Méthode de base réduite pour des problèmes linéaires dépendants de paramètres. Application aux problèmes harmoniques en électromagnétisme et en aéroacoustique

Philip Edel

▶ **To cite this version:**

Philip Edel. Méthode de base réduite pour des problèmes linéaires dépendants de paramètres. Application aux problèmes harmoniques en électromagnétisme et en aéroacoustique. Equations aux dérivées partielles [math.AP]. Sorbonne Université, 2022. Français. NNT : 2022SORUS270 . tel-03852900

HAL Id: tel-03852900
https://theses.hal.science/tel-03852900

Submitted on 15 Nov 2022

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

ÉCOLE DOCTORALE DE SCIENCES MATHÉMATIQUES DE PARIS CENTRE

Thèse de Doctorat

en vue de l'obtention du grade de

Docteur ès Sciences de Sorbonne Université

Discipline : Mathématiques Appliquées

présentée par

Philip EDEL

Reduced basis method for parameter-dependent linear equations. Application to time-harmonic problems in electromagnetism and in aeroacoustics.

---

dirigée par

Yvon MADAY

Soutenue le 24 octobre 2022
devant le jury présidé par M. Bruno Després, composé de

M. Yvon Maday	Sorbonne Université	Directeur de thèse
M. Ludovic Chamoin	ENS Paris Saclay	Rapporteur
Mme Virginie Ehrlacher	École des Ponts ParisTech	Rapporteure
M. Bruno Després	Sorbonne Université	Examinateur
M. Gianluigi Rozza	Scuola Internazionale Superiore di Studi Avanzati (SISSA)	Examinateur
Mme Anne-Sophie Bonnet-Ben Dhia	ENSTA Paris	Examinatrice
M. Anthony Patera	Massachusetts Institute of Technology (MIT)	Examinateur
M. François-Xavier Roux	ONERA	Co-encadrant de thèse

Abstract/Résumé

Reduced basis method for parameter-dependent linear equations. Application to time-harmonic problems in electromagnetism and in aeroacoustics.

Abstract

Many engineering applications require the solutions of a partial differential equation (PDE) for a vast set of parameter configurations. Despite the use of efficient numerical methods and algorithms to solve the PDE, the computational costs associated with repeated solves for different parameter configurations can be prohibitive. In this thesis, we explore the use of the reduced basis method (RBM) to accelerate parametric simulation campaigns with linear PDEs. The first part of this thesis is mainly focused on error estimation strategies. We propose an easy-to-implement heuristic method for problems with smooth and slow-varying inf-sup stability constants. For close-to-degenerate and potentially resonant problems, we introduce a rigorous error estimator based on the dual natural-norm of the residual. We generalize the error estimation approach to problems with multiple sources and derive a block version of the RBM. The second part of this thesis is mostly concerned with applications of the RBM to frequency-parametrized time-harmonic Maxwell's equations in electromagnetism and impedance-parametrized time-harmonic linearized Euler equations in aeroacoustics. We propose a non-intrusive RBM specifically tailored for frequency-sweeps with surface integral equations discretized with the boundary element method. Numerical illustrations confirm the benefits of the RBM, in particular when applied to real-world industrial problems.

Key words: Reduced basis method, model order reduction, finite element method, boundary element method, computational aeroacoustics, computational electromagnetism

Méthode de base réduite pour des problèmes linéaires dépendants de paramètres. Application aux problèmes harmoniques en éléctromagnétisme et en aéroacoustique.

Résumé

De nombreuses applications en sciences appliquées nécessitent la résolution successive d'une équation aux dérivées partielles (EDP) pour un vaste ensemble de valeurs de paramètres. Malgré la mise en œuvre de méthodes numériques et d'algorithmes efficaces pour résoudre l'EDP, les coûts de calcul associés à de nombreuses résolutions successives pour des paramètres différents peuvent être prohibitifs. Dans cette thèse, nous considérons la méthode de base réduite pour accélérer les campagnes de résolution paramétrique des EDPs linéaires. Dans la première partie de la thèse, nous nous focalisons sur la problématique d'estimation d'erreur. Nous proposons une méthode heuristique d'estimation d'erreur facile à implémenter et pertinente pour des problèmes caractérisés par une constante de stabilité inf-sup régulière et peu dépendante des paramètres. Pour les problèmes potentiellement résonants, nous introduisons un estimateur d'erreur rigoureux, basé sur la norme naturelle duale du résidu. Nous généralisons l'estimation d'erreur au cas des problèmes multi-sources et dérivons une version block de la méthode de base réduite. Dans la deuxième partie de la thèse, nous nous intéressons aux applications de la méthode aux équations de Maxwell harmoniques en contexte multi-fréquences et aux équations d'Euler linéarisées harmoniques en contexte multi-impédances. Pour les problèmes multi-fréquences en diffraction électromagnétique résolus par des équations intégrales de surface discrétisées par la méthode des éléments de frontière, nous proposons une version non-intrusive originale de la méthode de base réduite. Des exemples numériques illustrent l'intérêt de la méthode, en particulier pour des problèmes de taille industrielle.

Mots clefs : Méthode base réduite, réduction d'ordre, réduction de modèle, éléments finis, éléments finis de frontière, aéroacoustique, électromagnétisme

Remerciements

Tout d'abord, je tiens à remercier mon directeur de thèse Yvon Maday du Laboratoire Jacques-Louis Lions (LJLL). Merci de m'avoir transmis ton enthousiasme pour les méthodes de base réduite : je n'aurai pas pu trouver meilleur directeur de thèse que toi pour aborder ces méthodes là. Je te dois beaucoup de m'avoir aiguillé sur les bonnes pistes tout au long de ces trois années. Ton point de vue sur mes travaux m'a toujours été très précieux, ainsi que ton soutien. J'ai vraiment apprécié travailler avec toi. Merci !

Je tiens ensuite à remercier mon encadrant, François-Xavier Roux du LJLL et ingénieur de recherche à l'ONERA. Tu as réussi à rassembler une foule de collaborateurs autour de mon projet de thèse, ce qui en fait toute la richesse, je t'en remercie. J'ai réellement bénéficié des collaborations que tu as initiées à l'ONERA en électromagnétisme et en aéroacoustique. C'était une véritable chance pour moi de travailler dans un contexte non seulement de développement de méthodes numériques "en amont", mais aussi d'applications concrètes sur des problèmes réels. J'ai énormément progressé à tes côtés, tant en connaissances qu'en bonnes pratiques de développement. Enfin, tu m'as toujours soutenu et "pris sous ton aile" à l'ONERA, ce pour quoi je te suis très reconnaissant.

Je remercie mes deux rapporteurs: Virginie Ehrlacher (École des Ponts ParisTech) et Ludovic Chamoin (ENS Paris-Saclay) pour leur lecture attentive de ce manuscrit de thèse et pour avoir pris part au jury le jour de ma soutenance.

Je remercie également mes examinateurs: Anne-Sophie Bonnet-Ben Dhia (ENSTA Paris), Anthony Patera (MIT) Gianluigi Rozza (SISSA Trieste) et Bruno Després (LJLL) qui a aussi été Président du jury. C'était un honneur et un plaisir d'avoir soutenu ma thèse devant vous et d'avoir répondu à vos questions.

Je tiens enfin à remercier le LJLL pour m'avoir permis de soutenir dans ses locaux du campus Pierre et Marie Curie dans d'excellentes conditions, en particulier grâce à Corentin Maday qui m'a apporté une aide logistique et matérielle de grande qualité.

Je voudrais témoigner ma profonde reconnaissance envers l'ONERA qui a financé ce projet de thèse, m'a accueilli dans ses locaux de Palaiseau pendant ces trois années et a mis à ma disposition d'importants moyens de calculs. Fait extrêmement rare, j'ai eu l'opportunité d'interagir avec quatre départements différents de l'ONERA : le Départe-

ment Traitement de l'Information et Systèmes (DTIS) où j'étais accueilli, le Département Électromagnétisme et Radar (DEMR) avec qui j'ai collaboré sur les applications antennes et Radar, le Département Multi-Physique pour l'Énergétique (DMPE) et le Département Aérodynamique Aéroélasticité et Acoustique (DAAA) avec qui j'ai collaboré sur les applications liners en aéroacoustique.

Du DEMR, je remercie tout d'abord André Barka pour sa collaboration sur les applications antennes, notamment pour m'avoir fourni les modèles numériques d'antennes.

Côté radar, je tiens tout particulièrement à remercier Jérôme Simon : merci pour ton enthousiasme autour de la méthode base réduite, qui m'a franchement stimulé et motivé. Merci pour cette belle dynamique de travail que tu as créée, notamment en m'impliquant dans le PRF MONALISA aux côtés de Vincent Gobin (DEMR), Xavier Juvigny (DAAA), Sébastien Pernet (DTIS), David Levadou (DTIS), Mathieu Lorteau (DAAA) que je remercie tous pour les discussions enrichissantes autour de mes travaux de thèse. Je suis sincèrement reconnaissant envers toi Jérôme pour le soutien indéfectible que tu m'as apporté durant la dernière ligne droite de mon projet de thèse. Enfin, merci de m'avoir fait confiance pour co-encadrer avec toi le stage de Prisca LeDily. Merci d'avoir géré l'organisation de ce stage que j'ai eu beaucoup de plaisir à co-encadrer avec toi.

