathsf{A}\ } & Y' \\ {\scriptstyle \mathsf{M}^{-1}}\uparrow & & \downarrow{\scriptstyle \mathsf{N}^{-1}} \\ W' & \xleftarrow[\ \mathsf{C}\ ]{} & V \end{array} , \tag{2.3}$$

with $\mathsf{M} : X \to W'$ and $\mathsf{N} : V \to Y'$ linking the domain and range spaces for A and C, and leading to an endomorphism on X. Notice that in this case $\mathsf{P} := \mathsf{M}^{-1}\mathsf{C}\mathsf{N}^{-1}$ and that (2.3) reduces to (2.1) if $W = V$, $Y = X$ and $\mathsf{N} = \mathsf{M}^*$. In this regard, our main contribution is a theory for OP-PG for which we provide estimates for spectral and Euclidean condition numbers. For the latter, we make use of the synthesis operator linking the domain space and its basis expansion, thereby acknowledging the dimension dependence.

Yet, and despite leading to bounded spectral condition numbers, OP does not necessarily ensure convergence for iterative solvers such as GMRES or GMRES(m). Theoretically, one requires further assumptions on the induced problems, related primarily to the matrix FoV distribution (Starke, 1997; Liesen & Tichỳ, 2012), to obtain linear convergence results for GMRES. Still, these bounds are pessimistic (Liesen & Tichỳ, 2012; Kirby, 2010), with convergence radius for GMRES close to one. This justifies the derivation of sharper convergence results at the expense of tighter assumptions on the operators. For instance, one can observe a super-linear convergence of the iterative scheme for systems derived from second-kind Fredholm operator equations. (Moret, 1997; Winther, 1980; Campbell, Ipsen, Kelley, & Meyer, 1996; Axelsson & Karátson, 2018).

[1] Recently, (Stevenson & van Venetië, 2021) proposed a construction with $\mathsf{M}, \mathsf{N} \neq \mathsf{I}$, whose discretization leads to $\mathbf{M}$ and $\mathbf{N}$ being diagonal matrices.

Furthermore, though OP general properties are retained as the linear system dimension increases, it can quickly become impractical. A well-known example is the dual mesh-based OP—also known as multiplicative Calderón preconditioning—for boundary element methods (Hiptmair, 2006; Steinbach & Wendland, 1998; Andriulli et al., 2008). Indeed, due to barycentric grid refinement, the standard method entails a dramatic increase in memory and computational costs. To counter this, low-accuracy Calderón preconditioners have been recently proposed with promising results (Bebendorf, 2008; Escapil-Inchauspé & Jerez-Hanckes, 2019; Fierro & Jerez-Hanckes, 2020). Indeed, iterative solvers' performance is seen to remain stable when building relatively coarse approximations of a given operator preconditioner. Clearly, this has no impact over the solution accuracy as this is only induced by the numerical approximation of the original problem, estimated by Strang's lemma (Strang, 1972) and its variants (Ern & Guermond, 2013; Di Pietro & Droniou, 2018). Accordingly, we recently proposed the idea of systematically "*combining distinct precision orders of magnitude inside the resolution scheme*" (Escapil-Inchauspé & Jerez-Hanckes, 2019) with successful numerical results for boundary element methods in electromagnetics (Escapil-Inchauspé & Jerez-Hanckes, 2019; Kleanthous et al., 2020) and acoustics (Fierro & Jerez-Hanckes, 2020), despite hitherto the lack of rigorous proof.

Thus, we aim to provide theoretical grounds for the above observations by considering parameter-dependent perturbed problems and introducing the *bi-parametric* OP paradigm (Theorem 2.2), with bounds on spectral and Euclidean condition numbers with respect to perturbations. We further deduce linear (resp. super-linear) convergence results for GMRES(m) (resp. GMRES), and present exhaustive new convergence bounds for iterative solvers when working on Hilbert spaces. Due to their generality, our results apply to diverse research areas: equivalent operators theory (Faber, Manteuffel, & Parter, 1990; Axelsson & Karátson, 2009; Kirby, 2010), opposite-order OP (Winther, 1980; Stevenson & van Venetië, 2021), compact equivalent OP (Axelsson, Karátson, & Magoulès, 2018; Axelsson & Karátson, 2018) and (fast) Calderón preconditioning (Andriulli et al., 2008; Escapil-Inchauspé & Jerez-Hanckes, 2019; Antoine & Darbas, 2021; Fierro & Jerez-Hanckes, 2020; Hiptmair & Urzúa-Torres, 2020). Furthermore, these ideas could be also

applied on high frequency wave propagation problems (Graham, Spence, & Vainikko, 2017; Galkowski, Müller, & Spence, 2019), Schwarz preconditioning (Sarkis & Szyld, 2007; Feischl, Führer, Praetorius, & Stephan, 2017) and second-kind Fredholm operator equations (Atkinson, 1976; Colton & Kress, 2012).

This manuscript is structured as follows. In Section 2.2, we present the abstract PG setting. In Section 2.3 we introduce perturbed forms and state the first Strang's lemma for completeness. Next, we arrive at the bi-parametric OP framework and state our main result in Section 2.4. Finally, we investigate the performance of iterative solvers in Section 2.5, and discuss new research avenues in Section 2.6. Figure 2.1 summarizes constants and problems defined throughout this work.

2.2. Continuous, Discrete and Matrix Problem Statements

Let X and Y be two reflexive Banach spaces and let $\mathsf{a} \in \mathcal{L}(X \times Y; \mathbb{C})$ be a continuous complex sesqui-linear—weak—form with norm $\|\mathsf{a}\|$. We tag dual spaces by prime ($'$) and adjoint operators by asterisk (*). For a linear form $b \in Y'$, the *weak continuous problem* is

$$\text{seek} \quad u \in X \quad \text{such that} \quad \mathsf{a}(u, v) = b(v), \quad \forall\, v \in Y. \tag{2.4}$$

Throughout, we assume for each $b \in Y'$ the existence of a unique continuous solution u to (2.4). The form a induces a—strong—bounded linear operator $\mathsf{A} \in \mathcal{L}(X; Y')$ defined through the *dual pairing* in Y as follows (Ern & Guermond, 2013, Proposition A.21)

$$\langle \mathsf{A}u, v \rangle_{Y' \times Y} := a(u, v), \quad \forall\, u \in X,\ \forall\, v \in Y. \tag{2.5}$$

Hence, (2.4) is equivalent to the *strong continuous problem*:

$$\text{seek} \quad u \in X \quad \text{such that} \quad \mathsf{A}u = b. \tag{2.6}$$

Notation	Value	Eq.
K_{A}	$\frac{\|\mathsf{a}\|}{\gamma_{\mathsf{A}}}$	(2.12)
K_{Λ_h}	$\frac{\|\Lambda\|}{\gamma_{\Lambda}}$	(2.16)
$\mathrm{K}_{\star}$	$\frac{\|\mathsf{m}\|\|\mathsf{n}\|\|\mathsf{c}\|\|\mathsf{a}\|}{\gamma_{\mathsf{M}}\gamma_{\mathsf{N}}\gamma_{\mathsf{C}}\gamma_{\mathsf{A}}}$	(2.36)
$\mathrm{K}_{\star,\mu,\nu}$	$\mathrm{K}_{\star}\left(\frac{1+\mu}{1-\mu}\right)\left(\frac{1+\nu}{1-\nu}\right)$	(2.45)
$\overline{\sigma}_k(\mathsf{K})$	$\frac{1}{k}\sum_{j=1}^{k}\sigma_j(\mathsf{K})$	(2.86)

Problem	Problem in matrix form	Eq.
$(\!(\mathrm{A})\!)$	$\mathbf{A}\mathbf{u} = \mathbf{b}$	(2.13)
$(\!(\mathrm{A})\!)_{\nu}$	$\mathbf{A}_{\nu}\mathbf{u}_{\nu} = \mathbf{b}_{\nu}$	(2.26)
$(\!(\mathrm{CA})\!)$	$\mathbf{M}^{-1}\mathbf{C}\mathbf{N}^{-1}\mathbf{A}\mathbf{u} = \mathbf{M}^{-1}\mathbf{C}\mathbf{N}^{-1}\mathbf{b}$	(2.34)
$(\!(\mathrm{CA})\!)_{\mu,\nu}$	$\mathbf{M}^{-1}\mathbf{C}_{\mu}\mathbf{N}^{-1}\mathbf{A}_{\nu}\mathbf{u}_{\nu} = \mathbf{M}^{-1}\mathbf{C}_{\mu}\mathbf{N}^{-1}\mathbf{b}_{\nu}$	(2.43)
$(\!(\mathrm{CA})\!)^{p}_{\mu,\nu}$		(2.83)
$(\!(\mathrm{A})\!)^{p}$	$\mathbf{N}^{-1}\mathbf{A}\mathbf{u} = \mathbf{N}^{-1}\mathbf{b}$	(2.82)

FIGURE 2.1. Comprehensive review of the constants (left) and problems (right) defined throughout this manuscript, along with their corresponding introductions. Convergence radius $\Theta_k^{(m)}$ and $\widetilde{\Theta}_k^{(m)}$ for the preconditioned GMRES are defined in (2.64).

Given an index $h > 0$, we introduce finite-dimensional *conforming spaces*, i.e. $X_h \subset X$ and $Y_h \subset Y$, and assume that $\dim(X_h) = \dim(Y_h) =: N$, with $N \to \infty$ as $h \to 0$. Customarily, h relates to the mesh-size of finite or boundary elements approximations.[2]

The counterpart of (2.4) is the *weak discrete problem*:

$$\text{find} \quad u_h \in X_h \quad \text{such that} \quad \mathsf{a}(u_h, v_h) = b(v_h), \quad \forall\, v_h \in Y_h, \tag{2.7}$$

[2]For the sake of simplicity, the problems under consideration are defined for a given $h > 0$ although asymptotic considerations are key in proving properties such as *h-independent* condition numbers, i.e. remaining bounded as $h \to 0$ (cf. Corollary 2.3).

The above admits a unique solution u_h (Ern & Guermond, 2013, Theorem 2.22) if a satisfies the discrete inf-sup—Banach-Nečas-Babuška (BNB)—condition, for a constant $\gamma_{\mathsf{A}} > 0$:

$$\sup_{v_h \in Y_h \setminus \{\mathbf{0}\}} \frac{|\mathsf{a}(u_h, v_h)|}{\|v_h\|_Y} \geq \gamma_{\mathsf{A}} \|u_h\|_X > 0, \ \forall\, u_h \in X_h. \tag{2.8}$$

Assumption 1. Throughout, we assume that a is continuous and satisfies the BNB condition (2.8).

Equivalently, we define the discrete operator $\mathsf{A}_h : X_h \to Y_h'$:

$$\langle \mathsf{A}_h u_h, v_h \rangle_{Y_h' \times Y_h} := \mathsf{a}(u_h, v_h), \quad \forall\, u_h \in X_h, \ \forall v_h \in Y_h, \tag{2.9}$$

and $b_h \in Y_h'$ such that $b_h(v_h) := b(v_h)$ for all $v_h \in Y_h$, wherein the norms of b_h and A_h are given by (refer to (Sauter & Schwab, 2010, Section 4.2.3)):

$$\|b_h\|_{Y_h'} := \sup_{v_h \in Y_h \setminus \{\mathbf{0}\}} \frac{|\mathsf{a}(u_h, v_h)|}{\|v_h\|_{Y_h}} \quad \text{and} \quad \|\mathsf{A}_h\|_{X_h \to Y_h'} := \sup_{u_h \in X_h \setminus \{\mathbf{0}\}} \frac{\|\mathsf{A}_h u_h\|_{Y_h'}}{\|u_h\|_{X_h}}. \tag{2.10}$$

Consequently, the *strong discrete problem* related to (2.6) reads

$$\text{seek} \quad u_h \in X_h \quad \text{such that} \quad \mathsf{A}_h u_h = b_h. \tag{2.11}$$

One can introduce the *discrete condition number*:

$$\kappa(\mathsf{A}_h) := \|\mathsf{A}_h\|_{X_h \to Y_h'} \|\mathsf{A}_h^{-1}\|_{Y_h' \to X_h} \leq \gamma_{\mathsf{A}}^{-1} \|\mathsf{a}\| =: \mathrm{K}_{\mathsf{A}}, \tag{2.12}$$

with K_{A} being referred to as *BNB condition number*, not to be confused with the BNB condition (2.8).

Pick bases such that $\operatorname{span}\{\varphi_i\}_{i=1}^N = X_h \subset X$ and $\operatorname{span}\{\phi_i\}_{i=1}^N = Y_h \subset Y$, and write the corresponding coefficient vectors in $\mathbb{C}^N$ for the basis expansion in bold letters, e.g.,

$$u_h \in X_h : \quad u_h = \sum_{i=1}^N u_i \varphi_i, \qquad \mathbf{u} := (u_i)_{i=1}^N \in \mathbb{C}^N,$$

$$v_h \in Y_h : \quad v_h = \sum_{i=1}^N v_i \phi_i, \qquad \mathbf{v} := (v_i)_{i=1}^N \in \mathbb{C}^N,$$

and build the (stiffness) Galerkin matrix and right-hand side

$$\mathbf{A} := (\mathsf{a}(\varphi_j, \phi_i))_{i,j=1}^N, \quad \mathbf{b} := (b_h(\phi_i))_{i=1}^N.$$

It holds that

$$\langle \mathsf{A} u_h, v_h \rangle_{Y' \times Y} = \langle \mathsf{A}_h u_h, v_h \rangle_{Y_h' \times Y_h} = (\mathbf{A}\mathbf{u}, \mathbf{v})_2,$$

where $(\mathbf{u}, \mathbf{v})_2$ denotes the Euclidean inner product in $\mathbb{C}^N$ with induced norm $\|\mathbf{u}\|_2 = \sqrt{(\mathbf{u}, \mathbf{u})_2}$. The matrix norm is

$$\|\mathbf{A}\|_2 := \max_{\mathbf{u} \in \mathbb{C}^N \setminus \{\mathbf{0}\}} \frac{\|\mathbf{A}\mathbf{u}\|_2}{\|\mathbf{u}\|_2}.$$

We set $\mathbf{A}^H := \overline{\mathbf{A}}^T$ the conjugate transpose of $\mathbf{A}$ and define vector and matrix norms induced by the Banach space setting as $\|\mathbf{u}\|_{X_h} := \|u_h\|_{X_h}$ and $\|\mathbf{A}\|_{X_h \to Y_h'} := \|\mathsf{A}_h\|_{X_h \to Y_h'}$, for A_h in (2.10). Notice that inclusion $X_h \subset X$ ensures that $\|\mathbf{u}\|_{X_h} = \|\mathbf{u}\|_X =: \|u_h\|_X$.

Consequently, (2.7) and (2.11) correspond to the *matrix problem* referred to[3] as (⟪A⟫):

$$(\!|\mathrm{A}|\!) : \quad \text{Seek} \quad \mathbf{u} \in \mathbb{C}^N \quad \text{such that} \quad \mathbf{A}\mathbf{u} = \mathbf{b}. \tag{2.13}$$

Next, we introduce Λ_h the *synthesis operator* for X_h:

$$\begin{aligned} \Lambda_h : \mathbb{C}^N &\to X_h \\ \mathbf{u} &\mapsto u_h, \end{aligned} \tag{2.14}$$

along with strictly positive constants for $h > 0$

$$\gamma_{\Lambda_h} := \inf_{u_h \in X_h \setminus \{\mathbf{0}\}} \frac{\|u_h\|_X}{\|\mathbf{u}\|_2} \quad \text{and} \quad \|\Lambda_h\| := \sup_{u_h \in X_h \setminus \{\mathbf{0}\}} \frac{\|u_h\|_X}{\|\mathbf{u}\|_2}. \tag{2.15}$$

Notice that, for any $u_h \in X_h$, it holds that (Ern & Guermond, 2006, Section 2.3)

$$\gamma_{\Lambda_h} \|\mathbf{u}\|_2 \le \|u_h\|_X \le \|\Lambda_h\| \|\mathbf{u}\|_2,$$

and set

$$\mathrm{K}_{\Lambda_h} := \frac{\|\Lambda_h\|}{\gamma_{\Lambda_h}}. \tag{2.16}$$

[3] In the following, notation ⟪·⟫ denotes matrix equations.

REMARK 2.1. *One should observe the explicit use of h-subscripts for the synthesis operator. Indeed, while discrete inf-sup conditions are generally bounded as h tends to zero, the bounds $\|\Lambda_h\|$ and γ_{Λ_h} are not. For example, let $\Omega \subset \mathbb{R}^d$, $d = 2, 3$ be a smooth bounded Lipschitz domain (Steinbach, 2007, Section 2) with boundary $\Gamma := \partial\Omega$. For D being either Γ or Ω and $s \in [0, 1]$, we introduce the Sobolev space $H^s(D)$ (Steinbach, 2007) and let $X := H^s(D)$. We assume that D is decomposed into a shape regular, locally quasi-uniform mesh $\mathcal{T}$ (Steinbach, 2007, Section 9.1) with elements $\tau \in \mathcal{T}$. Set h_τ as the diameter of each element $\tau \in \mathcal{T}$, along with $h_{\min} := \min_{\tau\in\mathcal{T}} h_\tau$ and $h \equiv h_{\max} := \max_{\tau\in\mathcal{T}} h_\tau$, and introduce a nodal $\mathcal{C}^0$-Lagrangian basis (Ainsworth, McLean, & Tran, 1999) on $\mathcal{T}$ as* $\operatorname{span}\{\phi_i\}_{i=1}^N = X_h \subset X$, *for any $N(h) \in \mathbb{N}$. For all $u_h \in X_h$, there holds that (Sauter & Schwab, 2010, Sections 4.4 and 4.5):*

$$Ch_{\min}^{\frac{d}{2}}\|\mathbf{u}\|_2 \leq C\|u_h\|_{L^2(D)} \leq \|u_h\|_{H^s(D)} \leq Ch_{\min}^{-s}\|u_h\|_{L^2(D)} \leq Ch_{\min}^{-s}h_{\max}^{\frac{d}{2}}\|\mathbf{u}\|_2. \quad (2.17)$$

Consequently, one obtains

$$\gamma_{\Lambda_h} \geq Ch_{\min}^{\frac{d}{2}}, \quad \|\Lambda_h\| \leq Ch_{\min}^{-s}h_{\max}^{\frac{d}{2}} \quad \textit{and} \quad \mathrm{K}_{\Lambda_h} \leq C\left(\frac{h_{\max}}{h_{\min}}\right)^{\frac{d}{2}} h_{\min}^{-s}. \quad (2.18)$$

For $D = \Gamma$ and $H^s(\Gamma)$, with $s \in [-1, 1]$, one has

$$\mathrm{K}_{\Lambda_h} \leq C\left(\frac{h_{\max}}{h_{\min}}\right)^{\frac{d}{2}} h_{\min}^{-|s|}. \quad (2.19)$$

In this case, one can see the synthesis operator's explicit h-dependence via (2.18) *and* (2.19)*. A similar situation holds in the case of Nédélec and Raviart-Thomas (Rao-Wilton-Glisson) elements applied in electromagnetic scattering (cf. (Hiptmair, Jerez-Hanckes, & Mao, 2015) and references therein).*

For the remainder of this work, we will make extensive use of the spectral and Euclidean condition numbers, $\kappa_S(\mathbf{A})$ and $\kappa_2(\mathbf{A})$, respectively, defined as

$$\kappa_S(\mathbf{A}) := \varrho(\mathbf{A})\varrho(\mathbf{A}^{-1}) = \frac{|\lambda_{\max}(\mathbf{A})|}{|\lambda_{\min}(\mathbf{A})|} \quad \text{and} \quad \kappa_2(\mathbf{A}) := \|\mathbf{A}\|_2\|\mathbf{A}^{-1}\|_2, \quad (2.20)$$

with $\varrho(\mathbf{A}) := |\lambda_{\max}(\mathbf{A})|$ being the spectral radius of $\mathbf{A}$. We denote the spectrum of $\mathbf{A}$ by $\mathfrak{S}(\mathbf{A})$. Since the spectral radius is bounded by any norm on $\mathbb{C}^N$, we set, for any $\mathsf{Q}_h : X_h \to X_h$ with matrix representation $\mathbf{Q}$, the Banach space induced norm $\|\mathbf{Q}\|_X := \|\mathbf{Q}\|_{X\to X}$, with $\|\mathbf{Q}\|_X = \|\mathbf{Q}\|_{X_h} =: \|\mathbf{Q}\|_{X_h\to X_h}$ by inclusion $X_h \subset X$, leading to $\varrho(\mathbf{Q}) \leq \|\mathbf{Q}\|_X$. The latter is key in proving the operator preconditioning result in Theorem 2.1.

As mentioned in Section 2.1, we are concerned with the consequences of perturbing the above sesqui-linear and linear forms over discretization spaces as it occurs when employing finite-arithmetic, numerical integration or compression algorithms. To this end, we give a notion of admissible perturbations needed for the ensuing analysis.

Definition 2.1 ((h,ν)-perturbation)**.** *Let $\nu \in [0,1)$ and $h > 0$ be given. We say that $\mathsf{a}_\nu \in \mathcal{L}(X \times Y;\mathbb{C})$ is a (h,ν)-perturbation of a if it belongs to the set $\Phi_{h,\nu}(\mathsf{a})$:*

$$\mathsf{a}_\nu \in \Phi_{h,\nu}(\mathsf{a}) \iff \gamma_{\mathsf{A}}^{-1}\,|\mathsf{a}(u_h,v_h) - \mathsf{a}_\nu(u_h,v_h)| \leq \nu\|u_h\|_X\|v_h\|_Y,\ \forall\, u_h \in X_h,\ \forall\, v_h \in Y_h. \tag{2.21}$$

Similarly, $b_\nu \in Y'$ is called a (h,ν)-perturbation of the linear form b if it belongs to the set $\Upsilon_{h,\nu}(b)$ defined as

$$b_\nu \in \Upsilon_{h,\nu}(b) \iff |b(v_h) - b_\nu(v_h)| \leq \nu\|b_h\|_{Y_h'}\|v_h\|_{Y_h}, \quad \forall\, v_h \in Y_h.$$

We identify a_0 and b_0 with a and b, respectively.

The (h,ν)-perturbation formalism allows to control precisely the perturbed sesqui-linear (resp. linear) form.

PROPOSITION 2.1. *Consider $\mathsf{a}_\nu \in \Psi_{h,\nu}(\mathsf{a})$. Then, a_ν has a discrete inf-sup condition and is continuous, with corresponding constants γ_{A_ν}, $\|\mathsf{a}_\nu\|$, satisfying*

$$\gamma_{\mathsf{A}_\nu} \geq \gamma_{\mathsf{A}}(1-\nu) \quad \textit{and} \quad \|\mathsf{a}_\nu\| \leq \|\mathsf{a}\| + \nu\gamma_{\mathsf{A}} \leq \|\mathsf{a}\|(1+\nu). \tag{2.22}$$

PROOF. For any $u_h \in X_h$, it holds that

$$\begin{aligned}\sup_{v_h \in Y_h \setminus \{\mathbf{0}\}} \frac{|\mathsf{a}_\nu(u_h, v_h)|}{\|v_h\|_Y} &\geq \sup_{v_h \in Y_h \setminus \{\mathbf{0}\}} \left(\frac{|\mathsf{a}(u_h, v_h)|}{\|v_h\|_Y} - \frac{|\mathsf{a}(u_h, v_h) - \mathsf{a}_\nu(u_h, v_h)|}{\|v_h\|_Y} \right) \\ &\geq \gamma_{\mathsf{A}} \|u_h\|_X - \gamma_{\mathsf{A}} \nu \|u_h\|_X = \gamma_{\mathsf{A}} (1 - \nu) \|u_h\|_X \end{aligned} \tag{2.23}$$

by Assumption 1 and Definition 2.1. Similarly, for any $u_h \in X_h$ and $v_h \in Y_h$, one has

$$\begin{aligned}|\mathsf{a}_\nu(u_h, v_h)| &\leq |\mathsf{a}(u_h, v_h)| + |\mathsf{a}_\nu(u_h, v_h) - \mathsf{a}(u_h, v_h)| \\ &\leq (\|\mathsf{a}\| + \nu \gamma_{\mathsf{A}}) \|u_h\|_X \|v_h\|_Y \leq \|\mathsf{a}\| (1 + \nu) \|u_h\|_X \|v_h\|_Y,\end{aligned}$$

as stated. □

REMARK 2.2. *Though the sets of admissible perturbations $\Phi_{h,\nu}(\mathsf{a})$ and $\Upsilon_{h,\nu}(b)$ depend on h, the perturbed forms remain continuous. Also, for a given h, one may choose different parameters for each set.*

Set $\nu \in [0, 1)$ and introduce perturbations $\mathsf{a}_\nu \in \Phi_{h,\nu}(\mathsf{a})$ and $b_\nu \in \Upsilon_{h,\nu}(b)$. We arrive at the *perturbed weak discrete problem*:

$$\text{seek} \quad u_{h,\nu} \in X_h \quad \text{such that} \quad \mathsf{a}_\nu(u_{h,\nu}, v_h) = b_\nu(v_h), \quad \forall\, v_h \in Y_h, \tag{2.24}$$

with strong discrete counterpart

$$\text{find} \quad u_{h,\nu} \in X_h \quad \text{such that} \quad \mathsf{A}_{h,\nu} u_{h,\nu} = b_{h,\nu}, \tag{2.25}$$

and matrix form

$$(\!|\mathrm{A}|\!)_\nu : \quad \text{Seek} \quad \mathbf{u}_\nu \in \mathbb{C}^N \quad \text{such that} \quad \mathbf{A}_\nu \mathbf{u} = \mathbf{b}_\nu. \tag{2.26}$$

Notice that $(\!|\mathrm{A}|\!)_0 = (\!|\mathrm{A}|\!)$. Moreover, one can combine Proposition 2.1 with (Ern & Guermond, 2013, Theorem 2.22) to obtain the next result.

PROPOSITION 2.2. *For $\nu \in [0, 1)$, $(\!|\mathrm{A}|\!)_\nu$ admits a unique solution.*

2.3. First Strang's lemma for perturbed forms

We start by characterizing the error between continuous and discrete solutions for the unperturbed version of (A) recalling Céa's lemma (Céa, 1964; Ern & Guermond, 2013).

Lemma 2.1 (Céa's Lemma (Ern & Guermond, 2013, Lemma 2.28))**.** *Let $u \in X$ and $u_h \in X_h$ be the solutions to* (2.4) *and* (2.7)*, respectively. Then, one has*

$$\|u - u_h\|_X \leq (1 + \mathrm{K_A}) \inf_{w_h \in X_h} \|u - w_h\|_X, \tag{2.27}$$

with $\mathrm{K_A}$ *defined in* (2.12)*.*

REMARK 2.3. *This fundamental result highlights the importance of the BNB condition number. It shows that if the problem has poor intrinsic conditioning for either continuous or discrete settings, then the quasi-optimality constant* $(1 + \mathrm{K_A})$ *will be large and the solution* u_h *far from the best approximation error. Observe that in Lemma 2.1 both sesqui-linear and linear forms are computed exactly.*

Next, we present a modified version of the above lemma for perturbed problems $(\mathrm{A})_\nu$.

Lemma 2.2 (First Strang's Lemma)**.** *Set $\nu \in [0, 1)$ and let $u_{h,\nu} \in X_h$ and $u \in X$ be the unique solutions to* (2.25) *and* (2.4)*, respectively. It holds that*

$$\begin{aligned}\|u - u_{h,\nu}\|_X &\leq \inf_{w_h \in X_h} \left(\left(1 + \frac{\mathrm{K_A}}{1 - \nu} \right) \|u - w_h\|_X + \frac{\nu}{1 - \nu} \|w_h\|_X \right) + \frac{\nu}{\gamma_{\mathsf{A}}(1 - \nu)} \|b_h\|_{Y_h'} \\ &\leq (1 + \mathrm{K_A}) \left(1 + \frac{\mathrm{K_A}}{1 - \nu} \right) \inf_{w_h \in X_h} \|u - w_h\|_X + \frac{2\nu}{\gamma_{\mathsf{A}}(1 - \nu)} \|b_h\|_{Y_h'}.\end{aligned}$$

PROOF. For any $w_h \in X_h$ and for all $v_h \in Y_h$, it holds that

$$\begin{aligned}\mathsf{a}_\nu(u_{h,\nu} - w_h, v_h) &= b_\nu(v_h) - \mathsf{a}_\nu(w_h, v_h) + \mathsf{a}(w_h, v_h) + \mathsf{a}(u - w_h, v_h) - b(v_h) \\ &= \mathsf{a}(u - w_h, v_h) + (\mathsf{a}(w_h, v_h) - \mathsf{a}_\nu(w_h, v_h)) + (b_\nu(v_h) - b(v_h)),\end{aligned}$$

leading to

$$\gamma_{\mathsf{A}_\nu} \|u_{h,\nu} - w_h\|_X \leq \|\mathsf{a}\| \|u - w_h\|_X + \nu \gamma_{\mathsf{A}} \|w_h\|_X + \|b_h - b_{h,\nu}\|_{Y_h'} \tag{2.28}$$

by Assumption 1 and Definition 2.1. Next, by combining the triangle inequality, and (2.28), one derives

$$\begin{aligned}\|u-u_{h,\nu}\|_X &\le \|u-w_h\|_X + \|w_h-u_{h,\nu}\|_X \\ &\le \left(1+\frac{\|\mathsf{a}\|}{\gamma_{\mathsf{A}_\nu}}\right)\|u-w_h\|_X + \nu\frac{\gamma_{\mathsf{A}}}{\gamma_{\mathsf{A}_\nu}}\|w_h\|_X + \frac{1}{\gamma_{\mathsf{A}_\nu}}\|b_h-b_{h,\nu}\|_{Y_h'},\end{aligned}$$

and, since w_h is arbitrary in X_h, there holds

$$\begin{aligned}\|u-u_{h,\nu}\|_X &\le \frac{1}{\gamma_{\mathsf{A}_\nu}}\|b_h-b_{h,\nu}\|_{Y_h'} + \inf_{w_h\in X_h}\left(\left(1+\frac{\|\mathsf{a}\|}{\gamma_{\mathsf{A}_\nu}}\right)\|u-w_h\|_X + \frac{\gamma_{\mathsf{A}}}{\gamma_{\mathsf{A}_\nu}}\nu\|w_h\|_X\right) \\ &\le \frac{\nu}{\gamma_{\mathsf{A}}(1-\nu)}\|b_h\|_{Y_h'} + \inf_{w_h\in X_h}\left(\left(1+\frac{\mathrm{K}_{\mathsf{A}}}{1-\nu}\right)\|u-w_h\|_X + \frac{\nu}{1-\nu}\|w_h\|_X\right) \\ &\le \frac{\nu}{\gamma_{\mathsf{A}}(1-\nu)}\|b_h\|_{Y_h'} + \left(1+\frac{\mathrm{K}_{\mathsf{A}}}{1-\nu}\right)\|u-u_h\|_X + \frac{\nu}{1-\nu}\|u_h\|_X \\ &\le \frac{2\nu}{\gamma_{\mathsf{A}}(1-\nu)}\|b_h\|_{Y_h'} + (1+\mathrm{K}_{\mathsf{A}})\left(1+\frac{\mathrm{K}_{\mathsf{A}}}{1-\nu}\right)\inf_{w_h\in X_h}\|u-w_h\|_X,\end{aligned}$$

as stated, by recalling the continuous dependence on b for u_h solution of (2.7), i.e. $\|u_h\|_X \le \frac{1}{\gamma_{\mathsf{A}}}\|b_h\|_{Y_h'}$, and by application of Lemma 2.1. □

REMARK 2.4. *Since*

$$\kappa(\mathsf{A}_{h,\nu}) \le \frac{\|\mathsf{a}_\nu\|}{\gamma_{\mathsf{A}_\nu}} =: \mathrm{K}_{\mathsf{A}_\nu} \le \mathsf{K}_{\mathsf{A}}\frac{1+\nu}{1-\nu},$$

one can expect the discrete (resp. BNB) condition number of a_ν *to be stable with respect to small perturbations, as* $\mathrm{K}_{\mathsf{A}_\nu} = \mathrm{K}_{\mathsf{A}}(1-2\nu+o(\nu))$ *for* $\nu \ll 1$. *For* $\nu \ll 1$, *Lemma 2.2 shows that the perturbation implies: a best approximation error term with quasi-optimality constant* $(1+\mathrm{K}_{\mathsf{A}})^2$, *and* $\mathcal{O}(\nu)$ *errors induced by the perturbed sesqui-linear form and right-hand side (cf. (Escapil-Inchauspé & Jerez-Hanckes, 2019, Sections 2 and 3)).*

REMARK 2.5. *Observe that contrary to Céa's lemma, Lemma 2.2 does not invoke the solution to the continuous perturbed problem:*

$$\textit{seek}\quad u_\nu \in X \quad \textit{such that}\quad \mathsf{a}_\nu(u_\nu, v) = b_\nu(v), \quad \forall\, v \in Y. \tag{2.29}$$

Assuming the existence of a unique continuous solution u_ν to (2.29)*, Lemma 2.1 and Remark 2.4 lead to the following quasi-optimal bound:*

$$\|u_\nu - u_{h,\nu}\|_X \leq \left(1 + \mathrm{K}_{\mathsf{A}} \frac{1+\nu}{1-\nu}\right) \inf_{w_h \in X_h} \|u_\nu - w_h\|_X. \tag{2.30}$$

2.4. Bi-parametric Operator Preconditioning

We complete the setting in Section 2.2 by introducing preconditioners. To this end, let V and W be two reflexive Banach spaces. We consider an operator $\mathsf{c} \in \mathcal{L}(V \times W; \mathbb{C})$ as well as pairings $\mathsf{n} \in \mathcal{L}(V \times Y; \mathbb{C})$ and $\mathsf{m} \in \mathcal{L}(X \times W; \mathbb{C})$. These forms induce operators $\mathsf{C} : V \to W'$, $\mathsf{N} : V \to Y'$ and $\mathsf{M} : X \to W'$. With these, we state the preconditioned version of the operator equation (2.6):

$$\text{seek} \quad u \in X \quad \text{such that} \quad \mathsf{P}\mathsf{A}u = \mathsf{P}b, \quad \text{with} \quad \mathsf{P} := \mathsf{M}^{-1}\mathsf{C}\mathsf{N}^{-1}. \tag{2.31}$$

We refer the readers to Figure 2.1 and to the previous diagram in (2.3) for an overview of domain mappings and functional spaces for OP-PG.

For our new spaces, we set conforming finite-dimensional spaces $V_h \subset V$ and $W_h \subset W$ of the same dimension N as for X_h and Y_h.

Assumption 2. We assume that c, n and m satisfy a discrete inf-sup condition (cf. (2.8)) over the approximations spaces, with constants $\gamma_{\mathsf{C}}, \gamma_{\mathsf{N}}$ and γ_{M}, respectively.

Consequently, the *strong discrete preconditioned problem*

$$\text{seek} \quad u_h \in X_h \quad \text{such that} \quad \mathsf{P}_h\mathsf{A}_h u_h = \mathsf{P}_h b_h, \quad \text{with} \quad \mathsf{P}_h := \mathsf{M}_h^{-1}\mathsf{C}_h\mathsf{N}_h^{-1}, \tag{2.32}$$

is well posed, by the same arguments as in Proposition 2.2. As in Section 2.2, we now pick bases $\{\psi_i\}_{i=1}^N \subset V_h$ and $\{\xi_i\}_{i=1}^N \subset W_h$ of V_h and W_h, and build the Galerkin matrices

$$\mathbf{C} := ((\mathsf{c}(\psi_j, \xi_i))_{i,j=1}^N, \quad \mathbf{M} := ((\mathsf{m}(\varphi_j, \xi_i))_{i,j=1}^N \quad \text{and} \quad \mathbf{N} := ((\mathsf{n}(\psi_j, \phi_i))_{i,j=1}^N. \tag{2.33}$$

	Operator	sesqui-linear form	Matrix	Constants
Impedance	$\mathsf{A} : X \to Y'$	$\mathsf{a} : X \times Y$	$\mathbf{A} : [Y_h \times X_h]$	$\gamma_{\mathsf{A}}, \|\mathsf{a}\|$
Preconditioner	$\mathsf{C} : V \to W'$	$\mathsf{c} : V \times W$	$\mathbf{C} : [W_h \times V_h]$	$\gamma_{\mathsf{C}}, \|\mathsf{c}\|$
Pairing A	$\mathsf{N} : V \to Y'$	$\mathsf{n} : V \times Y$	$\mathbf{N}^{-1} : [V_h \times Y_h]$	$\gamma_{\mathsf{N}}, \|\mathsf{n}\|$
Pairing C	$\mathsf{M} : X \to W'$	$\mathsf{m} : X \times W$	$\mathbf{M}^{-1} : [X_h \times W_h]$	$\gamma_{\mathsf{M}}, \|\mathsf{m}\|$

TABLE 2.1. Overview of functional spaces for OP-PG. We specify spaces for the corresponding continuous operators and sesqui-linear forms, along with their induced discrete matrices, continuity and discrete-inf sup constants. Brackets for matrices indicate the spaces associated to rows $\times$ columns.

Therefore, we arrive at the matrix problem:

$$(\!(\text{CA})\!): \quad \text{find} \quad \mathbf{u} \in \mathbb{C}^N \quad \text{such that} \quad \mathbf{PAu} = \mathbf{Pb}, \quad \text{with} \quad \mathbf{P} := \mathbf{M}^{-1}\mathbf{C}\mathbf{N}^{-1}. \tag{2.34}$$

REMARK 2.6. *As hinted in (Betcke et al., 2020), OP allows to obtain an equivalent representation for both the discrete and matrix settings, referred to as Galerkin product algebra. Indeed, introduce a unique $v_h \in V_h$ such that $\mathsf{N}_h v_h = b_h$, $w_h := \mathsf{C}_h v_h \in W_h'$, and a unique $q_h \in X_h$ such that $\mathsf{M}_h q_h = w_h$. We obtain that*

$$\mathsf{A}_h u_h = b_h = \mathsf{N}_h v_h \quad \Rightarrow \quad \mathsf{N}_h^{-1}\mathsf{A}_h u_h = v_h \in V_h,$$
$$\mathsf{C}_h v_h = w_h = \mathsf{M}_h q_h \quad \Rightarrow \quad \mathsf{M}_h^{-1}\mathsf{C}_h v_h = q_h \in X_h,$$

leading to matrix counterparts

$$\mathbf{Au} = \mathbf{Nv} \quad \Rightarrow \quad \mathbf{N}^{-1}\mathbf{Au} = \mathbf{v},$$
$$\mathbf{Cv} = \mathbf{Mq} \quad \Rightarrow \quad \mathbf{M}^{-1}\mathbf{Cv} = \mathbf{q}.$$

Hence

$$q_h = \mathsf{P}_h\mathsf{A}_h u_h \quad \textit{with basis expansion} \quad \mathbf{q} = \mathbf{PAu}. \tag{2.35}$$

Consequently, $u_h = (\mathsf{P}_h\mathsf{A}_h)^{-1} q_h$ is with basis expansion $\mathbf{u} = (\mathbf{PA})^{-1}\mathbf{q}$.

We state the following estimates for the condition numbers of $\mathbf{PA}$.

Theorem 2.1 (Estimates for OP-PG)**.** *For problem* ((CA)) *given in* (2.34)*, the spectral condition number is bounded as*

$$\kappa_S(\mathbf{PA}) \leq \kappa(\mathsf{P}_h\mathsf{A}_h) \leq \frac{\|\mathsf{m}\|\|\mathsf{n}\|\|\mathsf{c}\|\|\mathsf{a}\|}{\gamma_{\mathsf{M}}\gamma_{\mathsf{N}}\gamma_{\mathsf{C}}\gamma_{\mathsf{A}}} =: \mathrm{K}_\star\,. \tag{2.36}$$

Furthermore, the Euclidean condition number satisfies

$$\kappa_2(\mathbf{PA}) \leq \mathrm{K}_\star \left(\frac{\|\Lambda_h\|}{\gamma_{\Lambda_h}}\right)^2 = \mathrm{K}_\star\,\mathrm{K}_{\Lambda_h}^2, \tag{2.37}$$

with K_{Λ_h} *introduced in* (2.16)*.*

PROOF. Remark that, for any $u_h \in X_h$, it holds that

$$\frac{\gamma_{\mathsf{C}}\gamma_{\mathsf{A}}}{\|\mathsf{m}\|\|\mathsf{n}\|}\|u_h\|_X \leq \|\mathsf{P}_h\mathsf{A}_h u_h\|_X \leq \frac{\|\mathsf{c}\|\|\mathsf{a}\|}{\gamma_{\mathsf{N}}\gamma_{\mathsf{M}}}\|u_h\|_X. \tag{2.38}$$

Let us introduce $\mathbf{u}$ linked to u_h so as to deduce that

$$\|\mathsf{P}_h\mathsf{A}_h\|_X = \|\mathbf{PA}\|_X \leq \frac{\|\mathsf{c}\|\|\mathsf{a}\|}{\gamma_{\mathsf{N}}\gamma_{\mathsf{M}}} \quad \text{and} \quad \|(\mathsf{P}_h\mathsf{A}_h)^{-1}\|_X = \|(\mathbf{PA})^{-1}\|_X \leq \frac{\|\mathsf{m}\|\|\mathsf{n}\|}{\gamma_{\mathsf{C}}\gamma_{\mathsf{A}}}, \tag{2.39}$$

which leads to the stated result for the spectral condition number given in (2.20), since $\varrho(\mathbf{PA}) \leq \|\mathbf{PA}\|_X$ and $\varrho((\mathbf{PA})^{-1}) \leq \|(\mathbf{PA})^{-1}\|_X$.

For the Euclidean condition number, we employ the synthesis operator Λ_h, introduced in (2.14), and (2.38), to derive

$$\frac{1}{\|\Lambda_h\|}\left(\frac{\gamma_{\mathsf{C}}\gamma_{\mathsf{A}}}{\|\mathsf{m}\|\|\mathsf{n}\|}\right)\|u_h\|_X \leq \|\mathbf{PAu}\|_2 \leq \frac{1}{\gamma_{\Lambda_h}}\left(\frac{\|\mathsf{c}\|\|\mathsf{a}\|}{\gamma_{\mathsf{M}}\gamma_{\mathsf{N}}}\right)\|u_h\|_X, \tag{2.40}$$

yielding

$$\frac{\gamma_{\Lambda_h}}{\|\Lambda_h\|}\left(\frac{\gamma_{\mathsf{C}}\gamma_{\mathsf{A}}}{\|\mathsf{m}\|\|\mathsf{n}\|}\right)\|\mathbf{u}\|_2 \leq \|\mathbf{PAu}\|_2 \leq \frac{\|\Lambda_h\|}{\gamma_{\Lambda_h}}\left(\frac{\|\mathsf{c}\|\|\mathsf{a}\|}{\gamma_{\mathsf{M}}\gamma_{\mathsf{N}}}\right)\|\mathbf{u}\|_2, \tag{2.41}$$

providing the second result. □

As mentioned in Section 2.1, the abstract formulation in Theorem 2.1 for OP-PG encompasses the following important cases:

(i) OP-BG (Hiptmair, 2006; S. H. Christiansen & Nédélec, 2000): $X = Y$, $V = W$ and $\mathsf{N} := \mathsf{M}^*$ (cf. (2.1));

(ii) Opposite-order OP (Andreev, 2013): $Y = X'$ and $W = V'$ and $\mathsf{M} = \mathsf{N} := \mathsf{I}$ (cf. (2.2)).

We are now ready to introduce perturbed sesqui-linear forms and their preconditioners. In the spirit of (2.24), we consider the family of *bi-parametric* perturbed preconditioned problems.

For two parameters $\mu, \nu \in [0, 1)$, we define $\mathsf{c}_\mu \in \Phi_{h,\mu}(\mathsf{c})$, $\mathsf{a}_\nu \in \Phi_{h,\nu}(\mathsf{a})$, and $b_\nu \in \Upsilon_{h,\nu}(b)$. The *perturbed preconditioned* problem reads

$$\text{find} \quad u_{h,\nu} \in X_h \quad \text{such that} \quad \mathsf{P}_{h,\mu}\mathsf{A}_{h,\nu}u_{h,\nu} = \mathsf{P}_{h,\mu}b_{h,\nu}, \quad \text{with} \quad \mathsf{P}_{h,\mu} := \mathsf{M}_h^{-1}\mathsf{C}_{h,\mu},\mathsf{N}_h^{-1}, \tag{2.42}$$

with corresponding matrix form

$$(\!|\text{CA}|\!)_{\mu,\nu}: \quad \text{seek} \quad \mathbf{u}_\nu \in \mathbb{C}^N \quad \text{such that} \quad \mathbf{P}_\mu\mathbf{A}_\nu\mathbf{u}_\nu = \mathbf{P}_\mu\mathbf{b}_\nu, \quad \text{with} \quad \mathbf{P}_\mu := \mathbf{M}^{-1}\mathbf{C}_\mu\mathbf{N}^{-1}. \tag{2.43}$$

Naturally, $(\!|\text{CA}|\!)_{0,0} =$(CA). In practice, one seeks the preconditioner parameter μ to be much larger than the original system's accuracy ν while retaining the convergence properties. Indeed, we can now state our main result.

Theorem 2.2 (Bi-Parametric Operator Preconditioning)**.** *For the problem* $(\!|\text{CA}|\!)_{\mu,\nu}$*, given in* (2.43) *for* $\mu, \nu \in [0, 1)$ *and* $h > 0$*, the spectral condition number is bounded as*

$$\kappa_S(\mathbf{P}_\mu\mathbf{A}_\nu) \leq \mathrm{K}_\star \left(\frac{1+\mu}{1-\mu}\right)\left(\frac{1+\nu}{1-\nu}\right) =: \mathrm{K}_{\star,\mu,\nu} \tag{2.44}$$

and the Euclidean condition number satisfies

$$\kappa_2(\mathbf{P}_\mu\mathbf{A}_\nu) \leq \mathrm{K}_{\star,\mu,\nu}\,\mathrm{K}_{\Lambda_h}^2, \tag{2.45}$$

with $\mathrm{K}_\star$ *and* K_{Λ_h} *defined in* (2.16) *and* (2.36)*, respectively.*

PROOF. Application of Proposition 2.1 to c_μ and a_ν leads to:

$$\forall\, u_h \in X_h, \quad (1-\mu)(1-\nu)\frac{\gamma_{\mathsf{C}}\gamma_{\mathsf{A}}}{\|\mathsf{m}\|\|\mathsf{n}\|}\|u_h\|_X \leq \|\mathsf{P}_{h,\mu}\mathsf{A}_{h,\nu}u_h\|_X \leq \frac{\|\mathsf{c}\|\|\mathsf{a}\|}{\gamma_{\mathsf{A}}\gamma_{\mathsf{C}}}(1+\mu)(1+\nu)\|u_h\|_X, \tag{2.46}$$

from where one derives the result for the spectral condition number following the proof of Theorem 2.1. For the Euclidean condition number, the proof is similar modulo the term K_{Λ_h} due to the synthesis operator. □

REMARK 2.7. *Theorem 2.2 provides bounds for both spectral and Euclidean condition numbers. Notice that* (2.45) *involves the synthesis operators in* X_h *(see Remark 2.1). Moreover, it holds that* $\mathrm{K}_{\star,\mu,\nu} = \mathrm{K}_{\star,\nu,\mu}$*, and* $\mathrm{K}_{\star,\mu,\nu}$ *does not involve cross-terms in* μ *and* ν*. Remark that* (2.44) *is a sharper estimate than the previous bound in (Escapil-Inchauspé & Jerez-Hanckes, 2019, Proposition 1). Also, we have assumed* M *and* N *to be exact or unperturbed but one could also extend the above results to account for perturbed pairings.*

Theorem 2.2 constitutes the formal proof of the effectiveness of preconditioning with low-accuracy approximations hinted, for instance, by Bebendorf in (Bebendorf, 2008, Section 3.6). To illustrate this, assume that the best approximation error in Lemma 2.1 converges at a rate $\mathcal{O}(h^r)$, $r > 0$. First, Theorem 2.2 shows that one can set $\nu = \mathcal{O}(h^r)$ to preserve the convergence rate. Second, one can relax μ by setting a bounded $\mu = \mathcal{O}(1)$ guaranteeing a bounded spectral condition number. Consequently, the result suggests using different parameters for the assembly of $\mathbf{P}_\mu$ and $\mathbf{A}_\nu$. For example, one can keep standard Galerkin methods for building stiffness matrices with preconditioners built using coarser Galerkin approximations (Escapil-Inchauspé & Jerez-Hanckes, 2019; Kleanthous et al., 2020; Fierro & Jerez-Hanckes, 2020), collocation methods (Atkinson, 1976), compression techniques (Bebendorf, 2008; Bebendorf & Kunis, 2009), or feedforward neural networks (Meade Jr & Fernandez, 1994; Sappl, Seiler, Harders, & Rauch, 2019).