Quant à toi Prisca, merci de nous avoir choisis Jérôme et moi pour ton stage de fin d'étude : tu as été une excellente stagiaire, tu as fourni un travail remarquable et j'ai été très heureux de t'encadrer pendant tes six mois de stage. Ton stage a été une véritable source de motivation et d'idées pour moi et je t'en remercie.

Je tiens à remercier Christophe Peyret du DAAA, qui m'a ouvert la voie vers les applications aéroacoustiques. Merci Christophe de m'avoir proposé le modèle numérique de l'entrée d'air de nacelle d'avion, qui a vraiment permis de montrer l'intérêt de la méthode base réduite sur un cas industriel. J'ai apprécié ta grande disponibilité et ta patience : je ne sais pas combien de fois je t'ai demandé de re-compiler le code sur Spiro ou Topaze, mais à chaque fois tu étais d'une grande efficacité pour répondre à mes besoins ce qui m'a fait gagner un temps précieux. J'ai pris beaucoup de plaisir à interagir avec toi, tant dans les phases de développement que pour faire tourner les campagnes de simulation.

Toujours en lien avec l'aéroacoustique mais côté DMPE de Toulouse, je remercie Rémi Roncen. Merci pour tout l'entrain que tu as mis dans la méthode base réduite, c'était réellement stimulant pour moi. J'ai apprécié ta casquette d'expérimentateur/concepteur de liners aéroacoustiques, qui m'a permis de replacer mes travaux dans un contexte beaucoup plus large. Merci d'avoir donné tant d'élan à mes travaux sur les bases réduites, c'est notamment par ton élan que le papier AIAA a pu se concrétiser.

Je voudrais exprimer ma gratitude envers Christophe Peyret et Rémi Roncen pour m'avoir confié la responsabilité de co-encadrer avec eux le stage d'Ancelin Rocamora. J'aimerais tout particulièrement remercier Ancelin : tu m'as apporté un soutien indispensable pour porter à maturation l'application de la méthode base réduite au calcul des cartes d'atténuation acoustique sur le cas de l'entrée d'air de nacelle d'avion. Les six mois que tu as passés à mes côtés ont été formidables. Je te remercie pour le très bon travail que tu as effectué dans la bonne humeur et toujours avec le sourire.

Je remercie le DTIS qui m'a accueilli. J'aimerais surtout remercier le troisième étage du batiment N avec qui j'ai passé une grande partie de mon quotidien (quand il ne fallait pas télétravailler pour cause covid). A commencer par l'unité Image Vision Apprentissage (IVA) du DTIS, dont je ne faisais pas partie mais qui m'a accueilli comme si. Merci au chef d'unité, Martial Sanfourche, pour son aide précieuse en fin de thèse. Et bien sûr, merci aux membres d'IVA qui sont venus en force depuis Palaiseau pour m'encourager le jour de ma soutenance !

Merci à ma chère co-bureau : Pelin. Tu as suivi ma dernière année de thèse de près en m'apportant beaucoup d'écoute et de soutien, vraiment merci, tu as été une excellente co-bureau que je regrette déjà ! Merci à toi Anthelme pour ta compagnie : j'étais heureux d'occuper le bureau à côté du tien pendant ces trois années, car tu es bien le seul à être resté fidèle à l'aile A pendant tout ce temps !

Enfin, merci à tous les doctorants et permanents de la salle de pause d'avoir partagé tous ces moments de détente avec moi. Merci aux anciens: Guillaume "le petit", Guillaume "le grand", Rodolphe, Rodrigo, Javiera, Pierre, Alexis, Louis, Gaston... c'était un plaisir de partager toutes ces jongleries et ces parties de tamalou avec vous ! Merci à Rémi, Maxime, Nathan, Thomas, Quentin, Marius, Kévin, Adrien, Pol, Dao, Flora, Pierre, Baptiste, Clara, Mathias, Louis,... pour ces innombrables parties de tarot, de coinche, ces excellents gâteaux et tous ces moments conviviaux qui m'ont permis de recharger mes batteries !

Pour finir merci à mes parents, qui n'ont cessé de me soutenir pendant ces trois années et merci à mes frères et sœurs. Enfin, merci à toi Alice : tu as partagé toutes les difficultés de cette thèse avec moi ; aujourd'hui, tu en partages aussi toute la réussite.

Table of contents

Résumé/Abstract **ii**

Remerciements **v**

Introduction **1**

1 Basic principles and properties of the reduced basis method **5**
1.1 Introduction to parametrized PDEs 6
1.1.1 Two parametrized problems 6
1.1.2 Weak forms and well-posedness 7
1.1.3 High-fidelity approximation 8
1.1.4 Mathematical framework for parametrized PDEs 10
1.2 The reduced basis method 12
1.2.1 Model order reduction and Kolmogorov width 12
1.2.2 Choice of subspace 13
1.2.3 Choice of approximation 15
1.3 Efficiency of the reduced basis method 18
1.3.1 Affine operators and right-hand sides 18
1.3.2 Computational strategy in the affine case 19
1.3.3 The non-affine case 22
1.4 Numerical illustration 24
1.4.1 Model problem 1: Helmholtz 24
1.4.2 Model problem 2: Laplace 27
1.5 Conclusions 29

2 The classical *a posteriori* error estimation approach **31**
2.1 The need for *a posteriori* error estimation 32
2.1.1 The need for certification 32
2.1.2 Greedy RB approach 33
2.2 Classical inf-sup based error estimates 34
2.2.1 Error bound 34

2.2.2 Efficient computation of the residual norm . . . 36
2.2.3 Efficient computation of the inf-sup constant . . . 39
2.2.4 A heuristic approach . . . 44
2.3 Numerical illustration . . . 46
2.3.1 Model problem 1: Helmholtz . . . 46
2.3.2 Model problem 2: Laplace . . . 48
2.4 Conclusions . . . 51

3 The natural-norm *a posteriori* error estimation approach 52
3.1 Natural-norm error bounds . . . 53
3.1.1 Reminders . . . 53
3.1.2 Error bound using the primal natural-norm . . . 53
3.1.3 Error bound using the dual natural-norm . . . 55
3.1.4 Re-interpretation of the primal natural-norm approach as a right-preconditionning approach . . . 58
3.2 Practical natural-norm *a posteriori* error estimators . . . 59
3.2.1 Practical inf-sup based and primal natural-norm error estimators . 59
3.2.2 Practical error estimator based on dual natural-norm . . . 61
3.2.3 The self-adjoint case . . . 63
3.3 Computational strategy . . . 64
3.3.1 Offline/online strategy . . . 64
3.3.2 Procedure for selecting anchor points . . . 66
3.4 Numerical results . . . 67
3.4.1 Problem setting . . . 67
3.4.2 Self-adjoint case . . . 68
3.4.3 Non self-adjoint case . . . 72
3.5 Conclusions . . . 75

4 The reduced basis method in the context of multiple sources 77
4.1 Introduction to parametrized problems with multiple sources . . . 78
4.1.1 The parametrized problem with ℓ independent sources . . . 78
4.1.2 Outline of two reduced basis strategies . . . 80
4.2 Multiple RBs strategy . . . 81
4.2.1 Motivation . . . 81
4.2.2 Reduced basis approximation in a glance . . . 82
4.2.3 Greedy construction . . . 82
4.3 Unique RB strategy . . . 83
4.3.1 A short review . . . 83
4.3.2 The block formulation . . . 84
4.3.3 Block reduced basis approximations . . . 85
4.3.4 Greedy construction . . . 87
4.3.5 Offline/Online strategy and complexity analysis . . . 88
4.4 The parametrized problem with parametrized source . . . 90
4.4.1 Problem formulation . . . 90

4.4.2 Block approximation . . . 92
4.4.3 Affine approximation of the block RHS . . . 93
4.5 Numerical illustration . . . 95
4.5.1 The parametrized problem with ℓ independent source terms . . . 95
4.5.2 The parametrized problem with parametrized source . . . 95
4.6 Conclusions . . . 103

5 Reduced basis method for frequency sweeps with edge finite elements 105
5.1 Strong formulation and high-fidelity approximation . . . 106
5.1.1 Governing equations . . . 106
5.1.2 High-fidelity discretization . . . 108
5.1.3 Numerical solver: FETI-2LM . . . 110
5.2 The RBM for the frequency sweep problem . . . 112
5.2.1 The frequency-parametrized problem . . . 112
5.2.2 Results on the Horn Antenna test case . . . 113
5.3 Application to antenna arrays . . . 117
5.3.1 RBM with a single right-hand side . . . 118
5.3.2 RBM with multiple right-hand sides . . . 119
5.4 Conclusions . . . 121

6 Reduced basis method for frequency sweeps with the boundary element method 122
6.1 Strong formulation and high-fidelity approximation . . . 123
6.1.1 The Stratton-Chu integral representation formulas . . . 123
6.1.2 Scattering by a perfect electric conductor . . . 124
6.1.3 Discretization using the BEM . . . 128
6.1.4 Numerical solvers . . . 129
6.2 A non-intrusive RBM for frequency sweep analysis . . . 130
6.2.1 Affine approximations for the frequency-parametrized problem . . . 130
6.2.2 Non-intrusive local affine approximations . . . 132
6.2.3 Non-intrusive RB approximation . . . 135
6.2.4 Greedy construction . . . 136
6.2.5 Monolithic construction . . . 138
6.3 Numerical illustrations . . . 139
6.3.1 EFIE vs CFIE: tests on the unit sphere . . . 139
6.3.2 Approximation of the CFIE operator on the geometry of a fighter aircraft . . . 142
6.3.3 RB approximations on the geometry of a fighter aircraft . . . 143
6.3.4 Broadband frequency-sweeps . . . 146
6.4 Conclusions . . . 150

7 Reduced basis method for aeroacoustic liner optimization in a discontinuous-Galerkin framework 151
7.1 Strong formulation and high-fidelity approximation . . . 152
7.1.1 The time-harmonic linearized Euler equations . . . 152

7.1.2 Discontinuous Galerkin scheme 155
7.2 The RBM for liner optimization 157
7.2.1 The impedance-parametrized problem 157
7.2.2 Results on the test case ALIAS 159
7.2.3 Validation campaign on the test case ALIAS 163
7.3 The RBM applied to a 3D nacelle engine 164
7.3.1 Problem description 164
7.3.2 Reduced basis approach 165
7.3.3 Validation campaign 168
7.4 Conclusions 169

Conclusions **171**

A Methods for solving large-scale generalized eigenvalue problems **175**
A.1 Context 175
A.2 Inverse iteration 176
A.3 Lanczos method 176

B Implementation details **180**
B.1 Four possible offline phases 180
B.2 Reduced basis updates throughout the greedy iterations 181

Introduction

Context

Many fields in engineering or applied sciences rely on numerical simulation for their predictive power or for system design and optimization [107]. Clearly, the ever-increasing available computational resources cannot answer the growing demand for fast and high-fidelity numerical simulations alone. Indeed, the development of efficient numerical methods and algorithms is crucial not only to ensure the reliability of the simulations, but also to make the best use of the available computational power.

This thesis deals with numerical simulation of partial differential equations (PDEs) in time-harmonic electromagnetism and aeroacoustics. Using *ad-hoc* discretization techniques such as the finite element method [38], the computation of the solution field (*i.e.*, the electric and magnetic fields in electromagnetism or the acoustic pressure and velocity fields in aeroacoustics) at given geometry, material properties and frequency is brought down to solving a large-scale linear system. A plethora of dedicated methods can be used to efficiently solve such a large-scale linear system on parallel architectures [106]. However, when the user is interested in varying the material properties or the frequency, which typically occurs in an optimization context, the cost of repeatedly solving large-scale linear systems can be computationally demanding and may involve several hours or days on supercomputers.

Model order reduction is a possible, well-known approach to relieve the computational burden associated with numerical simulations with varying parameters. It consists in replacing the large-scale, high-fidelity numerical model by a reduced model, featuring much less degrees of freedom [109]. Evaluating the solution of the reduced model at any parameter query is very fast. However, this approach only makes sense if a reduced model producing reliable solutions is able to be built. Typically, it is expected from the reduced model to provide cheap reduced solutions which are ”adequately close” to the costly high-fidelity solutions. Of course, the costly high-fidelity solutions should not have to be computed in order to check whether or not the reduced solutions are ”adequately

close", hence the key notion of *a posteriori* error estimation [96, 123, 83].

Among the various model order reduction techniques, the focus of this thesis is on the reduced basis method (RB method or RBM) [97, 54]. It consists in seeking the reduced solutions as linear combinations of a small number of high-fidelity solutions computed for a small set of parameter values. The overall goal of this thesis is to investigate all aspects of the reduced basis method (both theoretical and algorithmic) in order to set up the best strategies for parameter-dependent linear equations. The objectives are:

- *efficiency*, for the best computational performance and
- *reliability*, in order to certify the quality of the output of the reduced models.

In terms of applications, this thesis is specifically concerned with time-harmonic problems in electromagnetism and aeroacoustics. In particular, we will focus on three applications of the reduced basis method:

- for antenna applications in electromagnetism, we want to solve the pattern of radiating antennas over a frequency band,
- for RADAR applications in electromagnetism, we are interested in resolving the scattering of a plane wave on a metallic object over a frequency band,
- finally, for liner applications in aeroacoustics, we seek to vary the material properties of an aircraft engine nacelle in order to identify the best material properties such that the noise attenuation is maximized.

Notice that the two applications in electromagnetism are concerned with the efficient prediction of input/output responses, while the last application in aeroacoustics deals with the efficient optimization of a system. Different discretization techniques are used for each application: the antenna applications rely on edge finite elements, the scattering applications employ the boundary element method and finally the aeroacoustic applications use the discontinuous-Galerkin method. Our aim is to show that the reduced basis method is relevant in various discrete frameworks.

Content of the thesis

This thesis is divided into seven chapters. The first four chapters are general and not necessarily thought for any particular application, while the chapters 5, 6 and 7 tackle the specific applications motivating this thesis.

Chapters 1 and 2

The first two chapters provide a quick introduction to high fidelity discretization techniques and a short survey of reduced basis methods. The focus is on the finite element

approximation of coercive and weakly coercive PDEs, which is the proper mathematical framework to address the time-harmonic problems in electromagnetism and aeroacoustics targeted in this thesis. Two contributions are to be found in Chapter 2: (i) a robust formulation of the least-squares reduced basis problem and (ii) a heuristic method for approximating the inf-sup stability factor. For illustration, we propose academic numerical examples on a conductivity-parametrized Laplace equation and on a frequency-parametrized Helmholtz equation.

Chapter 3

Chapter 3 is an original contribution, where we introduce a *dual natural-norm* for measuring the reduced basis residual. We derive *a posteriori* error estimators approach based the dual natural-norm which drastically reduces the amount of overestimation compared to the classical *a posteriori* error estimation approach based on the inf-sup stability constant. Numerical examples on a resonant Helmholtz problem confirm the potential of the dual-natural norm for estimating the reduced basis approximation error.

Chapter 4

In chapter 4, we explore the reduced basis method in the context of parametrized problems with multiple sources. The originality of our work is that we successively enrich the reduced basis not with one basis function (corresponding to the PDE solution at a given parameter value and with given source term), but with multiple basis functions (corresponding to the PDE solution at a given parameter value and with multiple source terms). We show on an academic Laplace problem that this strategy can bring down the number of operator factorizations when building a reduced basis.

Chapters 5 and 6

Chapters 5 and 6 are concerned with applications of the reduced basis method to frequency sweep analysis in electromagnetism. In chapter 5, we consider radiating sources in a bounded domain (with absorbing boundary conditions to simulate an unbounded domain). We illustrate the reduced basis method on numerical examples of industrial interest in antenna applications using edge finite elements. We use the FETI-2LM domain decomposition in order to make the best use of parallel computing architectures.

In chapter 6, we consider electromagnetic scattering problems in unbounded domain brought to integral equations on the surface of the scattering object and discretized using the boundary element method. We introduce the concept of *non-intrusive local affine approximations* of the discretized integral operators and use this concept in an original

non-intrusive reduced basis method for frequency analysis with discretized surface integral equations. Our approach is illustrated on numerical examples of industrial interest in electromagnetic scattering applications. Our work uses the state-of-the-art Fast-Multipole Method (FMM) for accelerating the matrix-vector operations.

Chapter 7

The last chapter of this thesis is devoted to the application of the reduced basis method to accelerate simulation campaigns in acoustic liner optimization in a discontinuous-Galerkin framework with domain decomposition. Such simulation campaigns usually consist in successively solving the time-harmonic linearized Euler equations at different acoustic impedance values (typically, thousands of different acoustic impedance values). In this context, the reduced basis method builds fast and reliable approximations of the fluid pressure and velocity fields, which significantly accelerates the process of finding the optimal impedance value (*i.e.*, such that the noise attenuation is maximized). We provide numerical illustrations on an industrial aircraft engine nacelle configuration.

Chapter 1

Basic principles and properties of the reduced basis method

Summary. In this chapter, we recall the notion of parametrized linear equation, their high-fidelity discretization and their reduced basis approximation. We review the Galerkin and Least-Squares reduced basis approximations and detail the so-called *offline/online* computational strategy, for which the concept of *affine* operator and right-hand side is key. We present the Empirical Interpolation Method (EIM), which is an indispensable tool to recover approximate affine operators and right-hand sides when these are non-affine. Two academic model problems serve as illustrations: a wavenumber-parametrized 1-dimensional Helmholtz equation and a 2-dimensional conductivity-parametrized Laplace equation.