2.5. Iterative Solvers Performance: Hilbert space setting

Throughout Section 2.5, we restrict ourselves to $X \equiv H$ with H being a Hilbert space with inner product $(\cdot,\cdot)_H$ and $\|\cdot\|_H = \sqrt{(\cdot,\cdot)_H}$. We set $\mathbf{H} := ((\varphi_j,\varphi_i)_H)_{i,j=1}^N$, being Hermitian positive definite with $\{\varphi_i\}_{i=1}^N$ defined in Section 2.2, satisfying

$$\forall\, u_h, v_h \in X_h, \quad (u_h, v_h)_H = \langle \mathsf{R} u_h, v_h\rangle_{H'\times H} = (\mathbf{H}\mathbf{u},\mathbf{v})_2 =: (\mathbf{u},\mathbf{v})_H, \tag{2.47}$$

where R is the isometric Riesz-isomorphism $H \to H'$ (Hiptmair, 2006, Section 3).

We aim at detailing how the context of Theorem 2.1 and Theorem 2.2 transfers onto the behavior of iterative solvers such as GMRES under the above Hilbertian setting. To this end, the following matrix properties will prove useful.

2.5.1. Matrix properties: H-FoV

For any $\mathbf{Q} \in \mathbb{C}^{N\times N}$, $N \in \mathbb{N}$, we introduce $\mathcal{F}_H(\mathbf{Q})$, the matrix H-FoV of $\mathbf{Q}$—also referred to as H-numerical range

$$\mathcal{F}_H(\mathbf{Q}) := \left\{ \frac{(\mathbf{Q}\mathbf{u},\mathbf{u})_H}{(\mathbf{u},\mathbf{u})_H} : \mathbf{u} \in \mathbb{C}^N \setminus \{\mathbf{0}\} \right\} \tag{2.48}$$

and $\mathcal{V}_H(\mathbf{Q})$, the distance of $\mathcal{F}_H(\mathbf{Q})$ from the origin

$$\mathcal{V}_H(\mathbf{Q}) := \min_{z\in\mathcal{F}_H(\mathbf{Q})} |z| = \min_{\mathbf{u}\in\mathbb{C}^N\setminus\{\mathbf{0}\}} \frac{|(\mathbf{Q}\mathbf{u},\mathbf{u})_H|}{(\mathbf{u},\mathbf{u})_H}. \tag{2.49}$$

Likewise, we introduce $\mathcal{F}_2(\mathbf{Q})$, or equivalently 2-FoV, and $\mathcal{V}_2(\mathbf{Q})$. Moreover, for any $\mathsf{Q}_h : X_h \to X_h$, we set the discrete H-FoV and $\mathcal{V}_H(\mathsf{Q}_h)$:

$$\mathcal{F}_H(\mathsf{Q}_h) := \left\{ \frac{(\mathsf{Q}_h u_h, u_h)_H}{(u_h,u_h)_H} : u_h \in X_h \setminus \{0\} \right\} \quad \text{and} \quad \mathcal{V}_H(\mathsf{Q}_h) := \inf_{u_h\in X_h\setminus\{0\}} \frac{|(\mathsf{Q}_h u_h, u_h)_H|}{(u_h,u_h)_H}. \tag{2.50}$$

We recall that the H-adjoint of $\mathbf{Q}$ is $\mathbf{Q}^\star := \mathbf{H}^{-1}\mathbf{Q}^H\mathbf{H}$, and that $\mathbf{Q}$ is said to be *H-normal* if $\mathbf{Q}$ commutes with $\mathbf{Q}^\star$ (Axelsson & Karátson, 2009, Section 2.2.1.1).

The matrix H-FoV (and 2-FoV) being key in describing the linear convergence of GMRES(m), we aim at giving a further insight on these sets. Following (Benzi, 2016, Section 4), we state some useful properties of the matrix H-FoV.

Lemma 2.3 (Properties of the matrix H-FoV $\mathcal{F}_H(\mathbf{Q})$)**.** *Consider any* $\mathbf{Q}, \mathbf{H} \in \mathbb{C}^{N\times N}$, $N \in \mathbb{N}$, *with* $\mathbf{H}$ *being a Hermitian positive definite matrix. The following properties hold:*

(i) $\mathcal{F}_H(\mathbf{Q}) = \mathcal{F}_2(\mathbf{H}^{\frac{1}{2}}\mathbf{Q}\mathbf{H}^{-\frac{1}{2}})$*;*

(ii) Spectral containment: $\mathfrak{S}(\mathbf{Q}) \subset \mathcal{F}_H(\mathbf{Q})$*;*

(iii) H*-normal matrices: If* $\mathbf{Q}$ *is* H*-normal, then* $\mathcal{F}_H(\mathbf{Q}) = \mathrm{Conv}(\mathfrak{S}(\mathbf{Q}))$ *the convex hull of* $\mathfrak{S}(\mathbf{Q})$*;*

(iv) $\mathcal{F}_H(\mathbf{Q})$ *is contained in a disk centered at* 0 *with radius* $\|\mathbf{Q}\|_H$*;*

(v) $\mathcal{F}_H(\mathbf{Q})$ *is compact and convex.*

PROOF. Set $\widehat{\mathbf{Q}} := \mathbf{H}^{\frac{1}{2}}\mathbf{Q}\mathbf{H}^{-\frac{1}{2}}$.

(i) For any $\mathbf{u} \in \mathbb{C}^N \setminus \{\mathbf{0}\}$, one can define $\widehat{\mathbf{u}} := \mathbf{H}^{\frac{1}{2}}\mathbf{u}$ such that

$$\frac{(\mathbf{Q}\mathbf{u},\mathbf{u})_H}{(\mathbf{u},\mathbf{u})_H} = \frac{(\mathbf{H}^{\frac{1}{2}}\widehat{\mathbf{Q}}\mathbf{H}^{\frac{1}{2}}\mathbf{u},\mathbf{u})_2}{(\mathbf{H}^{\frac{1}{2}}\mathbf{H}^{\frac{1}{2}}\mathbf{u},\mathbf{u})_2} = \frac{(\widehat{\mathbf{Q}}\widehat{\mathbf{u}},\widehat{\mathbf{u}})_2}{(\widehat{\mathbf{u}},\widehat{\mathbf{u}})_2},$$

proving that $\mathcal{F}_H(\mathbf{Q}) = \mathcal{F}_2(\widehat{\mathbf{Q}})$, and that $\|\widehat{\mathbf{Q}}\|_2 = \|\mathbf{Q}\|_H$.

(ii) By (Benzi, 2016, Section 4, Item 1), one has $\mathfrak{S}(\widehat{\mathbf{Q}}) \subset \mathcal{F}_2(\widehat{\mathbf{Q}})$. Clearly, $\mathbf{Q}$ and $\widehat{\mathbf{Q}}$ share the same spectrum.

(iii) If $\mathbf{Q}$ is H-normal, there holds that $\mathbf{Q}\mathbf{H}^{-1}\mathbf{Q}^H\mathbf{H} = \mathbf{H}^{-1}\mathbf{Q}^H\mathbf{H}\mathbf{Q}$, hence

$$\mathbf{H}^{\frac{1}{2}}\mathbf{Q}\mathbf{H}^{-1}\mathbf{Q}^H\mathbf{H}\mathbf{H}^{-\frac{1}{2}} = \mathbf{H}^{\frac{1}{2}}\mathbf{H}^{-1}\mathbf{Q}^H\mathbf{H}\mathbf{Q}\mathbf{H}^{-\frac{1}{2}},$$

leading to $\widehat{\mathbf{Q}}\widehat{\mathbf{Q}}^H = \widehat{\mathbf{Q}}^H\widehat{\mathbf{Q}}$, proving that $\widehat{\mathbf{Q}}$ is normal. By (Benzi, 2016, Section 4, Item 10), we deduce that $\mathcal{F}_2(\widehat{\mathbf{Q}}) = \mathrm{Conv}(\mathfrak{S}(\widehat{\mathbf{Q}}))$.

(iv) $\mathcal{F}_2(\widehat{\mathbf{Q}})$ is contained in a disk centered at zero with radius $\|\widehat{\mathbf{Q}}\|_2$ (Benzi, 2016, Section 4, Item 3). Moreover, $\mathcal{F}_H(\mathbf{Q}) = \mathcal{F}_2(\widehat{\mathbf{Q}})$ and $\|\mathbf{Q}\|_H = \|\widehat{\mathbf{Q}}\|_2$.

(v) $\mathcal{F}_2(\widehat{\mathbf{Q}})$ is compact and convex by (Benzi, 2016, Section 4, Items 7 and 12).

□

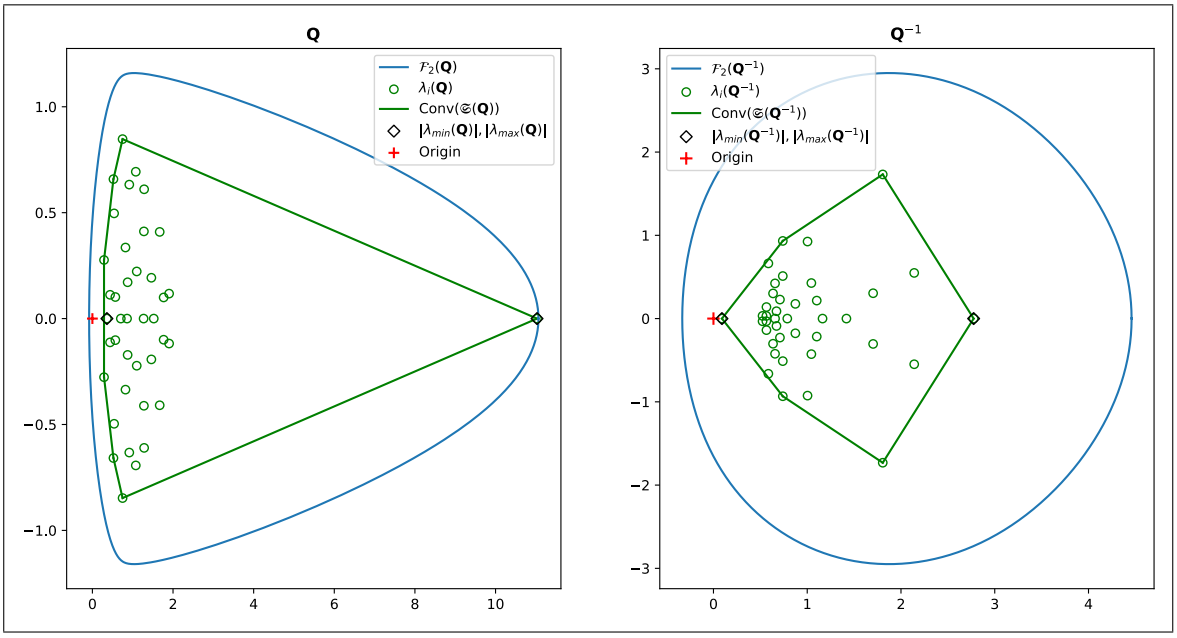


FIGURE 2.2. 2-FoV boundary (blue line), eigenvalues (green circles), convex hull for eigenvalues (green line) and $|\lambda_{\min}|, |\lambda_{\max}|$ (black diamonds) for a matrix $\mathbf{Q} := \mathbf{I} + 0.5\mathbf{E} \in \mathbb{R}^{40\times 40}$ (left) and its inverse $\mathbf{Q}^{-1}$ (right). $\mathbf{E}$ is a random matrix with $\mathbf{E}_{i,j}$ uniformly distributed random numbers in $[0, 1]$ for $0 \leq i, j \leq 40$. Remark that $0 \neq \text{Conv}(\mathfrak{S}(\mathbf{Q}))$ (resp. $0 \neq \text{Conv}(\mathfrak{S}(\mathbf{Q}^{-1}))$) while $0 \in \mathcal{F}_2(\mathbf{Q})$ (resp. $0 \in \mathcal{F}_2(\mathbf{Q}^{-1})$).

Figure 2.2 illustrates the above definitions for a random matrix. Remark that: (i) $\mathbf{Q}$ is invertible, as $|\lambda_{\min}(\mathbf{Q})| > 0$; (ii) $\mathcal{F}_2(\mathbf{Q}) \not\subset \text{Conv}(\mathfrak{S}(\mathbf{Q}))$ and $\mathcal{F}_2(\mathbf{Q}^{-1}) \not\subset \text{Conv}(\mathfrak{S}(\mathbf{Q}^{-1}))$; (iii) $\text{Conv}(\mathfrak{S}(\mathbf{Q}))$ and $\text{Conv}(\mathfrak{S}(\mathbf{Q}^{-1}))$ are bounded away from the origin, whereas $0 \in \mathcal{F}_2(\mathbf{Q})$ and $0 \in \mathcal{F}_2(\mathbf{Q}^{-1})$. Moreover, one has $\kappa_S(\mathbf{Q}) = 30.6$ while $\kappa_2(\mathbf{Q}) = 58.3$, evidencing the non-normality of $\mathbf{Q}$.

2.5.2. General linear convergence estimates for GMRES(m)

Following (Graham et al., 2017, Chapter 5), let us recall the application of the weighted (resp. Euclidean) GMRES to a linear system $\mathbf{Qx} = \mathbf{d}$ in $\mathbb{C}^N$, where $\mathbf{Q} \in \mathbb{C}^{N\times N}$ is a complex nonsingular matrix. For an initial guess $\mathbf{x}_0 \neq \mathbf{x}$, we introduce the residual $\mathbf{r}_0 = \mathbf{d} - \mathbf{Qx}_0$ such that $\mathbf{r}_0 \neq 0$ as well as Krylov spaces

$$\mathcal{K}^k(\mathbf{Q}, \mathbf{r}_0) := \text{span}\{\mathbf{Q}^j\mathbf{r}_0 : j = 0, \ldots, k-1\}, \quad 1 \leq k \leq N. \tag{2.51}$$

For any step $1 \leq k \leq N$, we define $\mathbf{x}_k$ and $\tilde{\mathbf{x}}_k$ to be the unique elements of $\mathcal{K}^k(\mathbf{Q}, \mathbf{r}_0)$ satisfying the minimal residual property under the energy and Euclidean norms:

$$\begin{aligned} \|\mathbf{r}_k\|_H &:= \|\mathbf{d} - \mathbf{Q}\mathbf{x}_k\|_H = \min_{\mathbf{x} \in \mathcal{K}^k(\mathbf{Q}, \mathbf{r}_0)} \|\mathbf{d} - \mathbf{Q}\mathbf{x}\|_H, \\ \|\tilde{\mathbf{r}}_k\|_2 &:= \|\mathbf{d} - \mathbf{Q}\tilde{\mathbf{x}}_k\|_2 = \min_{\mathbf{x} \in \mathcal{K}^k(\mathbf{Q}, \mathbf{r}_0)} \|\mathbf{d} - \mathbf{Q}\mathbf{x}\|_2, \end{aligned} \tag{2.52}$$

respectively. We refer to either weighted or Euclidean GMRES collectively as GMRES. We add an (m) superscript to signal restarted GMRES(m), for any natural number $1 \leq m \leq N$. Notice that GMRES and GMRES(m) coincide up to iteration m.

Lemma 2.4 (Weighted GMRES(m): Linear bounds)**.** *Let* $\mathbf{Q} \in \mathbb{C}^{N \times N}$*, with* $0 \neq \mathcal{F}_H(\mathbf{Q})$ *in* (2.48) *and set* $1 \leq m \leq N$*. Then, the* k*-th residual of weighted GMRES*(m) *for* $1 \leq k \leq N$ *satisfies:*

$$\frac{\|\mathbf{r}_k\|_H}{\|\mathbf{r}_0\|_H} \leq \left(1 - \mathcal{V}_H(\mathbf{Q})\mathcal{V}_H\left(\mathbf{Q}^{-1}\right)\right)^{\frac{k}{2}}. \tag{2.53}$$

PROOF. We first remark that Lemma 2.4 for the Euclidean GMRES, i.e. for $\mathbf{H} = \mathbf{I}$, is proved in (Liesen & Tichý, 2012). Thus, we focus on the extension to weighted GMRES. Following (Graham et al., 2017, Theorem 5.1), we set $\widehat{\mathbf{Q}} := \mathbf{H}^{\frac{1}{2}}\mathbf{Q}\mathbf{H}^{-\frac{1}{2}}$, $\widehat{\mathbf{d}} := \mathbf{H}^{\frac{1}{2}}\mathbf{d}$, $\widehat{\mathbf{x}} := \mathbf{H}^{\frac{1}{2}}\mathbf{x}$ and $\widehat{\mathbf{r}}_0 := \mathbf{H}^{\frac{1}{2}}\mathbf{r}_0$. Application of the Euclidean GMRES to $\widehat{\mathbf{Q}}\widehat{\mathbf{x}} = \widehat{\mathbf{d}}$ yields

$$\frac{\|\widehat{\mathbf{r}}_k\|_2}{\|\widehat{\mathbf{r}}_0\|_2} \leq \left(1 - \mathcal{V}_2(\widehat{\mathbf{Q}})\mathcal{V}_2(\widehat{\mathbf{Q}}^{-1})\right)^{\frac{k}{2}}. \tag{2.54}$$

By Lemma 2.3, we obtain that $\mathcal{F}_H(\mathbf{Q}) = \mathcal{F}_2(\widehat{\mathbf{Q}})$ and $\mathcal{F}_H(\mathbf{Q}^{-1}) = \mathcal{F}_2(\mathbf{H}^{1/2}\mathbf{Q}^{-1}\mathbf{H}^{-1/2}) = \mathcal{F}_2(\widehat{\mathbf{Q}}^{-1})$. Consequently, by (2.54) we derive the final bound for the weighted GMRES

$$\frac{\|\mathbf{r}_k\|_H}{\|\mathbf{r}_0\|_H} \leq \left(1 - \mathcal{V}_H(\mathbf{Q})\mathcal{V}_H(\mathbf{Q}^{-1})\right)^{\frac{k}{2}} = \rho^k, \tag{2.55}$$

with $\rho := \left(1 - \mathcal{V}_H(\mathbf{Q})\mathcal{V}_H(\mathbf{Q}^{-1})\right)^{\frac{1}{2}}$ and $\rho < 1$ by (Liesen & Tichý, 2012, Section 1). Finally, we remark that ρ above does not depend on k: it provides a one-step bound. Therefore, we set a restart $1 \leq m \leq N$ and define $k =: im + t$ with $0 \leq t < m$ and $0 \leq i \leq \lfloor \frac{N}{m} \rfloor$. By application of (2.55), there holds that:

$$\|\mathbf{r}_{im+t}\|_H \leq \rho^t \|\mathbf{r}_{im}\|_H, \tag{2.56}$$

and thus, $\|\mathbf{r}_k\|_H \leq \rho^k \|\mathbf{r}_0\|_H$, leading to the expected result for GMRES(m). □

REMARK 2.8. *Lemma 2.4 provides a linear convergence bound for weighted GMRES*(m) *with respect to* $\mathcal{V}_H(\mathbf{Q})$ *and* $\mathcal{V}_H(\mathbf{Q}^{-1})$ *and constitutes a sharper version of the classic result for weighted GMRES in (Graham et al., 2017; Sarkis & Szyld, 2007). Indeed, for* $\mathbf{Q}$ *in Lemma 2.4 there holds that (Liesen & Tichý, 2012, Section 1):*

$$1 - \mathcal{V}_H(\mathbf{Q})\mathcal{V}_H(\mathbf{Q}^{-1}) \leq 1 - \frac{\mathcal{V}_H(\mathbf{Q})^2}{\|\mathbf{Q}\|_H^2} < 1.$$

2.5.3. Discrete $\langle X_h \rangle$- and (X_h)-coercivity

For the ensuing GMRES analysis of our preconditioned problem (CA), we need precise definitions for coercivity that relate to the BNB condition for Hilbert spaces. We introduce the notion of (discrete) $\langle X_h \rangle$-coercivity, with angle brackets referring to the dual pairing in (2.5) (refer to (Chandler-Wilde, Graham, Langdon, & Spence, 2012, Section 5)).

Definition 2.2 ($\langle X_h \rangle$-coercivity)**.** *Consider* $\mathsf{A} : X \to X'$ *as in the BG case with* X *being a Hilbert space. For* $h > 0$ *given,* A *is said to be* $\langle X_h \rangle$*-coercive if there exists* α_{A} *such that*

$$0 < \alpha_{\mathsf{A}} \leq \frac{|\mathsf{a}(u_h, u_h)|}{\|u_h\|_X^2} = \frac{|\langle \mathsf{A}u_h, u_h \rangle_{X' \times X}|}{\|u_h\|_X^2} \quad \forall\, u_h \in X_h \setminus \{\mathbf{0}\}. \tag{2.57}$$

Thus, discrete $\langle X_h \rangle$*-ellipticity* refers to self-adjoint operators satisfying the $\langle X_h \rangle$-coercivity condition. These definitions extend naturally to continuous $\langle X \rangle$-coercivity and -ellipticity.

REMARK 2.9 (BNB condition and $\langle X_h \rangle$-coercivity). *As pointed out for* $\langle X \rangle$*-coercivity in (Ern & Guermond, 2013, Lemma 2.8) ,* $\langle X_h \rangle$*-coercivity for* A *in Definition 2.2 provides a BNB constant* $\gamma_{\mathsf{A}} = \alpha_{\mathsf{A}}$*, since for any* $u_h \in X_h \setminus \{\mathbf{0}\}$*, it holds that*

$$0 < \alpha_{\mathsf{A}} \|u_h\|_X \leq \frac{|\mathsf{a}(u_h, u_h)|}{\|u_h\|_X} \leq \sup_{v_h \in X_h \setminus \{\mathbf{0}\}} \frac{|\mathsf{a}(u_h, v_h)|}{\|v_h\|_X}. \tag{2.58}$$

REMARK 2.10. $\langle X_h \rangle$*-coercivity is a common property for BG methods in Hilbert spaces. For example, under suitable assumptions on the discretization scheme, operators with a Gårding inequality on* X*—of the form* $\mathsf{A} = \mathsf{A}_0 + \mathsf{K} : X \to X'$ *with* A_0 $\langle X \rangle$*-coercive*

and K *compact (Sauter & Schwab, 2010, Section 2.1)— admit a* $h_0 > 0$ *such that* A *is* $\langle X_h \rangle$*-coercive for all* $0 < h \leq h_0$ *(Sauter & Schwab, 2010, Section 4.2.3).*

Similarly to Definition 2.2, we introduce the discrete (X_h)-coercivity, the difference being the use of inner products.

Definition 2.3 ((X_h)-coercivity)**.** *Consider* $\mathsf{A} : X \to X$ *for the PGE case with* $X =: H$ *being a Hilbert space.* A *is said to be* (X_h)*-coercive if, for any* $u_h \in X_h \setminus \{\mathbf{0}\}$*, there exists* $\alpha_{\mathsf{A}} > 0$ *such that*

$$\alpha_{\mathsf{A}} \leq \frac{|(\mathsf{A}u_h, u_h)_H|}{(u_h, u_h)_H}, \tag{2.59}$$

or equivalently

$$\alpha_{\mathsf{A}} \leq \inf_{u_h \in X_h \setminus \{\mathbf{0}\}} \frac{|(\mathsf{A}u_h, u_h)_H|}{(u_h, u_h)_H} = \mathcal{V}_H(\mathsf{A}_h). \tag{2.60}$$

Definition 2.3 via (2.60) shows the strong connection between (X_h)-coercivity and discrete $\mathcal{V}_H$. Furthermore, under the OP setting, the discrete and matrix H-FoVs coincide.

Lemma 2.5. *Consider* (CA) *with* $X =: H$ *a Hilbert space with inner product* $(\cdot, \cdot)_H$*. There holds that*

(i) $\mathcal{F}_H(\mathsf{P}_h\mathsf{A}_h) = \mathcal{F}_H(\mathbf{PA})$ *and* $\mathcal{F}_H((\mathsf{P}_h\mathsf{A}_h)^{-1}) = \mathcal{F}_H((\mathbf{PA})^{-1})$*;*
(ii) $\mathcal{V}_H(\mathsf{P}_h\mathsf{A}_h) = \mathcal{V}_H(\mathbf{PA})$ *and* $\mathcal{V}_H((\mathsf{P}_h\mathsf{A}_h)^{-1}) = \mathcal{V}_H((\mathbf{PA})^{-1})$.

PROOF. Following (2.35), $q_h := \mathsf{P}_h\mathsf{A}_h u_h \in X_h$ for any $u_h \in X_h$ has basis expansion $\mathbf{q} = \mathbf{PAu}$. Therefore, there holds by (2.47) that for any $u_h \in X_h$:

$$(\mathsf{P}_h\mathsf{A}_h u_h, u_h)_H = (\mathbf{PAu}, \mathbf{u})_H \quad \text{and} \quad (u_h, u_h)_H = (\mathbf{u}, \mathbf{u})_H, \tag{2.61}$$

yielding $\mathcal{F}_H(\mathsf{P}_h\mathsf{A}_h) = \mathcal{F}_H(\mathbf{PA})$, and thus the expected result for $\mathsf{P}_h\mathsf{A}_h$. Similarly, one deduces the same result for the inverse operator since for any $q_h \in X_h$, $(\mathsf{P}_h\mathsf{A}_h)^{-1}q_h$ has basis expansion $(\mathbf{PA})^{-1}\mathbf{q}$. □

2.5.4. Linear convergence estimates for GMRES(m) applied to (CA)

Following Section 2.5.2, application of the weighted (resp. Euclidean) preconditioned GMRES(m), for $1 \leq m \leq N$, to $(\mathrm{CA})_{\mu,\nu}$ and initial guess $\mathbf{x}_0 \neq \mathbf{u}_\nu$, has the iterates $\mathbf{x}_k$ (resp. $\tilde{\mathbf{x}}_k$) for any step $1 \leq k \leq N$ with minimal residual properties:

$$\begin{aligned} \|\mathbf{P}_\mu \mathbf{r}_k\|_H &:= \|\mathbf{P}_\mu \mathbf{b}_\nu - \mathbf{P}_\mu \mathbf{A}_\nu \mathbf{x}_k\|_H = \min_{\mathbf{x} \in \mathcal{K}^k(\mathbf{P}_\mu \mathbf{A}_\nu, \mathbf{r}_0)} \|\mathbf{P}_\mu \mathbf{b}_\nu - \mathbf{P}_\mu \mathbf{A}_\nu \mathbf{x}\|_H, \\ \|\mathbf{P}_\mu \tilde{\mathbf{r}}_k\|_2 &:= \|\mathbf{P}_\mu \mathbf{b}_\nu - \mathbf{P}_\mu \mathbf{A}_\nu \tilde{\mathbf{x}}_k\|_2 = \min_{\mathbf{x} \in \mathcal{K}^k(\mathbf{P}_\mu \mathbf{A}_\nu, \mathbf{r}_0)} \|\mathbf{P}_\mu \mathbf{b}_\nu - \mathbf{P}_\mu \mathbf{A}_\nu \mathbf{x}\|_2, \end{aligned} \tag{2.62}$$

with $\mathcal{K}^k$ introduced in (2.51) and obvious construction for (CA). By (2.62), the minimal residuals satisfy

$$\|\mathbf{P}_\mu \tilde{\mathbf{r}}_k\|_2 \leq \|\mathbf{P}_\mu \mathbf{r}_k\|_2 \quad \text{and} \quad \|\mathbf{P}_\mu \mathbf{r}_k\|_H \leq \|\mathbf{P}_\mu \tilde{\mathbf{r}}_k\|_H. \tag{2.63}$$

We set convergence rates for the weighted (resp. Euclidean) preconditioned GMRES(m):

$$\Theta_k^{(m)} := \left(\frac{\|\mathbf{P}_\mu \mathbf{r}_k\|_H}{\|\mathbf{P}_\mu \mathbf{r}_0\|_H} \right)^{\frac{1}{k}} \quad \text{and} \quad \widetilde{\Theta}_k^{(m)} := \left(\frac{\|\mathbf{P}_\mu \tilde{\mathbf{r}}_k\|_2}{\|\mathbf{P}_\mu \mathbf{r}_0\|_2} \right)^{\frac{1}{k}}. \tag{2.64}$$

Finally, we define convergence rates for non-restarted weighted (resp. Euclidean) preconditioned GMRES $\Theta_k := \Theta_k^{(N)}$ (resp. $\widetilde{\Theta}_k := \widetilde{\Theta}_k^{(N)}$).

The bounds in Theorem 2.1 and Theorem 2.2 are related to the spectral radius, and rely on the continuity and discrete inf-sup constants: they do not supply information on the eigenvalue or FoV distributions, as pointed out in Figure 2.2, which are required to derive convergence results for iterative solvers to (CA)or $(\mathrm{CA})_{\mu,\nu}$. Thus, more specific conditions are required. For instance, for (CA) we enforce the following (X_h)-coercivity condition for $\mathsf{P}_h\mathsf{A}_h$.

Assumption 3 (X_h)-coercivity for (CA)**.** For problem (CA) with $X := H$ being a Hilbert space with inner product $(\cdot,\cdot)_H$, we assume that $\mathsf{P}_h\mathsf{A}_h$ and its inverse are (X_h)-coercive satisfying

$$\frac{\gamma_{\mathsf{C}}\gamma_{\mathsf{A}}}{\|\mathsf{m}\|\|\mathsf{n}\|} \leq \mathcal{V}_H(\mathsf{P}_h\mathsf{A}_h) \quad \text{and} \quad \frac{\gamma_{\mathsf{M}}\gamma_{\mathsf{N}}}{\|\mathsf{c}\|\|\mathsf{a}\|} \leq \mathcal{V}_H((\mathsf{P}_h\mathsf{A}_h)^{-1}). \tag{2.65}$$

REMARK 2.11. *The (X_h)-coercivity constants in Assumption 3 emerge naturally, as they are related to the BNB constants for both $\mathsf{P}_h\mathsf{A}_h$ and its inverse (cf. proof of Theorem 2.1). Alternatively, one can write Assumption 3 as:*

$$0 < \Gamma_0 \leq \mathcal{V}_H(\mathsf{P}_h\mathsf{A}_h) \quad \textit{and} \quad 0 < \Gamma_1 \leq \mathcal{V}_H((\mathsf{P}_h\mathsf{A}_h)^{-1}), \tag{2.66}$$

with Γ_0, Γ_1 constants depending on the discrete inf-sup and continuity constants for the induced operators and eventually for $(\cdot,\cdot)_H$ in (2.47)*.*

We are ready to state the linear convergence result for GMRES(m) for (CA) for the Hilbertian case.

Theorem 2.3 (GMRES(m): Linear convergence estimates for (CA))**.** *Consider* (CA) *with $X =: H$ Hilbert and $(\cdot,\cdot)_H$ such that Assumption 3 holds. Then, GMRES(m) for $1 \leq k, m \leq N$ leads to*

$$\Theta_k^{(m)} \leq \left(1 - \frac{1}{\mathrm{K}_\star}\right)^{\frac{1}{2}} \quad \text{and} \quad \widetilde{\Theta}_k^{(m)} \leq \mathrm{K}_{\Lambda_h}\left(1 - \frac{1}{\mathrm{K}_\star}\right)^{\frac{1}{2}}, \tag{2.67}$$

with $\mathrm{K}_\star$ as defined in (2.36) *and K_{Λ_h} in* (2.16)*.*

PROOF. By combining Assumption 3, Lemma 2.5 and definition of $\mathrm{K}_\star$, there holds that

$$\mathcal{V}_H(\mathbf{PA})\mathcal{V}_H((\mathbf{PA})^{-1}) \geq \frac{1}{\mathrm{K}_\star} \tag{2.68}$$

and thus

$$1 - \mathcal{V}_H(\mathbf{PA})\mathcal{V}_H((\mathbf{PA})^{-1}) \leq 1 - \frac{1}{\mathrm{K}_\star}$$

Application of Lemma 2.4 to the preconditioned system with residuals (2.62) provides the first bound in (2.67), namely

$$\frac{\|\mathbf{Pr}_k\|_H}{\|\mathbf{Pr}_0\|_H} \leq \left(1 - \frac{1}{\mathrm{K}_\star}\right)^{\frac{k}{2}}, \quad 1 \leq k \leq N. \tag{2.69}$$

Next, we follow the steps in (Sarkis & Szyld, 2007, Section 4) to arrive at the second bound in (2.67). First, the minimal residual property (2.63) yields

$$\|\mathbf{P}\tilde{\mathbf{r}}_k\|_2 \leq \|\mathbf{P}\mathbf{r}_k\|_2. \tag{2.70}$$

By virtue of the synthesis operator in (2.14), one has

$$\|\mathbf{P}\mathbf{r}_k\|_2 \leq \frac{1}{\gamma_{\Lambda_h}}\|\mathbf{P}\mathbf{r}_k\|_H \quad \text{and} \quad \|\mathbf{P}\mathbf{r}_0\|_H \leq \|\Lambda_h\|\|\mathbf{P}\mathbf{r}_0\|_2. \tag{2.71}$$

Therefore, (2.69) combined with (2.70) and (2.71) lead to the final result, as

$$\|\mathbf{P}\tilde{\mathbf{r}}_k\|_2 \leq \|\mathbf{P}\mathbf{r}_k\|_2 \leq \frac{1}{\gamma_{\Lambda_h}}\|\mathbf{P}\mathbf{r}_k\|_H \leq \frac{1}{\gamma_{\Lambda_h}}\left(1-\frac{1}{\mathrm{K}_\star}\right)^{\frac{k}{2}}\|\mathbf{P}\mathbf{r}_0\|_H \leq \mathrm{K}_{\Lambda_h}\left(1-\frac{1}{\mathrm{K}_\star}\right)^{\frac{k}{2}}\|\mathbf{P}\mathbf{r}_0\|_2,$$

as stated. □

REMARK 2.12. *This result provides extensive convergence bounds for GMRES*(m)*. It will guarantee* h*-independent convergence for weighted GMRES*(m) *for* ((CA)) *in the Hilbert setting (cf. Corollary 2.3). Also, the synthesis operator enters as an offset factor in* $\widetilde{\Theta}_k^{(m)} = \mathrm{K}_{\Lambda_h}\,\rho$ *with* $\rho := (1-1/\mathrm{K}_\star)^{1/2} < 1$*. One should observe that* $\widetilde{\Theta}_k^{(m)}$ *could be larger than 1 for* $\mathrm{K}_{\Lambda_h} > 1$*, an impractical bound for Euclidean GMRES*(m)*. The latter supports theoretically the use of weighted GMRES*(m) *as a solver (Feischl et al., 2017).*

To illustrate the application of the above results, we provide a case of interest where Assumption 3 is satisfied. Therein, notice the extra K_A-term in (2.72) below, justifying Remark 2.11.

Corollary 2.1 (Preconditioner-induced norm (Feischl et al., 2017; Starke, 1997; Kirby, 2010))**.** *Consider* ((CA)) *for OP-BG for Hilbert spaces* $X =: H$ *and* V*,* A *being* $\langle X_h\rangle$*-coercive, and* C *being* $\langle V_h\rangle$*-elliptic, with* $\gamma_\mathsf{A} := \alpha_\mathsf{A}$ *and* $\gamma_\mathsf{C} := \alpha_\mathsf{C}$*. Then,* P^{-1} *is Hermitian and yields an inner product on* X_h*, denoted by* $(\cdot,\cdot)_{P^{-1}}$*, and*

$$\frac{\gamma_\mathsf{C}\gamma_\mathsf{A}}{\|\mathsf{m}\|^2} \leq \mathcal{V}_{P^{-1}}(\mathsf{P}_h\mathsf{A}_h) \quad \textit{and} \quad \frac{\gamma_\mathsf{M}^2}{\|\mathsf{c}\|\|\mathsf{a}\|}\,\mathrm{K}_\mathsf{A} \leq \mathcal{V}_{P^{-1}}((\mathsf{P}_h\mathsf{A}_h)^{-1}). \tag{2.72}$$

PROOF. First, notice that since C is $\langle V_h \rangle$-elliptic, $\mathbf{C}$ and $\mathbf{C}^{-1}$ are Hermitian positive definite. Therefore, we deduce that $\mathbf{P} = \mathbf{M}^{-1}\mathbf{C}\mathbf{M}^{-H} = \mathbf{P}^H$ and $\mathbf{P}^{-1} = \mathbf{P}^{-H}$ are Hermitian positive definite. For any $\mathbf{u}, \mathbf{v} \in \mathbb{C}^N \setminus \{\mathbf{0}\}$ and $\mathbf{w} := \mathbf{P}^{-1}\mathbf{v}$, one has

$$\mathcal{V}_{P^{-1}}(\mathsf{P}_h\mathsf{A}_h) = \inf_{\mathbf{u}\in\mathbb{C}^N\setminus\{\mathbf{0}\}} \frac{|(\mathbf{P}\mathbf{A}\mathbf{u},\mathbf{u})_{P^{-1}}|}{(\mathbf{u},\mathbf{u})_{P^{-1}}} = \inf_{\mathbf{u}\in\mathbb{C}^N\setminus\{\mathbf{0}\}} \frac{|(\mathbf{A}\mathbf{u},\mathbf{u})_2|}{|(\mathbf{P}^{-1}\mathbf{u},\mathbf{u})_2|}$$

and

$$\mathcal{V}_{P^{-1}}((\mathsf{P}_h\mathsf{A}_h)^{-1}) = \inf_{\mathbf{v}\in\mathbb{C}^N\setminus\{\mathbf{0}\}} \frac{|(\mathbf{A}^{-1}\mathbf{P}^{-1}\mathbf{v},\mathbf{v})_{P^{-1}}|}{(\mathbf{v},\mathbf{v})_{P^{-1}}} = \inf_{\mathbf{w}\in\mathbb{C}^N\setminus\{\mathbf{0}\}} \frac{|(\mathbf{A}^{-1}\mathbf{w},\mathbf{w})_2|}{|(\mathbf{P}\mathbf{w},\mathbf{w})_2|}.$$

Next, using Equations 2.61 and 2.62 in Kirby (Kirby, 2010), we deduce that:

$$\gamma_{\mathsf{A}}\|u_h\|_X^2 \le |(\mathbf{A}\mathbf{u},\mathbf{u})_2| \le \|\mathsf{a}\|\|u_h\|_X^2 \quad \text{and} \quad \frac{\gamma_{\mathsf{A}}}{\|\mathsf{a}\|}\|w_h\|_{X_h'}^2 \le |(\mathbf{A}^{-1}\mathbf{w},\mathbf{w})_2| \le \|\mathsf{a}\|\|w_h\|_{X_h'}^2, \tag{2.73}$$

while for the preconditioner $\mathbf{P}$ one has by (Steinbach, 2007, Section 13.2) that

$$\frac{\gamma_{\mathsf{C}}}{\|\mathsf{m}\|^2}\|w_h\|_{X_h'}^2 \le |(\mathbf{P}\mathbf{w},\mathbf{w})_2| \le \frac{\|\mathsf{c}\|}{\gamma_{\mathsf{M}}^2}\|w_h\|_{X_h'}^2 \tag{2.74}$$

and

$$\frac{\gamma_{\mathsf{M}}^2}{\|\mathsf{c}\|}\|u_h\|_X^2 \le |(\mathbf{P}^{-1}\mathbf{u},\mathbf{u})_2| \le \frac{\|\mathsf{m}\|^2}{\gamma_{\mathsf{C}}}\|u_h\|_X^2. \tag{2.75}$$

Therefore,

$$\frac{\gamma_{\mathsf{A}}\gamma_{\mathsf{C}}}{\|\mathsf{m}\|^2} \le \frac{|(\mathbf{A}\mathbf{u},\mathbf{u})_2|}{|(\mathbf{P}^{-1}\mathbf{u},\mathbf{u})_2|} \quad \text{and} \quad \frac{\gamma_{\mathsf{A}}\gamma_{\mathsf{M}}^2}{\|\mathsf{a}\|^2\|\mathsf{c}\|} \le \frac{|(\mathbf{A}^{-1}\mathbf{w},\mathbf{w})_2|}{|(\mathbf{P}\mathbf{w},\mathbf{w})_2|}, \tag{2.76}$$

finalizing the proof. □

Corollary 2.2. *Consider* (CA) *for OP-BG for Hilbert spaces* $X =: H$ *and* V, A *being* $\langle X_h \rangle$*-coercive, and* C *being* $\langle V_h \rangle$*-elliptic. Then, GMRES*(m) *for* $1 \le k, m \le N$ *leads to*

$$\Theta_k^{(m)} \le \left(1 - \frac{1}{\mathrm{K}_\star\,\mathrm{K}_{\mathsf{A}}}\right) \quad \textit{and} \quad \widetilde{\Theta}_k^{(m)} \le \mathrm{K}_{\Lambda_h}\left(1 - \frac{1}{\mathrm{K}_\star\,\mathrm{K}_{\mathsf{A}}}\right)^{\frac{1}{2}}, \tag{2.77}$$

with $\mathrm{K}_\star$ *as defined in* (2.36), K_{Λ_h} *in* (2.16) *and* K_{A} *in* (2.12).

REMARK 2.13. *The* K_{A}*-term in Corollary 2.1 and Corollary 2.2 is removed if* A *is* $\langle X_h \rangle$*-elliptic, since* A^{-1} *is* $\langle X_h' \rangle$*-elliptic with constant* $1/\|\mathsf{a}\|$ *(Steinbach, 2007, Section 13.2).*

2.5.5. Linear convergence estimates for GMRES(m) applied to $(\!(\text{CA})\!)_{\mu,\nu}$

As in Section 2.5.4, we give counterparts to Assumption 3 and Theorem 2.3 for the bi-parametric preconditioned problem $(\!(\text{CA})\!)_{\mu,\nu}$.

Assumption 4 (X_h)-coercivity for $(\!(\text{CA})\!)_{\mu,\nu}$**.** For $(\!(\text{CA})\!)_{\mu,\nu}$ with $X := H$ being a Hilbert space with inner product $(\cdot,\cdot)_H$, assume that there holds that

$$\frac{\gamma_{\mathsf{C}_\mu}\gamma_{\mathsf{A}_\nu}}{\|\mathsf{m}\|\|\mathsf{n}\|} \leq \mathcal{V}_H(\mathsf{P}_{h,\mu}\mathsf{A}_{h,\nu}) \quad \text{and} \quad \frac{\gamma_{\mathsf{M}}\gamma_{\mathsf{N}}}{\|\mathsf{c}_\mu\|\|\mathsf{a}_\nu\|} \leq \mathcal{V}_H((\mathsf{P}_{h,\mu}\mathsf{A}_{h,\nu})^{-1}). \tag{2.78}$$

Remark 2.11 remains valid for Assumption 4. With this, we can extend Theorem 2.3 to $(\!(\text{CA})\!)_{\mu,\nu}$.

Theorem 2.4 (GMRES(m): Linear convergence estimates for $(\!(\text{CA})\!)_{\mu,\nu}$**).** *Consider* $(\!(\text{CA})\!)_{\mu,\nu}$ *along with Assumption 4. Then, the residuals for GMRES*(m) *for* $1 \leq k, m \leq N$ *are bounded as*

$$\Theta_k^{(m)} \leq \left(1 - \frac{1}{\mathrm{K}_{\star,\mu,\nu}}\right)^{\frac{1}{2}} \quad \text{and} \quad \widetilde{\Theta}_k^{(m)} \leq \mathrm{K}_{\Lambda_h}\left(1 - \frac{1}{\mathrm{K}_{\star,\mu,\nu}}\right)^{\frac{1}{2}}, \tag{2.79}$$

with $\mathrm{K}_{\star,\mu,\nu}$ *and* K_{Λ_h} *defined in* (2.44) *and* (2.16)*, respectively.*

PROOF. The result follows by direct application of Theorem 2.3 to $(\!(\text{CA})\!)_{\mu,\nu}$. □

REMARK 2.14. *The above result gives a controlled convergence rate for GMRES*(m) *with respect to bi-parametric* (μ,ν)*-perturbations. As in the discussion ensuing Theorem 2.2 and in order to illustrate its practical implications, assume that the best approximation error in Lemma 2.2 converges at a rate* $\mathcal{O}(h^r)$*,* $r > 0$*, and* $\nu = \mathcal{O}(h^r)$*. Therefore, provided that* $\mu = \mathcal{O}(1)$ *guarantees a bounded* $\mathrm{K}_{\star,\mu,\nu}$*, the bounds in* (2.79) *ensure linear convergence for the weighted GMRES*(m) *(resp. Euclidean GMRES*(m)*, for* $\mathrm{K}_{\Lambda_h} < 1$*).*

2.5.6. Compact and Carleman class operators

So far, we have focused on the linear convergence rates for GMRES(m). Yet, it is know that in many situations the bound in Lemma 2.4 "*may significantly overestimate the GMRES residual norms*" (Liesen & Tichỳ, 2012). To better understand this, we aim

to improve bounds for the case of second-kind Fredholm operators, which are known to display super-linear convergence results for GMRES, i.e. the radius of convergence tends to zero as $k \to \infty$. To this end, we introduce the concept of Carleman class operators.