Contents

1.1 Introduction to parametrized PDEs 6
1.1.1 Two parametrized problems 6
1.1.2 Weak forms and well-posedness 7
1.1.3 High-fidelity approximation 8
1.1.4 Mathematical framework for parametrized PDEs 10
1.2 The reduced basis method 12
1.2.1 Model order reduction and Kolmogorov width 12
1.2.2 Choice of subspace 13
1.2.3 Choice of approximation 15
1.3 Efficiency of the reduced basis method 18
1.3.1 Affine operators and right-hand sides 18
1.3.2 Computational strategy in the affine case 19

1.3.3 The non-affine case 22
1.4 Numerical illustration 24
1.4.1 Model problem 1: Helmholtz 24
1.4.2 Model problem 2: Laplace 27
1.5 Conclusions 29

1.1 Introduction to parametrized PDEs

1.1.1 Two parametrized problems

Let us introduce two model parametrized PDEs: a wavenumber-parametrized Helmholtz equation and a conductivity-parametrized Laplace equation. We consider a bounded and regular domain $\Omega \subset \mathbb{R}^d$, with $d = 1, 2, 3$ the spatial dimension. Let $C^0(\overline{\Omega})$ denote the space of all continuous functions defined on the closure of Ω and let $L^2(\Omega)$ denote the space of all square-integrable functions defined on Ω. Finally, we introduce the usual Sobolev space

$$H^1(\Omega) = \{v \in L^2(\Omega),\ \boldsymbol{\nabla} v \in L^2(\Omega)^d\}. \tag{1.1.1}$$

Model problem 1: the parametrized Helmholtz equation

The first model problem is the Helmholtz problem with homogeneous Dirichlet boundary conditions on the segment $\Omega =]0, 1[$: *find* $p \in H^1(]0, 1[)$, *such that*

$$\begin{cases} -\frac{d^2}{dx^2} p - \mu^2 p = S & \text{in }]0, 1[, \\ p(0) = p(1) = 0, \end{cases} \tag{1.1.2}$$

with a given source term $S \in L^2(]0, 1[)$. The unknown p typically represents the amplitude of a pressure perturbation in a static fluid. In this context, μ represents the wavenumber of the acoustic pressure wave.

Here, we do not want to simply solve eq. (1.1.2) for one given value of the wavenumber μ, but rather we want to solve this problem for all possible values of μ in a given interval $[\mu_{\min}, \mu_{\max}] \subset \mathbb{R}$. Denoting $p(\mu)$ the solution of eq. (1.1.2) at the wavenumber μ, the problem that we are interested in is finding the manifold $\{p(\mu),\ \mu \in [\mu_{\min}, \mu_{\max}]\}$.

Model problem 2: the parametrized Laplace equation

The second model problem is the Laplace problem with homogeneous Dirichlet boundary conditions on the 2D square domain $\Omega =]0,1[\times]0,1[$: *find* $T \in H^1(\Omega)$*, such that*

$$\begin{cases} -\text{div}\,(\kappa \boldsymbol{\nabla} T) = S & \text{in } \Omega, \\ T = 0 & \text{on } \partial\Omega, \end{cases} \tag{1.1.3}$$

with a given source term $S \in L^2(\Omega)$ and a given conductivity $\kappa \in L^\infty(\Omega)$. The unknown field T typically represents a temperature field.

In the fashion of [66], we consider a Gaussian conductivity, whose peak is localized at some point $\mu = (\mu_1, \mu_2)$ in the domain. Thus, the conductivity not only depends on the spatial variable $x = (x_1, x_2) \in \Omega$ but also on the choice of $\mu = (\mu_1, \mu_2)$, following the expression,

$$\kappa(x;\mu) = \exp(\mu_1 + \mu_2)\left[1 + 2\exp\left(-\frac{(x_1-\mu_1)^2 + (x_2-\mu_2)^2}{0.02}\right)\right]. \tag{1.1.4}$$

Clearly, since the conductivity depends on μ, then so does the PDE that we want to solve. Our original PDE is expressed as the parametrized problem: *find* $T(\mu) \in H^1(\Omega)$*, such that*

$$\begin{cases} -\text{div}\,(\kappa(\mu) \boldsymbol{\nabla} T(\mu)) = S & \text{in } \Omega, \\ T(\mu) = 0 & \text{on } \partial\Omega. \end{cases} \tag{1.1.5}$$

We are not interested in solving eq. (1.1.5) for just one given value of μ, but rather we are interested in the solutions $T(\mu)$ for all possible locations $\mu = (\mu_1, \mu_2)$ of the conductivity peak, say for all $\mu \in \mathcal{D} = [0.4,\ 0.6]^2$. The parametrized problem amounts to finding the manifold $\{T(\mu),\ \mu \in \mathcal{D}\}$.

1.1.2 Weak forms and well-posedness

Model problem 1: the parametrized Helmholtz equation

We recall that the solution $p(\mu)$ exists and is unique provided that μ is not a so-called *resonant* wavenumber. A resonant wavenumber is a wavenumber μ_{res} such that the homogeneous Helmholtz equation (*i.e.*, with source term $S = 0$) admits non-zero solutions. For the 1-dimensional problem, these wavenumbers can be obtained analytically, namely they are $\mu_{res} = \pi, 2\pi, 3\pi, 4\pi, \ldots$ and the associated non-zero solutions are $\sin(\pi x)$, $\sin(2\pi x)$, $\sin(2\pi x)$, $\sin(4\pi x), \ldots$.

The well-posedness for a non-resonant wavenumber can be obtained by invoking the Fredholm alternative (see [1, Chapter 4, §4.5]). To this end, introduce the Sobolev space

$H_0^1(]0,1[) = \{v \in H^1(]0,1[), v(0) = v(1) = 0\}$. Our PDE eq. (1.1.2) is equivalent to the weak form: *for given value of μ, find $p(\mu) \in H_0^1(]0,1[)$ such that*

$$\forall p' \in H_0^1(]0,1[), \quad a(p(\mu), p'; \mu) = f(p'), \tag{1.1.6}$$

where $a(\cdot,\cdot;\mu) : H_0^1(]0,1[) \times H_0^1(]0,1[) \to \mathbb{R}$ is the continuous bilinear form and $f : H_0^1(]0,1[) \to \mathbb{R}$ is the continuous linear form given by

$$a(p,p';\mu) = \int_{]0,1[} \left(\frac{dp}{dx}\frac{dp'}{dx} - \mu^2 pp' \right) \mathrm{d}x, \quad f(p') = \int_{]0,1[} Sp' \mathrm{d}x. \tag{1.1.7}$$

Recalling that $\|\frac{d}{dx}\cdot\|_{L^2(]0,1[)}$ defines a norm on $H_0^1(]0,1[)$ (which is a direct consequence of the Poincaré inequality, see [38, Appendix B, §3.7]), the bilinear form $(p,p') \in H_0^1(]0,1[) \times H_0^1(]0,1[) \mapsto \int_\Omega \frac{dp}{dx}\frac{dp'}{dx}\mathrm{d}x$ is clearly coercive. Next, we use the fact that the canonical imbedding $H^1(]0,1[) \to L^2(]0,1[)$ is compact (this result is known as the Rellich–Kondrachov theorem, [38, Appendix B, §3.3]). This shows that $a(\cdot,\cdot;\mu)$ is weakly coercive (*i.e.*, it consists of a coercive part and a compact part) and so the Fredholm alternative applies, thus establishing the existence and uniqueness of $p(\mu)$ for all $\mu \notin \{(n+1)\pi,\, n \in \mathbb{N}\}$.

Model problem 2: the parametrized Laplace equation

We briefly recall that the solution $T(\mu)$ exists and is unique by the Lax-Milgram theorem [38, Chapter 2, §2.1]. Denoting $H_0^1(\Omega) = \{v \in H^1(\Omega),\, v|_{\partial\Omega} = 0\}$, our parametrized PDE eq. (1.1.5) is equivalent to the weak form: *find $T(\mu) \in H_0^1(\Omega)$, such that*

$$\forall T' \in H_0^1(\Omega), \quad a(T(\mu), T'; \mu) = f(T'), \tag{1.1.8}$$

where $a(\cdot,\cdot;\mu) : H_0^1(\Omega) \times H_0^1(\Omega) \to \mathbb{R}$ is the continuous bilinear form and $f : H_0^1(\Omega) \to \mathbb{R}$ is the continuous linear form given by

$$a(T,T';\mu) = \int_\Omega \kappa(\mu) \boldsymbol{\nabla} T \cdot \boldsymbol{\nabla} T' \mathrm{d}\Omega, \quad f(T') = \int_\Omega ST' \mathrm{d}\Omega. \tag{1.1.9}$$

As already mentioned, $\|\boldsymbol{\nabla}\cdot\|_{L^2(\Omega)}$ defines a norm on $H_0^1(\Omega)$. Thus, we can easily show the following coercivity property

$$\forall T \in H_0^1(\Omega), \quad a(T,T;\mu) \geqslant \inf_{x\in\Omega} \kappa(x;\mu)\, \|\boldsymbol{\nabla} T\|_{L^2(\Omega)}. \tag{1.1.10}$$

In this context, the Lax-Milgram theorem provides existence and uniqueness of the solution $T(\mu)$ to the weak form eq. (1.1.8) and therefore to the PDE eq. (1.1.5).

1.1.3 High-fidelity approximation

In order to numerically solve the two parametrized problems, we first need to be able to provide numerical approximations for the solutions. In this thesis, this will be done by discretizing the weak form of the PDE using the finite element (FE) method. In this section, we briefly show how this works.

Model problem 1: the parametrized Helmholtz equation

The weak form eq. (1.1.6) is now discretized using the standard FE method. For this purpose, let $\mathcal{N} \geqslant 1$ and introduce a set of $\mathcal{N}+2$ uniformly distributed points $\{x_i\}_{1\leqslant i\leqslant \mathcal{N}+2}$ in $[0,1]$, that is $x_i = (i-1)h$, where $h = 1/(\mathcal{N}+1)$. Note that $x_0 = 0$ and $x_{\mathcal{N}+2} = 1$. The first order Lagrange finite element approximation space incorporating Dirichlet boundary conditions is defined as

$$X_h^0(]0,1[) = \{v \in C^0([0,1]),\ v|_{]x_i,x_{i+1}[} \in \mathbb{P}^1(]x_i,x_{i+1}[), 1 \leqslant i \leqslant \mathcal{N}+1, \\ v(0) = v(1) = 0\}. \tag{1.1.11}$$

Note that this FE approximation space is of dimension $\mathcal{N}$. For given value of μ, our FE approximation for $p(\mu) \in H_0^1(]0,1[)$ is the Galerkin approximation $p_h(\mu) \in X_h^0(]0,1[)$ satisfying the discrete weak form

$$\forall p_h' \in X_h^0(]0,1[), \quad a(p_h(\mu), p_h'; \mu) = f(p_h'). \tag{1.1.12}$$

We say that $p_h(\mu)$ is a *high-fidelity* approximation, because the discretization error $\|p_h(\mu) - p(\mu)\|_{H^1(]0,1[)}$ can be made arbitrarily small by adequately refining the partition $\mathcal{T}_h$. Since the exact PDE solution $p(\mu)$ is numerically unobtainable, we can forget about the exact PDE and focus on the high-fidelity approximation $p_h(\mu)$. For this reason, the high-fidelity approximation is also called the *truth* approximation. Unfortunately, the existence and uniqueness of the truth approximation cannot be straightforwardly established (noticing that the fact that $p(\mu)$ exists and is unique by the Fredholm alternative is not a sufficient condition). By the Banach-Nečas-Babuška theorem [38, Chapter 2, §2.1], the existence and uniqueness of $p_h(\mu)$ is equivalent to the existence of $\alpha(\mu) > 0$ such that the following (discrete) inf-sup condition is satisfied

$$\forall p_h \in X_h^0(]0,1[), \qquad \sup_{p_h' \in X_h^0(]0,1[)} \frac{a(p_h, p_h'; \mu)}{\|p_h'\|_{H_0^1(]0,1[)}} \geqslant \alpha(\mu)\|p_h\|_{H_0^1(]0,1[)}, \tag{1.1.13}$$

where $\|\cdot\|_{H_0^1(]0,1[)}$ denotes a norm on $H_0^1(]0,1[)$ (for instance the $\|\frac{d}{dx}\cdot\|_{L^2(]0,1[)}$ norm).

Model problem 2: the parametrized Laplace equation

As in the Helmholtz case, the weak form eq. (1.1.8) is discretized using the FE method using a triangulation $\mathcal{T}_h = \{T_e\}_{1\leqslant e\leqslant \mathcal{N}_{\text{elt}}}$ of Ω into $\mathcal{N}_{\text{elt}}$ triangle elements. We again use the first order Lagrange finite element approximation space incorporating Dirichlet boundary conditions, which is defined in this case as

$$X_h^0(\Omega) = \{v \in C^0(\overline{\Omega}),\ \forall T \in \mathcal{T}_h,\ v|_T \in \mathbb{P}^1(T),\ \gamma_0 v = 0\}. \tag{1.1.14}$$

Finally, for given value of μ, our FE approximation is the Galerkin approximation $T_h(\mu) \in X_h^0((\Omega))$ satisfying the discrete weak form

$$\forall T_h' \in X_h^0(\Omega), \quad a(T_h(\mu), T_h'; \mu) = f(T_h'). \tag{1.1.15}$$

Again, we can forget about the exact PDE solution $T(\mu)$ and focus on the high-fidelity approximation $T_h(\mu)$ also called the truth solution, since the two can be made indistinguishable by refining the triangulation. The truth solution exists an is unique thanks to the conforming property $X_h^0(\Omega) \subset H_0^1(\Omega)$. Indeed, the coercivity property on $H_0^1(\Omega) \times H_0^1(\Omega)$ directly implies a (discrete) coercivity property on $X_h^0(\Omega) \times X_h^0(\Omega)$.

1.1.4 Mathematical framework for parametrized PDEs

So far, we have reviewed the Helmholtz and Laplace equations in parametrized settings and have shown their high-fidelity discretization using finite elements. We now present the general mathematical framework that will be used throughout this thesis. Our Helmholtz and Laplace problems must be thought of as specific instances in this general mathematical framework.

Variational setting

Let $\mathcal{D} \subset \mathbb{R}^p$ be a compact set of parameters, where p is the number of independent parameters. Let $a^{\text{ex}}(\cdot,\cdot;\mu) : V^{\text{ex}} \times W^{\text{ex}} \to \mathbb{C}$ and $f^{\text{ex}}(\cdot;\mu) : W^{\text{ex}} \to \mathbb{C}$ be given μ-dependent sesquilinear and linear forms and consider the parametrized PDE: *find* $u^{\text{ex}}(\mu) \in V^{\text{ex}}$ *such that*

$$\forall w^{\text{ex}} \in W^{\text{ex}}, \quad a^{\text{ex}}(u^{\text{ex}}(\mu), w^{\text{ex}}; \mu) = f^{\text{ex}}(w^{\text{ex}}; \mu). \tag{1.1.16}$$

The spaces V^{ex} and W^{ex} are two infinite-dimensional Sobolev spaces such that the parametrized PDE is well-posed. For instance, V^{ex} could verify $H_0^1(\Omega) \subset V^{\text{ex}} \subset H^1(\Omega)$, where $\Omega \subset \mathbb{R}^d$ is the domain in which the PDE is solved and W^{ex} could be the same space as V^{ex} (c.f. our two model problems).

Since the exact PDE solution $u^{\text{ex}}(\mu)$ is numerically unobtainable, our focus is on a given high-fidelity approximation: *find* $u(\mu) \in V$ *such that*

$$\forall w \in W, \quad a(u(\mu), w; \mu) = f(w; \mu), \tag{1.1.17}$$

where $a(\cdot,\cdot;\mu) : V \times W \to \mathbb{C}$ and $f(\cdot;\mu) : W \to \mathbb{C}$ are given μ-dependent sesquilinear and linear forms and V, W denote two Hilbert spaces, with finite dimension $\mathcal{N}$.

The Hilbert space V is typically the FE trial space (where the FE approximation is sought), while W is the FE test space. The spaces V, W may be thought of as conforming approximation spaces verifying $V \subset V^{\text{ex}}$, $W \subset W^{\text{ex}}$, in which case $a(\cdot,\cdot;\mu)$ and $f(\cdot;\mu)$ coincide with $a^{\text{ex}}(\cdot,\cdot;\mu)$ and $f^{\text{ex}}(\cdot;\mu)$ and the norm on V (resp. W) is inherited from V^{ex} (resp. W^{ex}). Yet we emphasize that the present is also fit to non-conforming approximation spaces, in which case V, W must be equipped with adequate, usually mesh-dependent norms. For instance, this situation occurs when using Discontinuous-Galerkin methods [101].

The dimension $\mathcal{N}$ can be chosen large enough so that the difference between the exact PDE solution $u^{\text{ex}}(\mu)$ and the numerical approximation $u(\mu)$ is adequately small. Thus our focus is on the high-fidelity solution $u(\mu)$ also called the truth solution and we forget about the exact PDE solution.

Operator setting

We can re-write eq. (1.1.17) more simply in terms of operator and right-hand side. To this end, let us introduce V', the topological dual of V which is the Hilbert space comprising all linear forms $\ell : V \to \mathbb{C}$. We denote ${}_{V'}\langle \ell, v \rangle_V$ (or simply $\langle \ell, v \rangle$, when there is no ambiguity) the duality bracket between $v \in V$ and a member $\ell \in V'$ from the topological dual. We adopt the convention that the duality bracket is linear with respect to the first and anti-linear with respect to the second variable, that is $\langle \lambda\ell, \eta v \rangle = \lambda\overline{\eta}\langle \ell, v \rangle$ for all $\lambda, \eta \in \mathbb{C}$.