Again, assuming H to be a separable Hilbert space, we introduce $\mathcal{C}(H) \equiv \mathcal{C}(H;H)$ the space of compact operators on H. Given $\mathsf{T} \in \mathcal{L}(H;H)$, we denote the ordered singular values of T as $\sigma_j(\mathsf{T}) := \{\inf \|\mathsf{T} - \mathsf{T}_i\|_H : \mathsf{T}_i : H \to H,\ \text{rank}\mathsf{T}_i < j\}$. For any $k \geq 1$, the kth partial arithmetic mean for the singular values reads

$$\overline{\sigma}_k(\mathsf{K}) := \frac{1}{k}\sum_{j=1}^{k} \sigma_j(\mathsf{K}). \tag{2.80}$$

For $p > 0$, a compact operator $\mathsf{K} \in \mathcal{C}(H)$ is said to belong to the *Carleman class* $\mathcal{C}^p(H)$ (Dunford & Schwartz, 1963, Section XI.9) if it holds that

$$|||\mathsf{K}|||_p = \|\sigma(\mathsf{K})\|_p := \left(\sum_{i=1}^{\infty} \sigma_i(\mathsf{K})^p\right)^{1/p} < \infty. \tag{2.81}$$

Next, we identify $\mathcal{C}^0(H) \equiv \mathcal{C}(H)$ and for $p \geq 0$, we say that Q is a *p-class Fredholm operator of the second-kind*, $\mathsf{Q} \in \mathcal{FC}^p(H)$ if and only if $\mathsf{Q} - \mathsf{I} \in \mathcal{C}^p(H)$. Consequently, for $H =: X$ a separable Hilbert space, $p \geq 0$ and $\nu, \mu \in [0,1)$, we define the following problems:

$$(\!(\mathrm{A})\!)^p : \quad (\!(\mathrm{A})\!) \quad \text{for PGE (i.e. } \mathsf{A} : H \to H) \text{ with } \mathsf{A} \in \mathcal{FC}^p(H) \quad \text{and} \quad \mathsf{N} := \mathsf{I}, \tag{2.82}$$

and

$$(\!(\mathrm{CA})\!)^p_{\mu,\nu} : \quad (\!(\mathrm{CA})\!)_{\mu,\nu} \quad \text{with} \quad \mathsf{C}_\mu \mathsf{N}^{-1} \mathsf{A}_\nu \in \mathcal{FC}^p(H), \quad \text{and} \quad \mathsf{M} := \mathsf{I}, \tag{2.83}$$

whose diagram representation is

$$(\!(\mathrm{CA})\!)^p_{\mu,\nu} : \quad \begin{array}{ccc} H & \xrightarrow{\mathsf{A}_\nu} & Y' \\ {\scriptstyle \mathsf{I}^{-1}}\uparrow & & \downarrow{\scriptstyle \mathsf{N}^{-1}} \\ H & \xleftarrow[\mathsf{C}_\mu]{} & V \end{array}.$$

Finally, for $(\!|\mathsf{A}|\!)^p$ and $(\!|\mathsf{CA}|\!)^p_{\mu,\nu}$, the corresponding compact terms $\mathsf{K} := \mathsf{A} - \mathsf{I}$ and $\mathsf{K}_{\mu,\nu} := \mathsf{C}_\mu \mathsf{N}^{-1} \mathsf{A}_\nu - \mathsf{I}$) have discrete counterparts $\mathsf{K}_h := \mathsf{A}_h - \mathsf{I}_h$ and $\mathsf{K}_{h,\mu,\nu} := \mathsf{C}_h \mathsf{N}_h^{-1} \mathsf{A}_h - \mathsf{I}_h$ with Galerkin matrices defined as

$$\mathbf{K} := \mathbf{A} - \mathbf{N} \quad \text{and} \quad \mathbf{K}_{\mu,\nu} := \mathbf{C}_\mu \mathbf{N}^{-1} \mathbf{A}_\nu - \mathbf{M}, \tag{2.84}$$

respectively. In the sequel, we introduce ordered (matrix) singular values with respect to the H-norm (Axelsson et al., 2018, Proposition 4.2):

$$\sigma_j^H(\mathbf{Q}) := \lambda_j(\mathbf{Q}^\star \mathbf{Q})^{1/2} = \sigma_j(\mathbf{H}^{1/2} \mathbf{Q} \mathbf{H}^{-1/2}), \tag{2.85}$$

for any $\mathbf{Q} \in \mathbb{C}^{N \times N}$ and $\mathbf{Q}^\star = \mathbf{H}^{-1} \mathbf{Q}^H \mathbf{H}$ its H-adjoint.

2.5.7. Super-linear convergence estimates for GMRES applied to $(\!|\mathsf{A}|\!)^p$

We recall the classic super-linear convergence result for weighted GMRES on a (continuous) Hilbert setting level (cf. (Moret, 1997) and (Axelsson et al., 2018, Theorem 3.1)).

PROPOSITION 2.3 (Weighted GMRES: Classic super-linear convergence estimate (Axelsson et al., 2018, Theorem 3.1)). *Let H be a Hilbert space. Set $p \geq 0$ and consider the application of weighted GMRES on $\mathsf{Q}x = f$, for a bounded and invertible operator $\mathsf{Q} \in \mathcal{FC}^p(H)$ with $f \in H$. Introduce GMRES iterates $x_0 \neq x$, and x_k, along with $r_k := \mathsf{Q}x_k - f$, for any $k \geq 1$. Then, the residuals satisfy*

$$\left(\frac{\|r_k\|_H}{\|r_0\|_H} \right)^{\frac{1}{k}} \leq \|\mathsf{Q}^{-1}\|_H \overline{\sigma}_k(\mathsf{K}),$$

wherein $\mathsf{K} := \mathsf{Q} - \mathsf{I} \in \mathcal{C}^p(H)$ and $\overline{\sigma}_k(\mathsf{K})$ defined in (2.86).

Remark that $\overline{\sigma}_k(\mathsf{K}) \to 0$ as $k \to \infty$ evidencing the super-linear convergence rate for residuals of weighted GMRES in this particular case. Furthermore, the convergence rate depends directly on the singular values of the continuous operator K. The following result shows that the above is applicable to $(\!|\mathsf{A}|\!)^p$ as well.

Theorem 2.5 (GMRES: Super-linear convergence estimates for $(\!(\mathsf{A})\!)^p$)**.** *Consider the PGE problem* $(\!(\mathsf{A})\!)^p$ *in* (2.82) *for any* $p \geq 0$*. Then, for* $1 \leq k \leq N$*, it holds that*

$$\left(\frac{\|\mathbf{r}_k\|_H}{\|\mathbf{r}_0\|_H}\right)^{\frac{1}{k}} \leq \frac{\overline{\sigma}_k(\mathsf{K})}{\gamma_{\mathsf{A}}\gamma_{\mathsf{N}}} \quad \left(\leq \frac{|||\mathsf{K}|||_p}{\gamma_{\mathsf{A}}\gamma_{\mathsf{N}}} k^{-\frac{1}{p}} \quad \textit{if} \quad p > 0\right), \tag{2.86}$$

and

$$\left(\frac{\|\tilde{\mathbf{r}}_k\|_2}{\|\mathbf{r}_0\|_2}\right)^{\frac{1}{k}} \leq \mathrm{K}_{\Lambda_h} \frac{\overline{\sigma}_k(\mathsf{K})}{\gamma_{\mathsf{A}}\gamma_{\mathsf{N}}} \quad \left(\leq \mathrm{K}_{\Lambda_h} \frac{|||\mathsf{K}|||_p}{\gamma_{\mathsf{A}}\gamma_{\mathsf{N}}} k^{-\frac{1}{p}} \quad \textit{if} \quad p > 0\right), \tag{2.87}$$

wherein $\mathsf{K} := \mathsf{A} - \mathsf{I} \in \mathcal{C}^p(H)$ *and* $\overline{\sigma}_k(\mathsf{K})$ *in* (2.86)*.*

PROOF. By hypothesis, we have that $\mathbf{N}^{-1}\mathbf{A} = \mathbf{I} + \mathbf{N}^{-1}\mathbf{K}$, with $\mathbf{K}$ such as in (2.84). Following the same steps as in Axelsson (Axelsson et al., 2018), we deduce that the following relations hold (cf. proofs of Theorem 2.1 and Lemma 2.5):

$$\|(\mathbf{N}^{-1}\mathbf{A})^{-1}\|_H \leq \frac{\|\mathsf{n}\|}{\gamma_{\mathsf{A}}} = \frac{1}{\gamma_{\mathsf{A}}},$$

since $\mathsf{N} = \mathsf{I}$. Furthermore, it holds that the singular values (Axelsson et al., 2018, Proposition 4.2)

$$\sigma_j^H(\mathbf{N}^{-1}\mathbf{K}) \leq \frac{1}{\gamma_{\mathsf{N}}}\sigma_j(\mathsf{K}_h) \leq \frac{1}{\gamma_{\mathsf{N}}}\sigma_j(\mathsf{K}).$$

Therefore, for $1 \leq k \leq N$, following (Axelsson et al., 2018) and using Proposition 2.3, we can show that

$$\frac{\|\mathbf{r}_k\|_H}{\|\mathbf{r}_0\|_H} \leq \frac{\|(\mathbf{N}^{-1}\mathbf{A})^{-1}\|_H}{k} \sum_{j=1}^{k} \sigma_j^H(\mathbf{N}^{-1}\mathbf{K}) \leq \frac{1}{\gamma_{\mathsf{A}}} \sum_{j=1}^{k} \frac{\sigma_j^H(\mathbf{N}^{-1}\mathbf{K})}{k} \leq \frac{1}{\gamma_{\mathsf{A}}\gamma_{\mathsf{N}}} \sum_{j=1}^{k} \frac{\sigma_j(\mathsf{K})}{k}.$$

Now, if $\mathsf{K} \in \mathcal{C}^p(H)$ for any $p > 0$, we follow (Winther, 1980, Theorem 2.2) and derive

$$\sum_{j=1}^{k} \frac{\sigma_j(\mathsf{K})}{k} \leq |||\mathsf{K}|||_p k^{-\frac{1}{p}},$$

providing the final estimate in energy norm.

Finally, the bounds in Euclidean norm are deduced in the same fashion as in Theorem 2.3. □

REMARK 2.15. *This result appears to be new and it justifies the positive results of employing mass matrix preconditioning, i.e.* $\mathsf{N} := \mathsf{I}$*, to transfer the super-linear convergence bounds from the continuous to the discrete level. Indeed, the choice of* $\mathsf{N} = \mathsf{I}$ *guarantees a discrete system* $\mathbf{N}^{-1}\mathbf{A} = \mathbf{I} + \mathbf{N}^{-1}\mathbf{K}$ *of the form* $\mathbf{I}$ *plus discretization of a compact operator. The latter enables the application of the classical super-linear results for GMRES given in Proposition 2.3. Notice that the bounds in* (2.86) *and* (2.87) *depend on* k *via* $\overline{\sigma}_k(\mathsf{K})$*: they are not one-step bounds, and do not generalize to GMRES*(m)*, as the relative error at iteration* k *for* $2 \leq k \leq N$ *depends on previous iterations.*

REMARK 2.16. *The super-linear convergence rate depends on the decay rate of* $\overline{\sigma}_k(\mathsf{K})$*. For example, for trace class operators* ($p = 1$)*, it holds that* $\|\mathbf{r}_k\|_H / \|\mathbf{r}_0\|_H = \mathcal{O}(k^{-1})$ *while for Hilbert-Schmidt operators* ($p = 2$)*, one observes the faster rate* $\|\mathbf{r}_k\|_H / \|\mathbf{r}_0\|_H = \mathcal{O}(k^{-2})$ *(Dunford & Schwartz, 1963, Chapter XI). Results describing the Carleman class index for pseudo-differential operators (resp. the Laplace double-layer operator) can be found in (Sobolev, 2014) (resp. (Bessoud & Krasucki, 2006; Miyanishi & Suzuki, 2015)) and will be investigated elsewhere.*

2.5.8. Super-linear convergence estimates for GMRES applied to $(\!|\mathrm{CA}|\!)^p_{\mu,\nu}$

We next show that the reasoning in Theorem 2.5 can also be applied to $(\!|\mathrm{CA}|\!)^p_{\mu,\nu}$.

Theorem 2.6 (GMRES: Super-linear convergence estimates for $(\!|\mathrm{CA}|\!)^p_{\mu,\nu}$)**.** *Consider* $(\!|\mathrm{CA}|\!)^p_{\mu,\nu}$ *in* (2.83) *for any* $p \geq 0$ *and define* $\mathsf{K}_{\mu,\nu} := \mathsf{C}_\mu \mathsf{N}^{-1} \mathsf{A}_\nu - \mathsf{I} \in \mathcal{C}^p(H)$*. Then, for weighted and Euclidean GMRES, respectively, it holds that*

$$\Theta_k \leq \frac{\|\mathsf{n}\|}{\gamma_{\mathsf{C}}\gamma_{\mathsf{A}}\gamma_{\mathsf{M}}} \frac{\overline{\sigma}_k(\mathsf{K}_{\mu,\nu})}{(1-\mu)(1-\nu)} \quad \left(\leq \frac{\|\mathsf{n}\|}{\gamma_{\mathsf{C}}\gamma_{\mathsf{A}}\gamma_{\mathsf{M}}} \frac{|||\mathsf{K}_{\mu,\nu}|||_p}{(1-\mu)(1-\nu)} k^{-\frac{1}{p}} \quad \textit{if} \quad p > 0 \right), \tag{2.88}$$

and

$$\widetilde{\Theta}_k \leq \mathrm{K}_{\Lambda_h} \frac{\|\mathsf{n}\|}{\gamma_{\mathsf{C}}\gamma_{\mathsf{A}}\gamma_{\mathsf{M}}} \frac{\overline{\sigma}_k(\mathsf{K}_{\mu,\nu})}{(1-\mu)(1-\nu)} \quad \left(\leq \mathrm{K}_{\Lambda_h} \frac{\|\mathsf{n}\|}{\gamma_{\mathsf{C}}\gamma_{\mathsf{A}}\gamma_{\mathsf{M}}} \frac{|||\mathsf{K}_{\mu,\nu}|||_p}{(1-\mu)(1-\nu)} k^{-\frac{1}{p}} \quad \textit{if} \quad p > 0 \right), \tag{2.89}$$

with Θ_k *and* $\widetilde{\Theta}_k$ *defined in* (2.64) *and* $\overline{\sigma}_k(\cdot)$ *in* (2.86).

PROOF. Consider $(\!(\mathrm{CA})\!)^{p}_{\mu,\nu}$ and follow the proof of Theorem 2.5. First, we use Proposition 2.1 to deduce that

$$\|(\mathbf{P}_\mu\mathbf{A}_\nu)^{-1}\|_H = \|\mathbf{A}_\nu^{-1}\mathbf{N}\mathbf{C}_\mu^{-1}\mathbf{M}\|_H \leq \frac{\|\mathsf{n}\|}{\gamma_{\mathsf{A}}\gamma_{\mathsf{C}}}\frac{1}{(1-\mu)(1-\nu)}.$$

Next, for $1 \leq j \leq N$, one has

$$\sigma_j^H(\mathbf{M}^{-1}\mathbf{C}_\mu\mathbf{N}^{-1}\mathbf{A}_\nu - \mathbf{I}) = \sigma_j^H(\mathbf{M}^{-1}\mathbf{K}_{\mu,\nu}) \leq \frac{1}{\gamma_{\mathsf{M}}}\sigma_j(\mathsf{K}_{\mu,\nu}),$$

with $\mathbf{K}_{\mu,\nu} = \mathbf{C}_\mu\mathbf{N}^{-1}\mathbf{A} - \mathbf{M}$ as in (2.84). Therefore, we obtain

$$\Theta_k \leq \frac{\|(\mathbf{P}_\mu\mathbf{A}_\nu)^{-1}\|_H}{k}\sum_{j=1}^{k}\sigma_j^H(\mathbf{M}^{-1}\mathbf{K}_{\mu,\nu}) \leq \frac{\|\mathsf{n}\|}{\gamma_{\mathsf{C}}\gamma_{\mathsf{A}}\gamma_{\mathsf{M}}}\frac{\overline{\sigma}_k(\mathsf{K}_{\mu,\nu})}{(1-\mu)(1-\nu)}.$$

The second bound in (2.88) and (2.89) follows by the same arguments as in the proof of Theorem 2.5. □

Theorem 2.6 describes precisely the residual convergence behavior of GMRES for $(\!(\mathrm{CA})\!)^{p}_{\mu,\nu}$ for $p \geq 0$. In particular, (2.89) shows that the Euclidean GMRES converges super-linearly, up to a K_{Λ_h}-term as observed experimentally for the electric field integral equation on screens in (Hiptmair & Urzúa-Torres, 2020).

Corollary 2.3 (h-Asymptotics)**.** *Consider* $(\!(\mathrm{CA})\!)_{\mu,\nu}$ *in* (2.79)*, for* $\mu \to 0$ *and* $\nu \to 0$ *as* $h \to 0$*. Additionally, let us suppose that (i) the finite dimensional subspaces are dense in their function space, satisfying the approximability property (Ern & Guermond, 2013, Definition 2.14); and, (ii) the forms in* $(\!(\mathrm{CA})\!)$ *have a uniform discrete inf-sup condition with respect to* h*. Then, for vanishing* h*, the following statements hold:*

- *(i)* $\|u - u_h\|_X \to 0$ *in Lemma 2.2;*
- *(ii)* $\mathrm{K}_\star$ *in Theorem 2.1, and subsequently* $\mathrm{K}_{\star,\mu,\nu}$ *Theorem 2.2 remain bounded (*h*-independence);*
- *(iii) Under Assumption 4, the residual* $\Theta_k^{(m)}$ *in* (2.79) *remains bounded (*h*-independent linear convergence);*

(iv) For $(\!(\mathrm{CA})\!)^p_{\mu,\nu}$ *in* (2.83)*,* $p \geq 0$*, the residual* $\Theta_k \to 0$ *as* $k \to \infty$ *in Theorem 2.6 (*h*-independent super-linear convergence).*

REMARK 2.17. *Theorem 2.6 requires the operator* $\mathsf{K}_{\mu,\nu}$ *to be compact so as to ensure application of Proposition 2.3. Recent results by Bletcha (Blechta, 2021) allow to consider a more general Proposition 2.3 with* $\mathsf{A} : H \to H$ *of the form* $\mathsf{A} = \mathsf{Q} + \mathsf{K}$*, with* Q *a bounded invertible operator and* K *compact. The latter could allow to relax the compactness for* $\mathsf{K}_{\mu,\nu}$ *and to analyze* $(\!(\mathrm{CA})\!)_{\mu,\nu}$ *as a general bounded perturbation of* $(\!(\mathrm{CA})\!)^p$*. This will be investigated elsewhere.*

2.5.9. Elliptic Case

We give further insight on the bi-parametric operator preconditioning framework by considering the elliptic case for OP-BG for $X =: H$ and V being Hilbert spaces. To this end, we assume that A is $\langle X \rangle$-elliptic and C is $\langle V \rangle$-elliptic. Therefore, we have the ellipticity conditions

$$\mathsf{a}(u,u) \geq \alpha_{\mathsf{A}} \|u\|_X^2 \quad \text{and} \quad \mathsf{c}(v,v) \geq \alpha_{\mathsf{C}} \|v\|_V^2,$$

for all $u \in X$ and all $v \in V$. Notice that continuous ellipticity implies a discrete inf-sup condition for conforming discretization spaces, with $\gamma_{\mathsf{A}} = \alpha_{\mathsf{A}}$ and $\gamma_{\mathsf{C}} = \alpha_{\mathsf{C}}$, respectively, and allows to apply our previous analysis—without requiring Assumption 3.

For $p \geq 0$, problem $(\!(\mathrm{CA})\!)^p$ leads to $\mathsf{C}\mathsf{N}^{-1}\mathsf{A} = \mathsf{I} + \mathsf{K}$ with K compact and self-adjoint. Thus, we introduce the ordered eigenvalues $|\lambda_{i+1}(\mathsf{K})| \leq |\lambda_i(\mathsf{K})|$ for $i \geq 1$. By (Winther, 1980, Section 2), $|\lambda_i(\mathsf{K})| = \sigma_i(\mathsf{K})$ and the Carleman class in (2.81) simplifies to the Neumann-Schatten class

$$|||\mathsf{K}|||_p := \left(\sum_{i=1}^{\infty} |\lambda_i(\mathsf{K})| \right)^{1/p} < \infty. \tag{2.90}$$

As ellipticity allows for more refined bounds, one can examine the use of preconditioned CG solvers (Steinbach, 2007, Section 13.1).

Corollary 2.4 (Elliptic Case). *Consider* $(\!|\mathrm{CA}|\!)_{\mu,\nu}$ *with* $\mathsf{c}_\mu \in \Phi_{h,\mu}(\mathsf{c})$, $\mathsf{a}_\nu \in \Phi_{h,\nu}(\mathsf{a})$, *such that* A_ν *is* $\langle X\rangle$*-elliptic and* C_μ *is* $\langle V\rangle$*-elliptic.Then, the continuous and perturbed problems have a unique solution,* u *and* $u_{h,\nu}$*, respectively, with the following error bound*

$$\begin{aligned}\|u-u_{h,\nu}\|_X &\le \inf_{w_h\in X_h}\left(\frac{\mathrm{K}_\mathsf{A}}{1-\nu}\|u-w_h\|_X+\frac{\nu}{1-\nu}\|w_h\|_X\right)+\frac{\nu}{\gamma_\mathsf{A}(1-\nu)}\|b_h\|_{Y_h'}\\ &\le \left(\frac{\mathrm{K}_\mathsf{A}^2}{1-\nu}\right)\inf_{w_h\in X_h}\|u-w_h\|_X+\frac{2\nu}{\gamma_\mathsf{A}(1-\nu)}\|b_h\|_{Y_h'}.\end{aligned}$$

with K_A *defined in* (2.12)*. Furthermore, it holds that*

$$\kappa_S(\mathbf{P}_\mu\mathbf{A}_\nu)=\kappa_2(\mathbf{P}_\mu\mathbf{A}_\nu)\le \mathrm{K}_{\star,\mu,\nu}, \tag{2.91}$$

with $\mathrm{K}_{\star,\mu,\nu}$ *in* (2.45)*. Therefore, for* $\mathbf{x}_0\neq\mathbf{u}_\nu$ *and* $1\le k\le N$*, the* k*-th iterate* $\mathbf{x}_k$ *of CG with an error* $\mathbf{e}_k:=\mathbf{x}_k-\mathbf{u}_\nu$ *is bounded in the* A_ν*-norm as*

$$\Theta_k^{\mathrm{CG}}:=\left(\frac{\|\mathbf{e}_k\|_{A_\nu}}{\|\mathbf{e}_0\|_{A_\nu}}\right)^{\frac{1}{k}}\le 2^{\frac{1}{k}}\left(1-\frac{2}{\sqrt{\mathrm{K}_{\star,\mu,\nu}}+1}\right). \tag{2.92}$$

Finally, consider $(\!|\mathrm{CA}|\!)^p_{\mu,\nu}$ *for* $p\ge 0$*. It holds that*

$$\Theta_k^{\mathrm{CG}}\le\frac{2\|\mathsf{n}\|}{\gamma_\mathsf{C}\gamma_\mathsf{A}\gamma_\mathsf{M}}\frac{1}{(1-\mu)(1-\nu)}\cdot\frac{1}{k}\sum_{j=1}^{k}|\lambda_j(\mathsf{K}_{\mu,\nu})| \tag{2.93}$$

and, if $p>0$*, one retrieves*

$$\Theta_k^{\mathrm{CG}}\le\frac{2\|\mathsf{n}\|}{\gamma_\mathsf{C}\gamma_\mathsf{A}\gamma_\mathsf{M}}\frac{|||\mathsf{K}_{\mu,\nu}|||_p}{(1-\mu)(1-\nu)}k^{-\frac{1}{p}}. \tag{2.94}$$

PROOF. By the ellipticity hypothesis on the sesqui-linear form a, Lemma 2.1 is replaced by the Lax-Milgram lemma (Ern & Guermond, 2013, Section 2.1.2), providing the sharper quasi-optimality constant K_A. Since the resulting system is Hermitian positive definite, the spectral and Euclidean condition numbers coincide. Next, we set $\varkappa:=\kappa_S(\mathbf{P}_\mu\mathbf{A}_\nu)$ and introduce the linear bound for the preconditioned CG with respect to the condition number (Kurics, 2010, Theorem 1.8):

$$\Theta_k^{\mathrm{CG}}\le 2^{\frac{1}{k}}\left(\frac{\sqrt{\varkappa}-1}{\sqrt{\varkappa}+1}\right). \tag{2.95}$$

Observe that

$$\left(\frac{\sqrt{\varkappa}-1}{\sqrt{\varkappa}+1}\right) = \left(1-\frac{2}{\sqrt{\varkappa}+1}\right) \leq \left(1-\frac{2}{\sqrt{\mathrm{K}_{\star,\mu,\nu}}+1}\right),$$

leading to (2.92). Since, $(\!(\mathrm{CA})\!)^{p}_{\mu,\nu}$ entails a self-adjoint compact perturbation $\mathsf{K}_{\mu,\nu} := \mathsf{C}_{\mu}\mathsf{N}^{-1}\mathsf{A}_{\nu} - \mathsf{I}$, one has an ordered eigenvalue decomposition, and the application of super-linear result for CG (Kurics, 2010, Theorem 1.9):

$$\Theta_{k}^{\mathrm{CG}} \leq 2\|(\mathbf{P}_{\mu}\mathbf{A}_{\nu})^{-1}\|_{H}\left(\frac{1}{k}\sum_{j=1}^{k}|\lambda_{j}(\mathbf{M}^{-1}\mathbf{K}_{\mu,\nu})|\right).$$

Finally, one can show that (cf. proof of Theorem 2.6):

$$\|\mathbf{A}_{\nu}^{-1}\mathbf{P}_{\mu}^{-1}\|_{H} \leq \frac{\|\mathsf{n}\|}{\gamma_{\mathsf{C}}\gamma_{\mathsf{A}}}\frac{1}{(1-\mu)(1-\nu)} \quad \text{and} \quad |\lambda_{j}(\mathbf{M}^{-1}\mathbf{K}_{\mu,\nu})| \leq \frac{1}{\gamma_{\mathsf{M}}}|\lambda_{j}(\mathsf{K}_{\mu,\nu})|,$$

proving the final result. □

2.6. Conclusion

For general Petrov-Galerkin methods, we considered their operator preconditioning and introduced the novel bi-parametric framework. Several results were derived including bounds in Euclidean norm for the convergence of iterative solvers when preconditioning, with GMRES as a reference. These results pave the way toward new paradigms for preconditioning, as they allow to craft robust preconditioners, better understand the efficiency of existing ones and relate them to experimental results. We see direct applications in a variety of research areas including wave propagation problems (Gander, Graham, & Spence, 2015), singular perturbation theory (Axelsson & Karátson, 2009, Section 3), fast numerical methods (Bebendorf, 2008; Bebendorf & Kunis, 2009) and iterative solvers (Saad & Schultz, 1986).

Future work avenues we foresee are: further analysis of second-kind Fredholm integral equations, with applications to acoustics and electromagnetics; deep learning of preconditioners for GMRES, and wavenumber asymptotic analysis for preconditioners. Also, we mention two promising research areas: (i) extension of $(\!(\mathrm{CA})\!)^{p}_{\mu,\nu}$ to bounded perturbations

of $(\!(\mathrm{CA})\!)^p$ via (Blechta, 2021); and (ii) characterization of Carleman class for compact operators using elliptic regularity theorems (Bessoud & Krasucki, 2006).

3. FAST CALDERÓN PRECONDITIONING FOR THE ELECTRIC FIELD INTEGRAL EQUATION

This chapter was published in IEEE Transactions on Antennas and Propagation in January, 2019.

3.1. Introduction

Prowess in computational electromagnetism (EM) has shown to be key in supporting the overwhelming pace of technological innovation seen for several decades. Indeed, as the EM spectrum is continuously exploited for ever more complex and varied purposes, the need for robust, fast and efficient simulators becomes all the more relevant. To measurement accuracy, Maxwell equations depict the physical phenomena of interest but their solution generally calls for numerical methods. Among these, common choices are finite differences, finite elements or boundary element methods (BEM). All these methods rely on a domain discretization over which fields/currents are approximated by easily computable bases and solutions derived from a linear system (Buffa & Hiptmair, 2003; Buffa, Hiptmair, von Petersdorff, & Schwab, 2003).

In this chapter, we consider time-harmonic EM waves scattered by a Perfect Electric Conductor (PEC) embedded in an exterior unbounded domain. To approximate fields, we reduce the original volume problem to the obstacle's boundary via Green's formulas and functions, incorporating implicitly radiation conditions at infinity. This leads to an integral equation for surface electric and magnetic currents called the Electric Field Integral Equation (EFIE) discretized with Rao-Wilton-Glisson (RWG) elements (Rao, Wilton, & Glisson, 1982). Due to the non-locality of the integral kernel, the BEM originates dense indefinite matrices with large and often impractical requirements in memory and computational work. This motivates the development of so-called *fast approximation techniques*. Among these, the Fast Multipole Method (FMM) clusters matrix terms according to interaction distances and constitutes one of the first and more widely spread techniques (Dembart & Yip, 1998; Gumerov & Duraiswami, 2004; Darve, 2000). Alternatively, Hierarchical Matrices

($\mathcal{H}$-mat) take a purely algebraic approach approximating the operator by low-rank matrices (Bautista, Francavilla, Vipiana, & Vecchi, 2014; Steinbach, 2007). This technique can be enhanced by incorporating a nested cluster structure to achieve linear complexity in some cases ($\mathcal{H}^2$-mat) (Omar & Jiao, 2014; Bebendorf, 2008). Regardless of the paradigm, all these methods share several features: only matrix-vector products are performed; interactions are partitioned into far and near ones; cluster decompositions; tolerance parameters set approximation accuracy; and, commonly exhibit log-linear complexity and memory consumption.

Yet, solving the resulting large EFIE matrices remains a hard task for traditional direct solvers, and thus one resorts to iterative ones such as GMRES (Saad & Schultz, 1986; Saad, 2003). Moreover, the spectral properties of EFIE matrices often lead to bad convergence rates, requiring preconditioning. Calderón Multiplicative Preconditioners (CMP) are a particular case of operator-based ones that lead to provable mesh-independent condition numbers (Hiptmair, 2006; Nédélec, 2001; S. Christiansen & Nédélec, 2001; Andriulli et al., 2008; Buffa & Christiansen, 2007). They employ matching Galerkin discretizations of operators with complementary mapping properties: using an opposite order operator so as to generate an endormorphism with a ratio between continuity and coercivity constants independent of mesh size. Calderón preconditioners are applicable even for Lipschitz scatterers and optimal if the surface is closed, i.e. has no boundary (Hiptmair, 2006, Section 4). Still, their building cost can be prohibitive as dual functions defined on a barycentric mesh lead to a six-fold increase in the size of the considered matrices. For instance, Zhang *et al.* (Xu, Bo, & Zhang, 2016) used $\mathcal{H}^2$-mat for discretizing the EFIE operator on the barycentric grid while Guo *et al.* (Guo, Hu, Yin, & Nie, 2009) appeal to the Adaptive Cross Approximation (ACA) algorithm (Bebendorf, 2008). Still, even when associated to efficient resolution techniques, Calderón preconditioning remains very expensive in terms of memory, assembly time and time per iteration for iterative solvers.

Inspired by (Bebendorf, 2008, Section 3.6), we present a reduced cost Calderón preconditioning strategy, dubbed *bi-parametric*, which considers the splitting of (i) solution accuracy and (ii) preconditioner quality, combining distinct precision orders of magnitude

inside the resolution scheme. The proposed bi-parametric paradigm is suitable for any preconditioner, provided tolerance parameters are accesible. Indeed, the ideas presented in this chapter extend verbatim to generic CMP-based applications (see e.g., (Cools, Andriulli, & Michielssen, 2011; Gossye, Huynen, Ginste, De Zutter, & Rogier, 2018; Beghein, Mitharwal, Cools, & Andriulli, 2017; Niino, Akagi, & Nishimura, 2017)). Here, we show similar performance to that of algebraic techniques such as the Near-Field (NF) preconditioner (Bunse-Gerstner & Gutiérrez-Cañas, 2006; Malas & Gurel, 2007; Carpentieri, Duff, & Giraud, 2000) and a significant reduction in memory requirements, assembly time and time per iteration with respect to the standard Calderón one.

The chapter is organized as follows. In Section 3.2, we recall the fundamentals of the EFIE and its RWG discretization. Details on its implementation are provided in Section 3.3, where the original approach is referred to as ε-Calderón. The bi-parametric Calderón is introduced in Section 3.4. Numerical tests for different configurations are given in Section 3.5 and Section 3.6 elaborates on future work.

3.2. The Electric Field Integral Equation

Consider $D \subset \mathbb{R}^3$ to be an open bounded Lipschitz PEC domain with boundary $\Gamma := \partial D$ and exterior unit normal $\mathbf{n}$. Its complement $D^c := \mathbb{R}^3 \backslash \overline{D}$ is a purely dielectric unbounded domain with real permeability and permittivity constants $\mu, \epsilon > 0$, yielding a wavenumber $k := \omega\sqrt{\mu\epsilon} \equiv \frac{2\pi}{\lambda}$ with ω being the angular frequency for a time dependence $e^{-\imath\omega t}$, $\imath$ being the imaginary unit, and λ the associated wavelength. Let us consider an incident field $\mathbf{U}^{\text{inc}}$ that $\mathbf{curl}\,\mathbf{curl}\,\mathbf{U}^{\text{inc}} - k^2\mathbf{U}^{\text{inc}} = 0$. By linearity, the total field $\mathbf{U} = \mathbf{U}^{\text{inc}} + \mathbf{U}^{\text{sc}}$, wherein the scattered field $\mathbf{U}^{\text{sc}}$ solves

$$\begin{aligned} &\mathbf{curl}\,\mathbf{curl}\,\mathbf{U}^{\text{sc}} - k^2\mathbf{U}^{\text{sc}} = 0 \quad &&\text{in } D^c, \\ &\mathbf{n} \times \mathbf{U}^{\text{sc}} = -\mathbf{n} \times \mathbf{U}^{\text{inc}} &&\text{on } \Gamma, \end{aligned} \tag{3.1}$$

satisfying the Silver-Müller radiation condition:

$$\left| \mathbf{curl}\,\mathbf{U}^{\text{sc}}(\mathbf{x}) \times \frac{\mathbf{x}}{\|\mathbf{x}\|_2} - \imath k \mathbf{U}^{\text{sc}}(\mathbf{x}) \right| = o\left(\|\mathbf{x}\|_2^{-1}\right), \tag{3.2}$$

for $\|\mathbf{x}\|_2 \to \infty$, $\mathbf{x} \in \mathbb{R}^3$, with $\|\cdot\|_2$ being the Euclidean l^2-norm. The induced electrical current density $\mathbf{j}$ on Γ satisfies the EFIE:

$$\mathcal{T}(\mathbf{j}) := \mathcal{T}_{\mathscr{S}}(\mathbf{j}) + \mathcal{T}_{\mathscr{H}}(\mathbf{j}) = -\mathbf{n} \times \mathbf{U}^{\text{inc}}, \tag{3.3}$$

where

$$\begin{aligned} \mathcal{T}_{\mathscr{S}}(\mathbf{j}) &:= \imath\omega\mu\mathbf{n} \times \int_\Gamma \frac{\mathrm{e}^{-\imath k\|\mathbf{x}-\mathbf{y}\|_2}}{4\pi\|\mathbf{x}-\mathbf{y}\|_2}\mathbf{j}(\mathbf{y})d\mathbf{y}, \\ \mathcal{T}_{\mathscr{H}}(\mathbf{j}) &:= -\frac{1}{\imath\omega\epsilon}\mathbf{n} \times \nabla \int_\Gamma \frac{\mathrm{e}^{-\imath k\|\mathbf{x}-\mathbf{y}\|_2}}{4\pi\|\mathbf{x}-\mathbf{y}\|_2}\nabla_\Gamma \cdot \mathbf{j}(\mathbf{y})d\mathbf{y}. \end{aligned}$$

The kernel of these operators decays as $\|\mathbf{x}-\mathbf{y}\|_2^{-1}$ and becomes singular whenever $\mathbf{x} = \mathbf{y}$ but their integration in variational form is valid. Thus, $\mathcal{T}$ is non-local and its discretization leads to dense indefinite complex matrices with complex eigenvalues.

To solve the EFIE, we generate meshes Γ_h representing the boundary Γ, consisting of a subdivision of Γ into a set of planar triangular non-overlapping elements τ_l such that $\Gamma_h := \cup_l \tau_l$. Following (Sauter & Schwab, 2010, Section 4.1.2), we define $h_{\tau_l} := \sup_{\mathbf{x},\mathbf{y}\in\tau_l} \|\mathbf{x}-\mathbf{y}\|_2$ and inscribed circles diameters ρ_{τ_l} per element. We then introduce the meshwidths $h \equiv h_{max} := \max_{\tau_l\in\Gamma_h} h_{\tau_l}$, $h_{min} := \min_{\tau_l\in\Gamma_h} h_{\tau_l}$ and the number of elements per wavelength r such that $h \leq \frac{\lambda}{r}$ for a given λ. Mesh quality information is condensed in the vector:

$$\mathbf{h} := \left[h, \frac{h_{max}}{h_{min}}, \max_{\tau_l\in\Gamma_h} \frac{h_{\tau_l}}{\rho_{\tau_l}}\right],$$

which contains grading and shape-regularity measures.

The EFIE (3.3) is solved numerically by approximating $\mathbf{j} \in X := \mathbf{H}_\times^{-1/2}(\mathrm{div}_\Gamma, \Gamma)$ (Buffa & Hiptmair, 2003) by $\mathbf{j}_h \in X_h \subset X$, using div-conforming RWG basis functions $\{\phi_n\}_{n=1}^N \subset X_h$ defined on Γ_h:

$$\mathbf{j}_h = \sum_{n=1}^{N} u_n \phi_n, \tag{3.4}$$

where $\mathbf{u} := (u_n) \in \mathbb{C}^N$ is a vector of unknown coefficients. By testing (3.3) with functions ϕ_i for $i = 1, \ldots, N$, we obtain the linear system:

$$\mathbf{Z}^{RWG}\mathbf{u} = \mathbf{b}, \tag{3.5}$$

wherein $\mathbf{Z}^{RWG}$ and $\mathbf{b}$ are called the impedance matrix and load vector, each with elements defined as

$$Z_{ij}^{RWG} := \int_{\Gamma_h} \mathcal{T}(\phi_j) \cdot (\mathbf{n} \times \overline{\phi_i}), \; b_i := -\int_{\Gamma_h} (\mathbf{n} \times \overline{\phi_i}) \cdot (\mathbf{n} \times \mathbf{U}^{\text{inc}}).$$

3.3. Standard Calderón preconditioner

Multiplicative Calderón preconditioning using a dual mesh was introduced by Andriulli *et al.* (Andriulli et al., 2008). Calderón preconditioning techniques exploit the self-regularizing property of the EFIE for smooth surfaces, i.e. the square of the EFIE operator does not have eigenvalues accumulating at zero or infinity. Furthermore, for closed Lipschitz surfaces, Calderón preconditioning yields a mesh-independent sequence of linear systems, i.e. whose condition number is bounded independently of h. Still, in Figure 3.1 we summarize the drawbacks of the original method and explain several steps that lead to the proposed fast bi-parametric Calderón preconditioning.

3.3.1. Dense Calderón preconditioner

A regularization of the EFIE is obtained by leveraging on the Calderón identity:

$$\mathcal{T}^2(\mathbf{j}) = -\frac{1}{4}\mathbf{j} + \mathcal{K}^2(\mathbf{j}), \tag{3.6}$$

where the operator

$$\mathcal{K}(\mathbf{j}) := \mathbf{n} \times \nabla \times \int_\Gamma \frac{\mathrm{e}^{-\imath k\|\mathbf{x}-\mathbf{y}\|_2}}{4\pi\|\mathbf{x}-\mathbf{y}\|_2}\mathbf{j}(\mathbf{y})d\mathbf{y} \tag{3.7}$$

is compact on smooth surfaces (Nédélec, 2001). In other words, on smooth surfaces, $\mathcal{T}^2$ is a second kind Fredholm operator whose spectrum accumulates at $-1/4$. Besides, it is an endomorphism on Lipschitz domains (Hiptmair, 2006, Section 4). Thus, we precondition

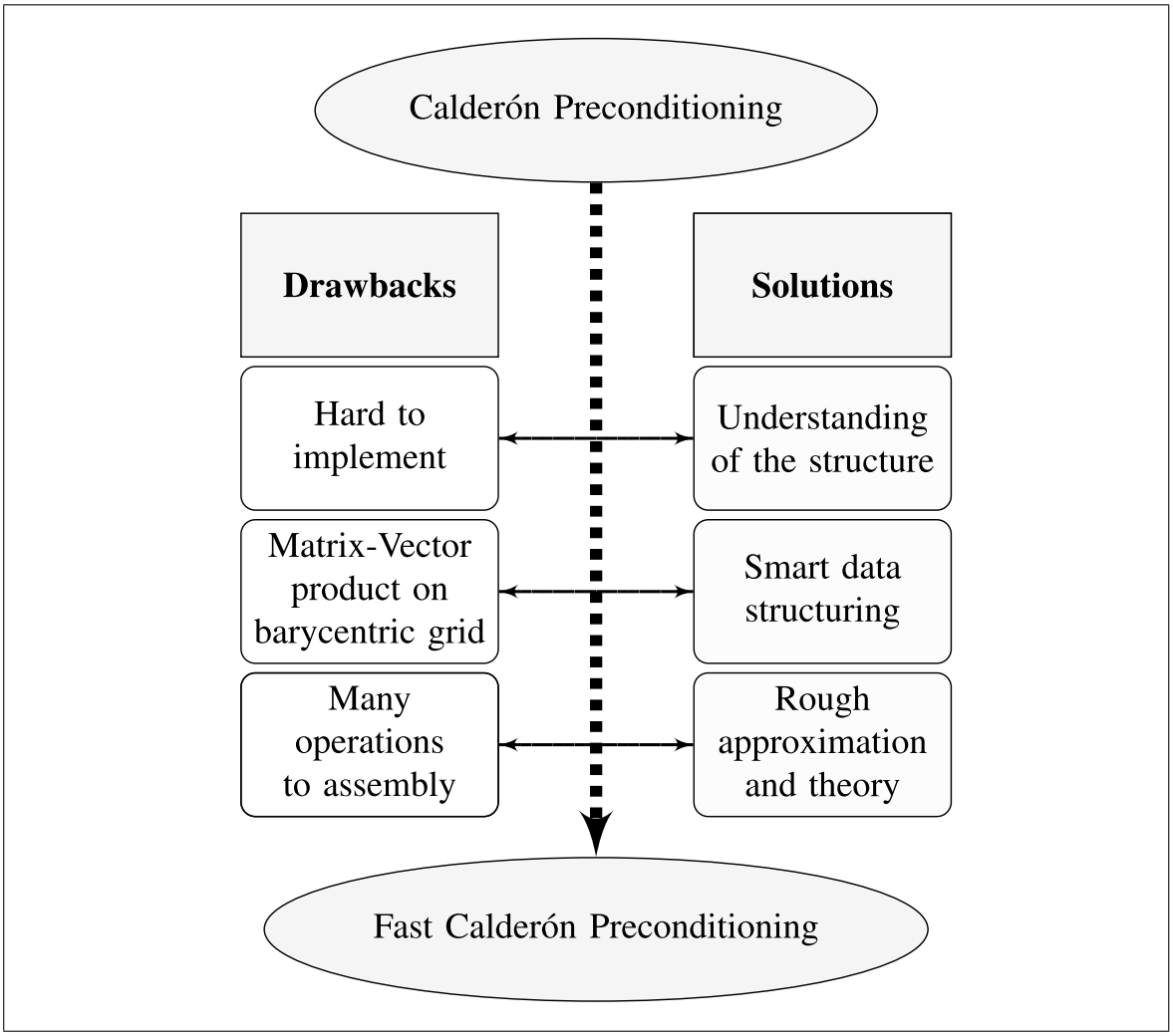


FIGURE 3.1. Summary of Calderón preconditioning drawbacks (left) and specific solutions (right). Observe the sequential improvement.

$\mathcal{T}$ by itself and solve

$$\mathcal{T}^2(\mathbf{j}) = \mathcal{T}(-\mathbf{n} \times \mathbf{U}^{\text{inc}}). \tag{3.8}$$

For the EFIE on a primal mesh Γ_h, one introduces Buffa-Christiansen (BC) div- and quasi-curl- conforming basis functions ϕ_{BC} (Buffa & Christiansen, 2007). This allows discretization of $\mathcal{T}^2$ in proper function spaces and ensures that the pairing between $\mathcal{T}$ and itself is invertible. BC functions are defined on the dual mesh (see Figure 3.2) but are built from a barycentrically refined mesh, denoted Γ_h^b, leading to a six-fold increase in computational and memory requirements with respect to those of the primal mesh. For each edge of the

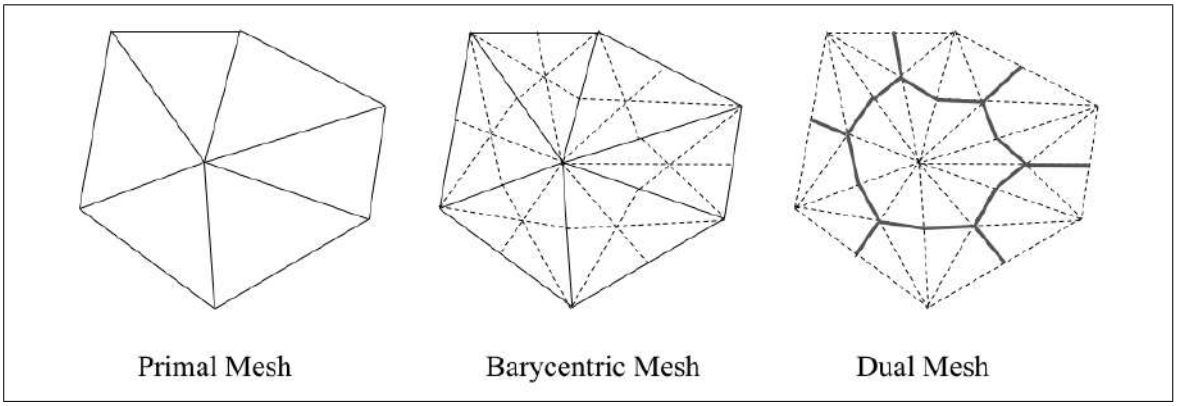


FIGURE 3.2. Meshes used for Calderón preconditioning. From the primal mesh Γ_h are generated barycentric and dual ones, denoted by Γ_h^b and Γ_h^d, respectively.

primal mesh, one considers the two hexagons of the dual mesh that cross it. They form the support of the BC function associated to this edge (Andriulli et al., 2008).

Input: Γ_h, problem parameters

(1) Define barycentric mesh $\leftarrow \Gamma_h^b$

(2) Assembly EFIE operator and right-hand side on barycentric mesh $\leftarrow \mathbf{Z}^b, \mathbf{b}^b$

(3) Assembly mapping and Gram matrices $\leftarrow \mathbf{R}, \mathbf{P}, \mathbf{G}$

Output: $\mathbf{Z}^b, \mathbf{b}^b, \mathbf{R}, \mathbf{P}, \mathbf{G}$

Algorithm 1: Calderón Preconditioning

The classical multiplicative Calderón preconditioning approach –based on discretization by RWG functions over the barycentric mesh Γ_h^b and leading to a $(6N) \times (6N)$ matrix $\mathbf{Z}^b \equiv \mathbf{Z}^{RWG,b}$– applied to (3.5) is summed up in Algorithm 1, inducing the linear system[1]:

$$\mathbf{C}\mathbf{Z}^p\mathbf{u} = \mathbf{C}\mathbf{b}, \tag{3.9}$$

with $\mathbf{C} \equiv \mathbf{G}^{-T}\mathbf{Z}^d\mathbf{G}^{-1}$, $\mathbf{Z}^d := \mathbf{P}^T\mathbf{Z}^b\mathbf{P}$, $\mathbf{Z}^p := \mathbf{R}^T\mathbf{Z}^b\mathbf{R}$ and $\mathbf{b} := \mathbf{R}^T\mathbf{b}^b$.