Let $R_V \in \mathcal{L}(V, V')$ denote the inverse Riesz operator such that the our inner product on V is

$$\forall v_1, v_2 \in V, \quad (v_1, v_2)_V = \langle R_V v_1, v_2 \rangle. \tag{1.1.18}$$

and the norm V denoted $\|\cdot\|_V$ verifies by $\|v\|_V^2 = (v, v)_V = \langle R_V v, v \rangle$. The norm $\|\cdot\|_{V'}$ on the topological dual V' is given by

$$\|\ell\|_{V'} = \sup_{v \in V} \frac{|\langle \ell, v \rangle|}{\|v\|_V} = \left(\langle \ell, R_V^{-1} \ell \rangle \right)^{1/2}. \tag{1.1.19}$$

Of course, we can introduce similar objects in the case of the Hilbert space W. This being set, it is clear from the Riesz representation theorem that there exists a unique linear operator $A(\mu) \in \mathcal{L}(V, W')$ such that

$$\forall (u, w) \in V \times W, \quad \langle A(\mu)u, w \rangle = a(u, w; \mu). \tag{1.1.20}$$

The parametrized linear form $f(\cdot; \mu) : W \to \mathbb{C}$ on the right-hand side of eq. (1.1.17) can be seen a μ-dependent member $f(\mu)$ of W', as

$$\forall w \in W, \quad \langle f(\mu), w \rangle = f(w; \mu). \tag{1.1.21}$$

With these new notations, eq. (1.1.17) can be equivalently re-expressed in operator form as: *find* $u(\mu) \in V$ *such that*

$$A(\mu)u(\mu) = f(\mu) \quad \text{in } W'. \tag{1.1.22}$$

Well-posedness

The parametrized problem eq. (1.1.22) is well-posed if and only if $A(\mu)$ is a weakly coercive operator satisfying the assumptions of the Banach-Nečas-Babuška theorem, *i.e.*, for all $v \in V$

$$\alpha(\mu)\|v\|_V \leqslant \|A(\mu)v\|_{W'} \leqslant \gamma(\mu)\|v\|_V, \tag{1.1.23}$$

with the so-called (strictly positive) *inf-sup constant*,

$$\alpha(\mu) = \inf_{v\in V}\sup_{w\in W}\frac{|\langle A(\mu)v, w\rangle|}{\|v\|_V\|w\|_W} = \inf_{v\in V}\frac{\|A(\mu)v\|_{W'}}{\|v\|_V} > 0, \tag{1.1.24a}$$

and with the (bounded) *continuity constant*,

$$\gamma(\mu) = \sup_{v\in V}\sup_{w\in W}\frac{|\langle A(\mu)v, w\rangle|}{\|v\|_V\|w\|_W} = \sup_{v\in V}\frac{\|A(\mu)v\|_{W'}}{\|v\|_V} < \infty. \tag{1.1.24b}$$

Remark (Coercive case)**.** *We can easily show that coercivity implies inf-sup stability. For this purpose, assume $V = W$ and that $A(\mu) \in \mathcal{L}(V, V')$ is coercive. Thus there exists $c(\mu) > 0$ such that the following coercivity property is satisfied*

$$\forall v \in V, \quad |\langle A(\mu)v, v\rangle| \geqslant c(\mu)\|v\|_V^2. \tag{1.1.25}$$

In this situation, we can bound the inf-sup constant from below using the coercivity constant as follows

$$\alpha(\mu) = \inf_{v\in V}\sup_{\hat{v}\in V}\frac{|\langle A(\mu)v, \hat{v}\rangle|}{\|v\|_V\|\hat{v}\|_V} \geqslant \inf_{v\in V}\frac{|\langle A(\mu)v, v\rangle|}{\|v\|_V^2} = c(\mu) > 0. \tag{1.1.26}$$

Under the assumptions of the Banach-Nečas-Babuška theorem, for any $\mu \in \mathcal{D}$ the solution $u(\mu)$ to eq. (1.1.22) exists and is unique. In this work, we are interested in the manifold $\mathscr{M} = \{u(\mu),\ \mu \in \mathcal{D}\}$, comprised of all truth solutions under variation of the parameter.

1.2 The reduced basis method

1.2.1 Model order reduction and Kolmogorov width

Recalling that the truth solution $u(\mu) \in V$ for given value of the parameter $\mu \in \mathcal{D}$ is actually a function of the spatial variable $x \in \Omega$, we can view u as the function $(x, \mu) \mapsto u(x, \mu)$ defined on $\Omega \times \mathcal{D}$. Model order reduction techniques aim at constructing an efficient approximation u_N of u under the following separated form

$$u_N(x, \mu) = \sum_{n=1}^{N} \beta_n(\mu)\xi_n(x), \tag{1.2.1}$$

for some functions $\beta_1, \ldots, \beta_N$ of the μ variable and some functions $\xi_1, \ldots, \xi_N$ of the x variable. Assuming that the N functions $\xi_1, \ldots, \xi_N$ are linearly independent functions of V; which is a reasonable assumption; then $V_N = \mathrm{Span}\{\xi_1, \ldots, \xi_N\}$ is a N-dimensional subspace of V. In this context, u_N given by eq. (1.2.1) is an approximation of u in the subspace V_N.

Intuitively, the approximation error decreases as the subspace dimension N increases. Model order reduction techniques are successful if the subspace dimension N does not increase too rapidly with decreasing approximation error [3]. The Kolmogorov N-width is the key notion in order to obtain *a priori* knowledge of the success of reduced order models [97, Chapter 5].

Let $V_N \subset V$ be a N-dimensional subspace of V. We define the angle between the solution manifold $\mathscr{M} = \{u(\mu),\ \mu \in \mathcal{D}\}$ and the subspace V_N as

$$d(\mathscr{M}, V_N) = \sup_{u \in \mathscr{M}} \inf_{v_N \in V_N} \|u - v_N\|_V. \tag{1.2.2}$$

In other words, when the angle between $\mathscr{M}$ and V_N is ϵ, this means that any truth solution $u(\mu) \in \mathscr{M}$ can be approximated by a member of V_N with an error smaller than ϵ. Thus, the smaller the angle between $\mathscr{M}$ and V_N, the better truth solutions can be approximated using members of V_N. The Kolmogorov N-width corresponds to the smallest angle between $\mathscr{M}$ and any N-dimensional subspace, that is

$$d_N(\mathscr{M}, V) = \inf_{V_N \subset V,\ \dim(V_N)=N} d(\mathscr{M}, V_N). \tag{1.2.3}$$

Thus, the Kolmogorov N-width is a theoretical lower bound for the approximation error when approximating the truth solution $u(\mu)$ for any $\mu \in \mathcal{D}$ using the subspace approximation eq. (1.2.1) [31].

Model order reduction techniques are particularly well suited for parametrized problems with fast-decaying Kolmogorov N-width. Indeed, assuming a fast-decaying Kolmogorov N-width, we can build a subspace with small dimension N in which the truth solutions can be approximated with small approximation errors. It has been proved that the Kolmogorov N-width decays exponentially fast for certain linear, coercive parametrized problems [120, 18, 31], therefore model order reduction techniques are expected to perform very well on this type of problems. On the other hand, model order reduction techniques are expected to be challenged on linear transport problems, for which the decay of the Kolmogorov N-width has been proven to be rather slow [88, 48, 86].

1.2.2 Choice of subspace

We briefly review two popular model order reduction techniques for constructing the approximation subspace $V_N \subset V$: the proper orthogonal decomposition (POD) method and the Reduced basis (RB) method.

Proper orthogonal decomposition (POD)

Given a discrete set $\mathcal{S}_M = \{\mu_1, \ldots, \mu_M\} \subset \mathcal{D}$ of M distinct parameter points (with $M \geqslant N$), the N-dimensional POD approximation subspace is the solution to the following

optimization problem

$$\inf_{\widetilde{V}_N \subset \mathrm{Span}\{u(\mu),\, \mu \in \mathcal{S}_M\},\, \dim(\widetilde{V}_N)=N} \sqrt{\sum_{m=1}^{M} \|u(\mu_m) - \Pi_{\widetilde{V}_N} u(\mu_m)\|_V^2}, \tag{1.2.4}$$

where $\Pi_{\widetilde{V}_N} : V \to \widetilde{V}_N$ denotes the orthogonal projection onto the subspace $\widetilde{V}_N$. We recall that given $u \in V$ the orthogonal projection $\Pi_{\widetilde{V}_N} u$ of u onto $\widetilde{V}_N$ is characterized by

$$\forall \widetilde{v}_N \in \widetilde{V}_N, \quad (\widetilde{v}_N, \Pi_{\widetilde{V}_N} u - u)_V = 0. \tag{1.2.5}$$

In order to build a POD approximation subspace one must proceed in two steps:

1. compute the M truth solutions $u(\mu_1), \ldots, u(\mu_M)$;
2. solve the optimization problem eq. (1.2.4).

The first step is known as the *exploration phase*, while the second consists in a *compression phase*, which retains only $N \leqslant M$ basis functions from the subspace spanned by the M truth solutions. For a more comprehensive review of the POD, we refer to the dedicated book [69].