In the above, the restriction matrix $\mathbf{P}$ maps div- and quasi-curl- conforming BC functions on Γ_h^d to RWG functions defined on Γ_h^b while $\mathbf{R}$ takes RWG functions defined on Γ_h to Γ_h^b. They are sparse matrices with $\mathcal{O}(N)$ non-zero values and can be evaluated **exactly** by geometrical considerations in $\mathcal{O}(N)$ steps and memory. Finally, the Gram matrix

[1]Superscripts p, d for matrices refer to both primal and dual meshes, and are used to underline the projective character of associated discretizations related on barycentric grids.

is given by $(\mathbf{G})_{ij} := \int_\Gamma (\mathbf{n} \times \overline{\phi}_i) \cdot \phi_{BC,j}$ and is sparse and invertible. Remark that $\mathbf{Z}^d$ and $\mathbf{Z}^{BC}$ (resp. $\mathbf{Z}^p$ and $\mathbf{Z}^{RWG}$) represent the same operator discretization but imply different assembly work for matrix-vector products. Finally, notice that dual BC functions are strictly reserved for building the preconditioner, as this low for an efficient bi-parametric implementation as shown in Section 3.4.

Theorem 3.1 (cf. Theorem 2.1)**.** *Consider matrices* $\mathbf{Z}^d, \mathbf{Z}^p, \mathbf{G}$ *induced by Algorithm 1 for a closed Lipschitz domain. Then, the spectral condition number*

$$\kappa_S(\mathbf{G}^{-T}\mathbf{Z}^d\mathbf{G}^{-1}\mathbf{Z}^p) \leq K, \tag{3.10}$$

where K *is a positive constant independent of* h *but dependent on* Γ *and* k*.*

3.3.2. ε-Calderón preconditioning

For a matrix $\mathbf{A}$, let $\mathbf{A}_\varepsilon$ denote its approximation for a given tolerance ε and similarly for a vector $\mathbf{v}(\varepsilon) := \mathbf{v}_\varepsilon$. This approximation can be obtained by FMM, $\mathcal{H}$-mat, etc. Application of compressed Calderón preconditioning leads to

$$\mathbf{C}_\varepsilon \mathbf{Z}^p_\varepsilon \mathbf{u}_\varepsilon = \mathbf{C}_\varepsilon \mathbf{b} \tag{3.11}$$

with

$$\begin{aligned}
\mathbf{C}_\varepsilon &:= \mathbf{G}^{-T}\mathbf{P}^T\mathbf{Z}^b_\varepsilon\mathbf{P}\mathbf{G}^{-1} \equiv \mathbf{C} + \delta\mathbf{C}_\varepsilon, \\
\mathbf{Z}^p_\varepsilon &:= \mathbf{R}^T\mathbf{Z}^b_\varepsilon\mathbf{R} \equiv \mathbf{Z}^p + \delta\mathbf{Z}^p_\varepsilon, \\
\mathbf{u}_\varepsilon &:= \mathbf{u} + \delta\mathbf{u}_\varepsilon.
\end{aligned}$$

Set $\mathbf{C}_\varepsilon\mathbf{Z}^p_\varepsilon = \mathbf{C}\mathbf{Z}^p + \delta(\mathbf{C}_\varepsilon\mathbf{Z}^p_\varepsilon)$ where $\delta(\mathbf{C}_\varepsilon\mathbf{Z}^p_\varepsilon)$ is a perturbation depending on ε.

The next result, proved in Section 3.7.1, connects tolerance, condition number and solution accuracy.

Theorem 3.2 (ε-Calderón)**.** *Let $\varepsilon \geq 0$ and assume that, for the linear system* (3.11)*, it holds*[2]*:*

$$\left\|(\mathbf{C}\mathbf{Z}^p)^{-1}\delta(\mathbf{C}_\varepsilon\mathbf{Z}_\varepsilon^p)\right\|_X \leq \varepsilon < 1. \tag{3.12}$$

Then,

$$\kappa_S(\mathbf{C}_\varepsilon\mathbf{Z}_\varepsilon^p) \leq K\frac{1+\varepsilon}{1-\varepsilon} \quad \textit{and} \quad [\mathbf{u}_\varepsilon]_X \leq \frac{\varepsilon}{1-\varepsilon}. \tag{3.13}$$

Theorem 3.2 states that for a relatively small tolerance parameter, the condition number remains approximately constant while the solution error grows like $\mathcal{O}(\varepsilon)$. Indeed, solution accuracy does not depend on neither $\mathbf{C}$ nor $\delta\mathbf{C}_\varepsilon$ as $\delta\mathbf{u}_\varepsilon = -(\mathbf{Z}_\varepsilon^p)^{-1}\delta\mathbf{Z}_\varepsilon^p\mathbf{u}$ (*cf.* Section 3.7.1). Since both preconditioner ($\mathbf{C}_\varepsilon$) and impedance ($\mathbf{Z}_\varepsilon^p$) matrices are built upon $\mathbf{Z}_\varepsilon^b$ they share the same compression. As K is $h-$independent, the condition number behaves well when increasing tolerance. The theorem remains valid in the case of non-preconditioned EFIE but with $K = \mathcal{O}(h^{-2})$ (Andriulli, Tabacco, & Vecchi, 2010).

The above hints at decoupling tolerances of preconditioner and impedance matrix. Since the condition number remains relatively constant while accuracy grows linearly, one could rather define different tolerances to both matrix compressions. As Calderón preconditioning exhibits h-independent convergence, one can increase solver iteration counts in order to accelerate assembly and matrix-vector time on the barycentric grid. Ultimately, for cases where the impedance matrix is given, this splitting would allow to implement a specific black-box to obtain a preconditioner. This is the gist of the proposed method shown in Section 3.4.

Theorem 3.2 follows the perturbation analysis found in matrix compression (*cf.* (Bebendorf et al., 2013, Section 5) or (Bebendorf, 2008, Definition 2.39)). In practice, Theorem 3.2 characterizes a spectral tolerance which is linked to $\mathcal{H}$-mat tolerance or block sizes (*cf.* (Faustmann, Melenk, & Praetorius, 2015; Bebendorf, 2008)). Finally, we should mention that though

[2]Using the correspondences between $j_h \in X_h \subset X$ and $\mathbf{u} \in \mathbb{C}^N$ in (3.4), we set $\|\mathbf{u}\|_X := \|j_h\|_X$. Accordingly, for a bounded linear operator $\mathcal{B}_h : X_h \to X_h$ with induced discretization matrix $\mathbf{B}$, we introduce $\|\mathcal{B}_h\|_X := \sup_{v_h \in X_h} \|\mathcal{B}_h u_h\|_X / \|u_h\|_X$ and set $\|\mathbf{B}\|_X := \|\mathcal{B}_h\|_X$. For vectors and matrices, we introduce $\|\cdot\|_2$ and $\|\cdot\|_F$ as Euclidean and Frobenius norms, respectively. Brackets $[\cdot]_\cdot$ refer to relative norm errors, e.g. $[\mathbf{u}_\varepsilon]_X := \|\mathbf{u}_\varepsilon - \mathbf{u}\|_X / \|\mathbf{u}\|_X$.

NF-based preconditioners are widely spread, they lack theoretical results such as the one discussed.

3.4. Bi-parametric Calderón preconditioner

Define a pair of tolerance parameters $\mu, \nu > 0$ and assembly the EFIE matrices on primal and barycentric meshes with different tolerances, $\mathbf{Z}_\nu$ and $\mathbf{Z}_\mu^b$, respectively. We now analyze the linear system:

$$\left(\mathbf{G}^{-T}\mathbf{P}^T\mathbf{Z}_\mu^b\mathbf{P}\mathbf{G}^{-1}\right)\mathbf{Z}_\nu\mathbf{u}_\nu = \left(\mathbf{G}^{-T}\mathbf{P}^T\mathbf{Z}_\mu^b\mathbf{P}\mathbf{G}^{-1}\right)\mathbf{b}$$

rewritten as:

$$\mathbf{C}_\mu\mathbf{Z}_\nu\mathbf{u}_\nu = \mathbf{C}_\mu\mathbf{b}. \tag{3.14}$$

with $\mathbf{C}_\mu \equiv \mathbf{C} + \delta\mathbf{C}_\mu$ and $\mathbf{Z}_\nu \equiv \mathbf{Z} + \delta\mathbf{Z}_\nu$. Observe that now we just need to assembly $\mathbf{P}$ and $\mathbf{G}$.

PROPOSITION 3.1 (Bi-parametric Calderón). *For $\mu, \nu \geq 0$, assume that for the linear system* (3.14)*,*

$$\|(\mathbf{C}\mathbf{Z})^{-1}(\delta\mathbf{C}_\mu\mathbf{Z})\|_X \leq \mu,$$
$$\|(\mathbf{Z})^{-1}(\delta\mathbf{Z}_\nu)\|_X \leq \nu,$$

with $\mu + \nu + \mu\nu < 1$. Then,

$$\kappa_S(\mathbf{C}_\mu\mathbf{Z}_\nu) \leq K\frac{1+\mu+\nu+\mu\nu}{1-\mu-\nu-\mu\nu} \tag{3.15}$$

and

$$[\mathbf{u}_\nu]_X \leq \frac{\nu}{1-\nu}. \tag{3.16}$$

The proof is detailed in Section 3.7.2 and relies on splitting the inequality

$$\|(\mathbf{C}\mathbf{Z})^{-1}\delta\left(\mathbf{C}_\varepsilon\mathbf{Z}_\varepsilon\right)\|_X \leq \varepsilon.$$

One term is the classical inequality on the EFIE operator, $\|(\mathbf{Z})^{-1}(\delta\mathbf{Z}_\nu)\|_X \leq \nu$, which is independent of the preconditioner and represents the tolerance of any approximated EFIE linear system. The second term relates to the preconditioner tolerance, given by $\|(\mathbf{CZ})^{-1}(\delta\mathbf{C}_\mu\mathbf{Z})\|_X \leq \mu$, similar to the ε one. By decoupling these terms, the proposed technique reduces the work on the barycentric grid, since this is reserved for $\mathbf{C}_\mu$. Notice that: (i) for small parameters μ and ν, the condition number asymptotically behaves as $\mathcal{O}(1)$ and the relative error $[\mathbf{u}_\nu]_X$ grows as $\mathcal{O}(\nu)$; (ii) for small ν, the condition number behaves as $K\frac{1+\mu}{1-\mu}$ (*cf.* Figure 3.3). The theorem states that the condition number of the induced system is very resilient to perturbations.

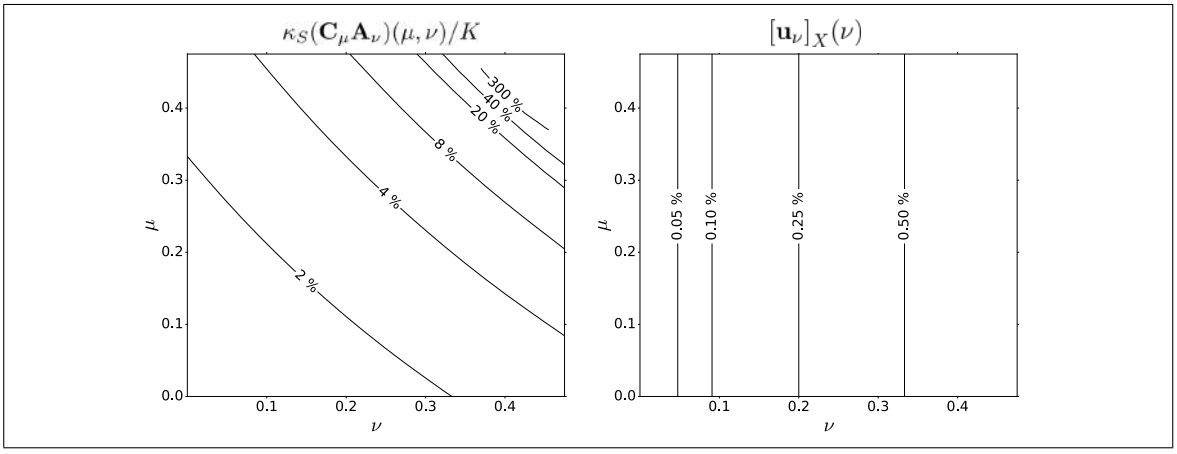


FIGURE 3.3. Behavior in change percentage of $\kappa_S(\mathbf{C}_\mu\mathbf{Z}_\nu)/K$ (left) and $[\mathbf{u}_\nu]_X$ (right) as a function of (μ,ν) for $\mu,\nu \in [0, 0.5]$.

We present the steps of the bi-parametric technique in Algorithm 2 with a dependency graph in Figure 3.4. Based on Proposition 3.1, one can confidently use a classical $\mathcal{H}$-mat approximation for the EFIE and assembly the preconditioner with $\mathcal{H}^2$-mat, to reduce error constants coming from the barycentric grid as much as possible. Moreover, one can perform a rough approximation of the preconditioner by relaxing the fast resolution scheme tolerance and/or simplifying quadrature rules. This is the approach followed successfully in Section 3.5.

3.5. Numerical Experiments

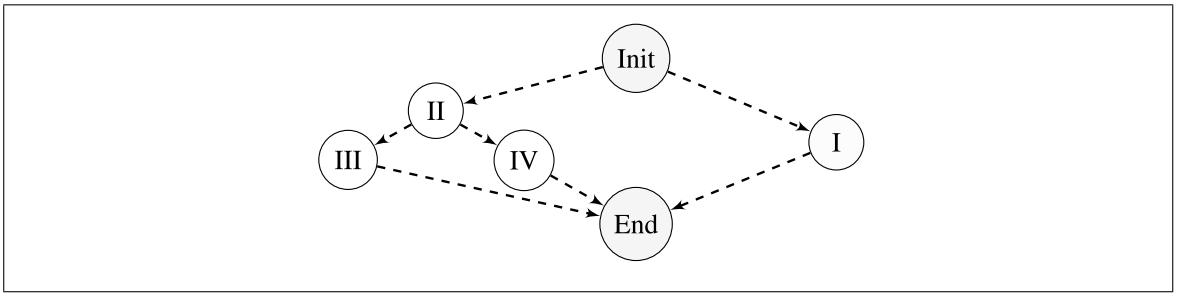


FIGURE 3.4. Dependency tree for bi-parametric Calderón preconditioning (*cf.* Algorithm 2). Observe independent processes for preconditioner (left leaf) and impedance matrix (right leaf) assemblies.

Input: Γ_h, problem parameters, a tolerance pair μ, ν
(I) Assembly EFIE operator with tolerance ν and right-hand side on primal mesh $\leftarrow \mathbf{Z}_\nu, \mathbf{b}$
(II) Define barycentric mesh $\leftarrow \Gamma_h^b$
(III) Assembly mapping and Gram matrices $\leftarrow \mathbf{R}, \mathbf{P}, \mathbf{G}$
(IV) Assembly EFIE operator with tolerance μ on barycentric mesh $\mathbf{Z}_\mu^b$
Output: $\mathbf{Z}_\mu^b, \mathbf{Z}_\nu, \mathbf{b}, \mathbf{R}, \mathbf{P}, \mathbf{G}$

Algorithm 2: Bi-parametric Calderón Preconditioning

3.5.1. Methodology

We consider an incident plane wave $\mathbf{U}^{\text{inc}} := [0, 0, e^{\imath k x_1}]$ scattered by objects in vacuum. In what follows, we solve the EFIE using $\mathcal{H}$-mat based on ACA (Steinbach, 2007, Subsection 14.2.3). This has been implemented in the open-source Galerkin boundary element library Bempp 3.2 (Śmigaj, Arridge, Betcke, Phillips, & Schweiger, 2015), along with an efficient implementation of Calderón preconditioning (*cf.* (Scroggs, Betcke, Burman, Śmigaj, & van't Wout, 2017; Kleanthous, Betcke, Hewett, Scroggs, & Baran, 2018) for interesting results on ε-Calderón). Tests were executed on a 32 core, 4 GB RAM per core, 64-bit Linux server using Python 2.7.6. Linear systems are solved with restarted GMRES(m) (Saad & Schultz, 1986; Saad, 2003).

Given an operator discretization $\mathbf{Z}$, we introduce quadrature orders

$$\mathbf{q}_\mathbf{Z} := [o_{near}, o_{medium}, o_{far}, o_{sing}],$$

as well as the rules for partitioning zones as presented in Bempp[3]. We choose default settings for zones and recall default quadrature orders ($\mathbf{q_Z} := [4, 3, 2, 6]$). Tolerance parameters $\mu_{\mathcal{H}}, \nu_{\mathcal{H}}$ are those introduced in ACA –they characterize the accuracy of each submatrix within the block cluster tree, e.g., $[(\mathbf{Z}_\nu)_{i,j}]_F \leq \nu_{\mathcal{H}}$ for all admissible partitions (i, j) (*cf.* (Steinbach, 2007; Harbrecht & Peters, 2013)). In practice, the spectral tolerances, ε, ν, μ, incorporate $\mathcal{H}$-mat tolerance and quadrature orders. Hence, they will be defined equivalently as numbers or as pairs $\boldsymbol{\varepsilon} := (\varepsilon_{\mathcal{H}}, \mathbf{q}_{\mathbf{Z}^b})$, $\boldsymbol{\mu} := (\mu_{\mathcal{H}}, \mathbf{q}_{\mathbf{Z}^b})$ and $\boldsymbol{\nu} := (\nu_{\mathcal{H}}, \mathbf{q}_{\mathbf{Z}})$. Any other relevant parameter inducing an approximation error could also be taken in account in the characterization of spectral tolerances.

For three test geometries, namely the unit sphere, the Fichera cube and a destroyer, we compare solving the EFIE in the following ways:

- **(NONE)**$_\nu$ is the unpreconditioned system;
- **(DIAG)**$_\nu$ stands for Jacobi or diagonal preconditioner;
- **(CALD)**$_\varepsilon$ is the ε-Calderón;
- **(CALD)**$_{\mu,\nu}$ is the bi-parametric Calderón;
- **(NF)**$_{\mu,\nu}$ uses NF preconditioning. It consists of (i) choosing a sparse NF pattern with integer distance parameter $\delta_{\mathbf{NF}}$ leading to $\mathbf{Z}^{near}_{\delta_{\mathbf{NF}}}$; and, (ii) a sparse LU decomposition of $\mathbf{Z}^{near}_{\delta_{\mathbf{NF}}}$ leading to a preconditioner fully described by $\boldsymbol{\mu}_{\mathbf{NF}} := (\delta_{\mathbf{NF}})$.

In all cases but **(CALD)**$_\varepsilon$, the original impedance matrix $\mathbf{Z}_\nu$ is the same. The shorthand **(CALD)** will refer to both Calderón-based techniques while **(NONE)**$_0$ refers to a reference solution, evaluated in dense mode with $\boldsymbol{\nu}_{\text{ref}} := (0, [10, 10, 10, 12])$ and solved with a direct solver, leading to a surface current $\mathbf{j}_{h,\text{ref}}$. Gram matrix inverses are solved internally by sparse LU decomposition and present computational and memory requirements negligible when compared to those for computing boundary integral operators –same considerations hold for $\mathbf{R}, \mathbf{P}$ matrices. For each implementation, we first perform a sensibility analysis so as to optimize the $\mathbf{Z}_\nu$ matrix parameters, i.e. $\boldsymbol{\varepsilon}$ for **(CALD)**$_\varepsilon$ (resp. $\boldsymbol{\nu}$ in other cases),

[3]https://bempp.com/quadrature/

guaranteeing a given solution accuracy. Consequently, given an approximation $\mathbf{Z}_\nu$ we can optimize $\boldsymbol{\mu}, \boldsymbol{\mu}_{\mathbf{NF}}$ for **(CALD)**$_{\mu,\nu}$ and **(NF)**$_{\mu,\nu}$. A fully documented Python/Bempp plugin allowing for a complete reproduction of our results is available online[4].

3.5.2. Results for the unit sphere

We start by considering the scattering by the unit sphere with a boundary condition such that $\mathbf{U}^{\text{sc}} = h_1^{(1)}(kr)\mathbf{e}_\phi$ for $\mathbf{x} \in D^c$, with $h_1^{(1)}(\cdot)$ being the first order spherical Hankel function, (r, θ, ϕ) spherical coordinates anchored at the point $\mathbf{x}_0$ with Cartesian coordinates $[0.1, 0.1, 0.1]$, and $\mathbf{e}_\phi$ denoting the unit vector parallel to $d\mathbf{x}/d\phi$ (Śmigaj et al., 2015, Section 4.4). We set $k = 1$ and simulate seven uniform meshes associated to indices $l = \{0, \ldots, 6\}$ with increasing number of elements per wavelength $r_l = 10(2l + 1)$, leading to meshes whose coarsest has $N_0 = 175$ degrees of freedom (dofs) and densest $N_6 = 20,043$ dofs. We observe experimentally that $[\mathbf{j}_{h,\text{ref}}]_{\boldsymbol{L}^2(\Gamma)}$ behaves as $\mathcal{O}(h) = \mathcal{O}(N^{-1/2})$, and fix $\boldsymbol{\nu}, \boldsymbol{\varepsilon}$ such that the approximation error does not deteriorate the solution accuracy, leading to $\boldsymbol{\nu}_l = \big(0.005 \cdot 2^{-l}, [4, 3, 2, 6]\big)$ and $\boldsymbol{\varepsilon}_l = \big(0.004 \cdot (2.5)^{-l}, [4, 3, 2, 6]\big)$. A stronger ACA compression was needed for $\boldsymbol{\varepsilon}$ to guarantee the same accuracy for all meshes due to work on baycentric grid. Indeed, for comparable results $\varepsilon_{\mathcal{H}}$ must be adapted to the $(6N) \times (6N)$ matrix $\mathbf{Z}_\varepsilon^b$. Then, we fix parameters for preconditioners to obtain mesh independent sequences for GMRES(1,000) chosen with a tolerance $tol = 10^{-8}$, hence $\boldsymbol{\mu}_{\text{NF},l} := \big(5 + l\big)$ and $\boldsymbol{\mu}_l := \big(0.1 \cdot 2^{-l}, [1, 1, 1, 2]\big)$. Figure 3.5 (a) shows that all solutions converge at the expected rate. Also, all techniques except for $(\mathbf{CALD})_\varepsilon$ led to errors almost equal to $(\mathbf{NONE})_\nu$ and so we only plot the latter. Figure 3.5 (b) displays the number of iterations for all methods, with behaviors of $\mathcal{O}(h^{0.77})$ and $\mathcal{O}(h^{0.75})$ for $(\mathbf{NONE})_\nu$ and $(\mathbf{DIAG})_\nu$, respectively. Notice also the strong mesh independence of $(\mathbf{CALD})$ and **(NF)**$_{\mu,\nu}$. Despite rough quadrature rules and ACA tolerance for $(\mathbf{CALD})_{\mu,\nu}$, the number of iterations remains exactly the same as for $(\mathbf{CALD})_\varepsilon$, both stabilizing at 8 iterations. Figure 3.5 (c) portrays total assembly time in seconds, and remark that both $(\mathbf{CALD})$ present an almost linear behavior, the bi-parametric approach taking half the time the standard version does. **(NF)**$_{\mu,\nu}$ presents

[4]https://github.com/pescap/cald

a similar behavior for coarse meshes but scales badly with N due to matrix inversion. Similar observations apply to solver time in seconds, presented in Figure 3.5 (d) arguing for the effectiveness of $(\mathbf{CALD})_{\mu,\nu}$ in this initial setting.

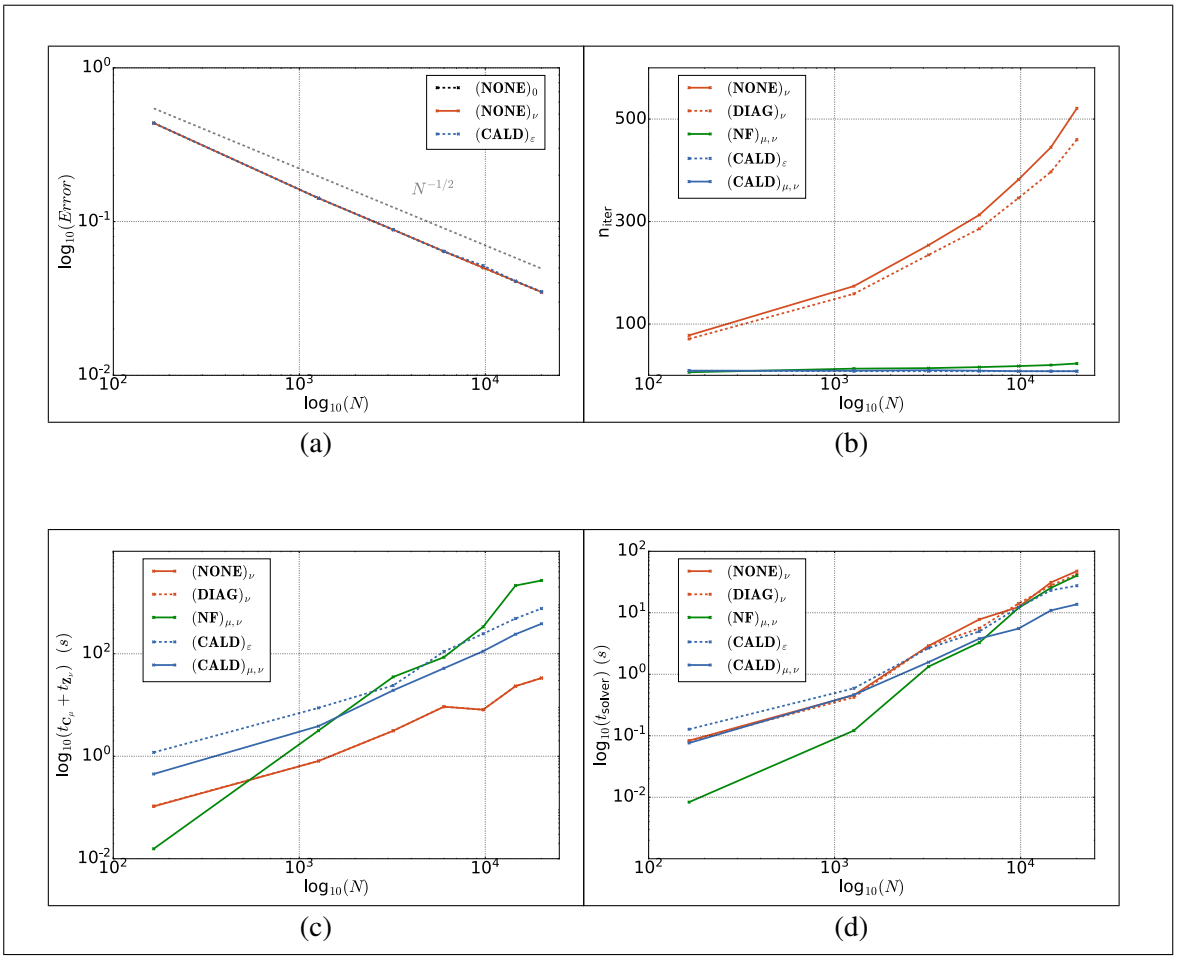


FIGURE 3.5. Results for unit sphere with known analytic solution and varying problem size. (a) Relative error in surface current versus N for increasing size problems. (b) Number of iterations for the relative error of GMRES to reach $tol = 10^{-8}$. (c) Total assembly times (in seconds) for all proposed formulations. (d) Solver times (in seconds).

3.5.3. Results for Fichera cube (reentrant corner)

For $k = 10$ ($f = 477.5$ MHz) and $r = 10$ elements per wavelength, Figure 3.6 shows primal and induced barycentric meshes graded towards corners with $\mathbf{h} = [0.0772, 5.67, 12.0]$ and $N = 16{,}113$ dofs. For the barycentric mesh $\mathbf{h}^b = [0.0460, 7.44, 19.7]$ and $N^b = 96{,}678$

dofs, showing how quality declines as grading and shape-regularity parameters increase by 33% and 64%, respectively. Tolerance for GMRES(m) is set to $tol = 10^{-5}$ in order to neglect iterative solver error.

We focus first on approximating the EFIE matrix $\mathbf{Z}_\nu$ yielding a vector $\mathbf{u}_\nu$ and associated current density $\mathbf{j}_{h,\nu}$. Table 3.1 shows relative errors for matrix approximation and resulting surface currents taking as reference $\mathbf{j}_{\text{ref},h}$ as explained before. Memory storage of operators –consisting in an inherited Scipy LinearOperator Class[5]– and assembly times $t_{\mathbf{Z}_\nu}$ are given in megabytes and seconds, respectively. We choose $\boldsymbol{\nu} := (\text{1e-03}, [4, 3, 2, 6])$, leading to a relative $\boldsymbol{L}^2$-error of 1.03% with limited memory consumption and assembly time. Using the parameters aforementioned, Figure 3.7 presents the resulting squared total electric field density.

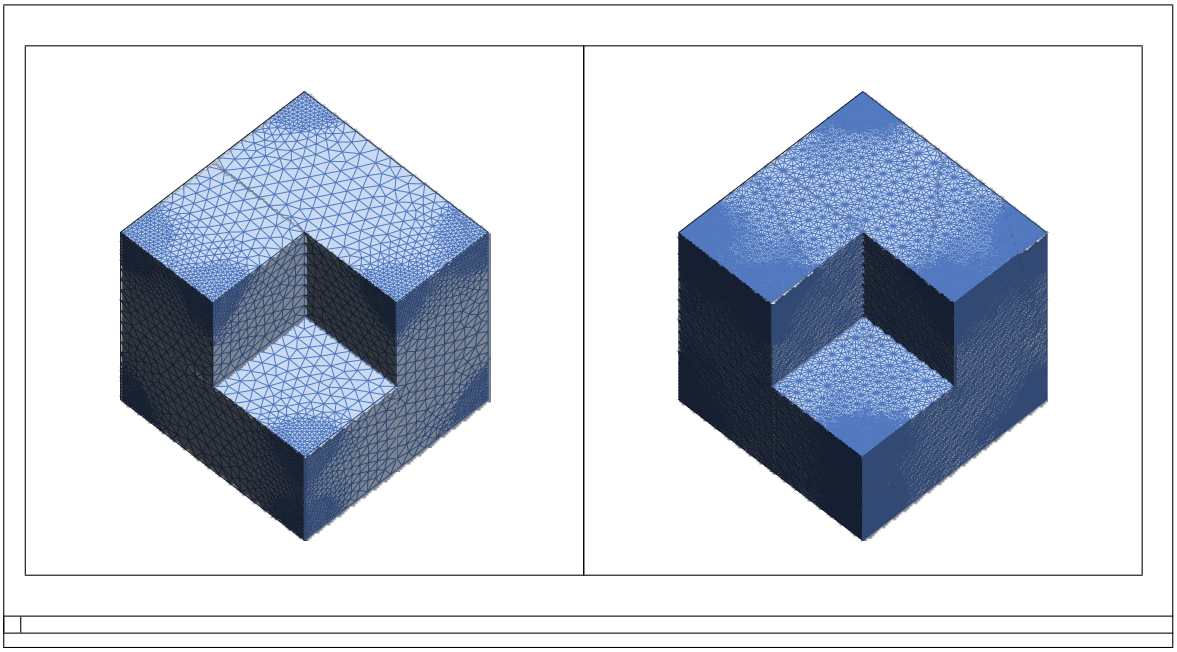


FIGURE 3.6. Fichera cube. Primal mesh Γ_h with $N = 16{,}113$ dofs (left) and induced barycentric mesh Γ_h^b with $N^b = 96{,}678$ dofs (right) used for a wavenumber $k = 10$ and $r = 10$ of elements per wavelength. The cube is of side one with bottom corner located at $[0, 0, 0]$.

A bi-parametric Calderón preconditioner $\mathbf{C}_\mu$ is chosen by analyzing different values of $\boldsymbol{\mu}$. An optimal choice of parameters leads to stable number of iterations and minimal

[5]https://bempp.com/operators/

TABLE 3.1. Fichera cube: Approximation results for $\mathbf{Z}_\nu$

$\boldsymbol{\nu}$		Relative errors			
$\nu_{\mathcal{H}}$	$\mathbf{q_Z}$	$[\mathbf{j}_{h,\nu}]_{\boldsymbol{L}^2(\Gamma)}$	$[\mathbf{Z}_\nu]_F$	$\text{mem}_{\mathbf{Z}_\nu}$	$t_{\mathbf{Z}_\nu}$
dense	[6,4,2,12]	7.262e-04	4.328e-05	4057	68.1
dense	[4,3,2,6]	1.501e-03	8.947e-05	4057	29.4
1e-06	[10,10,10,12]	8.685e-03	3.776e-05	1302	400
1e-04	[4,3,2,6]	7.500e-03	8.934e-04	944	29.0
1e-03	**[4,3,2,6]**	**1.030e-02**	**8.942e-04**	**696**	**25.0**
1e-03	[1,1,1,1]	4.157e-01	1.114e-02	764	13.0
1e-02	[4,3,2,6]	7.084e-02	9.720e-04	473	15.1
1e-01	[4,3,2,6]	9.255e-01	3.925e-03	268	10.6

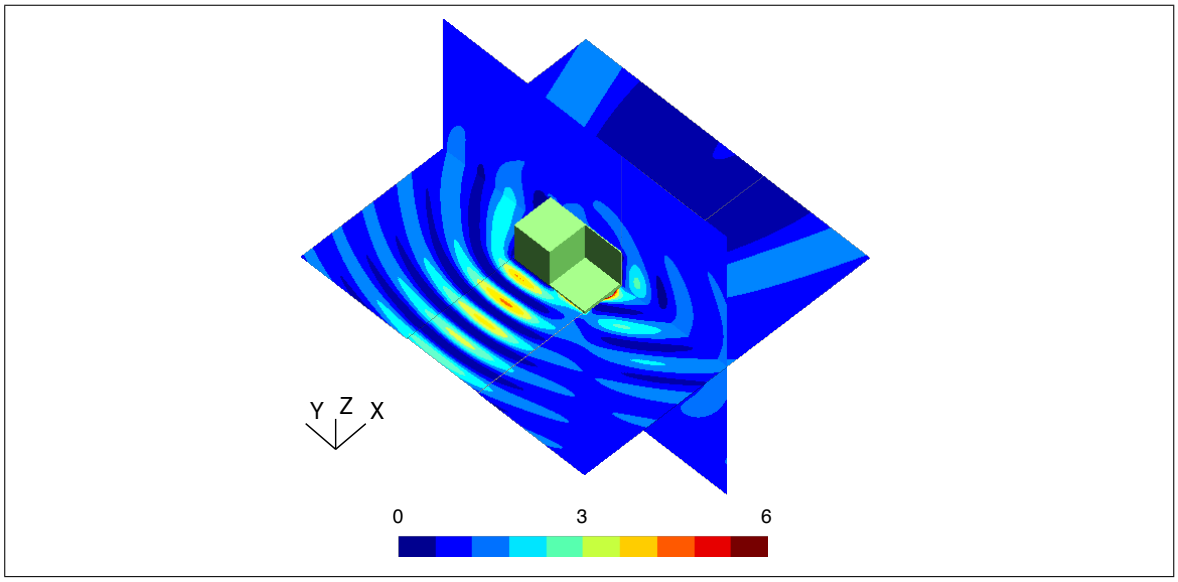


FIGURE 3.7. Fichera cube: Total electric density squared obtained with **(CALD)**$_{\mu,\nu}$ for the mesh in Figure 3.6, $\boldsymbol{\mu} = (\text{1e-01}, [1,1,1,2])$ and $\boldsymbol{\nu} = (\text{1e-03}, [4,3,2,6])$ with GMRES(200) and $tol = 10^{-5}$. Field evaluated on planes $X = 0.5$ and $Z = 0$ on a structured 200×200 square grid of side 10 by piecewise linear interpolation. Incident wave travels along X-axis and polarized along Z-axis.

total solver time and memory requirements. For several $\boldsymbol{\mu}$, Table 3.2 presents: number of iterations of GMRES(200), n_{iter}; preconditioner assembly ($t_{\mathbf{C}_\mu}$) and solver (t_{solve}) times; memory storage and relative $\boldsymbol{L}^2$-error for electric currents. For the impedance matrix, $\text{mem}_{\mathbf{Z}_\nu} = 696$ MB and $t_{\mathbf{Z}_\nu} = 25.0$ s. We see that the number of iterations remains stable despite crude preconditioner approximations. These values would lead to appalling accuracy if applied to $\mathbf{Z}_\nu$ (*cf.* Table 3.1).

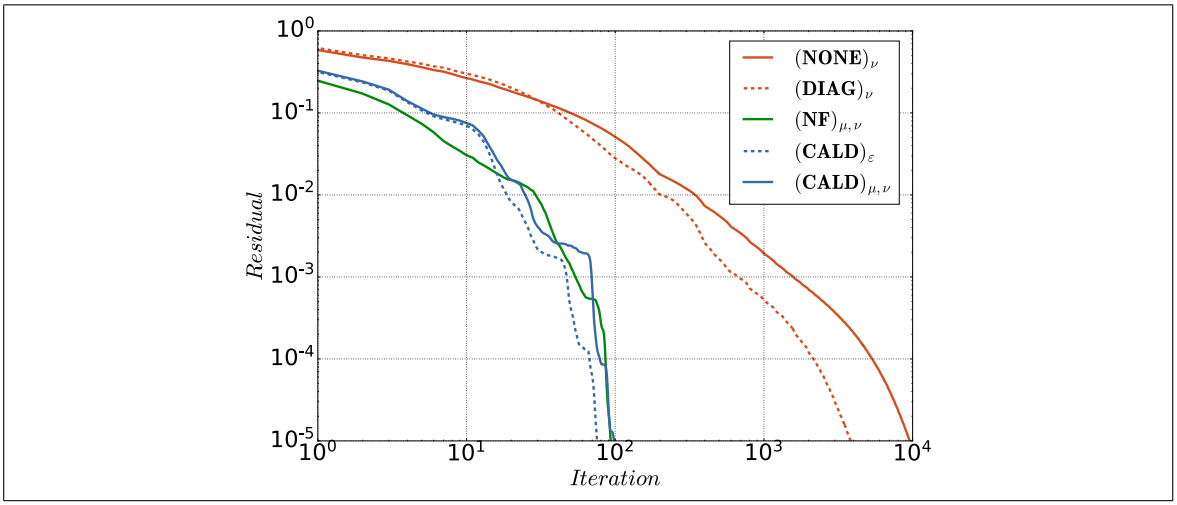


FIGURE 3.8. Fichera cube: GMRES(200) iterations for the parameters mentioned in Figure 3.7 along with $\boldsymbol{\varepsilon}$=$\boldsymbol{\nu}$=(1e-03, $[4, 3, 2, 6]$) and $\boldsymbol{\mu}_{\mathbf{NF}} = (5)$.

TABLE 3.2. Fichera cube: Approximation results for $\mathbf{C}_\mu$

$\boldsymbol{\mu}$		Performance parameters				
$\mu_{\mathcal{H}}$	$\mathbf{q_C}$	n_{iter}	$\text{mem}_{\mathbf{C}_\mu}$	$t_{\mathbf{C}_\mu}$	t_{solver}	$[\mathbf{j}_{h,\nu}]_{\boldsymbol{L}^2(\Gamma)}$
1e-04	[4,3,2,6]	76	7843	345.4	82.9	1.028e-02
1e-03	[4,3,2,6]	76	5783	313.4	66.75	1.030e-02
1e-02	[1,1,1,2]	100	3960	122.7	71.5	1.030e-02
1e-01	**[1,1,1,2]**	**99**	**2125**	**79.6**	**53.8**	**1.100e-02**
2e-01	[1,1,1,2]	118	1679	79.9	61.8	1.089e-02
3e-01	[1,1,1,2]	396	1458	69.1	200.8	1.035e-02

Based on the above, we choose $\boldsymbol{\mu} = (1\text{e-}01, [1, 1, 1, 2])$. This allows for a drastic reduction in memory, assembly time and time per matrix-vector product. Given $\boldsymbol{\mu}$, and following an analogous process, we build **(CALD)**$_\varepsilon$ and **(NF)**$_{\mu,\nu}$ in a bi-parametric fashion setting $\boldsymbol{\varepsilon} = \boldsymbol{\nu}$ and $\boldsymbol{\mu}_{\mathbf{NF}} = (5)$ respectively. Based on Figure 3.8, both **(CALD)** converge with low number of iterations and at similar rates despite of the rough approximation used in **(CALD)**$_{\mu,\nu}$. Also, **(NF)**$_{\mu,\nu}$ has similar n_{iter} as **(CALD)** with all three techniques surpassing **(NONE)**$_\nu$ and **(DIAG)**$_\nu$.

Table 3.3 presents a performance comparison for the techniques considered. To observe explicitly the mesh independence of both **(CALD)** in number of iterations ($n_{\text{iter}}^{\text{unif}}$), a uniform discretization using $r = 10$ without further grading is used leading to a mesh with $\mathbf{h}_{\text{unif}} := [0.0834, 1.72, 5.11]$ and 5,667 dofs. The number of iterations for **(NONE)**$_\nu$ and

(DIAG)$_\nu$ depends highly on the grading parameter. Also, for **(CALD)** the h-independence property is verified with a slight increase in number of iterations for **(CALD)**$_\varepsilon$ while no change is perceived for **(CALD)**$_{\mu,\nu}$. Observe the poor performance in terms of memory, assembly time and time per matrix-vector product[6] t_{mean} of **(CALD)**$_\varepsilon$ with a relative error growing more than two-fold when compared to the other ones (passing from around 1.1% to 2.476%). Indeed, in Section 3.5.2 we observed that a more drastic ACA compression was needed as computational work was carried out on the barycentric grid. In addition, the accuracy stabilizes at 1.56% for **(CALD)**$_\varepsilon$ when setting $\boldsymbol{\varepsilon} = (\text{1e-05}, [5, 4, 3, 7])$ or for more constrained ACA and quadrature parameters, despite uncompetitive computational requirements. This underlines the additional error induced by the deterioration of mesh shape regularity and grading, independently of compression parameter. Opposingly, the bi-parametric Calderón drastically reduces overall computational time and memory. To finish, remark that **(NF)**$_{\mu,\nu}$ and **(CALD)**$_{\mu,\nu}$ present similar behaviors in time but with a third of the memory cost for the **(NF)**$_{\mu,\nu}$ preconditioner.

TABLE 3.3. Fichera cube: Preconditioner performance comparison

Technique	$n_{\text{iter}}^{\text{unif}}$	n_{iter}	$\text{mem}_{\mathbf{C}_\mu}$	$\text{mem}_{\mathbf{Z}_\nu}$	$t_{\mathbf{C}_\mu}$	$t_{\mathbf{Z}_\nu}$	t_{solver}	t_{mean}	$[\mathbf{j}_{h,\nu}]_{\boldsymbol{L}^2(\Gamma)}$
(NONE)$_\nu$	1800	9572	-	696.4	-	31.75	520.9	3.645e-02	1.159e-02
(DIAG)$_\nu$	1838	3841	0.377	696.4	6.036e-03	31.75	228.8	3.560e-02	9.427e-03
(NF)$_{\mu,\nu}$	76	95	729.7	696.4	81.40	31.75	42.6	2.543e-01	1.011e-02
(CALD)$_\varepsilon$	73	77	5783		359.6		111.2	1.440	2.476e-02
(CALD)$_{\mu,\nu}$	**99**	**99**	**2125**	**696.4**	**79.6**	**31.75**	**53.8**	**4.986e-01**	**1.100e-02**

3.5.4. Results for a complex domain: destroyer

We now test how the preconditioning techniques perform in a complex scenario such as when solving the wave scattering by a destroyer. For $k = 0.19$ ($f = 9.07$ MHz) and $r = 10$, we obtain a mesh with $\mathbf{h} = [3.22, 21.2, 10.7]$, $N = 108{,}570$ dofs, $N^b = 651{,}420$ dofs and $\mathbf{h}^b = [1.99, 30.7, 25.2]$. We proceed as before and obtain $\boldsymbol{\nu} = \boldsymbol{\varepsilon} = (\text{1e-04}, [4, 3, 2, 6])$, $\boldsymbol{\mu}_{\mathbf{NF}} = (8)$, and $\boldsymbol{\mu} = (\text{1e-02}, [1, 1, 1, 2])$. Figures 3.11 and 3.12 show squared electric

[6] t_{mean} is estimated over 100 realizations of matrix-vector products.

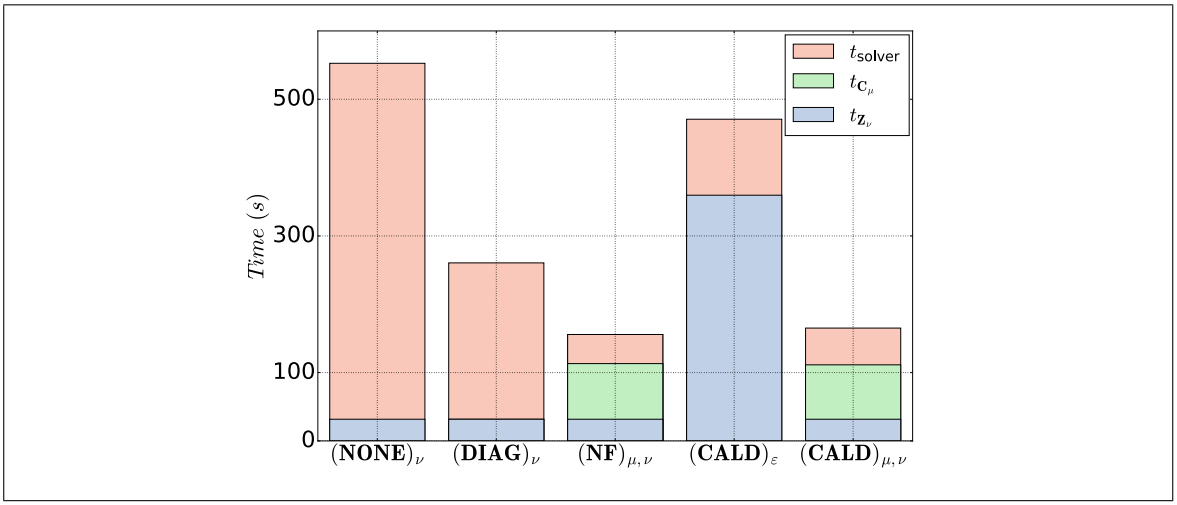


FIGURE 3.9. Fichera cube: CPU times for solving the EFIE using parameters specified in Figure 3.8. Blue (resp. green) boxes refer to assembly impedance matrix (resp. preconditioner) time while red boxes stand for total solver time.

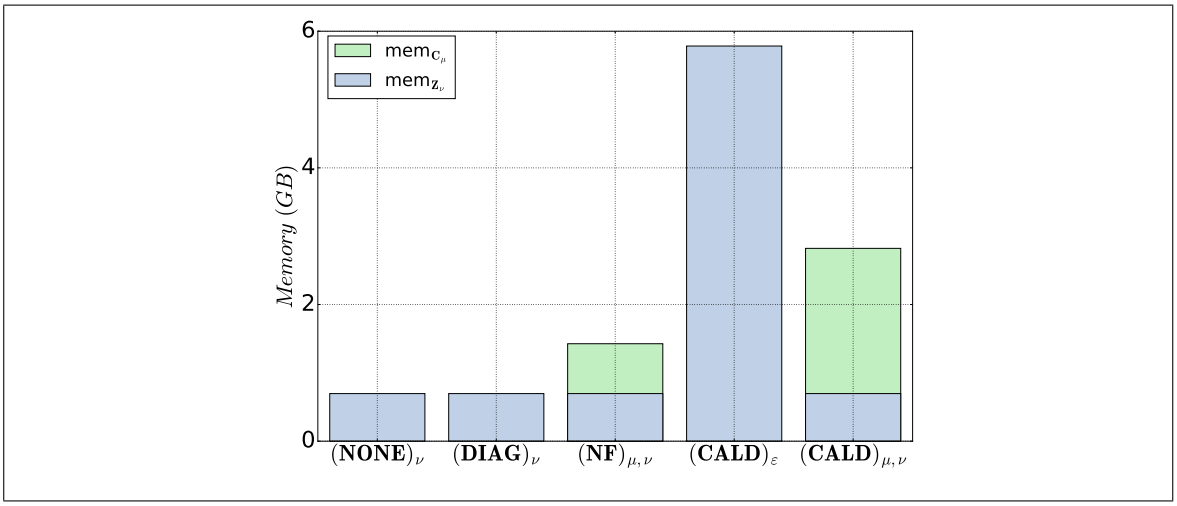


FIGURE 3.10. Fichera cube: Memory required to store impedance matrix (blue) and preconditioner (green) for parameters given in Figure 3.8.

surface current density in decibels. Observe the grading along small/elongated features such as the cannon (Figure 3.12).

Figure 3.13 presents residual iterative solver errors in the l^2-norm. We choose GMRES(1,500) with $tol = 10^{-5}$ and a maximum number of iterations of 10,000. Observe that **(NONE)**$_\nu$ and **(DIAG)**$_\nu$ do not converge, illustrating the need for a robust preconditioning technique. Concerning the other methods, similar remarks to those given for the Fichera

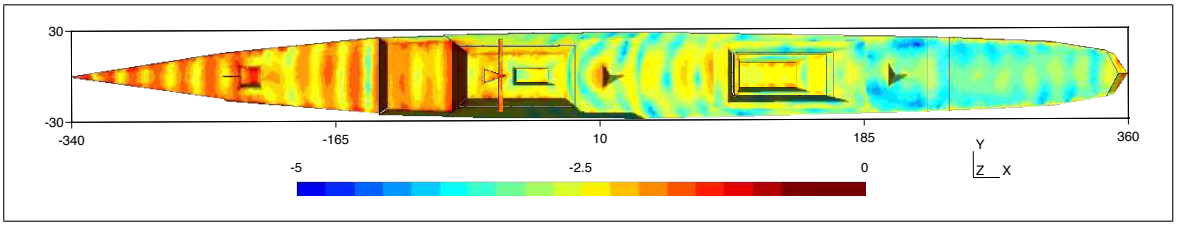


FIGURE 3.11. Destroyer: Top view of the squared current density $|\mathbf{j}_{h,\nu}|^2$ obtained with **(CALD)**$_{\mu,\nu}$ for $k = 0.19, r = 10$, $N =$ 108,570, $\boldsymbol{\mu} = (1\text{e-}02, [1, 1, 1, 2])$ and $\boldsymbol{\nu} = (1\text{e-}04, [4, 3, 2, 6])$. Solution obtained with GMRES(1,500) with $tol = 10^{-5}$. Evaluation is performed by piecewise linear interpolation on primal mesh nodes. Incident wave travels along X-axis and is polarized along the Z-axis.

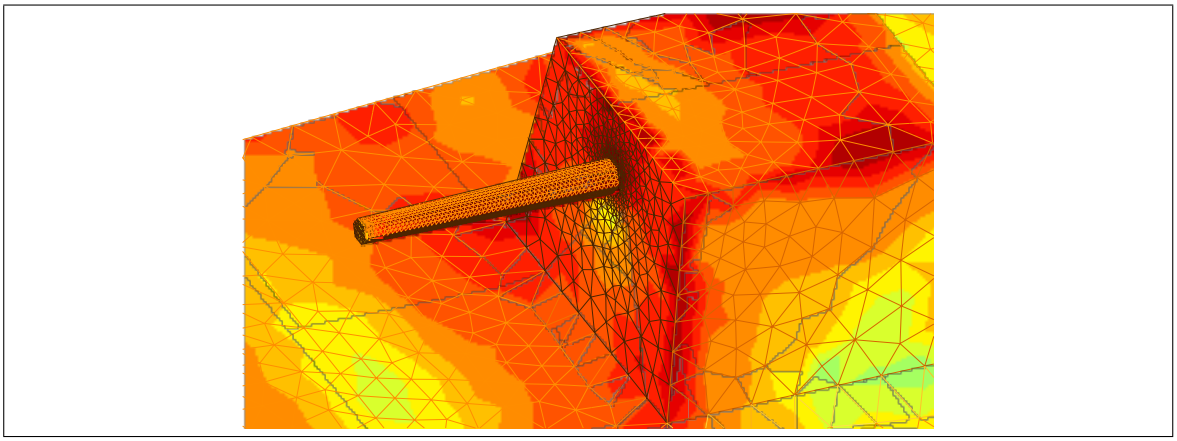


FIGURE 3.12. Destroyer: Squared current density of solution introduced in Figure 3.11. View centered on cannon, stressing a strong mesh grading.

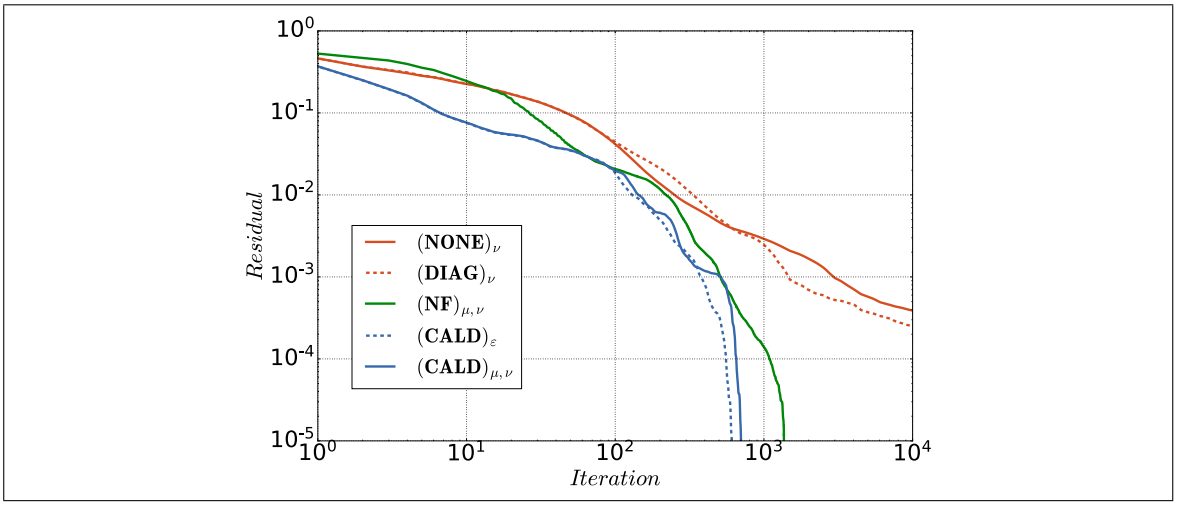


FIGURE 3.13. Destroyer: GMRES(1,500) iterations for parameters used in Figure 3.11 along with $\boldsymbol{\varepsilon} = \boldsymbol{\nu}$ and $\boldsymbol{\mu}_{\mathbf{NF}} = (8)$.

TABLE 3.4. Destroyer: Preconditioner performance comparison

Technique	$n_{\text{iter}}^{\text{unif}}$	n_{iter}	$\text{mem}_{\mathbf{C}_\mu}$	$\text{mem}_{\mathbf{Z}_\nu}$	$t_{\mathbf{C}_\mu}$	$t_{\mathbf{Z}_\nu}$	t_{solver}	t_{mean}
(NONE)$_\nu$	>10000	>10000	-	8107	-	400.5	>10108	0.8093
(DIAG)$_\nu$	>10000	>10000	2.544	8107	5.049e-02	400.5	>10391	0.8249
(NF)$_{\mu,\nu}$	485	1365	14497	8107	8698.7	400.5	9094.6	5.320
(CALD)$_\varepsilon$	624	610	63239		4231		11649	18.64
(CALD)$_{\mu,\nu}$	673	**703**	**34963**	**8107**	**1503**	**400.5**	**5664.0**	8.003

cube hold except that the optimal parameter for **(NF)**$_{\mu,\nu}$ led to a higher number of iterations than for **(CALD)**, attributable to the slow and expensive inversion step, limiting the acceptable range of distance parameters for the NF pattern.

Table 3.4 presents results for a uniform discretization leading to a mesh 100,665 dofs with $\mathbf{h}_{\text{unif}} := [3.30, 4.48, 10.72]$. Notice that when comparing the initial mesh to this uniform one, the grading parameter is multiplied by 4.73. When considering number of iterations, we find mesh-grading independence for **(CALD)**, while for **(NF)**$_{\mu,\nu}$, we observe an increase by a factor of 2.81 and that, at iteration 485, the residual is 186.2 times larger for the graded mesh case than in the uniform one. In the case of **(NONE)**$_\nu$ (resp. **(DIAG)**$_\nu$), the final residual in the graded mesh is 5.08 times larger (resp. 4.86) than for the uniform one.

Figure 3.14 shows: (i) poor performance of **(CALD)**$_\varepsilon$ and **(NF)**$_{\mu,\nu}$ in comparison to **(CALD)**$_{\mu,\nu}$ in total resolution time by a factor of 2.09 and 2.40 respectively; and, (ii) the significant increase in preconditioner assembly time for **(NF)**$_{\mu,\nu}$ when compared to results obtained in Section 3.5.3, due to the inherently sequential processing of LU factorization. Finally, Figure 3.15 shows that **(CALD)**$_\varepsilon$ requires 1.47 times more memory than **(CALD)**$_{\mu,\nu}$ while **(NF)**$_{\mu,\nu}$ is still less expensive in terms of memory (15.0GB vs. 35.0GB for preconditioner) and time per matrix-vector product (5.320s vs. 8.003s). This is again due to matrix-vector products on the barycentric mesh.

3.6. Conclusions

The presented bi-parametric splitting strategy provides an efficient and robust framework for preconditioning purposes. Its application to Calderón preconditioning leads to

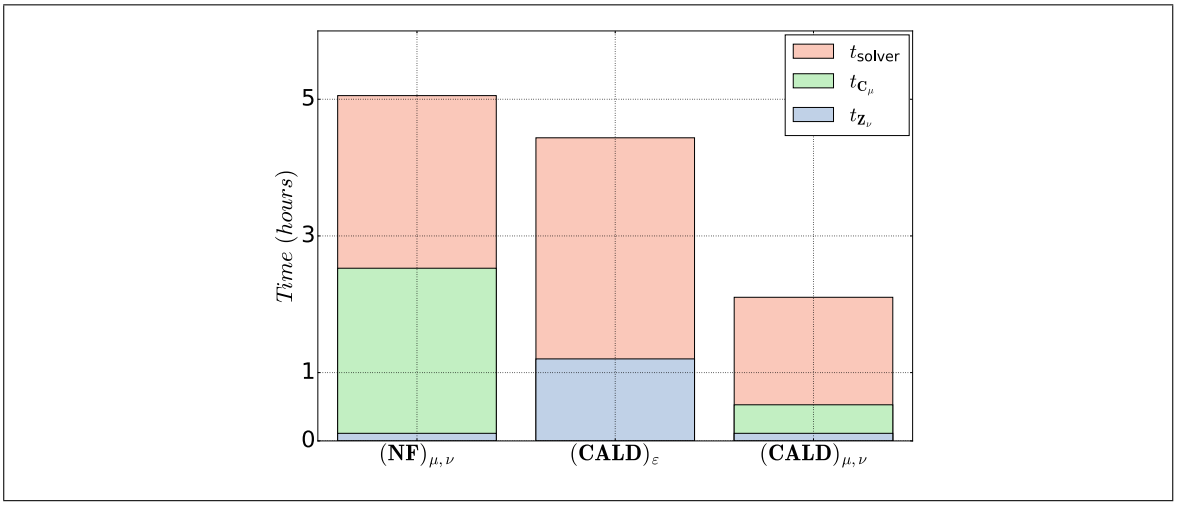


FIGURE 3.14. Destroyer: CPU times required for solving the EFIE using parameters given in Figure 3.11.

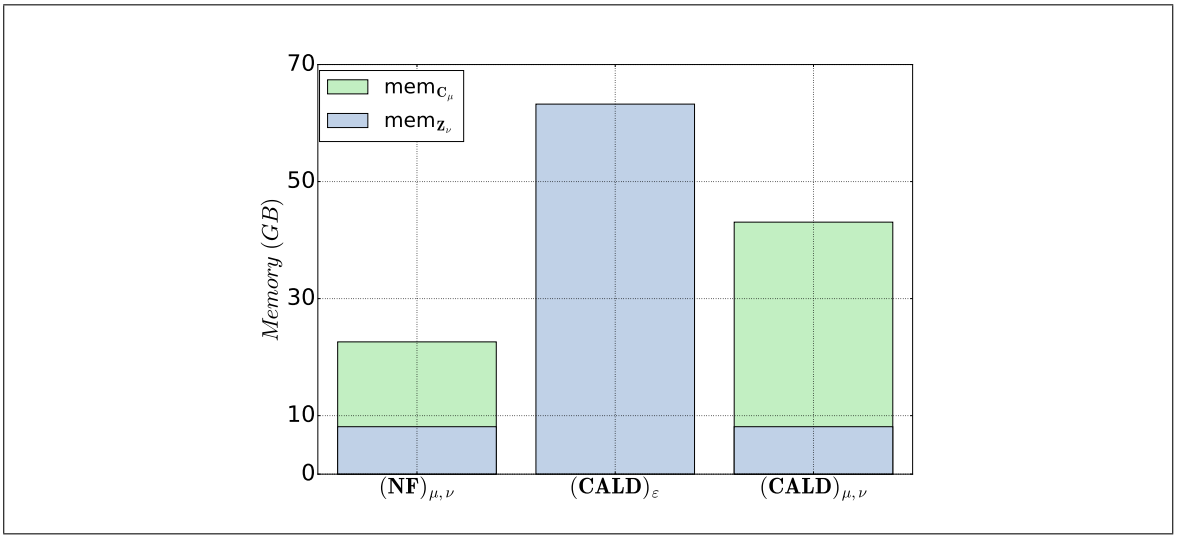


FIGURE 3.15. Destroyer: Memory required to store operators using parameters provided in Figure 3.11.

remarkable improvements in memory and computation times over the classical approach while retaining its stability properties. Several numerical tests attest to this with comparable and even better results than NF-preconditioning. Further work includes application to any type of operator-based preconditioning techniques. Also, given the stability displayed when performing rough preconditioner approximations, quadrature rules on dual meshes directly or partially, could discard the problematic use of barycentric meshes altogether.

3.7. Proof of theorems

3.7.1. ε-Calderón

By hypothesis

$$[\mathbf{C}_\varepsilon \mathbf{Z}_\varepsilon]_X \leq \|\mathbf{I} - (\mathbf{CZ})^{-1}\mathbf{C}_\varepsilon \mathbf{Z}_\varepsilon\|_X \leq \varepsilon. \tag{3.17}$$

Then,

$$\|\mathbf{C}_\varepsilon \mathbf{Z}_\varepsilon\|_X \leq \|\mathbf{CZ} - \mathbf{C}_\varepsilon \mathbf{Z}_\varepsilon\|_X + \|\mathbf{CZ}\|_X \leq (1+\varepsilon)\|\mathbf{CZ}\|_X.$$

Besides,

$$\begin{aligned}
\|(\mathbf{C}_\varepsilon \mathbf{Z}_\varepsilon)^{-1}\|_X &= \|(\mathbf{CZ} + \delta(\mathbf{C}_\varepsilon \mathbf{Z}_\varepsilon))^{-1}\|_X \\
&= \left\| \left(\mathbf{CZ}\left[\mathbf{I} + (\mathbf{CZ})^{-1}\delta(\mathbf{C}_\varepsilon \mathbf{Z}_\varepsilon)\right]\right)^{-1} \right\|_X \\
&\leq \left\| \left(\mathbf{I} + (\mathbf{CZ})^{-1}\delta(\mathbf{C}_\varepsilon \mathbf{Z}_\varepsilon)\right)^{-1} \right\|_X \left\|(\mathbf{CZ})^{-1}\right\|_X .
\end{aligned}$$

By Neumann series and application of Theorem 3.1 to $\mathbf{C}_\varepsilon \mathbf{Z}_\varepsilon$, we get

$$\kappa_S(\mathbf{C}_\varepsilon \mathbf{Z}_\varepsilon) \leq K\frac{1+\varepsilon}{1-\varepsilon}. \tag{3.18}$$

As per solution accuracy, we have $\delta\mathbf{u}_\varepsilon = -(\mathbf{Z} + \delta\mathbf{Z}_\varepsilon)^{-1}\delta\mathbf{Z}_\varepsilon\mathbf{u}$. Then,

$$\begin{aligned}
[\mathbf{u}_\varepsilon]_X &\leq \left\|(\mathbf{CZ} + \delta(\mathbf{C}_\varepsilon \mathbf{Z}_\varepsilon))^{-1}\delta(\mathbf{C}_\varepsilon \mathbf{Z}_\varepsilon)\right\|_X \\
&= \left\| \left[\mathbf{CZ}\left(\mathbf{I} + (\mathbf{CZ})^{-1}\delta(\mathbf{C}_\varepsilon \mathbf{Z}_\varepsilon)\right)\right]^{-1} \delta(\mathbf{C}_\varepsilon \mathbf{Z}_\varepsilon) \right\|_X \\
&= \left\| \left(\mathbf{I} + (\mathbf{CZ})^{-1}\delta(\mathbf{C}_\varepsilon \mathbf{Z}_\varepsilon)\right)^{-1} (\mathbf{CZ})^{-1}\delta(\mathbf{C}_\varepsilon \mathbf{Z}_\varepsilon) \right\|_X \\
&\leq \frac{\varepsilon}{1-\varepsilon}.
\end{aligned}$$

3.7.2. Bi-parametric Calderón

We notice that:

$$\begin{aligned}
&\|(\mathbf{CZ})^{-1}\delta(\mathbf{C}_\mu\mathbf{Z}_\nu)\|_X \\
&= \|(\mathbf{CZ})^{-1}[\mathbf{C}(\delta\mathbf{Z}_\nu) + (\delta\mathbf{C}_\mu)\mathbf{Z} + (\delta\mathbf{C}_\mu)(\delta\mathbf{Z}_\nu)]\|_X \\
&\leq \|(\mathbf{CZ})^{-1}\mathbf{C}(\delta\mathbf{Z}_\nu)\|_X + \|(\mathbf{CZ})^{-1}(\delta\mathbf{C}_\mu)\mathbf{Z}\|_X \\
&+ \|(\mathbf{CZ})^{-1}(\delta\mathbf{C}_\mu)(\delta\mathbf{Z}_\nu)\|_X \\
&= \|\mathbf{Z}^{-1}(\delta\mathbf{Z}_\nu)\|_X + \|\mathbf{Z}^{-1}\mathbf{C}^{-1}(\delta\mathbf{C}_\mu)\mathbf{Z}\|_X \\
&+ \|(\mathbf{CZ})^{-1}(\delta\mathbf{C}_\mu\mathbf{Z})(\mathbf{Z}^{-1}\delta\mathbf{Z}_\nu)\|_X \\
&\leq (\mu + \nu + \mu\nu).
\end{aligned}$$

Then, the proof for the condition number is derived if $\mu + \nu + \mu\nu < 1$. Concerning the accuracy, this time we have:

$$\delta\mathbf{u}_\nu = -(\mathbf{Z} + \delta\mathbf{Z}_\nu)^{-1}\delta\mathbf{Z}_\nu\mathbf{u}, \tag{3.19}$$

which gives

$$\begin{aligned}
[\mathbf{u}_\nu]_X &\leq \|(\mathbf{Z} + \delta\mathbf{Z}_\nu)^{-1}\delta\mathbf{Z}_\nu\|_X \\
&= \|(\mathbf{I} + (\mathbf{Z})^{-1}\delta\mathbf{Z}_\nu)^{-1}\mathbf{Z}^{-1}\delta\mathbf{Z}_\nu\|_X \leq \frac{\nu}{1-\nu}.
\end{aligned}$$

4. HELMHOLTZ SCATTERING BY RANDOM DOMAINS: FIRST-ORDER SPARSE BOUNDARY ELEMENT APPROXIMATION

This chapter was published in SIAM Journal on Scientific Computing in September, 2020

4.1. Introduction

Modeling wave scattering is key in numerous fields ranging from aeronautics to bioengineering or astrophysics. As applications become more complex, the ability to efficiently quantify the effects of random perturbations originated by actual manufacturing or operation conditions becomes ever more relevant for robust design. Under this setting, we consider standard time-harmonic wave scattering models with only aleatoric uncertainty, i.e. randomness in the shapes. More specifically, we aim at providing an accurate and fast uncertainty quantification (UQ) method for computing statistical moments of wave scattering solutions assuming small random perturbations or deviations from a nominal deterministic shape.

The model problems here considered involve solving Helmholtz equations in unbounded domains with constant coefficients supplemented by one or more different boundary conditions (BCs), namely, Dirichlet, Neumann, impedance and transmission ones. Under reasonable decay conditions at infinity, deterministic versions of such problems can be shown to be uniquely solvable even for Lipschitz scatterers (Nédélec, 2001; Sauter & Schwab, 2010). Considering Lipschitz parametrized transformations, the small perturbation assumption leads to diffeomorphisms between nominal and perturbed domains. This, in turn, gives rise to suitable shape Taylor expansions for the scattered fields, for which the corresponding shape derivatives (SDs) must be computed. Restricting ourselves to sufficiently smoother nominal domain, these SDs are solutions of homogeneous boundary value problems (BVPs) with boundary data depending on the normal component of the velocity field, allowed by the Hadamard structure theorem (see Theorem 2.27 in (Sokolowski & Zolesio, 1992)).

We will then approximate fields in the perturbed domains by quantities defined solely on the nominal shape. Indeed, for the cases considered –constant coefficients and unbounded domains–, one can conveniently reduce the volume problems associated to the scattered fields as well as to their SDs, onto the scatterers' boundaries by means of the integral representation formula (Sauter & Schwab, 2010). This involves solving boundary integral equations (BIEs) shown to be well posed.

The above described first order approximation (FOA) can be extended from the deterministic case to now random (but small) perturbations (Chernov, Pham, & Tran, 2015), giving birth to equations with deterministic operators with stochastic right-hand sides. Assuming separability of the underlying functional spaces as well as Bochner integrability, application of statistical moments on the linearized equation yields tensorized versions of the operator equations, thus parting from the multiple solves required by Monte Carlo (MC) methods. Yet, direct numerical approximation of these tensor systems gives rise to the infamous curse of dimensionality. This can be, in turn, remedied by applying the general sparse tensor approximation theory originally developed by von Petersdorff and Schwab (von Petersdorff & Schwab, 2006), and which has multiple applications ranging from diffraction by gratings (Silva-Oelker, Aylwin, Jerez-Hanckes, & Fay, 2018) to neutron diffusion (Fuenzalida, Jerez-Hanckes, & McClarren, 2019) problems. In our case, numerically, we will employ the Galerkin boundary element method (BEM) to solve the arising first kind BIEs. As both nominal solutions and SDs will be derived over the same surface, the FOA-BEM allows for substantial computational savings by employing the same matrix computations.

Depending on the regularity of solutions, statistical moments resulting from the FOA-BEM can be computed by sparse tensor approximations robustly. Harbrecht, Schneider and Schwab (Harbrecht, Schneider, & Schwab, 2008) studied the interior Laplace problem with Dirichlet BC whereas the Laplace transmission problem was analyzed in (Chernov et al., 2015). Jerez-Hanckes and Schwab (Jerez-Hanckes & Schwab, 2016) provide the numerical analysis of the method in the case of Maxwell scattering. Computationally, further acceleration can be achieved by employing the combination technique (CT), introduced by

Griebel and co-workers (Griebel et al., 1990; Harbrecht, Peters, & Siebenmorgen, 2013). Specifically, the method allows for simple and parallel implementation, which we will further detail in the chapter. Throughout, we apply the FOA-BEM-CT method –referred to as first-order sparse BEM (FOSB) method to alleviate notations– to the Helmholtz problem. To our knowledge, the case of the FOSB method for the Helmholtz-UQ remains untackled.

The chapter is structured in the following way. First, we introduce the mathematical tools used throughout in Section 4.2. Generic scattering problems formulations as well as the description of the BVPs solved by the SDs are given in Section 4.3. We then restrict ourselves to the associated BIEs in Section 4.4 and analyze their Galerkin solutions in Section 4.5. Implementation aspects of the FOSB method are given in Section 4.6 whereas numerical results are provided in Section 4.7. Finally, further research avenues are highlighted in Section 4.8.

4.2. Mathematical tools

We start by setting basic definitions as well as the functional space framework adopted for our analysis. As a reference, Table 4.3 beneath Section 4.5 provides a non-exhaustive list of the acronyms used throughout this chapter.

4.2.1. General notation

Throughout, vectors and matrices are expressed using bold symbols, $(\mathbf{a} \cdot \mathbf{b})$ denotes the classical Euclidean inner product, $\|\cdot\|_2 := \sqrt{\mathbf{a} \cdot \mathbf{a}}$ refers to the Euclidean norm, C is a generic positive constant and o, $\mathcal{O}$ are respectively the usual little-o and big-$\mathcal{O}$ notations. Also, we set $\imath^2 = -1$, $\mathbb{S}^1$ and $\mathbb{S}^2$ are the unit circle and sphere, respectively.

Let $D \subseteq \mathbb{R}^d$, with $d = 2, 3$, be an open set. For a natural number k, we set $\mathbb{N}_k := \{k, k+1, \ldots\}$. For $p \in \mathbb{N}_0 = \{0, 1, \ldots\}$, we denote by $C^p(D)$ the space of p-times differentiable functions over D, by $C^{p,\alpha}(D)$ the space of Hölder continuous functions with exponent α, where $0 < \alpha \leq 1$. Also, let $L^p(D)$ be the standard class of functions with bounded L^p-norm over D. For a Banach space X and an open set $T \subset \mathbb{R}$, we introduce

the usual Bochner space $C^p(T;X)$. Given $s \in \mathbb{R}$, $q \geq 0$, $p \in [1,\infty]$, we refer to (Sauter & Schwab, 2010, Chapter 2) for the definitions of function spaces $W^{s,p}(D)$, $H^s(D)$, $H^q_{\text{loc}}(D)$ and $H^q_{\text{loc}}(\Delta, D)$. Norms are denoted by $\|\cdot\|$, with subscripts indicating the associated functional space. Similarly, relative norms are denoted by brackets e.g., $[a-b] = \|a-b\|/\|b\|$ for a an approximation of a reference b.

For $k \in \mathbb{N}_1$, and $\mathbf{x}_i \in \mathbb{R}^d$, $i = 1, \cdots, k$, we set $\underline{\mathbf{x}} := (\mathbf{x}_1, \cdots, \mathbf{x}_k)$. Besides, k-fold tensors quantities are denoted with parenthesized subscripts, e.g., $f^{(k)} := f \otimes \cdots \otimes f$. This notation applies indifferently to functions, domains and function spaces. The diagonal terms of a k-fold tensor Σ^k at $\underline{\mathbf{x}}$ are denoted by $\operatorname{diag}\Sigma^k(\underline{\mathbf{x}}) := \Sigma^k|_{\mathbf{x}_1=\cdots=\mathbf{x}_k}$. Following (Jerez-Hanckes & Schwab, 2016, Section 4.1), for X, Y separable Hilbert spaces, we set $\mathsf{B} \in \mathcal{L}(X,Y)$ the space of linear continuous mapping from X to Y and define the unique continuous tensor product operator:

$$\mathsf{B}^{(k)} := \underbrace{\mathsf{B} \otimes \cdots \otimes \mathsf{B}}_{k\text{-times}} \in \mathcal{L}(X^{(k)}, Y^{(k)}).$$

4.2.2. Traces and surface operators

Let $D \subset \mathbb{R}^d$ with $d = 2, 3$ be open bounded with Lipschitz boundary $\Gamma := \partial D$ and complement exterior domain $D^c := \mathbb{R}^d \backslash \overline{D}$. Equivalently, we will write $D^0 \equiv D^c$ and $D^1 \equiv D$ to refer to exterior and interior domains, respectively. Accordingly, when defining scalar fields in $D^c \cup D$, we use notation $\mathrm{U} = (\mathrm{U}^0, \mathrm{U}^1)$. For $i = 0, 1$, we introduce the continuous and surjective trace mappings (Sauter & Schwab, 2010, Sections 2.6 and 2.7):

$$\begin{aligned} &\text{(Dirichlet trace)} \quad \gamma_0 : H^1_{\text{loc}}(D^i) \to H^{\frac{1}{2}}(\Gamma), \\ &\text{(Neumann trace)} \quad \gamma_1 : H_{\text{loc}}(\Delta, D^i) \to H^{-\frac{1}{2}}(\Gamma). \end{aligned}$$

For a suitable scalar field U^i, $i = 0, 1$, we refer to a pair of traces $\boldsymbol{\xi}^i$ as Cauchy data if

$$\boldsymbol{\xi}^i \equiv \begin{pmatrix} \lambda_i \\ \sigma_i \end{pmatrix} := \begin{pmatrix} \gamma_0 \mathrm{U}^i \\ \gamma_1 \mathrm{U}^i \end{pmatrix}. \tag{4.1}$$

Likewise, we introduce the second-order trace operator $\gamma_2 \mathrm{U}^i := (\nabla^2 \mathrm{U}_i|_\Gamma)\mathbf{n}\cdot\mathbf{n} = \frac{\partial^2 \mathrm{U}^i}{\partial \mathbf{n}^2}\big|_\Gamma$ along with the tangential gradient ∇_Γ and tangential divergence div_Γ (Nédélec, 2001, Section 2.5.6).

4.2.3. Random domains

Throughout, we consider an open bounded Lipschitz –nominal– domain $D \subset \mathbb{R}^d$, $d = 2, 3$, of class $C^{2,1}$ (McLean, 2000, Definition 3.28), with boundary $\Gamma := \partial D$ and exterior unit normal field $\mathbf{n} \in W^{2,\infty}(\Gamma)$ pointing by convention towards the exterior domain. Those domains are commonly referred to as domains with *Lyapunov boundary*. The mean curvature $\mathfrak{H} := \mathrm{div}\,\mathbf{n}$ belongs to $W^{1,\infty}(\Gamma)$.

Let $(\Omega, \mathcal{A}, \mathbb{P})$ be a suitable probability space and X a separable Hilbert space. For an index $k \in \mathbb{N}_1$ and $\underline{\mathbf{x}} := (\mathbf{x}_1, \cdots, \mathbf{x}_k)$, and for $\mathrm{U} : \Omega \to X$ a random field in the Bochner space $L^k(\Omega, \mathbb{P}; X)$ (Jerez-Hanckes & Schwab, 2016, Section 4.1), we introduce the statistical moments:

$$\mathcal{M}^k[\mathrm{U}(\omega)] := \int_\Omega \mathrm{U}(\mathbf{x}_1, \omega)\cdots\mathrm{U}(\mathbf{x}_k, \omega) d\mathbb{P}(\omega), \tag{4.2}$$

$$\mathbb{V}^k[\mathrm{U}(\omega)] =: \mathrm{diag}\,\mathcal{M}^k[\mathrm{U}(\omega)] - \mathbb{E}[\mathrm{U}(\omega)]^k, \tag{4.3}$$

with $\mathcal{M}^1 \equiv \mathbb{E}$ being the *expectation* and $\mathbb{V}^2$ the *pseudo-variance* (Silva-Oelker et al., 2018).

In a nutshell, the aim of the present chapter is as follows: given a random domain with realization $\mathcal{D}(\omega)$ specified later on, consider $\mathrm{U}(\mathbf{x}, \omega)$ defined over $\mathcal{D}(\omega)$ as the solution of a Helmholtz scattering problem (see Section 4.3). We seek at quantifying:

$$\mathbb{E}[\mathrm{U}(\omega)] \text{ and } \mathcal{M}^k[\mathrm{U}(\omega) - \mathbb{E}[\mathrm{U}(\omega)]] \text{ for } k \in \mathbb{N}_2. \tag{4.4}$$

REMARK 4.1 (Complex statistical moments). *Statistical moments for complex random fields induce several quantities of interest. As introduced in (Eriksson, Ollila, & Koivunen, 2010, Section V-A), for $k \in \mathbb{N}_2$ and $\mathrm{U} \in L^k(\Omega, \mathbb{P}; X)$, the kth statistical moments are defined as*

$$\alpha_{p;q} \equiv \alpha_{p;q}[\mathrm{U}(\omega)] := \mathbb{E}[\mathrm{U}^p \overline{\mathrm{U}}^q], \text{ for } p, q \in \mathbb{N}_0, \text{ such that } p + q = k. \tag{4.5}$$

Notice that symmetric moments are redundant, i.e. $\alpha_{p;q} = \overline{\alpha_{q;p}}$. *Also, for* $k = 2$, *complex moments are the pseudo-covariance* $\alpha_{2;0}$ *and covariance* $\alpha_{1;1} = \mathbb{E}[\mathrm{U}\overline{\mathrm{U}}]$. *In this chapter, we focus on*

$$\mathcal{M}^k[\mathrm{U}] = \alpha_{k;0} = \overline{\alpha_{0;k}}.$$

However, some applications involve other choices for p, q. *Still, our analysis applies verbatim to* $\alpha_{p;k}$ *up to conjugation of terms in the tensor deterministic formulation in Section 4.4.2 (see, for instance, (Silva-Oelker et al., 2018) for* $k = 2$*).*

We consider a centered random *velocity field* field $\mathbf{v} \in L^k(\Omega, \mathbb{P}; W^{2,\infty}(\Gamma; \mathbb{R}^d))$, i.e. such that $\mathbb{E}[\mathbf{v}(\cdot, \omega)] = \mathbf{0}$. Also, assume $\|\mathbf{v}(\cdot, \omega)\|_{W^{2,\infty}(\Gamma)} \lesssim 1$ uniformly for all $\omega \in \Omega$ and introduce a family of random surfaces $\{\Gamma_t\}_t$ via the mapping

$$\Omega \ni \omega \mapsto \Gamma_t(\omega) = \{\mathbf{x} + t\mathbf{v}(\mathbf{x}, \omega),\ \mathbf{x} \in \Gamma\} =: T_t(\Gamma)(\omega). \tag{4.6}$$

Following (Jerez-Hanckes & Schwab, 2016), we deduce that there exists $\varepsilon > 0$ such that, for each $|t| < \varepsilon$ and $\mathbb{P}$-a.s. ω, the collection $\{\Gamma_t(\omega)\}$ generates bi-Lipschitz diffeomorphisms and induces connected Lipschitz domain $D_t(\omega)$ by continuity of $\mathbf{v}(\omega)$ $\mathbb{P}$-a.s. on the compact surface Γ. Besides, we define $\mathfrak{D}_t(\omega)$ corresponding to either $D_t^c(\omega)$ or $D_t^c(\omega) \cup D_t(\omega)$ according to the problem considered. Finally, we notice that $(\mathbf{v}(\mathbf{x}, \omega) \cdot \mathbf{n}) \in W^{2,\infty}(\Gamma)$.

4.2.4. First-order approximations

With the domain transformation and velocity field defined in Section 4.2.3, we are ready to introduce the concept of random SD.

Definition 4.1 (Random SD (Harbrecht et al., 2008))**.** *For* $\omega \in \Omega$, *consider a random shape dependent scalar field* $\mathrm{U}_t(\omega)$ *defined in* $\mathfrak{D}_t(\omega)$ *for* $|t| < \varepsilon$ *and denote* $\mathrm{U} \equiv \mathrm{U}_0$ *for the nominal domain solution.* $\mathrm{U}_t(\omega)$ *is said to admit a SD* $\mathrm{U}'(\omega)$ *in* $\mathfrak{D}$ *along* $\mathbf{v}(\omega)$ *if the following (pointwise) limit exists*

$$\mathrm{U}'(\omega, \mathbf{x}) := \lim_{t \to 0} \frac{\mathrm{U}_t(\omega, \mathbf{x}) - \mathrm{U}(\mathbf{x})}{t}, \quad \mathbf{x} \in \mathfrak{D}_t(\omega) \cap \mathfrak{D}. \tag{4.7}$$

Assuming SD in Definition 4.1 belongs to $H^1_{\text{loc}}(\mathfrak{D})$ and a Lipschitz condition, then the following Taylor expansion holds for $|t| < \varepsilon$:

$$\mathrm{U}_t(\omega) = \mathrm{U} + t\mathrm{U}'(\omega) + \mathcal{O}(t^2) \text{ in } H^1(Q(\omega)),\, Q(\omega) \Subset \mathfrak{D} \cap \mathfrak{D}_t(\omega). \tag{4.8}$$

Finally, following (Dölz & Harbrecht, 2018, Section 2.1), we introduce K such that

$$K \Subset \mathfrak{D}_t^{\cap\Omega},\ \mathfrak{D}_t^{\cap\Omega} := \bigcap_{\omega\in\Omega} \mathfrak{D}_t(\omega). \tag{4.9}$$

Consequently, according to (4.8) and using the embedding arguments of (Dölz & Harbrecht, 2018, Lemma 5.9) for the variance, the quantities of interest can be accurately approximated for $k \geq 2$ by

$$\begin{aligned}
\mathbb{E}[\mathrm{U}_t(\omega)] &= \mathrm{U} + \mathcal{O}(t^2), && \text{in } H^1(K),\\
\mathcal{M}^k[\mathrm{U}_t(\omega) - \mathrm{U}] &= t^k\mathcal{M}^k[\mathrm{U}'(\omega)] + \mathcal{O}(t^{k+1}), && \text{in } H^1(K)^{(k)}, \text{ and}\\
\mathbb{V}^k[\mathrm{U}_t(\omega)] &= t^k \operatorname{diag}\mathcal{M}^k[\mathrm{U}'(\omega)] + \mathcal{O}(t^{k+1}), && \text{in } L^2(K).
\end{aligned} \tag{4.10}$$

Hence, for a random class of parametrized perturbations (see (4.6)), the statistical moments (refer to (4.4)) can be approximated accurately through U, $\mathcal{M}^k[\mathrm{U}'(\omega)]$, defined in $\mathfrak{D}$ and $\mathfrak{D}^{(k)}$: the FOA amounts to computing U and $\mathcal{M}^k[\mathrm{U}'(\omega)]$. Before proceeding, we decide to sum up the main points of the FOSB method in Figure 4.1. It describes the path followed throughout and, for each step, details the related section and the quantity of interest considered. The technique is sequential from top to bottom, and between each step we use arrows specify whether an approximation is done or an equivalent formulation is used. Notice that the two "equivalent" steps enclose the operations realized on the boundary of the nominal scatterer.

4.3. Deterministic Helmholtz scattering problems

Let us now describe the Helmholtz problems considered in two and three dimensions. We characterize physical domains by a positive bounded wave speed c and a material density constant μ –representing, for instance, the permeability in electromagnetics. For time-harmonic excitations of angular frequency $\omega > 0$, set the wavenumber $\kappa := \omega/c$ and define

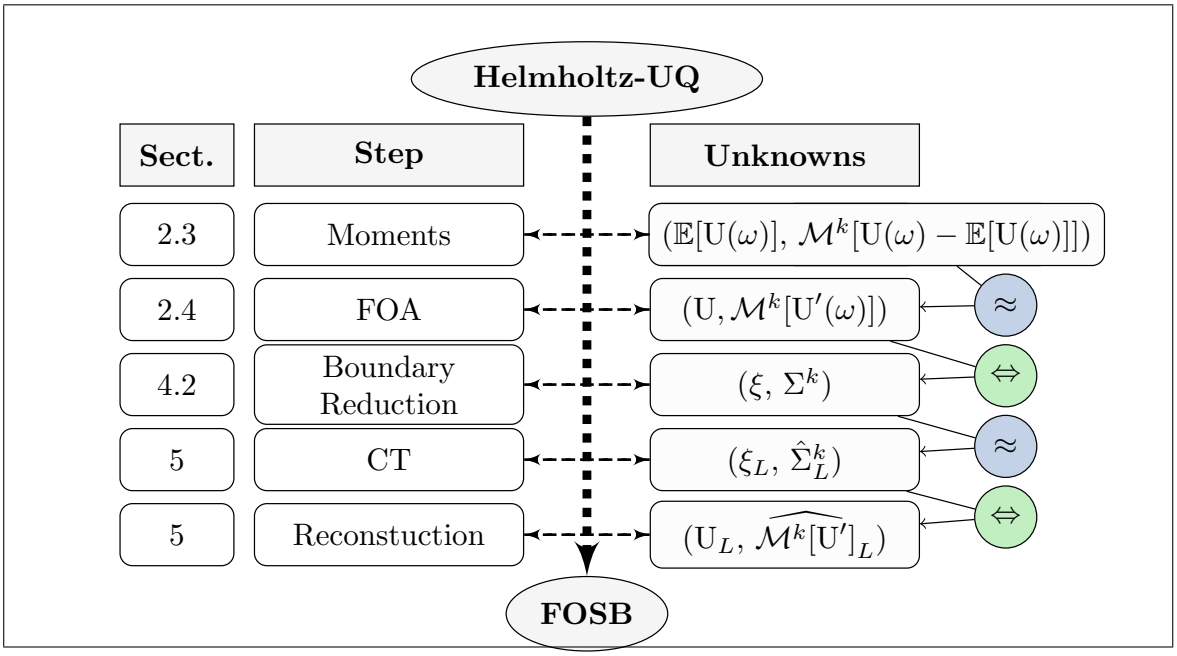


FIGURE 4.1. Sequential description of the FOSB. For each step, we detail the related section and unknwowns and precise whether it consists in an approximation (blue) or an equivalent (green) one.

the Helmholtz operator:

$$L_\kappa : \mathrm{U} \mapsto -\Delta \mathrm{U} - \kappa^2 \mathrm{U}.$$

The Sommerfeld radiation condition (SRC) (Nédélec, 2001, Section 2.2) for U defined over D^c and κ reads

$$\mathrm{SRC}(\mathrm{U},\kappa) \iff \left|\frac{\partial}{\partial r}\mathrm{U} - \imath\kappa\mathrm{U}\right| = o\left(r^{\frac{1-d}{2}}\right) \text{ for } r := \|\mathbf{x}\|_2 \to \infty, \tag{4.11}$$

for $d = 2, 3$. This condition will guarantee uniqueness of solutions Section 4.3.1 and the definition of $\mathrm{F} \in C^\infty(\mathbb{S}^{d-1})$ the far-field (Chandler-Wilde et al., 2012, Lemma 2.5) such that

$$\left|\mathrm{U} - \exp(\imath\kappa r) r^{\frac{1-d}{2}} \mathrm{F}(\mathbf{x}/r)\right| = \mathcal{O}\left(r^{-\frac{1+d}{2}}\right) \text{ for } r := \|\mathbf{x}\|_2 \to \infty. \tag{4.12}$$

REMARK 4.2 (Far-field Taylor expansions). *As the far-field does not depend on domain transformations, the Taylor expansion for* U_t, U *and* U' *in* (4.10) *transfers straightly to* F_t, F *and* F'*, respectively.*

4.3.1. Problem Formulations

We introduce the following BVPs corresponding to Dirichlet, Neumann, impedance and transmission BCs. By linearity, we can write the total wave as a sum of scattered and incident ones, i.e. $\mathrm{U} = \mathrm{U}^{\mathrm{sc}} + \mathrm{U}^{\mathrm{inc}}$. Notation ($\mathrm{P}_{\boldsymbol{\beta}}$) with bold $\boldsymbol{\beta}$ will refer to any of the problems considered. For the sake of clarity, we summarize these notations and illustrate domain perturbations in Figure 4.2.

PROBLEM 1 (P_β) ($\beta = 0, 1, 2$). *Given* $\kappa > 0$ *and* $\mathrm{U}^{\mathrm{inc}} \in H^1_{\mathrm{loc}}(D^c)$ *with* $L_\kappa \mathrm{U}^{\mathrm{inc}} = 0$ *in* D^c, *we seek* $\mathrm{U} \in H^1_{\mathrm{loc}}(D^c)$ *such that*

$$\begin{cases} \Delta \mathrm{U} + \kappa^2 \mathrm{U} = 0 & \text{in } D^c, \\ \gamma_\beta \mathrm{U} = 0 & \text{on } \Gamma, \quad \text{if } \beta \in \{0, 1\}, \text{ or} \\ \gamma_1 \mathrm{U} + \imath\eta\gamma_0 \mathrm{U} = 0,\ \eta > 0 & \text{on } \Gamma, \quad \text{if } \beta = 2, \\ \mathrm{SRC}(\mathrm{U}^{\mathrm{sc}}, \kappa). \end{cases}$$

PROBLEM 2 (P_3). *Let* $\kappa_i, \mu_i > 0$, $i = 0, 1$, *with either* $\kappa_0 \neq \kappa_1$ *or* $\mu_0 \neq \mu_1$, *and* $\mathrm{U}^{\mathrm{inc}} \in H^1_{\mathrm{loc}}(D^c)$ *with* $L_{\kappa_0} \mathrm{U}^{\mathrm{inc}} = 0$ *in* D^c. *We seek* $(\mathrm{U}^0, \mathrm{U}^1) \in H^1_{\mathrm{loc}}(D^c) \cup H^1(D)$ *such that*

$$\begin{cases} \Delta \mathrm{U}^i + \kappa_i^2 \mathrm{U}^i = 0 & \text{in } D^i, \text{ for } i = 0, 1, \\ [\gamma_0 \mathrm{U}]_\Gamma = 0 & \text{on } \Gamma, \\ [\mu^{-1} \gamma_1 \mathrm{U}]_\Gamma = 0 & \text{on } \Gamma, \\ \mathrm{SRC}(\mathrm{U}^{\mathrm{sc}}, \kappa_0). \end{cases}$$

Exterior problems (P_β), $\beta = 0, 1$, represent the sound-soft and -hard acoustic wave scattering while (P_2) and (P_3) describe the *exterior impedance* and *transmission* problems, respectively. Notice that ($\mathrm{P}_{\boldsymbol{\beta}}$) is known to be well posed (McLean, 2000, Chapter 4).

4.3.2. Shape derivatives for Helmholtz scattering problems

We summarize the BVPs, denoted by ($\mathrm{SP}_{\boldsymbol{\beta}}$), satisfied by the SD for each BC, as detailed in (Hiptmair & Li, 2017, Table 5.6).

β	Problem	$\mathfrak{D}$	BCs
0	Sound-soft	D^c	$\gamma_0 \mathrm{U} = 0$
1	Sound-hard	D^c	$\gamma_1 \mathrm{U} = 0$
2	Impedance	D^c	$\gamma_1 \mathrm{U} + \imath\eta\gamma_0 \mathrm{U} = 0$
3	Transmission	$D^c \cup D$	$[\gamma_0 \mathrm{U}]_\Gamma = [\mu^{-1}\gamma_1 \mathrm{U}]_\Gamma = 0$

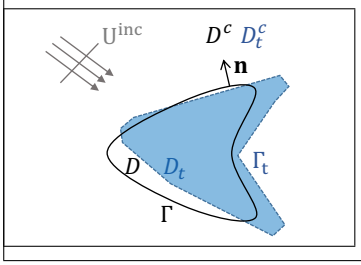


FIGURE 4.2. Overview of ($\mathrm{P}_{\boldsymbol{\beta}}$) (left) and representation of domain transformations (right).