Reduced Basis methods

The Lagrange reduced basis approximation subspace [94, 71, 105] is simply the span of some N truth solutions $u(\mu^1), \ldots, u(\mu^N)$ at some distinct parameters $\mu^{(1)}, \ldots, \mu^N \in \mathcal{D}$; that is,

$$V_N = \mathrm{Span}\{u(\mu^1), \ldots, u(\mu^N)\}. \tag{1.2.6}$$

A variant is the Taylor reduced basis approximation subspace [94], which is the span of the first derivatives of the truth solution $u(\mu)$ with respect to the parameter $\mu = (\mu_1, \ldots, \mu_p)^T$ evaluated at some parameter value $\mu^\star \in \mathcal{D}$; that is the span of $\frac{\partial u}{\partial \mu_\ell}(\mu^\star), \ell = 1, \ldots, p$. Note that this subspace is at most p-dimensional, with p the number of independent parameters. Another variant consists in combining the span of some N truth solutions $u(\mu^1), \ldots, u(\mu^N)$ at some distinct parameter points $\mu^1, \ldots, \mu^N \in \mathcal{D}$, with their first derivatives. One obtains the Hermite RB approximation subspace [62].

Discussion

Our goal in this thesis is to compute as few high-fidelity solutions as possible in order to obtain the best computational savings. In this context, the POD is not well suited, because the exploration phase requires computing a large number $M \gg N$ of high-fidelity solutions, the cost of which can be computationally significant.

Furthermore, we note that the use of Taylor or Hermite approximation subspaces incorporate information from the derivatives of $u(\mu)$ with respect to μ. In theory, this raises of course the question of the existence of these derivatives. In practice, when these derivatives exist, computing them requires solving for all $\ell = 1, \dots, p$ the following problem: *find* $v_\ell(\mu) = \frac{\partial u}{\partial \mu_\ell}(\mu) \in V$ *such that*

$$A(\mu)v_\ell(\mu) = \frac{\partial f}{\partial \mu_\ell}(\mu) - \frac{\partial A}{\partial \mu_\ell}(\mu)u(\mu) \quad \text{in } W'. \tag{1.2.7}$$

In terms of costs, solving eq. (1.2.7) is about as expensive as a high-fidelity solve. In terms of implementation, this is not an easy task: new assembly routines must be coded in order to assemble the derivatives of the operator and right-hand side.

For ease of implementation and thanks to its relative simplicity, we have chosen to work exclusively with Lagrange reduced basis approximation subspaces of the form of eq. (1.2.6).

1.2.3 Choice of approximation

Now that a N-dimensional approximation subspace $V_N \subset V$ is at hand, we have to define an efficient approximation $u_N(\mu)$ of $u(\mu)$ under the separated form eq. (1.2.1). Indeed, the *optimal* subspace approximation is given

$$u_N^\star(\mu) = \underset{v_N \in V_N}{\operatorname{argmin}} \, \|v_N - u(\mu)\|_V \,. \tag{1.2.8}$$

In general, the optimal reduced basis approximation $u_N^\star(\mu)$ cannot be computed without the knowledge of the truth solution $u(\mu)$. Indeed, $u_N^\star(\mu)$ is given by the orthogonal projection of $u(\mu)$ onto the subspace V_N, *i.e.*,

$$\forall v_N \in V_N, \quad (v_N, u_N^\star(\mu) - u(\mu))_V = 0. \tag{1.2.9}$$

Thus computing $u_N^\star(\mu)$ requires first computing $u(\mu)$ and then projecting it onto the subspace V_N. This strategy is of course inefficient! In order to obtain efficient approximations, we must turn to sub-optimal approximations. We now review the Galerkin and Least-squares approximations.

Remark. *There is one special case where the optimal subspace approximation $u_N^\star(\mu)$ can actually be computed without the knowledge of $u(\mu)$: we shall see this in the upcoming proposition 1.2.2.*

Galerkin approximation

Let $V_N \subset V$ be a N-dimensional subspace. The so-called Galerkin RB approximation is defined as follows [97].

Definition 1 (Galerkin RB approximation)**.** *For all $\mu \in \mathcal{D}$ the Galerkin RB approximation is given by $u_N(\mu) \in V_N$ satisfying the weak form*

$$\forall v_N \in V_N, \quad \langle A(\mu)u_N(\mu), v_N\rangle = \langle f(\mu), v_N\rangle.$$

The existence and uniqueness of the Galerkin RB approximation cannot be proven in the general Hilbert setting. However, it can be proven in the case $V = W$ (which corresponds to the situation where the high-fidelity approximation stems from a Galerkin projection) and under the hypothesis that $A(\mu) \in \mathcal{L}(V, V')$ is coercive.

Proposition 1.2.1 (Well-posedness of the Galerkin RB approximation)**.** *Assume that $W = V$ and that $A(\mu) \in \mathcal{L}(V, V')$ is coercive. Then the Galerkin RB approximation exists and is unique.*

Proof. By the Banach-Nečas-Babuška theorem, the Galerkin RB problem is well-posed if and only if the stability constant defined by

$$\alpha_N(\mu) = \inf_{v_N \in V_N} \sup_{\hat{v}_N \in V_N} \frac{|\langle A(\mu)v_N, \hat{v}_N\rangle|}{\|v_N\|_V \|\hat{v}_N\|_V} \tag{1.2.10}$$

is strictly positive. Choosing $\hat{v}_N = v_N$ as candidate supremizer in eq. (1.2.10) we get

$$\alpha_N(\mu) \geqslant \inf_{v_N \in V_N} \frac{|\langle A(\mu)v_N, v_N\rangle|}{\|v_N\|_V^2}$$

Since $A(\mu) \in \mathcal{L}(V, V')$ is coercive, there exists a coercivity constant $c(\mu) > 0$ such that the following coercivity property hold

$$\forall v \in V, \quad |\langle A(\mu)v, v\rangle| \geqslant c(\mu)\|v\|_V^2.$$

Using the subspace property $V_N \subset V$, we conclude that $\alpha_N(\mu) \geqslant c(\mu)$ and so $\alpha_N(\mu) > 0$; which concludes the proof. □

We now consider the adjoint operator $A(\mu)^* \in \mathcal{L}(W, V')$, defined by ${}_{W'}\langle A(\mu)v, w\rangle_W = {}_{V'}\langle A(\mu)^* w, v\rangle_V$ for all $v \in V$ and for all $w \in W$. The following result is classical, showing optimality of the Galerkin projection for self-adjoint operator and a specific choice for the norm on V.

Proposition 1.2.2 (Optimality of the Galerkin projection in self-adjoint case)**.** *Assume that $W = V$ and that $A(\mu) \in \mathcal{L}(V, V')$ is coercive and self-adjoint (i.e., $A(\mu) = A(\mu)^*$) and consider the norm on V such that the inverse Riesz operator is given by*

$$R_V = A(\mu).$$

Let $V_N \subset V$ be a N-dimensional RB approximation space. Then, the optimal RB approximation $u_N^\star(\mu) \in V_N$ defined by (1.2.8) *exists, is unique and is the given by the Galerkin RB approximation.*

Proof. Recall the characterization (1.2.9) for the optimal RB approximation $u_N^\star(\mu) \in V_N$,

$$\forall v_N \in V_N, \quad \langle R_V v_N, u_N^\star(\mu) - u(\mu)\rangle = 0. \tag{1.2.11}$$

Using $R_V = A(\mu) = A(\mu)^*$ in yields

$$\forall v_N \in V_N, \quad \langle A(\mu)u_N^\star(\mu), v_N\rangle = \langle A(\mu)u(\mu), v_N\rangle. \tag{1.2.12}$$

Recalling that $\langle A(\mu)u(\mu), v_N\rangle = \langle f(\mu), v_N\rangle$ we obtain that $u_N^\star(\mu) \in V_N$ is the Galerkin RB approximation defined in definition 1. Existence and uniqueness is provided by proposition 1.2.1. □

Remark. *Notice that $R_V = A(\mu)$ means that the norm on V is in fact a μ-dependent norm.*

The Least-Squares approximation

As we have seen, the well-posedness of the Galerkin RB approximation can only be shown in the situation $V = W$ and $A(\mu) \in \mathcal{L}(V, V')$ is coercive. The Least-squares RB approximation is an alternative RB approximation for which well-posed can be proven in the general case [97].

Definition 2 (Least-squares RB approximation)**.** *For all $\mu \in \mathcal{D}$ the least-squares RB approximation is given by $u_N(\mu) \in V_N$ solution to*

$$u_N(\mu) = \underset{v_N \in V_N}{\operatorname{argmin}} \|A(\mu)v_N - f(\mu)\|_{W'}^2 .$$

As its name suggests, the least-squares RB approximation minimizes the reduced basis residual norm. Contrary to the optimal reduced basis approximation, the least-squares reduced basis approximation is always computable without knowledge of the truth solution $u(\mu)$. It is further unconditionally well-posed, as shown by the following proposition.