PROBLEM 3 (SP_β) ($\beta = 0, 1, 2$). *We seek* $\mathrm{U}' \in H^1_{\mathrm{loc}}(D^c)$ *solution of*

$$\begin{cases} \Delta \mathrm{U}' + \kappa^2 \mathrm{U}' = 0 & \text{in } D^c, \\ \gamma_\beta \mathrm{U}' = g_\beta & \text{on } \Gamma, \quad \text{if } \beta \in \{0, 1\}, \text{ or} \\ \gamma_1 \mathrm{U}' + \imath\eta\gamma_0 \mathrm{U}' = g_2, \ \eta > 0 & \text{on } \Gamma, \quad \text{if } \beta = 2, \\ \mathrm{SRC}(\mathrm{U}', \kappa), \end{cases}$$

wherein, for U *being the respective solution of* (P_β)*, we have*

$$g_0 := -\gamma_1 \mathrm{U}(\mathbf{v} \cdot \mathbf{n}),$$

$$g_1 := \mathrm{div}_\Gamma \left((\mathbf{v} \cdot \mathbf{n}) \nabla_\Gamma \mathrm{U}\right) + \kappa^2 \gamma_0 \mathrm{U}(\mathbf{v} \cdot \mathbf{n}),$$

$$g_2 := \mathrm{div}_\Gamma \left((\mathbf{v} \cdot \mathbf{n}) \nabla_\Gamma \mathrm{U}\right) + \kappa^2 \gamma_0 \mathrm{U}(\mathbf{v} \cdot \mathbf{n}) + \imath\eta(\mathbf{v} \cdot \mathbf{n})(-\gamma_1 \mathrm{U} - \mathfrak{H}\gamma_0 \mathrm{U}).$$

PROBLEM 4 (SP_3). *We seek* $\mathrm{U}' = (\mathrm{U}'^0, \mathrm{U}'^1) \in H^1_{\mathrm{loc}}(D^c) \times H^1(D)$ *solution of*

$$\begin{cases} \Delta \mathrm{U}'^i + \kappa_i^2 \mathrm{U}'^i & = 0 \quad \text{in } D^i, \text{ for } i = 0, 1, \\ [\gamma_0 \mathrm{U}']_\Gamma & = h_0 \quad \text{on } \Gamma, \\ [\frac{1}{\mu}\gamma_1 \mathrm{U}']_\Gamma & = h_1 \quad \text{on } \Gamma, \\ \mathrm{SRC}(\mathrm{U}'^0, \kappa_0), & \end{cases}$$

with boundary data built using U *solution of (*P_3*), as follows*

$$h_0 := -[\gamma_1 \mathrm{U}]_\Gamma (\mathbf{v} \cdot \mathbf{n}),$$
$$h_1 := \left[\frac{1}{\mu}\right]_\Gamma \mathrm{div}_\Gamma \left((\mathbf{v} \cdot \mathbf{n}) \nabla_\Gamma \mathrm{U}\right) + [\kappa^2]_\Gamma \gamma_0 \mathrm{U} (\mathbf{v} \cdot \mathbf{n}).$$

In the proposed setting, ($\mathrm{SP}_{\boldsymbol{\beta}}$) is known to be well posed (*cf.* (Hiptmair & Li, 2017, Section 3.2)).

Having described the deterministic problems, we now consider the random domains described Section 4.2.3 and analyze, for each realization $\mathrm{U}_t(\omega)$, solutions of ($\mathrm{P}_{\boldsymbol{\beta}}$). The prior choice of random domains ensures wellposedness of the perturbed solution $\mathrm{U}_t(\omega)$ and of its shape derivative $\mathrm{U}'(\omega)$ for each realization. Therefore, we apply the FOA framework of Section 4.2.4 to $\mathrm{U}_t(\omega)$, allowing to obtain an accurate approximation of the statistical moments of $\mathrm{U}_t(\omega)$ through:

$$\mathrm{U} \text{ and } \mathcal{M}^k[\mathrm{U}'(\omega)],$$

defined over $\mathcal{D}$ and $\mathcal{D}^{(k)}$, respectively –check step 2 in Figure 4.1. In the same spirit as in (Dölz & Harbrecht, 2018), the domain and perturbations considered allow for a bounded shape Hessian in $H^1_{\mathrm{loc}}(\mathcal{D})$, hence the Lipschitz condition for the SD. As these domains are unbounded, we reduce the problem to the boundary Γ via BIEs. Notice that the randomness in $\mathrm{U}'(\omega)$ appears only through $(\mathbf{v} \cdot \mathbf{n})(\omega)$, which appears solely in BCs.

4.4. Boundary Reduction

In this section, we explain how to reduce the Helm-holtz boundary value problems described before as well as their SDs onto the boundary via the integral representation formula. Then, we consider the small random domain counterparts and show how the SDs are equivalently reduced to BIEs comprising deterministic operators with stochastic right-hand side. As mentioned initially, this will fit the general framework described in (von Petersdorff & Schwab, 2006) to compute statistical moments.

4.4.1. Boundary integral operators in scattering theory

First, we define the duality product between $\boldsymbol{\xi}_1 = (\lambda_1, \sigma_1)$ and $\boldsymbol{\xi}_2 = (\lambda_2, \sigma_2)$ both in the Cartesian product space $H^{1/2}(\Gamma) \times H^{-1/2}(\Gamma)$:

$$\langle \boldsymbol{\xi}_1, \boldsymbol{\xi}_2 \rangle_\Gamma := \langle \lambda_1, \sigma_1 \rangle_\Gamma + \langle \lambda_2, \sigma_2 \rangle_\Gamma.$$

Recall the fundamental solution $G_\kappa(\mathbf{x}, \mathbf{y})$ of the Helmholtz equation for $\kappa > 0$:

$$G_\kappa(\mathbf{x}, \mathbf{y}) := \begin{cases} \frac{\imath}{4} H_0^{(1)}(\kappa \|\mathbf{x} - \mathbf{y}\|_2) & \text{for } d = 2, \\ \frac{\imath}{4\pi} \frac{\exp(\imath\kappa \|\mathbf{x} - \mathbf{y}\|_2)}{\|\mathbf{x} - \mathbf{y}\|_2} & \text{for } d = 3, \end{cases} \tag{4.13}$$

where $H_0^{(1)}$ is the zeroth-order Hankel function of the first kind. With this, we introduce the single- and double-layer potentials for $\phi \in L^1(\Gamma)$:

$$\begin{aligned} \mathsf{SL}_\kappa(\phi)(\mathbf{x}) &:= \int_\Gamma G_\kappa(\mathbf{x} - \mathbf{y}) \phi(\mathbf{y}) d\Gamma(\mathbf{y}) && \mathbf{x} \in \mathbb{R}^d \backslash \Gamma, \\ \mathsf{DL}_\kappa(\phi)(\mathbf{x}) &:= \int_\Gamma \frac{\partial}{\partial \mathbf{n}_\mathbf{y}} G_\kappa(\mathbf{x} - \mathbf{y}) \phi(\mathbf{y}) d\Gamma(\mathbf{y}) && \mathbf{x} \in \mathbb{R}^d \backslash \Gamma. \end{aligned}$$

With this at hand, we introduce the block Green's potential:

$$\mathsf{R}_\kappa := (\mathsf{DL}_\kappa, -\mathsf{SL}_\kappa),$$

and for the sake of convenience, we identify implicitly the Cauchy data (4.1) with the domain index for $\beta = 2$, i.e.:

$$\mathsf{R}_\kappa(\boldsymbol{\xi})(\mathbf{x}) \equiv \mathsf{R}_{\kappa_i}(\boldsymbol{\xi}^i)(\mathbf{x}), \text{ if } \mathbf{x} \in D^i,\ i = 0, 1. \tag{4.14}$$

The identity operator Id and five continuous boundary integral operators in Lipschitz domains for $\kappa > 0$, $\eta > 0$ and $|s| \leq 1$ (Chandler-Wilde et al., 2012, Theorems 2.25 and 2.26):

$$\begin{aligned}
\mathsf{V}_\kappa &: H^{s-1/2}(\Gamma) \to H^{s+1/2}(\Gamma), \quad \mathsf{V}_\kappa := \{\!\{\gamma_0\}\!\}_\Gamma \circ \mathsf{SL}_\kappa, \\
\mathsf{K}_\kappa &: H^{s+1/2}(\Gamma) \to H^{s+1/2}(\Gamma), \quad \mathsf{K}_\kappa := \{\!\{\gamma_0\}\!\}_\Gamma \circ \mathsf{DL}_\kappa, \\
\mathsf{K}'_\kappa &: H^{s-1/2}(\Gamma) \to H^{s-1/2}(\Gamma), \quad \mathsf{K}'_\kappa := \{\!\{\gamma_1\}\!\}_\Gamma \circ \mathsf{SL}_\kappa, \\
\mathsf{W}_\kappa &: H^{s+1/2}(\Gamma) \to H^{s-1/2}(\Gamma), \quad \mathsf{W}_\kappa := -\{\!\{\gamma_1\}\!\}_\Gamma \circ \mathsf{DL}_\kappa, \\
\mathsf{B}'_{\kappa,\eta} &: H^{s+1/2}(\Gamma) \to H^{s-1/2}(\Gamma), \quad \mathsf{B}'_{\kappa,\eta} := \mathsf{W}_\kappa - \imath\eta \left(\tfrac{1}{2}\mathsf{Id} + \mathsf{K}'_\kappa\right).
\end{aligned} \tag{4.15}$$

Also, we introduce the following operator:

$$\mathsf{A}_\kappa := \begin{bmatrix} -\mathsf{K}_\kappa & \mathsf{V}_\kappa \\ \mathsf{W}_\kappa & \mathsf{K}'_\kappa \end{bmatrix},$$

along with

$$\widehat{\mathsf{A}}_{\kappa,\mu} := \begin{bmatrix} 1 & 0 \\ 0 & 1/\mu \end{bmatrix} \begin{bmatrix} -\mathsf{K}_\kappa & \mathsf{V}_\kappa \\ \mathsf{W}_\kappa & \mathsf{K}'_\kappa \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & \mu \end{bmatrix} = \begin{bmatrix} -\mathsf{K}_\kappa & \mu\mathsf{V}_\kappa \\ 1/\mu\mathsf{W}_\kappa & \mathsf{K}'_\kappa \end{bmatrix}.$$

Next, we consider a *radiating solution* U, i.e. $\mathsf{U} = (\mathsf{U}^0, \mathsf{U}^1)$ such that $L_\kappa \mathsf{U}^i = 0, i = 0, 1$, and U^0 with Sommerfeld radiation conditions (Sauter & Schwab, 2010, Section 3.6). Therefore, the following representation formula holds:

$$\mathsf{U} = \mathsf{DL}_\kappa([\gamma_0 \mathsf{U}]_\Gamma) - \mathsf{SL}_\kappa([\gamma_1 \mathsf{U}]_\Gamma) = \mathsf{R}_\kappa([\boldsymbol{\xi}]_\Gamma) \text{ in } D^0 \cup D^1. \tag{4.16}$$

Its Cauchy data $\boldsymbol{\xi}^i = (\gamma_0 \mathrm{U}^i, \gamma_1 \mathrm{U}^i) =: (\lambda^i, \sigma^i) \in H^{1/2}(\Gamma) \times H^{-1/2}(\Gamma)$ satisfy

$$\begin{aligned} &\text{(Interior)} \quad \boldsymbol{\xi}^1 = \left(\tfrac{1}{2}\mathsf{Id} + \mathsf{A}_\kappa\right) \boldsymbol{\xi}^1 =: \mathsf{P}^1_\kappa \boldsymbol{\xi}^1, \\ &\text{(Exterior)} \quad \boldsymbol{\xi}^0 = \left(\tfrac{1}{2}\mathsf{Id} - \mathsf{A}_\kappa\right) \boldsymbol{\xi}^0 =: \mathsf{P}^0_\kappa \boldsymbol{\xi}^0 = (\mathsf{Id} - \mathsf{P}^1_\kappa)\boldsymbol{\xi}^0. \end{aligned} \tag{4.17}$$

Notice that the above identities are also valid for $\widehat{\mathsf{A}}_{\kappa_0,\mu}$ and $\widehat{\mathsf{A}}_{\kappa_1,\mu}$. Operators $\mathsf{P}^0_\kappa, \mathsf{P}^1_\kappa$ are dubbed exterior and interior *Calderón projectors*. They share the interesting property that for $i = 0, 1$, $(\mathsf{P}^i_\kappa)^2 = \mathsf{I}$, allowing for Calderón-based operator preconditioning (see Section 4.6.2).

Lastly, we introduce $S_{\text{Dir}}(D) \equiv S_0(D)$ and $S_{\text{Neum}}(D) \equiv S_1(D)$ the countable set accumulating only at infinity of strictly positive eigenvalues of Helmholtz problem with homogeneous Dirichlet and Neumann boundary conditions (Sauter & Schwab, 2010, Section 3.9.2).

4.4.2. Tensor BIEs

We now show that the FOA analysis for (P_β) can be reduced to (B_β) defined further in Generic Problem 3, consisting in two deterministic first kind BIEs including a tensor one (refer to Figure 4.1). To begin with, we consider both deterministic problems (P_β) and (SP_β), and show that they can be reduced to two wellposed BIEs of the form:

GENERIC PROBLEM 1 (Deterministic BIEs). *Provided* $\mathsf{Z} \in \mathcal{L}(X, Y)$ *and* $\mathsf{B} \in \mathcal{L}(Y, Y)$ *for separable Hilbert spaces* X, Y, $f \in Y$ *and* $g \in Y$, *we seek* $\xi, \xi' \in X$ *such that:*

$$\begin{cases} \mathsf{Z}\xi &= f \quad \text{on } \Gamma, \\ \mathsf{Z}\xi' &= \mathsf{B}g \quad \text{on } \Gamma. \end{cases} \tag{4.18}$$

Equivalence between problems couple ((P_β),(SP_β)) and Generic Problem 1 is derived through the following steps:

(i) Using Section 4.4.1, we perform the boundary reduction for (P_β) and (SP_β), leading to Generic Problem 1 (see Section 4.9);

(ii) We prove that Generic Problem 1 is well-posed in adapted Sobolev spaces using Fredholm theory. Notice that we remove the spurious eigenvalues to guarantee injective boundary integral operators.

Step (ii) is extensively surveyed for $\beta = 0, 1, 2$ in (Chandler-Wilde & Monk, 2008); see Table 2.1 and Theorem 2.25 for $\beta = 0, 1$, and refer to Section 2.6 for $\beta = 2$. The transmission problem ($\beta = 2$) is analyzed in (Claeys, Hiptmair, & Jerez-Hanckes, 2012, Section 3).

After reducing the deterministic problem to the boundary, we retake Section 4.2.4 and consider the random counterparts of ($\mathrm{P}_{\boldsymbol{\beta}}$) and ($\mathrm{SP}_{\boldsymbol{\beta}}$), leading to U and $\mathrm{U}'(\omega)$, and perform correspondingly the boundary reduction. The generic random BIEs for the SD read:

GENERIC PROBLEM 2 (Random BIEs). *Provided* $\mathsf{Z} \in \mathcal{L}(X, Y)$ *and* $\mathsf{B} \in \mathcal{L}(Y, Y)$ *for separable Hilbert spaces* X, Y, $g \in L^k(\Omega, \mathbb{P}; Y)$, *for* $k \in \mathbb{N}_2$, *we seek* $\xi' \in L^k(\Omega, \mathbb{P}; X)$ *such that*

$$\mathsf{Z}\xi = \mathsf{B}g \text{ on } \Gamma. \tag{4.19}$$

Applying Theorem 6.1 in (von Petersdorff & Schwab, 2006), we deduce that the tensor operator equation admits a unique solution $\xi' \in L^k(\Omega, \mathbb{P}; Y)$ and that $\mathcal{M}^k[\xi'(\omega)] \in Y$. Therefore, we arrive at a tensor BIE with stochastic right-hand sides, providing the final form of the wellposed deterministic tensor operator BIEs ($\mathrm{B}_{\boldsymbol{\beta}}$):

GENERIC PROBLEM 3 ($\mathrm{B}_{\boldsymbol{\beta}}$) (Formulation for the BIEs). *Given* $\mathsf{Z} \in \mathcal{L}(X, Y)$, $\mathsf{B} \in \mathcal{L}(Y, Y)$ *for separable Hilbert spaces* X, Y, $k \in \mathbb{N}_2$, $f \in Y$ *and* $\mathcal{M}^k[g] \in Y^{(k)}$, *seek* $\xi \in X, \Sigma^k \in X^{(k)}$ *such that:*

$$\begin{cases} \mathsf{Z}\xi & = f & \text{on } \Gamma, \\ \mathsf{Z}^{(k)}\Sigma^k & = \mathsf{B}^{(k)}\mathcal{M}^k[g] & \text{on } \Gamma^{(k)}. \end{cases} \tag{4.20}$$

We now detail the resulting sets of BIEs for each problem ($\mathrm{B}_{\boldsymbol{\beta}}$) as well as for their statistical moments and their related potential reconstruction. As in (von Petersdorff &

Schwab, 2006, Section 6.2), notice that the statistical moments and the layer potentials commute by Fubini's theorem.

PROBLEM 5 (B_0). *If* $\kappa^2 \notin S_{\mathrm{Dir}}(D)$, $\gamma_0\mathrm{U}^{\mathrm{inc}} \in H^{1/2}(\Gamma)$ *and* $\mathcal{M}^k[g_0] \in H^{1/2}(\Gamma)^{(k)}$, $k \in \mathbb{N}_2$, *we seek* $\gamma_1\mathrm{U} \in H^{-1/2}(\Gamma)$ *and* $\mathcal{M}^k[\gamma_1\mathrm{U}'] \in H^{-1/2}(\Gamma)^{(k)}$ *such that:*

$$\begin{cases} \mathsf{V}_\kappa\gamma_1\mathrm{U} & = \gamma_0\mathrm{U}^{\mathrm{inc}} & \text{on } \Gamma, \\ \mathsf{V}_\kappa^{(k)}\mathcal{M}^k[\gamma_1\mathrm{U}'] & = \left(-\frac{1}{2}\mathsf{Id} + \mathsf{K}_\kappa\right)^{(k)}\mathcal{M}^k[g_0] & \text{on } \Gamma^{(k)}. \end{cases} \tag{4.21}$$

Then,

$$\begin{aligned} \mathrm{U} &= \mathrm{U}^{\mathrm{inc}} - \mathsf{SL}_\kappa\gamma_1\mathrm{U} \text{ in } D^c, \\ \mathcal{M}^k[\mathrm{U}'] &= \mathcal{M}^k[-\mathsf{SL}_\kappa\gamma_1\mathrm{U}' + \mathsf{DL}_\kappa g_0] \text{ in } (D^c)^{(k)} \\ &= \mathsf{R}_\kappa^{(k)}\mathcal{M}^k[(g_0, \gamma_1\mathrm{U}')] \text{ in } (D^c)^{(k)}. \end{aligned}$$

PROBLEM 6 (B_1). *If* $\kappa^2 \notin S_{\mathrm{Neum}}(D)$, $\gamma_1\mathrm{U}^{\mathrm{inc}} \in H^{-1/2}(\Gamma)$ *and* $\mathcal{M}^k[g_1] \in H^{-1/2}(\Gamma)^{(k)}$, $k \in \mathbb{N}_2$, *we seek* $\gamma_0\mathrm{U} \in H^{1/2}(\Gamma)$ *and* $\mathcal{M}^k[\gamma_0\mathrm{U}'] \in H^{1/2}(\Gamma)^{(k)}$ *such that:*

$$\begin{cases} \mathsf{W}_\kappa\gamma_0\mathrm{U} & = \gamma_1\mathrm{U}^{\mathrm{inc}} & \text{on } \Gamma, \\ \mathsf{W}_\kappa^{(k)}\mathcal{M}^k[\gamma_0\mathrm{U}'] & = \left(-\left(\frac{1}{2}\mathsf{Id} + \mathsf{K}'_\kappa\right)\right)^{(k)}\mathcal{M}^k[g_1] & \text{on } \Gamma^{(k)}. \end{cases} \tag{4.22}$$

Also,

$$\begin{aligned} \mathrm{U} &= \mathrm{U}^{\mathrm{inc}} + \mathsf{DL}_\kappa\gamma_0\mathrm{U} \text{ in } D^c, \\ \mathcal{M}^k[\mathrm{U}'] &= \mathcal{M}^k[-\mathsf{SL}_\kappa\gamma_1\mathrm{U}' + \mathsf{DL}_\kappa g_0] \text{ in } (D^c)^{(k)} \\ &= \mathsf{R}_\kappa^{(k)}\mathcal{M}^k[(\gamma_0\mathrm{U}', g_1)] \text{ in } (D^c)^{(k)}. \end{aligned}$$

PROBLEM 7 (B_2). *If* $\kappa^2 \notin S_{\mathrm{Neum}}(D)$, $\gamma_1\mathrm{U}^{\mathrm{inc}} \in H^{-1/2}(\Gamma)$ *and* $\mathcal{M}^k[g_2] \in H^{-1/2}(\Gamma)^{(k)}$, $k \in \mathbb{N}_2$, *we seek* $\gamma_0\mathrm{U} \in H^{1/2}(\Gamma)$ *and* $\mathcal{M}^k[\gamma_0\mathrm{U}'] \in H^{1/2}(\Gamma)^{(k)}$ *such that:*

$$\begin{cases} \mathsf{B}'_{\kappa,\eta}\gamma_0\mathrm{U} & = \gamma_1\mathrm{U}^{\mathrm{inc}} & \text{on } \Gamma, \\ (\mathsf{B}'_{\kappa,\eta})^{(k)}\mathcal{M}^k[\gamma_0\mathrm{U}'] & = \left(\frac{1}{2}\mathsf{Id} + \mathsf{K}'_\kappa\right)^{(k)}\mathcal{M}^k[g_2] & \text{on } \Gamma^{(k)}. \end{cases} \tag{4.23}$$

β	Problem	X^s	Y^s	Z	C
0	Soft	$H^{-1/2+s}(\Gamma)$	$H^{1/2+s}(\Gamma)$	V_κ	W_κ
1	Hard	$H^{1/2+s}(\Gamma)$	$H^{-1/2+s}(\Gamma)$	W_κ	V_κ
2	Impedance	$H^{1/2+s}(\Gamma)$	$H^{-1/2+s}(\Gamma)$	$\mathsf{W}_\kappa - \imath\eta\left(\frac{1}{2}\mathsf{Id} + \mathsf{K}'_\kappa\right)$	V_κ
3	Transmission	$H^{1/2+s}(\Gamma) \times H^{-1/2+s}(\Gamma)$	$H^{1/2+s}(\Gamma) \times H^{-1/2+s}(\Gamma)$	$(\widehat{\mathsf{A}}_{\kappa_0,\mu_0} + \widehat{\mathsf{A}}_{\kappa_1,\mu_1})$	$(\widehat{\mathsf{A}}_{\kappa_0,\mu_0} + \widehat{\mathsf{A}}_{\kappa_1,\mu_1})$

TABLE 4.1. Overview of the BIEs for (B_β) and associated operator preconditioner employed in Section 4.6.2.

Moreover,

$$\begin{aligned}\mathrm{U} &= \mathrm{U}^{\mathrm{inc}} + (\imath\eta\mathsf{SL}_\kappa + \mathsf{DL}_\kappa)\gamma_0\mathrm{U} \text{ in } D^c,\\ \mathcal{M}^k[\mathrm{U}'] &= \mathcal{M}^k[(\imath\eta\mathsf{SL}_\kappa + \mathsf{DL}_\kappa)\gamma_0\mathrm{U}' - \mathsf{SL}_\kappa g_2] \text{ in } (D^c)^{(k)}\\ &= \mathsf{R}_\kappa^{(k)}\mathcal{M}^k[(\gamma_0\mathrm{U}', g_2 - \imath\eta\gamma_0\mathrm{U}')] \text{ in } (D^c)^{(k)}.\end{aligned}$$

PROBLEM 8 (B_3). *For $\boldsymbol{\xi}^{\mathrm{inc}} := (\gamma_0\mathrm{U}^{\mathrm{inc}}, \gamma_1\mathrm{U}^{\mathrm{inc}}) \in \left[H^{1/2}(\Gamma) \times H^{-1/2}(\Gamma)\right]^{(k)}$ and $\mathcal{M}^k h = \mathcal{M}^k(h_0, h_1) \in (H^{1/2}(\Gamma) \times H^{-1/2}(\Gamma))^{(k)}$, for $k \in \mathbb{N}_2$, we seek $\boldsymbol{\xi}^0 := (\gamma_0\mathrm{U}^0, \gamma_1\mathrm{U}^0) \in H^{1/2}(\Gamma) \times H^{-1/2}(\Gamma)$ and $\mathcal{M}^k[\boldsymbol{\xi}'] \equiv \mathcal{M}^k[\boldsymbol{\xi}'^0] \in (H^{1/2}(\Gamma) \times H^{-1/2}(\Gamma))^{(k)}$ such that:*

$$\begin{cases} \left(\widehat{\mathsf{A}}_{\kappa_0,\mu_0} + \widehat{\mathsf{A}}_{\kappa_1,\mu_1}\right)\boldsymbol{\xi}^0 & = \boldsymbol{\xi}^{\mathrm{inc}} & \text{on } \Gamma,\\ \left(\widehat{\mathsf{A}}_{\kappa_0,\mu_0} + \widehat{\mathsf{A}}_{\kappa_1,\mu_1}\right)^{(k)}\mathcal{M}^k[\boldsymbol{\xi}'] & = \left(\frac{1}{2}\mathsf{Id} + \widehat{\mathsf{A}}_{\kappa_1,\mu_1}\right)^{(k)}\mathcal{M}^k[h] & \text{on } \Gamma^{(k)}. \end{cases} \tag{4.24}$$

Also,

$$\begin{aligned}\mathrm{U}(\mathbf{x}) &= \mathrm{U}^{\mathrm{inc}}(\mathbf{x}) - \mathsf{SL}_{\kappa_0}\gamma_1\mathrm{U}^0 + \mathsf{DL}_{\kappa_0}\gamma_0\mathrm{U}^0,\ \mathbf{x} \in D^c,\\ \mathrm{U}(\mathbf{x}) &= -\mathsf{SL}_{\kappa_1}\gamma_1\mathrm{U}^1 + \mathsf{DL}_{\kappa_1}\gamma_0\mathrm{U}^1,\ \mathbf{x} \in D,\\ \mathcal{M}^k[\mathrm{U}'] &= \mathcal{M}^k[\mathsf{R}_\kappa(\boldsymbol{\xi})] = \mathsf{R}_\kappa^{(k)}\mathcal{M}^k[\boldsymbol{\xi}] \text{ in } \mathcal{D}^{(k)}.\end{aligned}$$

Ultimately, we sum up the functional spaces and BIEs for (PB_β) in Table 4.1. Also, corresponding Sobolev spaces of higher regularity will be denoted X^s, Y^s, for $s \geq 0$, with $Y^0 \equiv Y$ and $X^0 \equiv X$. Operator C refers to the left-preconditioner that is used for operator preconditioning purposes, as detailed later on in Section 4.6.2.

4.5. Galerkin Method and Sparse Tensor Elements

We now aim to solve numerically the variational forms arising from the BIEs described in Generic Problem 3. Let us introduce a nested shape-regular and quasi-uniform family $\{\mathcal{M}_l\}_{l\in\mathbb{N}_0}$ of surface triangulations consisting of triangles or quadrilaterals, with each level l associated to a meshwidth $h_l > 0$. For $\beta = 0, 1$, we define the associated boundary element spaces $V_0^\beta \subset V_1^\beta \subset \cdots V_l^\beta \subset H^{\frac{1}{2}-\beta}(\Gamma)$:

$$V_l^0 := \{\lambda \in C(\Gamma) : \lambda|_K \in \mathsf{P}_p(K),\ \forall K \in \mathcal{M}_l,\ p \in \mathbb{N}_1\},$$
$$V_l^1 := \{\sigma \in L^2(\Gamma) : \sigma|_K \in \mathsf{P}_p(K),\ \forall K \in \mathcal{M}_l,\ p \in \mathbb{N}_0\}.$$

where $\mathsf{P}_p(K)$ stands for the space of polynomials of degree $\leq p$, $p \in \mathbb{N}_0$ on the cell K. Notice that under regular enough Neumann data, i.e. Neumann traces belong to $H^{(d-1)/2+\delta}(\Gamma)$ (Sauter & Schwab, 2010, Theorem 2.5.4) for any $\delta > 0$, we can also use piecewise continuous functions e.g., piecewise linear functions $\mathbb{P}^1$, as in Section 4.7. Afterwards, we introduce usual best approximation estimates for the h-version of boundary elements (Sauter & Schwab, 2010, Chapter 9).

Lemma 4.1 (Interpolation error for Dirichlet traces)**.** *For* $0 \leq t \leq s \leq p+1$ *and all* $\lambda \in H^s(\Gamma)$*, there holds*

$$\inf_{v_l \in V_l^0} \|\lambda - v_l\|_{H^t(\Gamma)} \leq C h^{s-t} \|\lambda\|_{H^s(\Gamma)}, \tag{4.25}$$

where $C > 0$ *is independent of* h *and* λ*.*

Lemma 4.2 (Interpolation error for Neumann traces)**.** *For* $0 \leq t \leq s \leq p+1$ *and all* $\sigma \in H^s(\Gamma)$*, there holds*

$$\inf_{v_l \in V_l^1} \|\sigma - v_l\|_{H^{-t}(\Gamma)} \leq C h^{s+t} \|\sigma\|_{H^s(\Gamma)}, \tag{4.26}$$

where $C > 0$ *is independent of* h *and* λ*.*

4.5.1. First-order statistical moments

Adopting the notation in Table 4.1, we define $X_L \subset X$ from V_L^0 and V_L^1, and arrive at the Galerkin formulation for ξ:

GENERIC PROBLEM 4 (Galerkin formulation). *Seek $\xi_L \in X_L \subset X$ such that:*

$$\langle \mathsf{Z}\xi_L, \phi_L \rangle_\Gamma = \langle f, \phi_L \rangle_\Gamma, \ \forall \phi_L \in X_L. \tag{4.27}$$

We define $N_L := \mathrm{card}(X_L)$. Classical results for coercive operators (Sauter & Schwab, 2010) ensure that there exists a minimum resolution L_0 such that the discrete solution is well defined and converges quasi-optimally in X. Thus, provided that $\xi \in X^s$ for any $0 \leq s \leq p+1$, by Lemmas 4.1 and 4.2 it holds

$$\|\xi - \xi_L\|_X \leq Ch^s \|\xi\|_{X^s}. \tag{4.28}$$

4.5.2. Higher-order statistical moments and CT

Having introduced the tensor L^2-product $\langle \cdot, \cdot \rangle_{\Gamma^{(k)}}$ (Harbrecht et al., 2013), we state the tensor deterministic variational forms of the BIEs:

GENERIC PROBLEM 5 (Tensor Galerkin). *Given $k \in \mathbb{N}_2$, seek $\Sigma_L^k \in X_L^{(k)}$ such that*

$$\langle \mathsf{Z}^{(k)} \Sigma_L^k, \Theta_L^k \rangle_{\Gamma^{(k)}} = \langle \mathsf{B}^{(k)} \mathcal{M}^k[g], \Theta_L^k \rangle_{\Gamma^{(k)}}, \ \forall \Theta_L^k \in X_L^{(k)}. \tag{4.29}$$

As shown in (von Petersdorff & Schwab, 2006, Section 3.5), there is a $L_0(\kappa) \in \mathbb{N}_0$ for which, for all $L \geq L_0(\kappa)$, the tensorized problem admits a discrete inf-sup, and has a unique solution converging quasi-optimally in $X^{(k)}$. From here, we deduce the following error estimates

$$\|\Sigma^k - \Sigma_L^k\|_{X^{(k)}} \leq Ch^s \|\Sigma^k\|_{(X^s)^{(k)}}. \tag{4.30}$$

provided that $\Sigma^k \in (X^s)^{(k)}$, for any $0 \leq s \leq p+1$. Now, we introduce the complement spaces:

$$W_0 := X_0, \ W_l := X_l \backslash X_{l-1}, \ l > 0,$$

and consider the sparse tensor product space:

$$\widehat{X}_L^{(k)}(L_0) = \bigoplus_{\|\underline{l}\|_1 \leq L+(k-1)L_0} W_{l_1} \otimes \cdots \otimes W_{l_k} \tag{4.31}$$

Then, we can state the following stability condition.

Lemma 4.3 ((von Petersdorff & Schwab, 2006, Theorem 5.2)). *For $k \in \mathbb{N}_2$, there exists $L_0(k)$ and $\hat{c}_S$ such that for all $L \geq L_0$, it holds*

$$\inf_{0\neq\hat{\Sigma}\in\widehat{X}_L^{(k)}} \sup_{0\neq\hat{\Theta}\in\widehat{X}_L^{(k)}} \frac{\langle \mathsf{Z}^{(k)}\hat{\Sigma}, \hat{\Theta}\rangle_{\Gamma^{(k)}}}{\|\hat{\Sigma}\|_{X^{(k)}}\|\hat{\Theta}\|_{X^{(k)}}} \geq \frac{1}{\hat{c}_S} > 0. \tag{4.32}$$

Therefore, we deduce that the problem is well posed and we deduce the following convergence error in sparse tensor spaces:

Lemma 4.4 ((von Petersdorff & Schwab, 2006, Theorem 5.3)). *Provided that $\Sigma^k \in (X^s)^{(k)}$ for any $0 \leq s \leq p+1$, the following error bound holds for $L \geq L_0(k)$:*

$$\|\Sigma^k - \hat{\Sigma}_L^k\|_{X^{(k)}} \leq Ch^s |\log h|^{(k-1)/2} \|\Sigma^k\|_{(X^s)^{(k)}}.$$

We solve the Galerkin system in the sparse tensor space applying the CT (Griebel et al., 1990). It consists in solving the full systems for $\underline{l}$ specified in (Harbrecht et al., 2013, Theorem 13) and for associated spaces $X_{\underline{l}}^k$ as described below.

GENERIC PROBLEM 6 (Tensor Galerkin - Subblocks). *Given $k \in \mathbb{N}_2$, seek $\Sigma_{\underline{l}}^k \in X_{\underline{l}}^{(k)}$ such that*

$$\langle (\mathsf{Z}_{l_1} \otimes \cdots \otimes \mathsf{Z}_{l_k})\Sigma_{\underline{l}}^k, \Theta_{\underline{l}}^k\rangle_{\Gamma^{(k)}} = \langle \mathsf{B}^{(k)}\mathcal{M}^k[g], \Theta_{\underline{l}}^k\rangle_{\Gamma^{(k)}}, \ \forall \Theta_{\underline{l}}^k \in X_{\underline{l}}. \tag{4.33}$$

Thus, following (Harbrecht et al., 2013, Lemma 12 and Theorem 13), the Galerkin orthogonality allows to rearrange the solution in the sparse tensor space as

$$\hat{\Sigma}_L^k(L_0) = \sum_{i=0}^{k-1} (-1)^i \binom{k-1}{i} \sum_{\|\underline{l}\|_1 = L+(k-1)L_0-i} \Sigma_{\underline{l}}^k. \tag{4.34}$$

	Dirichlet traces	Neumann traces
Norm	$\|\cdot\|_{H^{1/2}(\Gamma)}$	$\|\cdot\|_{H^{-1/2}(\Gamma)}$
ξ_L	$h^{3/2}$	h^2
Σ_L^k	$h^{3/2}$	h^2
$\hat{\Sigma}_L^k$	$h^{3/2}\|\log h\|^{(k-1)/2}$	$h^2\|\log h\|^{(k-1)/2}$

FOA	First-Order Approximation
SD	Shape Derivative
FOSB	First-Order Sparse Boundary
BIE	Boundary Integral Equation
MC	Monte-Carlo
CT	Combination Technique

Table 4.2 (left): Expected convergence rates for the quantities of interest for $k \in \mathbb{N}_2$ with $\mathbb{P}^1$ discretization and affine meshes. Table 4.3 (right): Non-exhaustive list of acronyms.

The total number of degrees of freedom (dofs) is of order $dofs = \mathcal{O}(N_L \log^{k-1} N_L)$.

Finally, we plug the unknowns $\xi_L, \hat{\Sigma}_L^k$ into the volume reconstruction formulas presented in Section 4.4.2 and obtain the couple

$$\mathrm{U}_L(\mathbf{x}), \text{ and } \widehat{\mathcal{M}^k[\mathrm{U}']}_L(\underline{\mathbf{x}}), \text{ for } \mathbf{x} \in \mathcal{D}, \underline{\mathbf{x}} \in \mathcal{D}^{(k)}, \tag{4.35}$$

being the final approximate delivered by the method.

REMARK 4.3 (Affine meshes). *Meshing by planar surface elements induces a geometrical error, which typically limits the order of convergence of Galerkin BEM to $\mathcal{O}(h^2)$. Following (Sauter & Schwab, 2010, Chapter 8), we present in Table 4.2 the conjectured convergence rates for $\mathbb{P}^1$ discretization with affine meshes for the mean field and two-point covariance for both Neumann and Dirichlet trace counterparts for (B_β).*

4.6. Implementation considerations

In what follows, we aim at understanding several technical aspects related to the implementation of the FOSB scheme.

4.6.1. Symmetric covariance kernels

Consider the case $k = 2$ for a solution $\Sigma^2 \equiv \Sigma$. In most applications, the right-hand side is a symmetric pseudo-covariance kernel, which entails a symmetric solution $\Sigma(\mathbf{x}_1, \mathbf{x}_2) = \Sigma(\mathbf{x}_2, \mathbf{x}_1)$. Therefore, the sparse tensor approximation or the CT allow for a

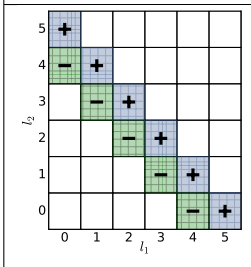
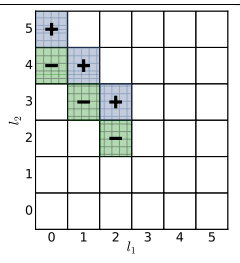
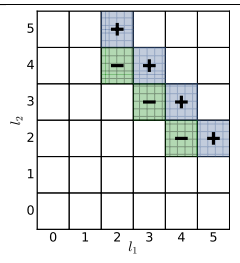
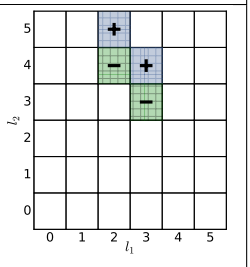
$L = 5$ and $L_0 = 0$		$L = 5$ and $L_0 = 2$	
			
$165,740$	$87,672$	$413,980$	$225,808$

TABLE 4.4. Subspaces used for the classical CT (left of each cell) and the symmetric CT (right of each cell) for $k = 2$. In last row, we detail dofs of each scheme. Notice that $N_L^2 = 595,984$.

two-fold reduction of the dofs for a given accuracy, since for any $l_1, l_2 \in \mathbb{N}_0$, the matrix representation of unknowns reads $\mathbf{\Sigma}_{l_1,l_2} = \mathbf{\Sigma}_{l_2,l_1}^T$, its transpose. We express the latter in Table 4.4, for $(L_0, L) = (0, 5)$ and $(2, 5)$ and for the test case that we detail further in Section 4.7.2, and giving $N_L^2 = 595,984$ in the full tensor space $V_L^{(2)}$, evidencing the efficiency of the CT and the benefits due to symmetry of the solution. Indeed, for $L, L_0 \in \mathbb{N}_0$, $L_0 \geq L$, the CT yields:

$$\hat{\Sigma}_L(L_0) = \sum_{l_1+l_2=L+L_0} \Sigma_{l_1,l_2} - \sum_{l_1+l_2=L+L_0-1} \Sigma_{l_1,l_2}, \tag{4.36}$$

while its symmetric counterpart uses a reduced number of subblock indices:

$$\hat{\Sigma}_L(L_0) = \sum_{\substack{l_1+l_2=L+L_0 \\ l_2<l_1}} (\Sigma_{l_1,l_2} + \Sigma_{l_2,l_1}) - \sum_{\substack{l_1+l_2=L+L_0-1 \\ l_2<l_1}} (\Sigma_{l_1,l_2} + \Sigma_{l_2,l_1})$$
$$+ \Sigma_{(L+L_0)/2} - \Sigma_{(L+L_0)/2-1}, \text{ if } L + L_0 \text{ is odd.}$$

The remark applies identically to complex Hermitian covariance matrices, as $\mathbf{\Sigma}_{l_1,l_2} = \overline{\mathbf{\Sigma}_{l_2,l_1}}^T = \mathbf{\Sigma}_{l_2,l_1}^H$ (see Remark 4.1) and can be directly generalized for higher moments.

4.6.2. Preconditioning

The CT allows to solve smaller subsystems by gathering the operators assembled over distinct levels –on the indices stated in (4.34). It is known that the condition number of

tensor operators grows with the dimension (*cf.* (Griebel & Knapek, 2009, Section 3) and (Fuenzalida et al., 2019)). Hence, the need to precondition with an adapted framework such that the linear systems remains at least mesh independent. We opt to apply operator-based preconditioners such as Calderón preconditioning (Hiptmair, 2006; Escapil-Inchauspé & Jerez-Hanckes, 2019) and assume for $\beta = 0, 1, 2$ that $\kappa^2 \notin \{S_{\text{Dir}}, S_{\text{Neum}}\}$. On each level, we apply the preconditioner C proposed in Table 4.1. We propose the following result for the induced linear system in Lemma 4.5.

Lemma 4.5 (Mesh independence result)**.** *For $k \in \mathbb{N}_1$ and for each $\underline{l} = (l_k)_k$ with $l_k \in \mathbb{N}_0$, $l_k \geq L_0(k)$ such as defined in Section 4.5, the discretized Galerkin system issued from operator $(\mathsf{CZ})^{(k)}$ has a spectral condition number κ_2 independent of the mesh size, i.e. remains bounded as $\|\underline{l}\|_1 \to \infty$.*

PROOF. The result is proved for $k = 1$ in (Hiptmair, 2006) and applies straightforwardly to $k \geq 2$ as the condition number of tensor operators is multiplicative (Fuenzalida et al., 2019). □

This last result shows mesh independence of the numerical scheme, i.e. it guarantees the h-stable (linear) convergence of GMRES (refer to (Galkowski, Müller, & Spence, 2019, Section 4)). Also, as the domains have a Lyapunov boundary, K_κ is compact in both $H^{1/2}(\Gamma)$ and $L^2(\Gamma)$ (see the discussion after Theorem 2.49 in (Chandler-Wilde & Monk, 2008)). Consequently, the induced operators are second-kind Fredholm operators of the form $\mathsf{I} + \mathsf{K} : X \to X$ (Antoine & Darbas, 2021) with X a separable Hilbert space, which entails fast asymptotic convergence (i.e. super-linear) of iterative solvers (Van der Vorst & Vuik, 1993). Additionally, L^2-compactness is advantageous as it is naturally suited to the euclidean norm-based GMRES (Campbell, Ipsen, Kelley, Meyer, & Xue, 1996). Still, second-kind Fredholmness is not transferred to $(\mathsf{I} + \mathsf{K})^{(k)}$, as the cross-terms are not compact, e.g., for $k = 2$, $\mathsf{I} \otimes \mathsf{K}$ and $\mathsf{K} \otimes \mathsf{I}$ are not compact at a continuous level.

REMARK 4.4 (Non-compactness of cross-terms). *As stated in (Zanni & Kubrusly, 2015, Corollary 1) for X a Hilbert complex space and $\mathsf{A}, \mathsf{B} \in \mathcal{L}(X)$:*

$$\mathsf{B} \otimes \mathsf{A} \text{ is nonzero and compact} \iff \mathsf{A} \text{ and } \mathsf{B} \text{ are both nonzero and compact.} \quad (4.37)$$

Suppose that K is nonzero and $\mathsf{I} \otimes \mathsf{K}$. Therefore, I is compact, which is a contradiction, proving that $\mathsf{I} \otimes \mathsf{K}$ is not compact. Similarly, we ensure that $\mathsf{K} \otimes \mathsf{K}$ is compact.

Despite the above, super-linear convergence for the higher moments is likely: clustering properties of A are surprisingly transferred to the tensor operator as hinted by the next result.

Theorem 4.1 (Clustering properties of the tensor matrix equation)**.** *Consider that $\mathsf{A} = \mathsf{I} + \mathsf{K} : X \to X$ with X a separable Hilbert space and K a compact operator. Therefore, for $k \in \mathbb{N}_2$, the discretized system of $\mathsf{A}^{(k)}$ has a cluster at 1.*

PROOF. As K is compact, its singular values $\sigma_j(\mathsf{K})$, $j = 1, \cdots$ with $\sigma_j(\mathsf{K}) \to 0$ as $j \to \infty$. Therefore, the singular values of $(\mathsf{I} \otimes \mathsf{K})$ give $\sigma_{j,l}(\mathsf{I} \otimes \mathsf{K}) = \sigma_j(\mathsf{K}) \to 0$. Therefore, $\mathsf{I} \otimes \mathsf{K}$ has a cluster at 1. The same proof applies to any cross-term. Finally, as the constants in the asymptotic bounds are independent of the mesh side, the clustering property transfers at discrete level. □

To quantify a possible super-linear behavior at iteration $m \in \mathbb{N}_1$, we introduce $\mathbf{r}_m$ the GMRES residual and the convergence factor given by the following m-th root:

$$Q_m := \left(\frac{\|\mathbf{r}_m\|_2}{\|\mathbf{r}_0\|_2} \right)^{1/m}. \quad (4.38)$$

Notice that the super-linear behavior shows up in the final phase of convergence of Krylov solvers and "is often not seen unless one iterates to very small relative errors and the condition number is large" ((Axelsson, 1996, Section 13.5)).

4.6.3. Wavenumber analysis

In this chapter, we focus on first-kind BIEs preconditioned via Calderón identities. Still, the proposed technique does not give results concerning the wavenumber dependence in the constants of: (i) the FOA; (ii) the quasi-optimality constant of the sparse tensor approximation; and, (iii) the condition number. Despite being out of the scope of this chapter, we aim at giving a few remarks about the analysis for high wavenumbers. The smoothness of the domains considered here hints at using non-resonant L^2-combined field formulations (Sauter & Schwab, 2010, Section 3.9.4), but would require a more complex analysis to prove enough regularity for the shape derivative, namely to prove that Cauchy data (λ_i, σ_i) for the SD $i = 0, 1$ belong to $H^1(\Gamma)$ and $L^2(\Gamma)$ respectively. Furthermore, extensive results were proved for ($\mathrm{P}_{\boldsymbol{\beta}}$), and the analysis could be carried on, under additional restrictive requirements on the domain such as star-shapedness (refer e.g., to (Spence, 2014; Galkowski, Müller, & Spence, 2019; Galkowski, Spence, & Wunsch, 2019). Those surveys can lead to elliptic formulations, allowing for application of Céa's lemma (Céa, 1964), with a simple characterization of the κ-dependence of the constants of (ii) and (iii). Concerning item (i), we expect the constant to be specified with the help of the BVP for the shape Hessian, provided sufficient regularity of both domain and transformations. Furthermore, star-shapedness is a classical assumption for the UQ by random domains, as it allows to handle the domain transformations more easily.

4.7. Numerical Results

We now apply the proposed technique to realistic applications. In order to investigate the accuracy of the first-order shape approximation, in Section 4.7.1 we analyze with the shape sensitivity analysis of sound-soft and -hard problems for a kite-shaped object. Thus, the transmission problem is set over the unit sphere and focus is set on the CT for the two-point covariance field i.e. $k = 2$. The error convergence rates for the CT relying on the Mie series are analyzed in Section 4.7.2. Finally, the behavior of GMRES is discussed in Section 4.7.3 and the FOSB is compared to MC simulation for a complex case in Section 4.7.4.

Domains are excited by a plane wave polarized along the x-direction, i.e. $\mathrm{U}^{\mathrm{inc}}(\mathbf{x}) = e^{\imath\kappa x}$, with $\mathbf{x} = (x, y, z) \in \mathbb{R}^3$.

All simulations are performed via the open-source Galerkin boundary element library Bempp 3.2 (Śmigaj et al., 2015).[1] The induced linear systems are preconditioned by strong-form multiplicative Calderón preconditioning (*cf.* (Kleanthous et al., 2018)). Tests are executed on a 32 core, 4 GB RAM per core, 64-bit Linux server using Python 2.7.6. Default parameters used throughout are the following: linear systems are solved with restarted GMRES(200) (Saad, 2003) with a tolerance of 10^{-4}. Simulations are accelerated with Hierarchical Matrices ($\mathcal{H}$-mat) (Bebendorf, 2008, Chapter 2) combined with the Adaptive Cross Approximation algorithm (ACA) (Bebendorf, 2008, Section 3.4). The relative tolerance for ACA is set to 10^{-5}. Meshes and simulations are fully reproducible using pioneering Bempp-UQ platform, a documented Python/Bempp-based plug-in including Python Notebooks.[2]

In our simulations, we shall represent the polar radar cross section (RCS) over the unit circle $\mathbb{S}^1$ and in decibels (dB) defined by:

$$\mathrm{RCS}_t(\theta) := 10\log_{10}\left(4\pi\frac{|\mathrm{F}_t^{\mathrm{scat}}|^2}{|\mathrm{F}^{\mathrm{inc}}|^2}\right),\ \theta := \mathrm{atan2}(y,x) \in [0, 2\pi].$$

4.7.1. Kite-shaped object: FOA analysis

First, we aim at evidencing FOA's accuracy. For this, we introduce a kite-shaped object perturbed according to $\Gamma_t := \{\mathbf{x} + t\mathbf{v},\ \mathbf{x} \in \Gamma\}$, with $\mathbf{v} := [(z^2-1)(\cos(\theta)-1), 0.25\sin(\theta)(1-z^2), 0]$ in Cartesian axes. In Figure 4.3, we represent the family of transformed boundaries considered here, corresponding to $t = \{0.01, 0.1, 0.25, 0.5, 1.0\}$. For wavenumbers $\kappa = 1$ and $\kappa = 8$, we illuminate the object for $\beta = 0, 1$ and solve (P_β) for all the values of t. We compare the far-field and RCS of U_t (in black) to:

U,	the zeroth order approximation (ZOA, in red), and
$(\mathrm{U} + t\mathrm{U}')$,	the FOA (in blue).

[1] https://bempp.com/download/
[2] https://github.com/pescap/Bempp-UQ

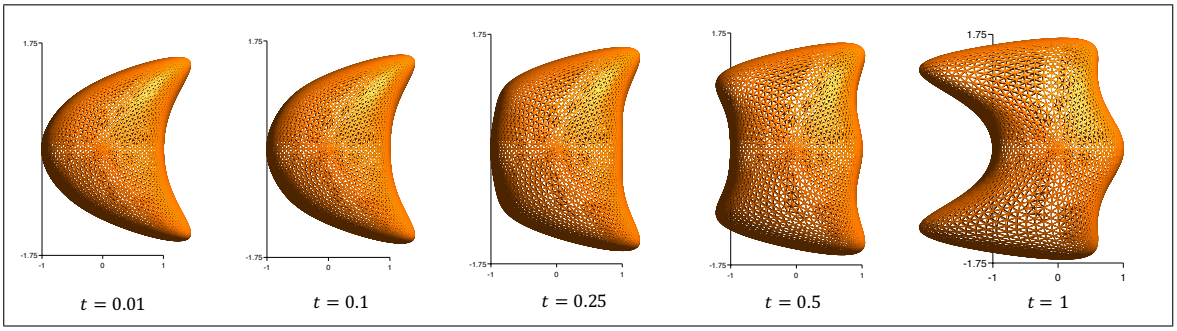


FIGURE 4.3. Transformed boundaries function to t, meshed with $3,249$ vertices.

Notice that the zeroth order approximation has no role in the FOA scheme, but will be used throughout as an additional reference for comparison purposes. The Galerkin discretization for $\kappa = 1$ (resp. $\kappa = 8$) was realized with a precision of 30 (resp. 20) triangular elements per wavelength and led to a Galerkin matrix of size $N = 3249$ (resp. $N = 9820$).

Table 4.5 presents RCSs for $\beta = 0, 1$ and $\kappa = 1, 8$. For different values of t, we plot RCS in dB of U_t on the left along with the one of the FOA $(\mathrm{U} + t\mathrm{U}')$. The x-axis represents the translated angle $(\theta + \pi)$ in radians. We remark that (i) as expected, the FOA gives a proper approximation for small values of t, (ii) the approximation seems less accurate for the shadow region, due to the oscillatory behavior of the latter and (iii) there is an evident dependence of the quality of the approximation function to the wavenumber. As an effect, we see that the FOA is accurate in a wider range of values of t for $\kappa = 1$ than for $\kappa = 8$.

To corroborate these remarks, we plot in Table 4.6 on the left-side of each cell: the error $[\cdot]_{L^2(\mathbb{S}^1)}$ for F and $(\mathrm{F} + t\mathrm{F}')$ the FOA for

$$t \in \{0.025, 0.05, 0.075, 0.1, 0.125, 0.25, 0.5, 0.75, 1.00\}.$$

These figures evidence the predicted linear and quadratic errors of both zero and first order approximations (see Remark 4.2). Besides, we observe that the FOA is indeed more accurate that F for small values of t. Still, the accuracy range of the FOA decreases strongly with κ. For instance, for $\beta = 0, 1$ and $\kappa = 1$, the FOA gives a 15% error for $t \leq 0.5$. Dissimilarly for $\kappa = 8$, the latter remains true only for $t \leq 0.1$ for $\beta = 0$ and even gives an error of 20% for $\beta = 1$.

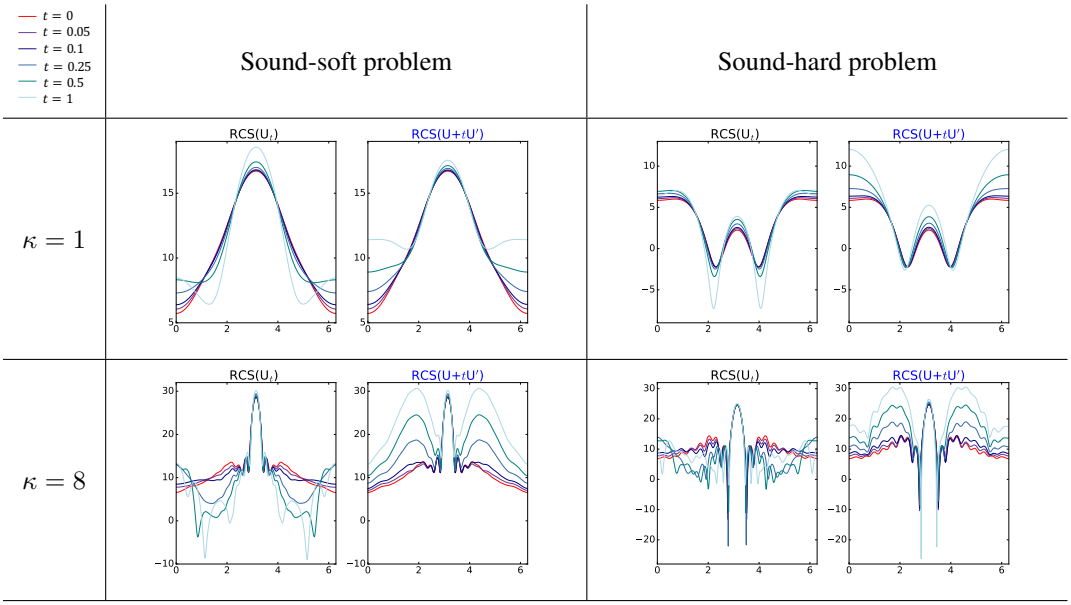


TABLE 4.5. RCS patterns (in dB) versus the angle $(\theta + \pi)$ in radians.

The right-side of each cell in Table 4.6 presents the RCS pattern of U_t, U and $(\mathrm{U}+t\mathrm{U}')$ for $(\kappa, t) = (1, 0.5)$ and $(\kappa, t) = (8, 0.1)$. Let us focus on $\kappa = 8$ and for $t \geq 0.25$: the FOA is clearly out of its admissible range. Next, we detail further the relative errors: in Table 4.7, we represent $[\cdot]_{L^2(\mathbb{S}^1)}$ in a log-log scale function to t and verify that for $\kappa = 1$, the error rate are as expected. For $\kappa = 8$, the FOA presents slightly reduced convergence rates for small values of t due to discretization error.

Henceforth, we aim at studying the wavenumber dependence of the approximates. We now fix $t = 0.1$ and solve the problem for $\kappa \in \{1, \cdots, 10\}$, with a precision of 20 elements per wavelength for each κ. In Figure 4.4, for $\beta = 0, 1$, we display the relative L^2-error of the approximates on $\mathbb{S}^1$ function to κ. We notice a linear dependence of the error for U with respect to κ and a dependence of order $\mathcal{O}(\kappa^{3/2})$ for the FOA for $\beta = 0$ and $\beta = 1$, respectively. The curves show a stable asymptotic behavior function to the wavenumber. The latter hints at using $\kappa^{3/2}t = \mathcal{O}(1)$ to keep an accuracy for the FOA bounded independently of the wavenumber. Notice that this estimate is more restrictive that the intuitive bound $\kappa t = \mathcal{O}(1)$ proposed in (Silva-Oelker et al., 2018), confirming the need for a proper wavenumber analysis for the FOA (Section 4.6.3).

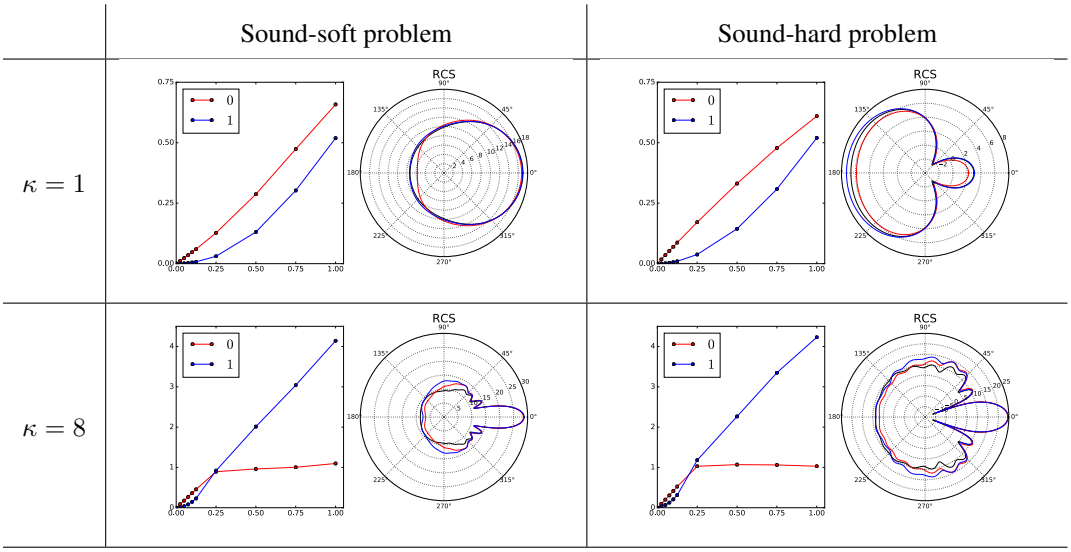


TABLE 4.6. ZOA (red) vs. FOA (blue): Relative L^2-error on $\mathbb{S}^1$ function to t (left) and RCS patterns (in dB) for $(\kappa, t) = (1, 0.25)$ and $(\kappa, t) = (8, 0.1)$ (right).

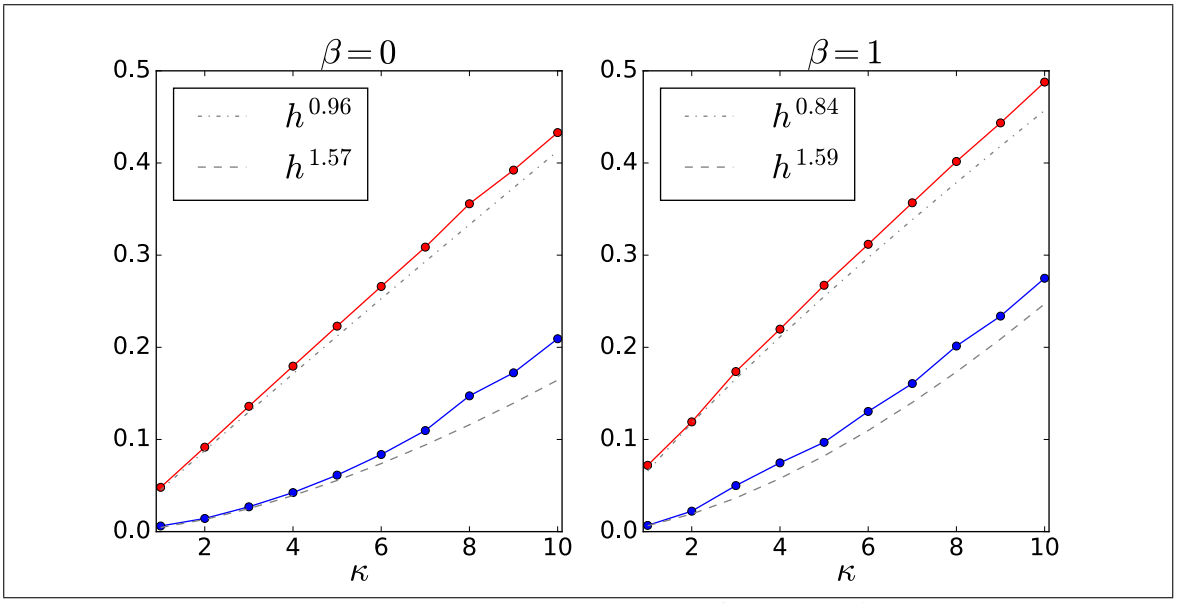


FIGURE 4.4. ZOA (red) vs. FOA (blue): Relative L^2-error on $\mathbb{S}^1$ function to κ for $\beta = 0$ (left) and $\beta = 1$ (right) and polynomial fit.

4.7.2. Unit sphere: convergence analysis

Consider the unit sphere $D := \{\mathbf{x} \in \mathbb{R}^3 : \|\mathbf{x}\|_2 \leq 1\}$ and focus on the convergence rates for the mean field and second order statistical moment. In order to inspect the behavior of both Dirichlet and Neumann traces separately for the second moment, we dispose of a

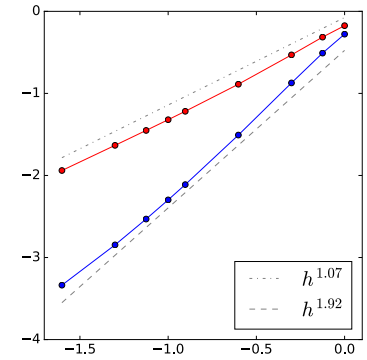
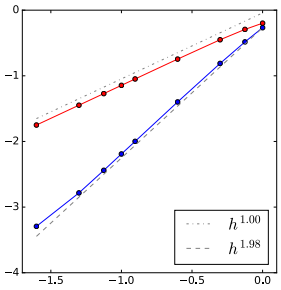
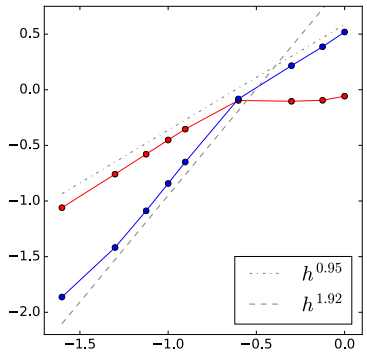
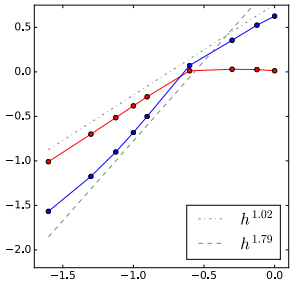
	Sound-soft problem	Sound-hard problem
$\kappa = 1$		
$\kappa = 8$		

TABLE 4.7. ZOA (red) vs. FOA (blue): Relative L^2-error on $\mathbb{S}^1$ function to t in log-log scale.

known solution, set $\boldsymbol{\xi}^{\text{inc}} = (\gamma_0 \mathrm{U}^{\text{inc}}, \gamma_1 \mathrm{U}^{\text{inc}})$, $\mu_0 = \mu_1 = 1$, and consider the following BIEs with $\boldsymbol{\xi} := \boldsymbol{\xi}^0 = (\xi_0, \xi_1)$:

$$(\widehat{\mathsf{A}}_{\kappa_0,\mu_0} + \widehat{\mathsf{A}}_{\kappa_1,\mu_1})\boldsymbol{\xi} = \boldsymbol{\xi}^{\text{inc}}, \qquad \text{on } \Gamma,$$
$$(\widehat{\mathsf{A}}_{\kappa_0,\mu_0} + \widehat{\mathsf{A}}_{\kappa_1,\mu_1})^{(2)}\Sigma = \boldsymbol{\xi}^{\text{inc}} \otimes \boldsymbol{\xi}^{\text{inc}}, \qquad \text{on } \Gamma^{(2)}.$$

Using the Mie series, we know exactly $\boldsymbol{\xi}$ as well as $\Sigma = \boldsymbol{\xi} \otimes \boldsymbol{\xi}$. Due to the domain regularity and the piecewise linear discretization on an affine mesh, we expect convergence rates of Table 4.2 to be verified. In particular here, the asymptotic error is limited by the

discretization error for Dirichlet counterpart of traces, i.e.

$$\|\Sigma - \hat{\Sigma}_L\|_{(H^{1/2}(\Gamma)\times H^{-1/2}(\Gamma))^{(2)}} = C|\log h|^{1/2}h^{3/2}\|\xi_0^{(2)}\|_{H^{1/2}(\Gamma)^{(2)}} + o(h^{3/2}).$$

For the sake of conciseness, we decide to focus on the error for the Dirichlet and Neumann counterparts of the solution and not on cross terms in $H^{1/2}(\Gamma)\otimes H^{-1/2}(\Gamma)$ and $H^{-1/2}(\Gamma)\otimes H^{1/2}(\Gamma)$ as they provide similar results to the problems (P_β), $\beta = 0, 1, 2$.

We set $\mu_0 = \mu_1 = 1$, and solve the transmission problem for two couples of wavenumbers, referred to as the low frequency (LF): $(\kappa_0, \kappa_1) = (0.4, 1)$ and high frequency (HF): $(\kappa_0, \kappa_1) = (8, 2)$ cases. We generate a sequence of meshes obtained by subdividing each triangle into two new ones, and by projecting the new vertices onto $\mathbb{S}^2$. We obtain a cardinality for V_l of $N_l = \mathcal{O}(2^l)$ and set $L = 9$. First, we study the convergence rates for the first moment and full tensor solutions in Table 4.8. We remark a $\mathcal{O}(h^{3/2})$ and $\mathcal{O}(h^2)$ convergence rate for the Dirichlet and Neumann traces respectively. Also, we notice an oscillatory behavior of the error for the Dirichlet trace for the HF case.

Next, we focus on the sparse tensor approximation and the minimal resolution level. For values of $L_0 \in \{0, 1, 2, 3, 4\}$, in Table 4.10 we represent the relative energy norm error of $\hat{\Sigma}_L(L_0)$ versus the full solution Σ_L function to $1/h$. We restrict the case $L_0 = 4$ to the HF case, as lower refinement levels give satisfactory results. As expected, we remark that for a sufficient minimal resolution level, the solution in the sparse tensor space converges with the same rate as the full solution Σ_L (in black).

We also represent the precision r, which represents the number of elements per wavelength in the x-axis. For the HF case: (i) the sparse solution inherits of limited convergence rate for small values of $1/h$; and, (ii) appears to oscillate less than in the full space.

In order to better assess the quality of the sparse approximate function to L_0, we present in Table 4.10 the same energy norm errors as in Table 4.9 function to dofs the number of dofs used to get the approximation. The optimal resolution level $\hat{L}_0$ depends on the type of trace and the frequency. In the sequel, we focus on the symmetric case (refer to Section 4.6.1). In Table 4.11 we corroborate that the symmetry of the right-hand side benefits

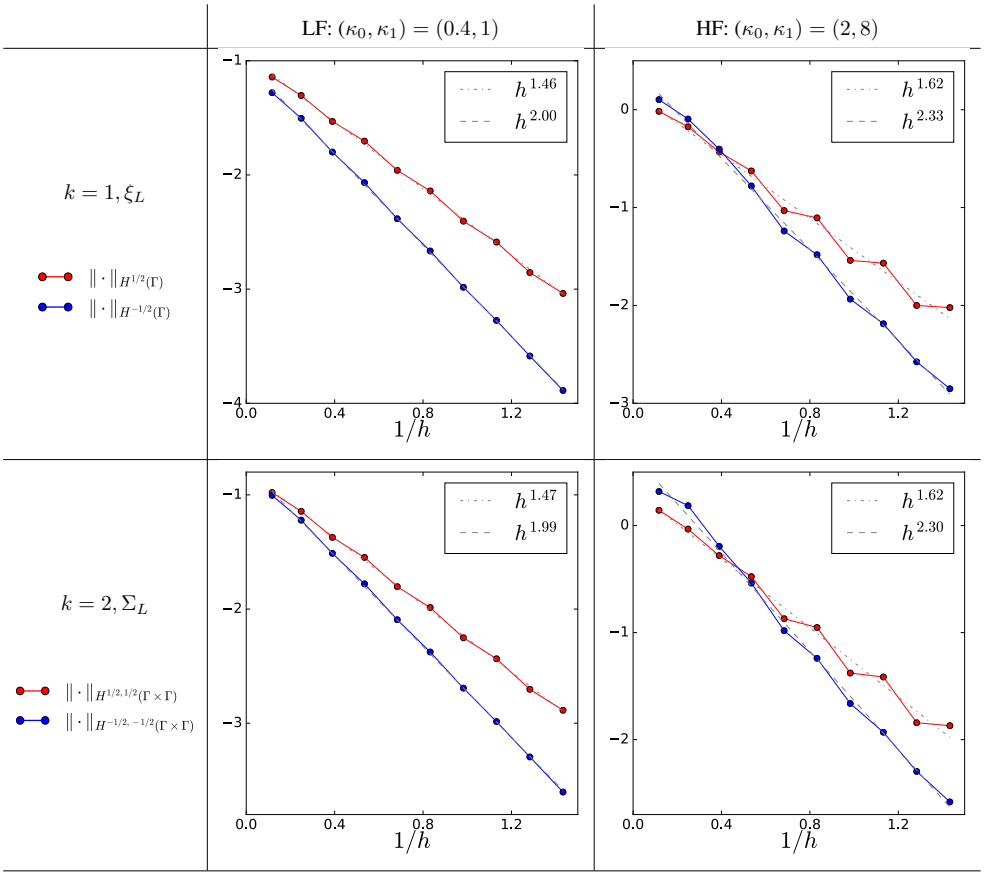


TABLE 4.8. Relative errors in energy norm of the Dirichlet and Neumann Traces on $\mathbb{S}^2$ and $(\mathbb{S}^2)^{(2)}$ for the LF and HF cases. Relative energy norm error for Dirichlet (red) and Neumann (blue) trace components with respect to the inverse mesh density $1/h$.

the sparse tensor approximation, as roughly half of the linear systems of the classical CT are needed.

4.7.3. Unit Sphere: Iterative solvers

We focus on the practical implementation of the CT. We solve the sub-blocks of the symmetric CT with GMRES and a tolerance of 10^{-8}. Figure 4.5 showcases the number of dofs and GMRES iterations needed to reach the prescribed tolerance of each sub-block for given indices l_1 and l_2. We highlight the case $L = 7$ and $L_0 = 0$ with bold (resp. italic) notation for the added (resp. subtracted) sub-blocks for the symmetric CT (*cf.* Table 4.4). Below, as a reference, we show the results for the first moment. The number of dofs on the

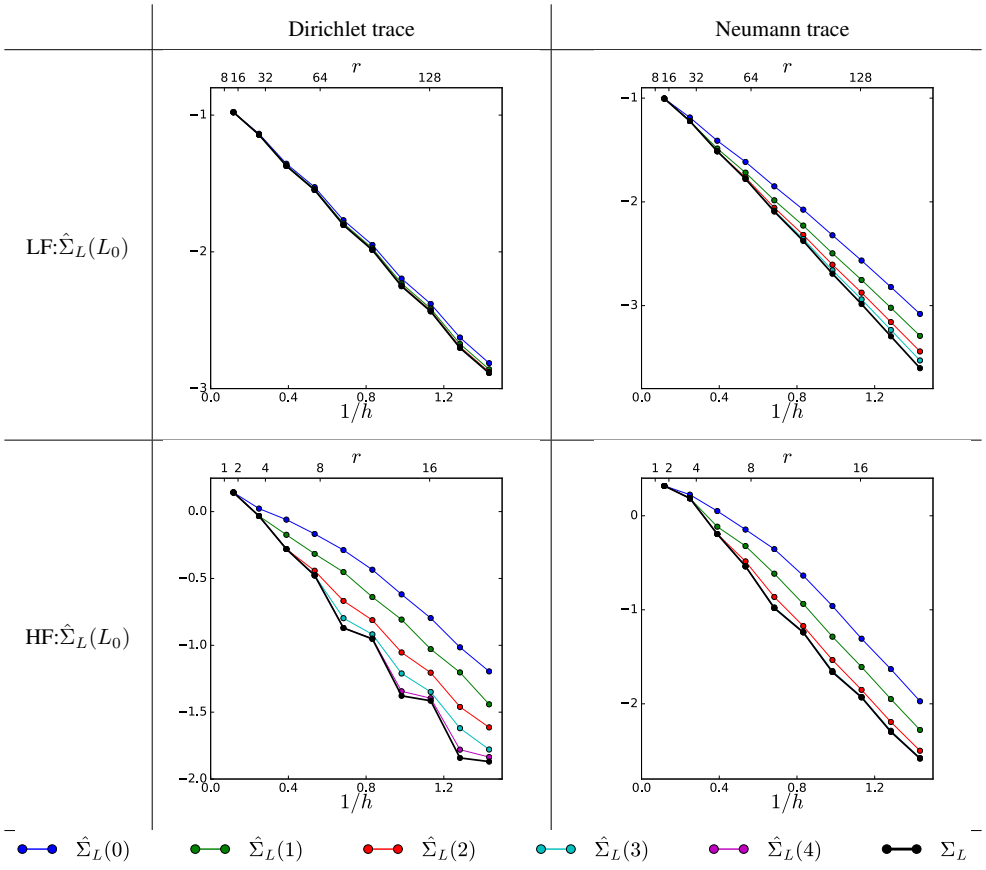


TABLE 4.9. Relative errors in energy norm function to h of the Dirichlet and Neumann Traces on $(\mathbb{S}^2)^{(2)}$ for the LF and HF cases.

diagonal (i.e. the bold and italic ones) are of size $N_0 \times N_L$ (resp. $N_0 \times N_{L-1}$). Thus, the resolution of subsystems of equivalent size when implementing the CT.

We also remark the effectiveness of Calderón preconditioning, as we notice that the number of iterations remains of 8 independently of l_1 and l_2. Also, the number of iterations passes from 3 for first moment to 8, likely due to $\kappa_2(\mathbf{A} \otimes \mathbf{B}) = \kappa_2(\mathbf{A})\kappa_2(\mathbf{B})$. Besides, we show the solver times in seconds in Table 4.12. Accordingly, we consider the HF case: we plot the relative residual l^2-error of GMRES in Figure 4.6 (in log-log scale). First, in black (resp. gray), we represent the iterations for $k = 1$ and $L = 0$ (resp. $l \in \{1, \cdots, 7\}$). We remark that: (i) the iteration count increases compared to the LF case; (ii) the relative error is reasonably resilient with the meshwidth, as the curves are close to each other; and, (iii) fast convergence of the residual towards zero. Still, mesh independence is key in reducing

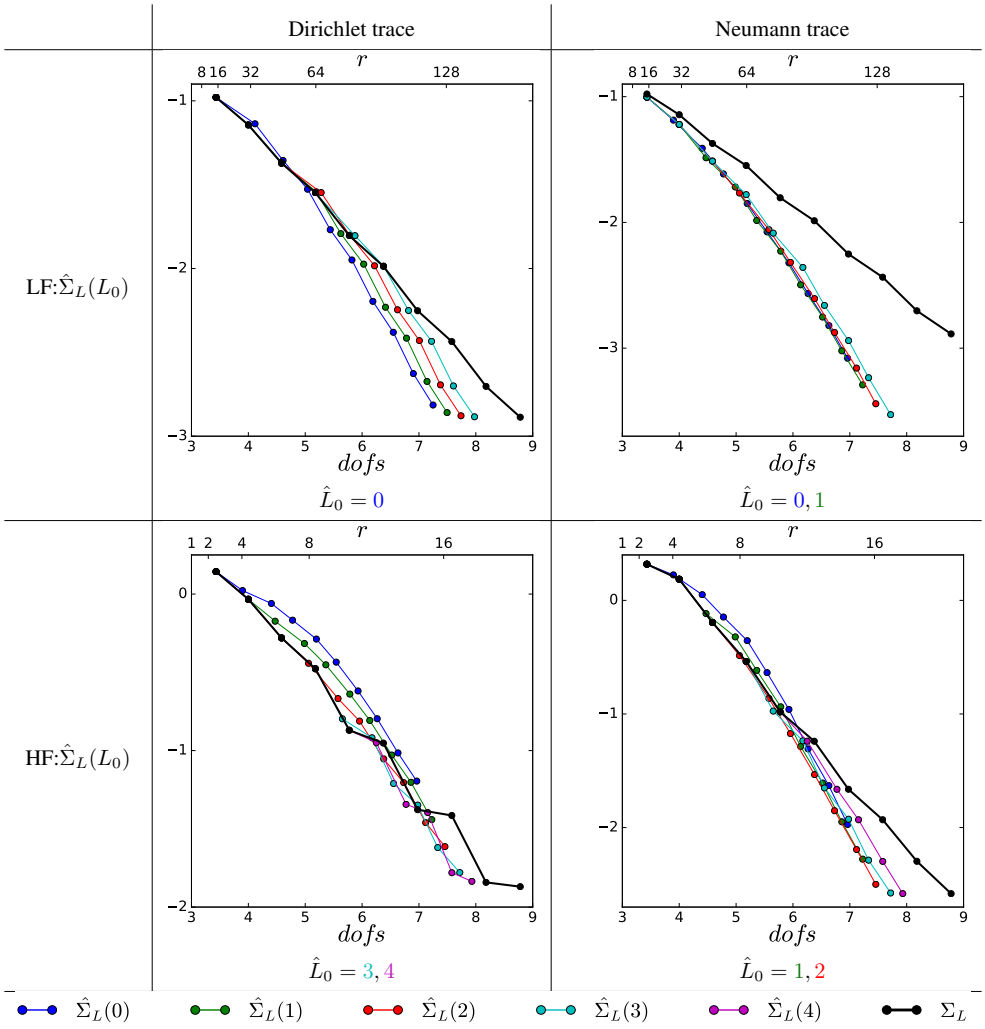


TABLE 4.10. Relative errors in energy norm function to dofs of the Dirichlet and Neumann Traces on $(\mathbb{S}^2)^{(2)}$ for the LF and HF cases.

the sensibility to the meshwidth but does not necessarily leads to faster convergence of GMRES, as the condition number remains bounded but can be large, as highlighted for the second moment. Indeed, for several values of (l_1, l_2) we add error convergence curves of GMRES for $k = 2$, and renew the previous remark, with a noticeable deterioration of the convergence results.

Based on the above, we further investigate the properties of the resulting linear systems and the convergence behavior. In Table 4.13, we portray again the GMRES residual

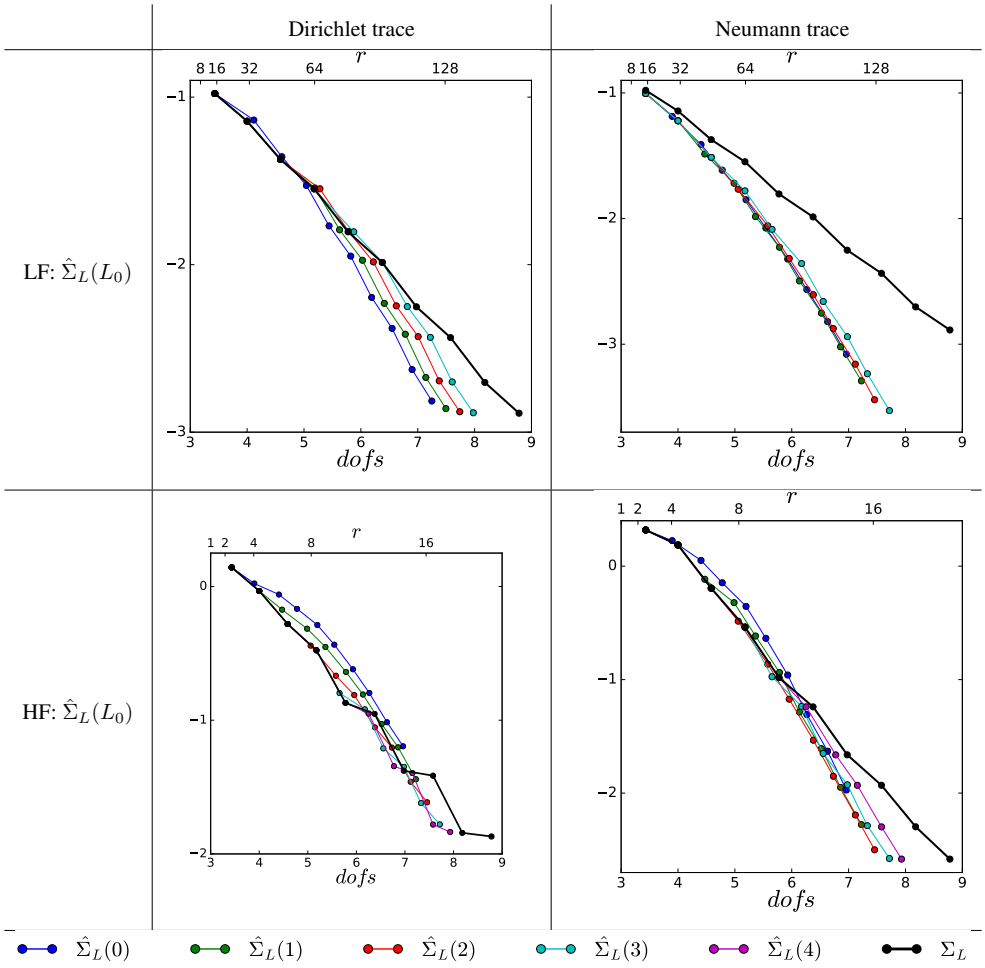


TABLE 4.11. Symmetric case: relative errors in energy norm function to dofs of the Dirichlet and Neumann Traces on $(\mathbb{S}^2)^{(2)}$ for the LF and HF cases.

error, in a semi-log scale (first row). Also, we present the convergence factor at each iteration (second row). The first row shows that all curves present at least a linear decrease, ensuring convergence of GMRES. Moreover, the convergence for the first moment is too fast to observe a super-linear phase. The second moment curves present poor convergence rates close to 1, with a very slow decrease (see the bottom-right figure), giving a moderate super-linear behavior, still noticeable for $\Sigma_{1,2}$ and $\Sigma_{0,3}$ (see the top-right figure). To finish, we introduce $\mathbf{M}$ the mass and $\mathbf{A}$ the impedance matrices. In Table 4.14, we plot the eigenvalues distributions of the resulting linear systems (in strong form, such as done in (Kleanthous et al., 2018)). We remark that the spectra present some clustering at one and

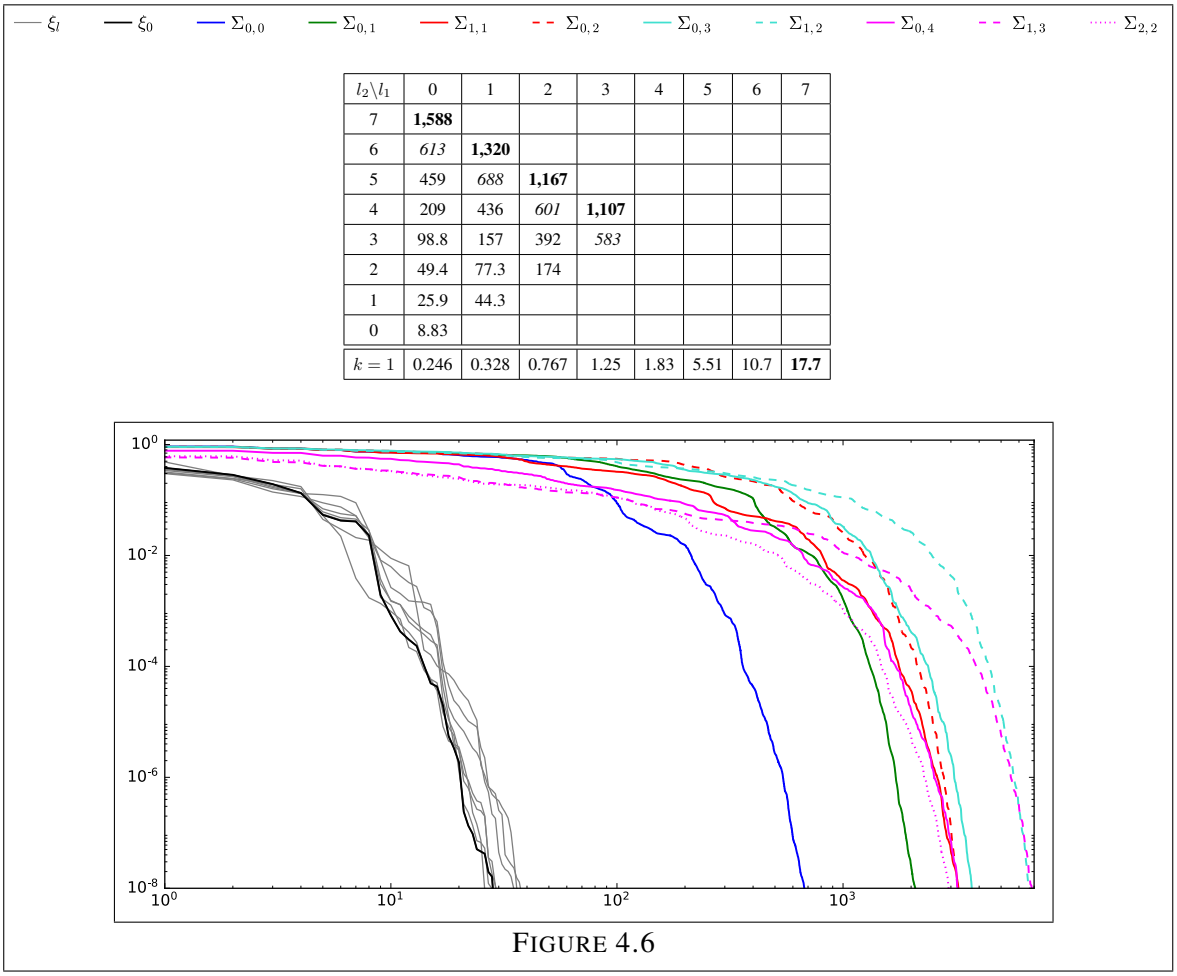
$l_2 \backslash l_1$	0	1	2	3	4	5	6	7
7	**319,696**							
6	*159,952*	**307,600**						
5	80,080	*154,000*	**301,840**					
4	40,144	77,200	*151312*	**299,536**				
3	20,176	38,800	76,048	*150,544*				
2	10,192	19,600	38,416					
1	5,200	10,000						
0	2,704							
$k = 1$	52	100	196	388	772	1,540	3,076	**6,148**

$l_2 \backslash l_1$	0	1	2	3	4	5	6	7
7	**8**							
6	*8*	**8**						
5	8	*8*	**8**					
4	8	8	*8*	**8**				
3	8	8	8	*8*				
2	8	8	8					
1	8	8						
0	8							
$k = 1$	3	3	3	3	3	3	3	**3**

FIGURE 4.5. Numbers of dofs for each subsystem (left) and GMRES iterations to reach prescribed tolerance (right).

have similar patterns. Also, we see that the tensor matrix for $k = 2$ has a more scattered cluster, and much more outliers. The latter emerges from the property of the tensor operator, and was expectable. To finish, despite the presence of non-compact terms at continuous level, we observe discrete clustering properties due to Theorem 4.1.

To conclude, the HF case shows that the limiting step of the CT is decisively the solver step, and justifies even more the use of efficient preconditioning techniques. In addition, the tensor operator structure of the CT and its speedup with hierarchical matrices allow for limited memory requirements for the impedance matrices and matrix-matrix product function to the number of dofs of the subsystems. These results motivate the use of hierarchical matrices to describe both unknown and right-hand side, in order to reduce matrix-matrix product execution times (*cf.* (Dölz, Harbrecht, & Schwab, 2017)).



$l_2 \backslash l_1$	0	1	2	3	4	5	6	7
7	**1,588**							
6	*613*	**1,320**						
5	459	*688*	**1,167**					
4	209	436	*601*	**1,107**				
3	98.8	157	392	*583*				
2	49.4	77.3	174					
1	25.9	44.3						
0	8.83							
$k = 1$	0.246	0.328	0.767	1.25	1.83	5.51	10.7	**17.7**

FIGURE 4.6

Table 4.12 (left) : solver times (in seconds) for the LF case. Figure 4.6 (right): relative l^2-error of GMRES in log-log scale for the HF case.

4.7.4. Real case: Non-smooth domain

To finish, we compare FOSB with Monte Carlo simulations for a complex case: the sound-soft scattering by a unit Fichera Cube with $\kappa = 5$. We perturb the boundary face located at the $z = 0.5$-plane –represented in red in Figure 4.10 later on– and use $\mathbb{P}^0$ elements. We set $L_0 = 0$ and $L = 2$ and generate a sequence of nested meshes associated with discrete spaces X_0, X_1, X_2 of cardinality $N_0 = 1140$, $N_1 = 2804$ and $N_2 = 6370$. The zeroth level corresponds to a precision of 10 elements per wavelength and related meshes are found in Figure 4.7.

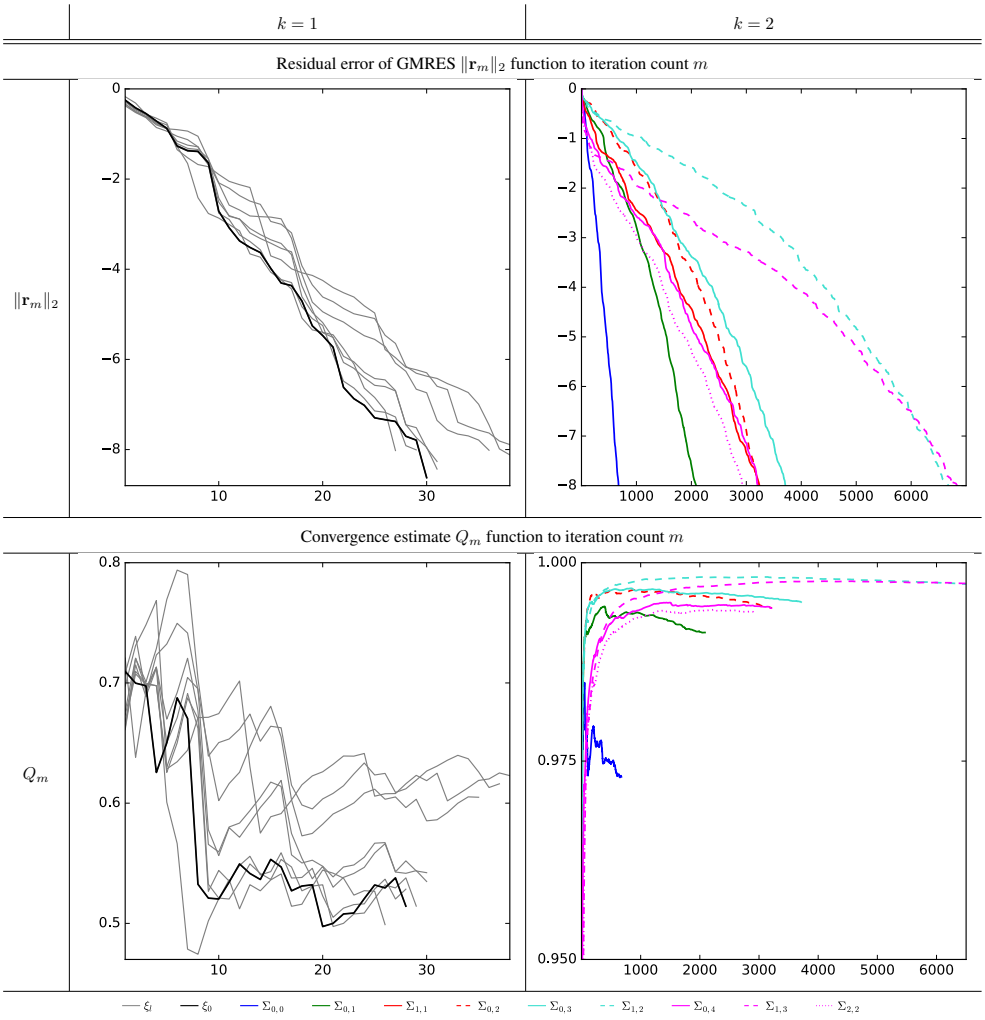


TABLE 4.13. HF case: Complete survey of the GMRES convergence.

Given uniformly distributed random variables $Y_{ij} \in \mathcal{U}[-1, 1], i = 0, \cdots, 5$, the perturbation field is given as:

$$\mathbf{v}(\mathbf{x}, \omega) := \sum_{i=0}^{5} \sum_{j=0}^{5} \Upsilon_i(x) \Upsilon_j(y) Y_{ij} \hat{\mathbf{e}}_z, \ \mathbf{x} \in \Gamma, \ z = 0.5, \tag{4.39}$$

with Υ_i denoting fundamental sine splines of the form $|\sin(q\pi x)|$, $x \in [0, 0.5]$, $q \in \{2, 4, 6\}$ with support of length $0.5/(q + 1)$ represented in Figure 4.8. Therefore, for $\mathbf{x}_1$

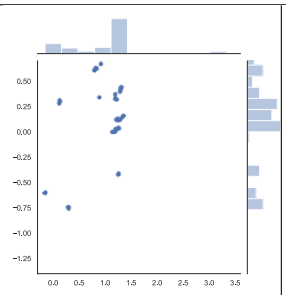
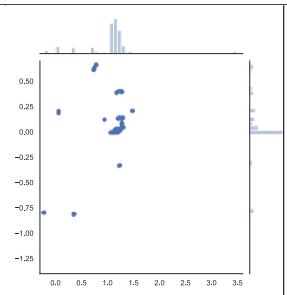
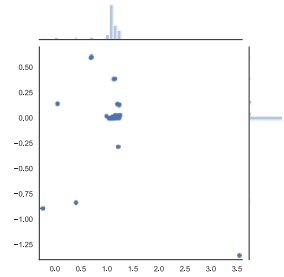
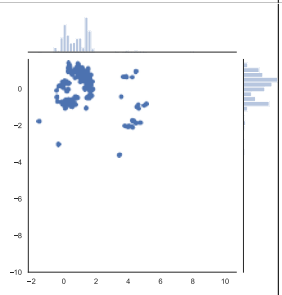
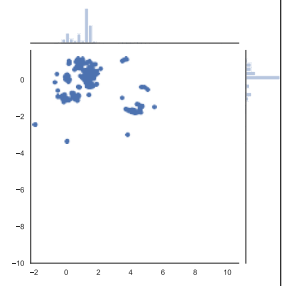
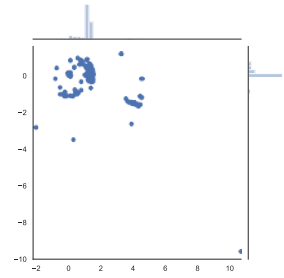
Case	$L = 2$	$L = 3$	$L = 4$
$k = 1$			
$k = 2$			

TABLE 4.14. HF case: Eigenvalues distribution dependence on L for the resulting preconditioned matrix $(\mathbf{M}^{-1}\mathbf{A}\mathbf{M}^{-1}\mathbf{A})^{(k)}$, $k = 1, 2$.

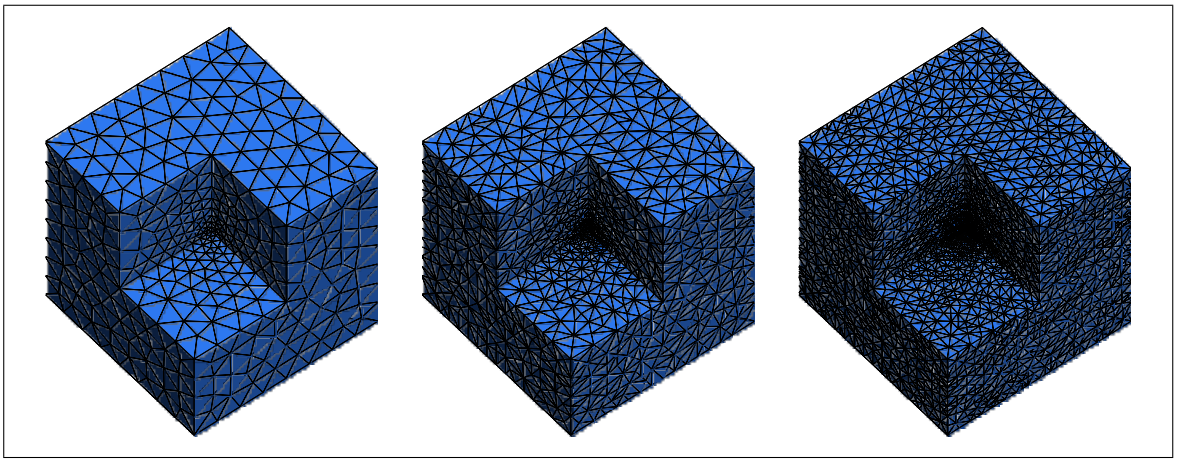


FIGURE 4.7. Sequence of nested meshes used to perform the FOSB.

and $\mathbf{x}_2$ in Γ, and $z_1 = z_2 = 0.5$, we have

$$\mathcal{M}^k[\mathbf{v} \cdot \mathbf{n}](\mathbf{x}_1, \mathbf{x}_2) = \sum_{i=0}^{5} \sum_{j=0}^{5} \frac{1}{3} \Upsilon_i(x_1) \Upsilon_j(y_1) \Upsilon_i(x_2) \Upsilon_j(y_2). \tag{4.40}$$

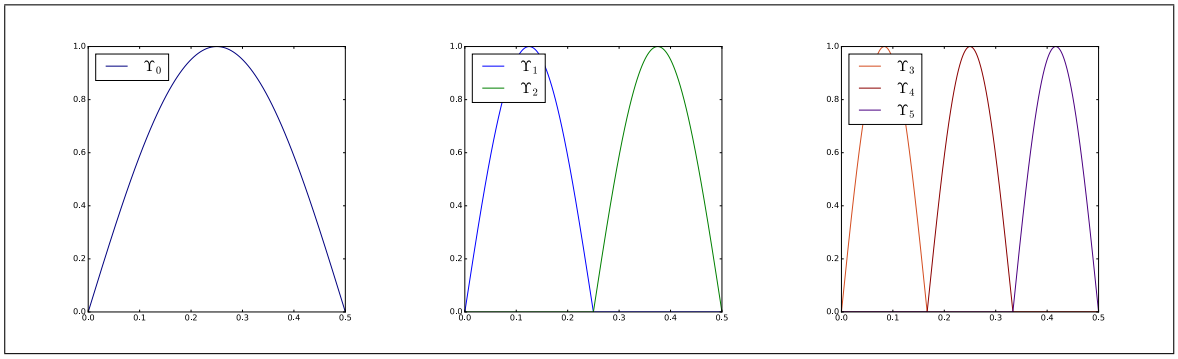


FIGURE 4.8. Splines sinusoidal functions used for random families of perturbed boundaries.

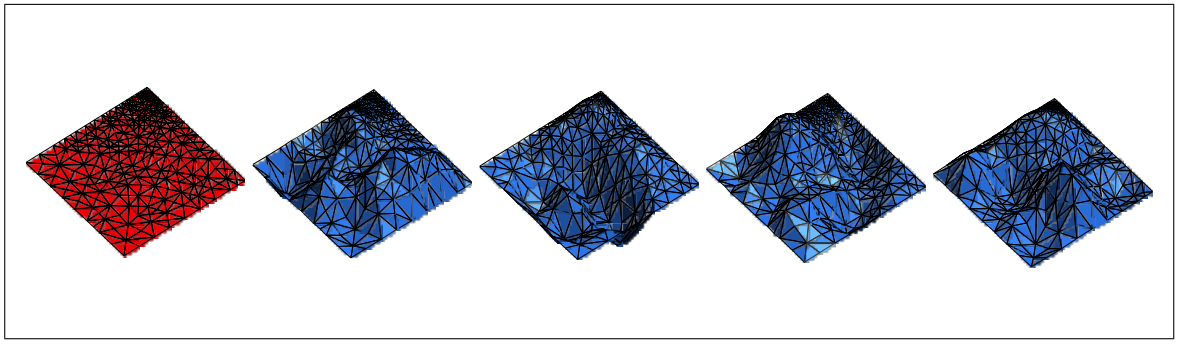


FIGURE 4.9. Nominal mesh (red) and transformed meshes corresponding to realizations of MC simulation (blue).

The perturbation parameter is set to $t = 0.075$. As a reference, we compute the mean field and variance through Monte Carlo method with $M = 5000$ simulations (see (Silva-Oelker et al., 2018, Section 5)). For each realization, we deform the mesh corresponding to level $L = 2$ and obtain $\mathrm{U}(\omega_m)$. Next, we evaluate:

$$\mathrm{U}^{\mathrm{MC}} := \frac{1}{M}\sum_{m=1}^{M}\mathrm{U}(\omega_m), \text{ and } \mathbb{V}^{\mathrm{MC}} := \frac{1}{M}\sum_{m=1}^{M}(\mathrm{U}(\omega_m) - \mathrm{U}^{\mathrm{MC}})\overline{(\mathrm{U}(\omega_m) - \mathrm{U}^{\mathrm{MC}})}. \quad (4.41)$$

In Figure 4.9, we plot mesh elements corresponding to $z = 0.5$ for the nominal shape (in red). Therefore, we plot deformed mesh issued from realizations of the perturbation field (in blue).

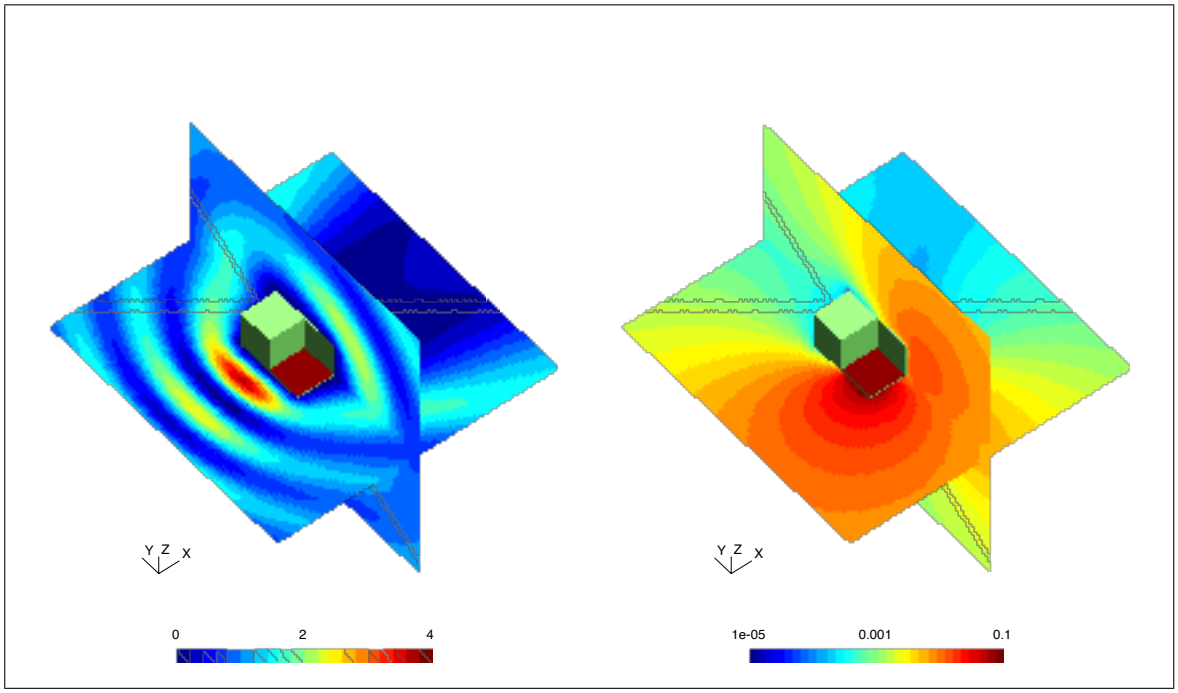


FIGURE 4.10. Volume plot of the squared density for U (left) and the standard deviation $\sqrt{\hat{\mathbb{V}}}$ obtained through the FOSB method.

For the implementation of a symmetric FOSB, we use an indirect formulation (Sauter & Schwab, 2010) for the tensor equation and choose a Near-Field preconditioner (Escapil-Inchauspé & Jerez-Hanckes, 2019) as it outperformed the Multiplicative Calderón preconditioner in solution times. In Figure 4.10, we plot the total squared density of U (left) and the standard deviation –square root variance– $\sqrt{\hat{\mathbb{V}}}$ (right). We remark that the area close to the perturbed boundary has a higher variance while the zone behind the Fichera Cube has low sensibility to shape variation. In Table 4.15 we compare the RCSs provided by both methods, namely MC (left column) and FOSB (right column) –we inspire ourselves of the plots in (Harbrecht, Ilić, & Multerer, 2019). In the first row, we represent the approximation of the mean field (red) and its sensibility (blue). Second row focuses on RCS for the squared root variance. We remark that both methods show similar patterns. Indeed, the relative L^2-error on $\mathbb{S}^1$ between the FFs differ by a 11.0% (resp. 18.4%) for U^{MC} and U (resp. $\sqrt{\mathbb{V}^{\mathrm{MC}}}$ and $\sqrt{\hat{\mathbb{V}}}$), evidencing accuracy of the FOSB scheme. The latter is interesting as it shows that the FOA behaves well for domains with corners, albeit lacking theoretical results on it. Finally, the symmetric CT led to a total $dofs = 18,420,424$ compared to

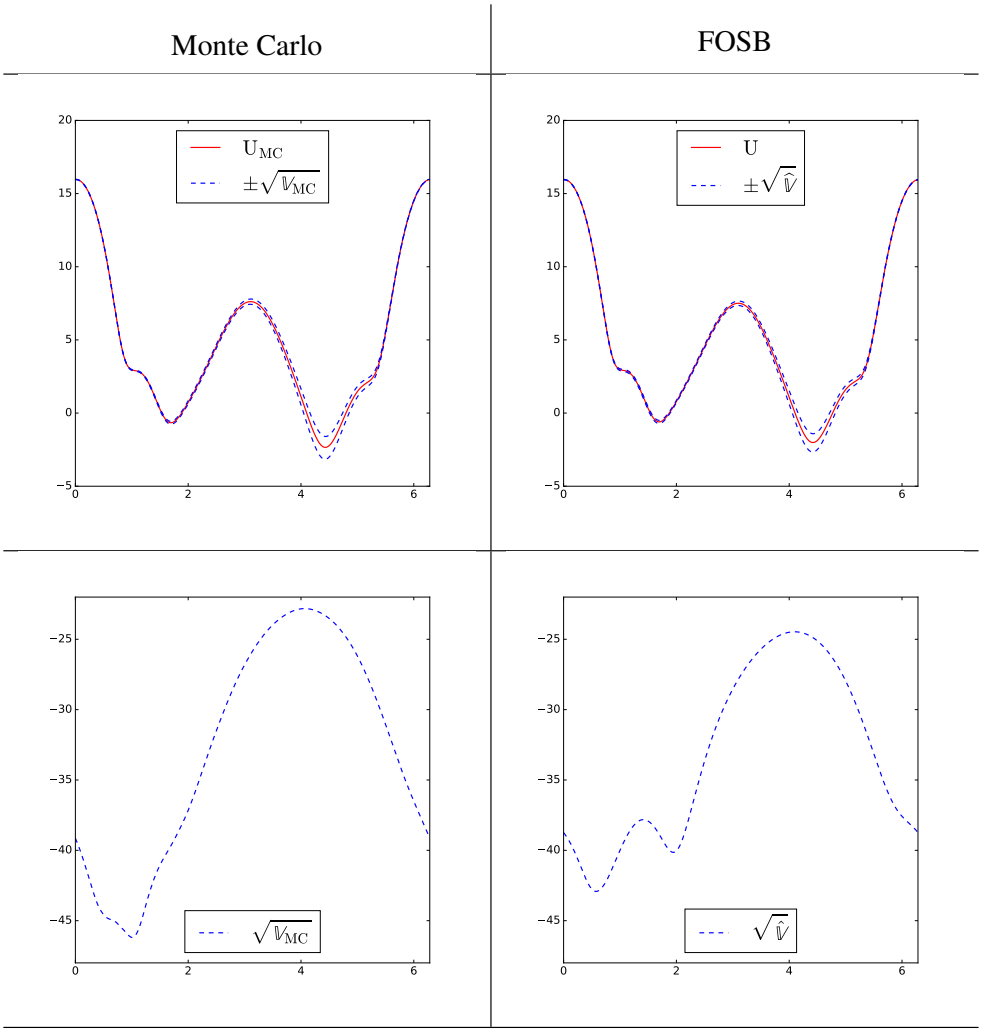


TABLE 4.15. Final comparative results between the MC (left) and FOSB (right) methods. First row shows the approximation for the mean RCS (red) and its standard deviation (blue) while second rows focuses on the standard deviation. RCSs are in represented (dB) versus the angle $(\theta + \pi)$ in radians.

$N_L^2 = 41,088,100$ for the full tensor space. As a comparison, $MN_L = 38,460,000$ for MC. The total execution time for MC was 13 hours 26 min. compared to 6 hours 19 min. for the FOSB.

4.8. Conclusion

In this chapter, we tackled UQ for random shape Helmholtz scattering problem. Under small perturbation assumptions, we applied the FOSB method and allowed for an accurate

approximation of statistical moments with an almost optimal memory and computational requirements. We provided the complete analysis for Helmholtz boundary value problems and added comments concerning the efficient implementation of the schemes. Numerical experiments evidenced the applicability of the technique and showed good scalability and robustness when coupled with fast resolution methods and efficient preconditioners. Observe that theory presented in Section 4.5 and Section 4.6.2 and numerical results of Sections 4.7.2 and 4.7.3 are of interest for general Helmholtz-based tensor operators BIEs, as they are developed aside from the FOA framework. Conversely, the FOA and the numerical results of Section 4.7.1 apply to low-rank approximation-based schemes to solve the deterministic equation (Dambrine, Harbrecht, & Puig, 2015).

Further research includes asymptotic wavenumber analysis of each specific boundary condition and under additional requirements would lead in some cases to elliptic formulations, simplifying greatly the Galerkin scheme –Céa's Lemma– and the sparse tensor approximation. We hope that the analysis carried on in (Galkowski, Müller, & Spence, 2019) can be extended to the FOSB method for the exterior sound-soft problem, and would provide κ-explicit estimates of the constants involved in the scheme along with bounds for the GMRES for both nominal solution and sub-blocks for the CT.

In parallel, this chapter suggests the use of more efficient tools such as: (i) multilevel matrix-matrix product and compression techniques for the covariance kernel (e.g., hierarchical matrices or low-rank approximations), (ii) efficient iterative solvers for multiple right-hand sides (Sun, Carpentieri, Huang, Jing, & Naveed, 2018), and (iii) fast preconditioning techniques (Escapil-Inchauspé & Jerez-Hanckes, 2019). Those improvements would allow to compute higher moments in a satisfactory number of operations. Also as a by-product of our study we have rendered available the Bempp-UQ plug-in for further improvement and usage. Our code currently supports the $\mathbb{P}^1$ projection between grids, the tensor GMRES, the CT for $k = 2$ and the FOA for all problems considered here. Current work seeks to implement the FOA for Maxwell equations; speed up the preconditioner matrix assembly for both Helmholtz and Maxwell cases; and incorporate high-order quadrature rules routines for UQ purposes.

4.9. Boundary reduction

4.9.1. Case: ($\mathrm{P}_{\boldsymbol{\beta}}$)

Set $\mathsf{A}_i \equiv \mathsf{A}_{\kappa_i}$ and recall the following identity:

Lemma 4.6. *Let* U *be the solution of problem ($P_{\boldsymbol{\beta}}$) with* $\boldsymbol{\xi}^{\text{inc}} := (\gamma_0 \mathrm{U}^{\text{inc}}, \gamma_1 \mathrm{U}^{\text{inc}})$. *Thus, for* $\boldsymbol{\xi} := (\gamma_0 \mathrm{U}^0, \gamma_1 \mathrm{U}^0)$, *we have*

$$\left(\frac{1}{2}\mathsf{Id} + \mathsf{A}_0\right) \boldsymbol{\xi} = \boldsymbol{\xi}^{\text{inc}}. \tag{4.42}$$

PROOF. Consider U the solution of ($\mathrm{P}_{\boldsymbol{\beta}}$). Since U^{inc} provides admissible Cauchy data inside the scatterer and U^{scat} is a radiative Helmholtz equation, the following identities hold:

$$\left(\frac{1}{2}\mathsf{Id} - \mathsf{A}_0\right) \boldsymbol{\xi}^{\text{scat}} = \boldsymbol{\xi}^{\text{scat}}, \text{ and } \left(\frac{1}{2}\mathsf{Id} + \mathsf{A}_0\right) \boldsymbol{\xi}^{\text{inc}} = \boldsymbol{\xi}^{\text{inc}}. \tag{4.43}$$

Summing both equations, we get,

$$\mathsf{A}_0(2\boldsymbol{\xi}^{\text{inc}} - \boldsymbol{\xi}) = \frac{1}{2}\boldsymbol{\xi} \Rightarrow \left(\mathsf{A}_0 + \frac{1}{2}\mathsf{Id}\right) \boldsymbol{\xi} = 2\mathsf{A}_0 \boldsymbol{\xi}^{\text{inc}} = \boldsymbol{\xi}^{\text{inc}}.$$

□

This last result allows us to determine directly the BIEs of U from the BCs for $\beta = 0, 1, 3$. Let us focus on (P_2). Second row of Lemma 4.6 and the BCs rewrite:

$$\left(\mathsf{W}_\kappa - \imath\eta \left(\frac{1}{2}\mathsf{Id} + \mathsf{K}'_\kappa\right)\right) \gamma_0 \mathrm{U} = \gamma_1 \mathrm{U}^{\text{inc}}. \tag{4.44}$$

Next, the integral representation formula (4.16) gives:

$$\begin{aligned} \mathrm{U}^{\text{scat}} &= -\mathsf{SL}_\kappa(\gamma_1 \mathrm{U}^{\text{scat}}) + \mathsf{DL}_\kappa(\gamma_0 \mathrm{U}^{\text{scat}}) \text{ in } D^c, \text{ and} \\ 0 &= -\mathsf{SL}_\kappa(\gamma_1 \mathrm{U}^{\text{inc}}) + \mathsf{DL}_\kappa(\gamma_0 \mathrm{U}^{\text{inc}}) \text{ in } D^c. \end{aligned}$$

Summing both equations yields

$$\mathrm{U} = \mathrm{U}^{\text{inc}} - \mathsf{SL}_\kappa(\gamma_1 \mathrm{U}) + \mathsf{DL}_\kappa(\gamma_0 \mathrm{U}) \text{ in } D^c, \tag{4.45}$$

giving the final results for the expected field.

4.9.2. Case: (SP_β)

Similarly, boundary reduction for the shape derivative is deduced from the following identities for $\boldsymbol{\xi}' := \boldsymbol{\xi}^0$:

$$\left(\frac{1}{2}\mathsf{Id} - \mathsf{A}_0\right)\boldsymbol{\xi}' = \boldsymbol{\xi}', \text{ and } \left(\frac{1}{2}\mathsf{Id} + \mathsf{A}_1\right)\boldsymbol{\xi}'^1 = \boldsymbol{\xi}'^1, \text{ if } \beta = 3, \tag{4.46}$$

by considering the boundary conditions. In particular, for $\beta = 2$:

$$\begin{aligned} &-\mathsf{W}_\kappa \gamma_0 \mathsf{U}' + \left(\frac{1}{2}\mathsf{Id} - \mathsf{K}'_\kappa\right)\gamma_1 \mathsf{U}' = \gamma_1 \mathsf{U}' \\ \Rightarrow &\left(\mathsf{W}_\kappa - \imath\eta\left(\frac{1}{2}\mathsf{Id} + \mathsf{K}'_\kappa\right)\right)\gamma_0 \mathsf{U}' = \left(\frac{1}{2}\mathsf{Id} + \mathsf{K}'_\kappa\right) g_2. \end{aligned}$$

Finally, we detail the potential reconstruction for the transmission problem. Representation formulas (4.16) yield:

$$\begin{aligned} \mathsf{U}'^i &= -\mathsf{SL}_{\kappa_i}\gamma_1 \mathsf{U}'^i + \mathsf{DL}_{\kappa_i}\gamma_0 \mathsf{U}'^i \text{ in } D^i,\ i = 0, 1 \\ &= \mathsf{R}_\kappa(\boldsymbol{\xi}) \text{ in } \mathcal{D} \quad (4.14). \end{aligned}$$

5. CONCLUSION AND FUTURE RESEARCH

In this thesis, we investigated the trade off between precision and efficiency in numerical schemes related to BVPs. We focused on resolution of scattering problems, being a challenging case of study. We provided new results regarding preconditioning, perturbation analysis, uncertainty quantification and convergence of iterative solvers.

To begin with, we introduced the bi-parametric OP. As a study case, it was applied successfully to the EFIE. Along with this, we considered the Helmholtz scattering by random objects and provided a detailed framework to apply the FOSB method, incorporating preconditioning issues.

As a consequence to the publication of (Escapil-Inchauspé & Jerez-Hanckes, 2019) (corresponding to Chapter 3), the bi-parametric OP framework was applied to Helmholtz Scattering Problems by Jerez-Hanckes and Fierro (Fierro & Jerez-Hanckes, 2020) and complex objects together with A. Kleanthous, C. Jerez-Hanckes, T. Betcke et al. in (Kleanthous et al., 2020), submitted to J. Comp. Phys., 2020. We also mention interesting results concerning preconditioning for the local multiple-trace formulation applied to electromagnetics (Ayala, Claeys, Escapil-Inchauspé, & Jerez-Hanckes, 2020).

Algonside the aforementioned, the chapter paves the way towards a number of novel applications. To begin with, bi-parametric OP can be a source of inspiration to (i) quantify more efficiently how current preconditioners behave and (ii) design new preconditioners, based on this result. Amongst others, we mention:

(i) artificial intelligence schemes—deep or shallow learning—to propose rough preconditioners;

(ii) compactness estimates—Carleman class—for differential and second-kind Fredholm operators;

(iii) asymptotic estimates for high-wavenumber and singular perturbation analysis.

To finish, FOSB framework's comprehension is allowing to easily apply it to EM. Ongoing work embrace numerical experiments for EM scattering by random objects.

References

Ainsworth, M., McLean, W., & Tran, T. (1999). The Conditioning of Boundary Element Equations on Locally Refined Meshes and Preconditioning by Diagonal Scaling. *SIAM Journal on Numerical Analysis*, *36*(6), 1901–1932.

Allaire, G., & Schoenauer, M. (2007). *Conception optimale de structures* (Vol. 58). Springer.

Andreev, R. (2013). Stability of sparse space-time finite element discretizations of linear parabolic evolution equations. *IMA Journal of Numerical Analysis*, *33*(1), 242–260.

Andriulli, F. P., Cools, K., Bagci, H., Olyslager, F., Buffa, A., Christiansen, S., & Michielssen, E. (2008). A Multiplicative Calderón Preconditioner for the Electric Field Integral Equation. *IEEE Transactions on Antennas and Propagation*, *56*(8), 2398–2412.

Andriulli, F. P., Tabacco, A., & Vecchi, G. (2010). Solving the EFIE at low frequencies with a conditioning that grows only logarithmically with the number of unknowns. *IEEE Transactions on Antennas and Propagation*, *58*(5), 1614-1624.

Antoine, X., & Darbas, M. (2021). An Introduction to Operator Preconditioning for the Fast Iterative Integral Equation Solution of Time-Harmonic Scattering Problems. *Multiscale Science and Engineering*, *3*, 1–35.

Atkinson, K. (1976). *A Survey of Numerical Methods for the Solution of Fredholm Integral Equations of the Second Kind*. Society for Industrial and Applied Mathematics (Philadelphia).

Axelsson, O. (1996). *Iterative Solution Methods*. Cambridge University Press.

Axelsson, O., & Karátson, J. (2009). Equivalent operator preconditioning for elliptic problems. *Numerical Algorithms*, *50*(3), 297–380.

Axelsson, O., & Karátson, J. (2018). Superlinear convergence of the GMRES for PDE-constrained optimization problems. *Numerical Functional Analysis and Optimization*, *39*(9), 921–936.

Axelsson, O., Karátson, J., & Magoulès, F. (2018). Superlinear convergence using block preconditioners for the real system formulation of complex Helmholtz equations. *Journal of Computational and Applied Mathematics*, *340*, 424–431.

Ayala, A., Claeys, X., Escapil-Inchauspé, P., & Jerez-Hanckes, C. (2020). Local Multiple Traces Formulation for Electromagnetics: Stability and Preconditioning for Smooth Geometries. *arXiv preprint arXiv:2003.08330*.

Aylwin, R., Jerez-Hanckes, C., Schwab, C., & Zech, J. (2020). Domain Uncertainty Quantification in Computational Electromagnetics. *SIAM/ASA Journal on Uncertainty Quantification*, *8*(1), 301–341.

Bautista, M. A. E., Francavilla, M. A., Vipiana, F., & Vecchi, G. (2014). A Hierarchical Fast Solver for EFIE-MoM Analysis of Multiscale Structures at Very Low Frequencies. *IEEE Transactions on Antennas and Propagation*, *62*(3), 1523-1528.

Bebendorf, M. (2008). *Hierarchical Matrices: A Means to Efficiently Solve Elliptic Boundary Value Problems* (1st ed.). Berlin: Springer.

Bebendorf, M., Bollhöfer, M., & Bratsch, M. (2013). Hierarchical matrix approximation with blockwise constraints. *BIT Numerical Mathematics*, *53*(2), 311-339.

Bebendorf, M., & Kunis, S. (2009). Recompression techniques for adaptive cross approximation. *The Journal of Integral Equations and Applications*, *21*(3), 331–357.

Beckermann, B., Goreinov, S. A., & Tyrtyshnikov, E. E. (2005). Some Remarks on the Elman estimate for GMRES. *SIAM Journal on Matrix Analysis and Applications*, *27*(3), 772-778.

Beghein, Y., Mitharwal, R., Cools, K., & Andriulli, F. P. (2017). On a Low-Frequency and Refinement Stable PMCHWT Integral Equation Leveraging the Quasi-Helmholtz Projectors. *IEEE Transactions on Antennas and Propagation*, *65*(10), 5365–5375.

Benzi, M. (2016). Localization in Matrix Computations: Theory and Applications. In *Exploiting Hidden Structure in Matrix Computations: Algorithms and Applications* (pp. 211–317). Springer.

Bessoud, A. L., & Krasucki, F. (2006). *Q*-superlinear convergence of the GMRES algorithm for multi-materials with strong interface. *Comptes Rendus Mathématique*, *343*(4), 279–282.

Betcke, T., Scroggs, M. W., & Śmigaj, W. (2020). Product algebras for Galerkin discretisations of boundary integral operators and their applications. *ACM Transactions on Mathematical Software (TOMS)*, *46*(1), 1–22.

Betcke, T., van't Wout, E., & Gélat, P. (2017). Computationally efficient boundary element methods for high-frequency Helmholtz problems in unbounded domains. In *Modern Solvers for Helmholtz Problems* (pp. 215–243). Springer.

Blechta, J. (2021). Stability of Linear GMRES Convergence with respect to Compact Perturbations. *SIAM Journal on Matrix Analysis and Applications*, *42*(1), 436–447.

Brenner, S., & Scott, R. (2007). *The Mathematical Theory of Finite Element Methods* (Vol. 15). Springer Science & Business Media.

Buffa, A., & Christiansen, S. (2007). A dual finite element complex on the barycentric refinement. *Mathematics of Computation*, *76*(260), 1743–1769.

Buffa, A., & Hiptmair, R. (2003). Galerkin Boundary Element Methods for Electromagnetic Scattering. *Topics in Computational Wave Propagation*, 83–124.

Buffa, A., Hiptmair, R., von Petersdorff, T., & Schwab, C. (2003). Boundary Element Methods for Maxwell Transmission Problem in Lipschitz Domains. *Numerische Mathematik*, *95*(3), 459–485.

Bunse-Gerstner, A., & Gutiérrez-Cañas, I. (2006). A preconditioned GMRES for complex dense linear systems from electromagnetic wave scattering problems. *Linear algebra and its applications*, *416*(1), 135–147.

Campbell, S. L., Ipsen, I. C., Kelley, C. T., Meyer, C., & Xue, Z. (1996). Convergence Estimates for Solution of Integral Equations with GMRES. *The Journal of Integral Equations and Applications*, 19–34.

Campbell, S. L., Ipsen, I. C., Kelley, C. T., & Meyer, C. D. (1996). GMRES and the minimal polynomial. *BIT Numerical Mathematics*, *36*(4), 664–675.

Carpentieri, B., Duff, I. S., & Giraud, L. (2000). Experiments With Sparse Preconditioning of Dense Problems from Electromagnetic Applications. *CERFACS, Toulouse, France, Tech. Rep. TR/PA/00/04*, 9.

Céa, J. (1964). Approximation variationnelle des problèmes aux limites. *Annales de l'Institut Fourier*, *14*(2), 345–444.

Chandler-Wilde, S. N., Graham, I. G., Langdon, S., & Spence, E. A. (2012). Numerical-asymptotic boundary integral methods in high-frequency acoustic scattering. *Acta Numerica*, *21*, 89–305.

Chandler-Wilde, S. N., & Monk, P. (2008). Wave-Number-Explicit Bounds in Time-Harmonic Scattering. *SIAM Journal on Mathematical Analysis*, *39*(5), 1428–1455.

Chernov, A., Pham, D., & Tran, T. (2015). A shape calculus based method for a transmission problem with a random interface. *Computers & Mathematics with Applications*, *70*(7), 1401-1424.

Christiansen, S., & Nédélec, J.-C. (2001). A Preconditioner for the Electric Field Integral Equation based on Calderón Formulas. *SIAM Journal on Numerical Analysis*, *40*(3), 1100–1135.

Christiansen, S. H., & Nédélec, J.-C. (2000). Des préconditionneurs pour la résolution numérique des équations intégrales de frontière de l'acoustique. *Comptes Rendus de l'Académie des Sciences-Series I-Mathematics*, *330*(7), 617–622.

Claeys, X., Hiptmair, R., & Jerez-Hanckes, C. (2012). Multi-trace boundary integral equations. *Direct and Inverse Problems in Wave Propagation and Applications*, *14*, 51–100.

Colton, D., & Kress, R. (2012). *Inverse Acoustic and Electromagnetic Scattering Theory* (Vol. 93). Springer Science & Business Media.

Cools, K., Andriulli, F. P., & Michielssen, E. (2011). A Calderón Multiplicative Preconditioner for the PMCHWT Integral Equation. *IEEE Transactions on Antennas and Propagation*, *59*(12), 4579.

Dambrine, M., Harbrecht, H., & Puig, B. (2015). Computing Quantities of Interest for Random Domains with Second Order Shape Sensitivity Analysis. *ESAIM: Mathematical Modelling and Numerical Analysis*, *49*(5), 1285–1302.

Darve, E. (2000). The Fast Multipole Method: Numerical Implementation. *Journal of Computational Physics*, *160*(1), 195–240.

Dembart, B., & Yip, E. (1998). The Accuracy of Fast Multipole Methods for Maxwell's Equations. *IEEE Computational Science and Engineering*, *5*(3), 48-56.

Di Pietro, D. A., & Droniou, J. (2018). A third Strang lemma and an Aubin–Nitsche trick for schemes in fully discrete formulation. *Calcolo*, *55*(3), 1–39.

Dölz, J., & Harbrecht, H. (2018). Hierarchical Matrix Approximation for the Uncertainty Quantification of Potentials on Random Domains. *Journal of Computational Physics*, *371*, 506–527.

Dölz, J., Harbrecht, H., & Schwab, C. (2017). Covariance Regularity and $\mathcal{H}$-matrix Approximation for Rough Random Fields. *Numerische Mathematik*, *135*(4), 1045–1071.

Dunford, N., & Schwartz, J. T. (1963). *Linear Operators, Part 2: Spectral Theory, Self Adjoint Operators in Hilbert space*. Wiley.

Engquist, B., Ying, L., et al. (2009). A fast directional algorithm for high frequency acoustic scattering in two dimensions. *Communications in Mathematical Sciences*, *7*(2), 327–345.

Eriksson, J., Ollila, E., & Koivunen, V. (2010). Essential Statistics and Tools for Complex Random Variables. *IEEE Transactions on Signal Processing*, *58*(10), 5400–5408.

Ern, A., & Guermond, J.-L. (2006). Evaluation of the condition number in linear systems arising in finite element approximations. *ESAIM: Mathematical Modelling and Numerical Analysis*, *40*(1), 29–48.

Ern, A., & Guermond, J.-L. (2013). *Theory and Practice of Finite Elements* (Vol. 159). Springer Science & Business Media.

Escapil-Inchauspé, P., & Jerez-Hanckes, C. (2019). Fast Calderón Preconditioning for the Electric Field Integral Equation. *IEEE Transactions on Antennas and Propagation*, *67*(4), 2555–2564.

Faber, V., Manteuffel, T. A., & Parter, S. V. (1990). On the Theory of Equivalent Operators and Application to the Numerical Solution of Uniformly Elliptic Partial Differential Equations. *Advances in Applied Mathematics*, *11*(2), 109–163.

Faustmann, M., Melenk, J. M., & Praetorius, D. (2015). Existence of $\mathcal{H}$-matrix approximants to the inverse of BEM matrices: the hyper-singular integral operator. *IMA Journal of Numerical Analysis*, *37*(3), 1211–1244.

Feischl, M., Führer, T., Praetorius, D., & Stephan, E. P. (2017). Optimal preconditioning for the symmetric and nonsymmetric coupling of adaptive finite elements and boundary elements. *Numerical Methods for Partial Differential Equations*, *33*(3), 603–632.

Fierro, I., & Jerez-Hanckes, C. (2020). Fast Calderón preconditioning for Helmholtz boundary integral equations. *Journal of Computational Physics*, *409*, 109355.

Fuenzalida, C., Jerez-Hanckes, C., & McClarren, R. G. (2019). Uncertainty Quantification for Multigroup Diffusion Equations using Sparse Tensor Approximations. *SIAM Journal on Scientific Computing*, *41*(3), B545-B575.

Galkowski, J., Müller, E. H., & Spence, E. A. (2019). Wavenumber-explicit analysis for the Helmholtz h-BEM: error estimates and iteration counts for the Dirichlet problem. *Numerische Mathematik*, *142*(2), 329–357.

Galkowski, J., Spence, E. A., & Wunsch, J. (2019). Optimal constants in nontrapping resolvent estimates and applications in numerical analysis. *Pure and Applied Analysis*, *2*(1), 157–202.

Gander, M. J., Graham, I. G., & Spence, E. A. (2015). 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.

Gossye, M., Huynen, M., Ginste, D. V., De Zutter, D., & Rogier, H. (2018). A Calderón Preconditioner for High Dielectric Contrast Media. *IEEE Transactions on Antennas and Propagation*, *66*(2), 808–818.

Graham, I., Spence, E., & Vainikko, E. (2017). Domain decomposition preconditioning for high-frequency Helmholtz problems with absorption. *Mathematics of Computation*, *86*(307), 2089–2127.

Griebel, M., & Harbrecht, H. (2014). On the Convergence of the Combination Technique. In *Sparse Grids and Applications-Munich 2012* (Vol. 97, pp. 55–74). Springer.

Griebel, M., & Knapek, S. (2009). Optimized general sparse grid approximation spaces for operator equations. *Mathematics of Computation*, *78*(268), 2223–2257.

Griebel, M., Schneider, M., & Zenger, C. (1990). *A Combination Technique for the Solution of Sparse Grid Problems*.

Gumerov, N. A., & Duraiswami, R. (2004). *Fast Multipole Method for the Helmholtz Equation in Three Dimension*. Elsevier Science.

Guo, H., Hu, J., Yin, J., & Nie, Z. (2009). An improved Calderón preconditioner for electric field integral equation. In *Microwave Conference, 2009. APMC 2009. Asia Pacific* (pp. 92–95).

Haldenwang, P., Labrosse, G., Abboudi, S., & Deville, M. (1984). Chebyshev 3-D Spectral and 2-D Pseudospectral Solvers for the Helmholtz Equation. *Journal of Computational Physics*, *55*, 115–128.

Harbrecht, H., Ilić, N., & Multerer, M. D. (2019). Rapid computation of far-field statistics for random obstacle scattering. *Engineering Analysis with Boundary Elements*, *101*, 243–251.

Harbrecht, H., & Peters, M. (2013). Comparison of fast boundary element methods on parametric surfaces. *Computer Methods in Applied Mechanics and Engineering*, *261*, 39–55.

Harbrecht, H., Peters, M., & Siebenmorgen, M. (2013). Combination technique based k-th moment analysis of elliptic problems with random diffusion. *Journal of Computational Physics*, *252*, 128–141.

Harbrecht, H., Schneider, R., & Schwab, C. (2008). Sparse second moment analysis for elliptic problems in stochastic domains. *Numerische Mathematik*, *109*(3), 385–414.

Hestenes, M. R., & Stiefel, E. (1952). *Methods of conjugate gradients for solving linear systems* (Vol. 49) (No. 1). NBS Washington, DC.

Hiptmair, R. (2006). Operator Preconditioning. *Computers and Mathematics with Applications*, *52*(5), 699–706.

Hiptmair, R., Jerez-Hanckes, C., & Mao, S. (2015). Extension by zero in discrete trace spaces: Inverse estimates. *Mathematics of Computation*, *84*, 2589–2615.

Hiptmair, R., & Li, J. (2017). Shape derivatives in Differential Forms II: Shape Derivatives for Scattering Problems. *Foundations of computational mathematics*.

Hiptmair, R., & Urzúa-Torres, C. (2020). Preconditioning the EFIE on screens. *Mathematical Models and Methods in Applied Sciences*, *30*(09), 1705–1726.

Hsiao, G. C., & Wendland, W. L. (2008). *Boundary Integral Equations*. Springer Berlin Heidelberg.

Jerez-Hanckes, C., & Schwab, C. (2016). Electromagnetic wave scattering by random surfaces: uncertainty quantification via sparse tensor boundary elements. *IMA Journal of Numerical Analysis*, *37*(3), 1175–1210.

Kirby, R. C. (2010). From Functional Analysis to Iterative Methods. *SIAM Review*, *52*(2), 269–293.

Kleanthous, A., Betcke, T., Hewett, D. P., Escapil-Inchauspé, P., Jerez-Hanckes, C., & Baran, A. J. (2020). Accelerated Calderón preconditioning for Maxwell transmission problems. *arXiv preprint arXiv:2008.04772*.

Kleanthous, A., Betcke, T., Hewett, D. P., Scroggs, M. W., & Baran, A. J. (2018). Calderón preconditioning of PMCHWT boundary integral equations for scattering by multiple absorbing dielectric particles. *Journal of Quantitative Spectroscopy and Radiative Transfer*.

Kurics, T. (2010). *Operator Preconditioning in Hilbert Space* (Unpublished doctoral dissertation). Eötvös Loránd University.

Liesen, J., & Tichỳ, P. (2012). The field of values bound on ideal GMRES. *arXiv preprint arXiv:1211.5969*.

Malas, T., & Gurel, L. (2007). Incomplete LU Preconditioning with the Multilevel Fast Multipole Algorithm for Electromagnetic Scattering. *SIAM Journal on Scientific Computing*, *29*(4), 1476–1494.

McLean, W. (2000). *Strongly Elliptic Systems and Boundary Integral Equations*. Cambridge University Press.

McLean, W., & Steinbach, O. (1999). Boundary element preconditioners for a hypersingular integral equation on an interval. *Advances in Computational Mathematics*, *11*(4), 271–286.

Meade Jr, A. J., & Fernandez, A. A. (1994). The numerical solution of linear ordinary differential equations by feedforward neural networks. *Mathematical and Computer Modelling*, *19*(12), 1–25.

Megginson, R. E. (2012). *An Introduction to Banach Space Theory* (Vol. 183). Springer Science & Business Media.

Miyanishi, Y., & Suzuki, T. (2015). Eigenvalues and Eigenfunctions of Double Layer Potentials. *Transactions of the American Mathematical Society*, *369*.

Monk, P., et al. (2003). *Finite Element Methods for Maxwell's Equations*. Oxford University Press.

Moret, I. (1997). A note on the superlinear convergence of GMRES. *SIAM Journal on Numerical Analysis*, *34*(2), 513–516.

Nédélec, J.-C. (2001). *Acoustic and Electromagnetic Equations: Integral Representations for Harmonic Problems* (Vol. 144). Springer Science & Business Media.

Nevanlinna, O. (1993). *Convergence of Iterations for Linear Equations*. Birkhauser Verlag.

Niino, K., Akagi, S., & Nishimura, N. (2017). A Discretization Method with the H_{div} Inner product for Electric Field Integral Equations. *IEEE Transactions on Antennas and Propagation*, *65*(6), 3102–3113.

Omar, S., & Jiao, D. (2014). $O(N)$ iterative and $O(N\log N)$ direct volume integral equation solvers for large-scale electrodynamic analysis. In *2014 International Conference on Electromagnetics in Advanced Applications (ICEAA)* (p. 593-596).

Rao, S., Wilton, D., & Glisson, A. (1982). Electromagnetic scattering by surfaces of arbitrary shape. *IEEE Transactions on Antennas and Propagation*, *30*(3), 409–418.

Saad, Y. (2003). *Iterative Methods for Sparse Linear Systems* (2nd ed.). USA: Society for Industrial and Applied Mathematics.

Saad, Y., & Schultz, M. H. (1986). GMRES: A generalized minimal residual algorithm for solving nonsymmetric linear systems. *SIAM Journal on Scientific and Statistical Computing*, *7*(3), 856–869.

Sappl, J., Seiler, L., Harders, M., & Rauch, W. (2019). Deep Learning of Preconditioners for Conjugate Gradient Solvers in Urban Water Related Problems. *arXiv preprint arXiv:1906.06925*.

Sarkis, M., & Szyld, D. B. (2007). Optimal left and right additive Schwarz preconditioning for minimal residual methods with Euclidean and energy norms. *Computer Methods in Applied Mechanics and Engineering*, *196*(8), 1612–1621.

Sauter, S. A., & Schwab, C. (2010). Boundary Element Methods. In *Boundary element methods* (pp. 183–287). Springer.

Scroggs, M. W., Betcke, T., Burman, E., Śmigaj, W., & van't Wout, E. (2017). Software frameworks for integral equations in electromagnetic scattering based on Calderón identities. *Computers & Mathematics with Applications*.

Shen, J., Tang, T., & Wang, L. L. (2011). Spectral methods: algorithms, analysis and applications. *Spectral methods: algorithms, analysis and applications*.

Silva-Oelker, G., Aylwin, R., Jerez-Hanckes, C., & Fay, P. (2018). Quantifying the Impact of Random Surface Perturbations on Reflective Gratings. *IEEE Transactions on Antennas and Propagation*, *66*(2), 838-847.

Śmigaj, W., Arridge, S., Betcke, T., Phillips, J., & Schweiger, M. (2015). Solving Boundary Integral Problems with BEM++. *ACM Trans. Math. Software*, *2*(41), 6:1–6:40.

Sobolev, A. V. (2014). On the Schatten–von Neumann properties of some pseudo-differential operators. *Journal of Functional Analysis*, *266*(9), 5886-5911.

Sokolowski, J., & Zolesio, J.-P. (1992). Introduction to Shape Optimization. In *Introduction to Shape Optimization* (pp. 5–12). Springer.

Spence, E. A. (2014). Wavenumber-Explicit Bounds in Time-Harmonic Acoustic Scattering. *SIAM Journal on Mathematical Analysis*, *46*(4), 2987–3024.

Starke, G. (1997). Field-of-values analysis of preconditioned iterative methods for nonsymmetric elliptic problems. *Numerische Mathematik*, *78*(1), 103–117.

Steinbach, O. (2007). *Numerical Approximation Methods for Elliptic Boundary Value Problems: Finite and Boundary Elements*. Springer New York.

Steinbach, O., & Wendland, W. L. (1998). The construction of some efficient preconditioners in the boundary element method. *Advances in Computational Mathematics*, *9*(1-2), 191–216.

Stevenson, R., & van Venetië, R. (2021). Uniform Preconditioners of Linear complexity for Problems of Negative Order. *Computational Methods in Applied Mathematics*, *21*(2), 469–478.

Strang, G. (1972). Variational crimes in the finite element method. In *The mathematical foundations of the finite element method with applications to partial differential equations* (pp. 689–710). Elsevier.

Sun, D.-L., Carpentieri, B., Huang, T.-Z., Jing, Y.-F., & Naveed, S. (2018). Variants of the Block-GMRES Method for Solving Linear Systems with Multiple Right-Hand Sides. In *2018 International Workshop on Computing, Electromagnetics, and Machine Intelligence (CEMi)* (pp. 13–14).

Thierry, B. (2014). A remark on the single scattering preconditioner applied to boundary integral equations. *Journal of Mathematical Analysis and Applications*, *413*(1), 212-228.

Van der Vorst, H. A., & Vuik, C. (1993). The superlinear convergence behaviour of GMRES. *Journal of Computational and Applied Mathematics*, *48*(3), 327–341.

von Petersdorff, T., & Schwab, C. (2006). Sparse Finite Element Methods for Operator Equations with Stochastic Data. *Applications of Mathematics*, *51*(2), 145–180.

Winther, R. (1980). Some Superlinear Convergence Results for the Conjugate Gradient Method. *SIAM Journal on Numerical Analysis*, *17*(1), 14–17.

Xu, H., Bo, Y., & Zhang, M. (2016). Combining Calderón preconditioner and $\mathcal{H}^2$-matrix method for solving electromagnetic scattering problems. In *2016 IEEE International Workshop on Electromagnetics: Applications and Student Innovation Competition (iWEM)* (p. 1-3).

Zanni, J., & Kubrusly, C. (2015). A note on compactness of tensor products. *Acta Mathematica Universitatis Comenianae*, *84*(1), 59–62.