Proposition 1.2.3 (Least-Squares Reduced Basis)**.** *The least-squares RB approximation $u_N(\mu) \in V_N$ defined in definition 2 is the unique solution to the Petrov-Galerkin weak formulation*

$$\forall w_N \in R_W^{-1}A(\mu)V_N, \quad \langle A(\mu)u_N(\mu), w_N\rangle = \langle f(\mu), w_N\rangle.$$

Or, equivalently, the unique solution to the Galerkin weak formulation

$$\forall v_N \in V_N, \quad \langle A(\mu)^* R_W^{-1} A(\mu)u_N(\mu), v_N\rangle = \langle A(\mu)^* R_W^{-1} f(\mu), v_N\rangle.$$

Proof. Define $\mathcal{J} : V_N \to \mathbb{R}$, $v_N \mapsto \|A(\mu)v_N - f(\mu)\|_V^2$. By direct differentiation, for all $h_N \in V_N$ there holds,

$$\langle \nabla \mathcal{J}(v_N), h_N\rangle = 2\Re\left\{\langle R_W^{-1}A(\mu)v_N, A(\mu)h_N\rangle - \langle R_W^{-1}A(\mu)h_N, f(\mu)\rangle\right\}.$$

Thus, the optimality condition

$$\forall h_N \in V_N, \quad \langle \nabla \mathcal{J}(u_N(\mu)), h_N \rangle = 0,$$

is equivalent to the Galerkin weak formulation

$$\langle A(\mu)^* R_W^{-1} A(\mu) u_N(\mu), v_N \rangle = \langle A(\mu)^* R_W^{-1} f(\mu), v_N \rangle \quad \forall v_N \in V_N.$$

Furthermore, the inf-sup constant of this Galerking weak formulation writes

$$\begin{aligned} \alpha_N(\mu) &= \inf_{v_N \in V_N} \sup_{w_N \in R_W^{-1} A(\mu) V_N} \frac{|\langle A(\mu) v_N, w_N \rangle|}{\|v_N\|_V \|w_N\|_W} \\ &= \inf_{v_N \in V_N} \sup_{\widehat{v}_N \in V_N} \frac{|\langle A(\mu)^* R_W^{-1} A(\mu) v_N, \widehat{v}_N \rangle|}{\|v_N\|_V \|\widehat{v}_N\|_V}. \end{aligned} \tag{1.2.13}$$

Taking the candidate supremizer $\widehat{v}_N = v_N$, there holds

$$\sup_{\widehat{v}_N \in V_N} \frac{|\langle A(\mu)^* R_W^{-1} A(\mu) v_N, \widehat{v}_N \rangle|}{\|v_N\|_V \|\widehat{v}_N\|_V} \geqslant \frac{|\langle A(\mu)^* R_W^{-1} A(\mu) v_N, v_N \rangle|}{\|v_N\|_V^2} = \frac{\|A(\mu) v_N\|_{W'}^2}{\|v_N\|_V^2}.$$

Thus, exploiting $V_N \subset V$, we get the following lower bound

$$\alpha_N(\mu) \geqslant \inf_{v_N \in V_N} \frac{\|A(\mu) v_N\|_{W'}^2}{\|v_N\|_V^2} \geqslant \inf_{v \in V} \frac{\|A(\mu) v\|_{W'}^2}{\|v\|_V^2} = \alpha(\mu)^2 > 0.$$

Thus the weak form is well-posed. □

1.3 Efficiency of the reduced basis method

We now explain how the RB approximations can be efficiently computed. In this context, the notion of *affinely parametrized operator* (or simply *affine operator* for brevity) is key [123, 105].

1.3.1 Affine operators and right-hand sides

Definition 3 (Affine operator)**.** *The μ-parametrized operator $A(\mu) \in \mathcal{L}(V, W')$ is affine if there exists*

- *Q^a functions $\theta_q^a : \mathcal{D} \to \mathbb{C}$, $1 \leqslant q \leqslant Q^a$ at least bounded (i.e., $\theta_q^a \in L^\infty(\mathcal{D})$, $1 \leqslant q \leqslant Q^a$);*
- *Q^a linear operators $A_q \in \mathcal{L}(V, W')$, $1 \leqslant q \leqslant Q^a$;*

such that

$$\forall \mu \in \mathcal{D}, \quad A(\mu) = \sum_{q=1}^{Q^a} \theta_q^a(\mu) A_q.$$

Let us illustrate the notion of affine operator with the our first model parametrized linear equation introduced in section 1.1. Namely, the operator stemming from the 1D Helmholtz equation with Dirichlet conditions writes, with $V = W = X_h^0(]0,1[)$,

$$\forall (v,w) \in V \times W, \quad \langle A(\mu)v, w\rangle = \int_{]0,1[} \left(\frac{dw}{dx}\frac{dw}{dx} - \mu^2 vw \right) \mathrm{d}x. \tag{1.3.1}$$

It is clearly affine with $Q^a = 2$ terms. Indeed, a possible affine parametrization (not unique) is $\theta_1^a(\mu) = 1$, $\theta_2^a(\mu) = -\mu^2$ and

$$\forall (v,w) \in X \times V, \quad \begin{cases} \langle A_1 v, w\rangle = \int_{]0,1[} \frac{dv}{dx}\frac{dw}{dx}\mathrm{d}x, \\ \langle A_2 v, w\rangle = \int_{]0,1[} vw\mathrm{d}x. \end{cases} \tag{1.3.2}$$

The concept of affine parametrization [123] can also be defined for the right-hand-side.

Definition 4 (Affine right-hand-side)**.** *The μ-parametrized linear form $f(\mu) \in W'$ is affine if there exists*

- *Q^f functions $\theta_q^f : \mathcal{D} \to \mathbb{C}$, $1 \leqslant q \leqslant Q^f$ at least bounded (i.e., $\theta_q^f \in L^\infty(\mathcal{D})$, $1 \leqslant q \leqslant Q^f$);*
- *Q^f linear forms $f_q \in W'$, $1 \leqslant q \leqslant Q^f$;*

such that

$$\forall \mu \in \mathcal{D}, \quad f(\mu) = \sum_{q=1}^{Q^a} \theta_q^f(\mu) f_q.$$

1.3.2 Computational strategy in the affine case

When the operator $A(\mu) \in \mathcal{L}(V, W')$ and right-hand-side $f(\mu) \in W'$ are affine, a very efficient computational strategy can be set up for the reduced basis method [84, 105, 54]. We detail this computational strategy in this section. For ease of implementation, we introduce an algebraic setting.

Algebraic setting

Since $\dim(V) = \dim(W) = \mathcal{N}$, we now introduce a basis $\{\phi_j^V\}_{1\leqslant j\leqslant \mathcal{N}}$ for V and $\{\phi_i^W\}_{1\leqslant i\leqslant \mathcal{N}}$ for W. For all $q = 1, \ldots, Q^a$ we define the matrix $\mathbf{A}_q \in \mathbb{C}^{\mathcal{N}\times\mathcal{N}}$, with entries

$$[\mathbf{A}_q]_{ij} = \langle A_q \phi_j^V, \phi_i^W\rangle, \quad 1 \leqslant i,j \leqslant \mathcal{N} \tag{1.3.3}$$

Using the affine decomposition, the operator $A(\mu) \in \mathcal{L}(V, W')$ is algebraically represented by the matrix $\mathbf{A}(\mu) \in \mathbb{C}^{\mathcal{N}\times\mathcal{N}}$ given by

$$\mathbf{A}(\mu) = \sum_{q=1}^{Q^a} \theta_q^a(\mu)\mathbf{A}_q. \tag{1.3.4}$$

Similarly, for all $q = 1, \ldots, Q^f$ we define the vector $\mathbf{f}_q \in \mathbb{C}^{\mathcal{N}}$ with entries

$$[\mathbf{f}_q]_i = \langle f_q, \phi_i^W \rangle, \quad 1 \leqslant i \leqslant \mathcal{N}. \tag{1.3.5}$$

Using the affine decomposition [123], the right-hand side $f(\mu) \in W'$ is algebraically represented by the vector $\mathbf{f}(\mu) \in \mathbb{C}^{\mathcal{N}}$ given by

$$\mathbf{f}(\mu) = \sum_{q=1}^{Q^f} \theta_q^f(\mu) \mathbf{f}_q. \tag{1.3.6}$$

This being set, it is clear that $\mathbf{u}(\mu) \in \mathbb{C}^{\mathcal{N}}$ solution to the linear system $\mathbf{A}(\mu)\mathbf{u}(\mu) = \mathbf{f}(\mu)$ holds the coordinates of the truth solution $u(\mu) \in V$ in the $\{\phi_j^V\}_{1 \leqslant j \leqslant \mathcal{N}}$ basis.

Let now $V_N \subset V$ denote a given reduced basis approximation space of dimension N and $\{\xi_1, \ldots, \xi_N\}$ denotes an orthonormal basis for V_N. Here, orthonormality is considered in the sense

$$\langle R_V \xi_i, \xi_j \rangle = \delta_{ij}, \quad 1 \leqslant i, j \leqslant N \tag{1.3.7}$$

with δ_{ij} the Kronecker symbol (1 if $i = j$, 0 otherwise). Each reduced basis function ξ_j ($1 \leqslant j \leqslant N$) can be decomposed in the basis $\{\phi_i^V\}_{1 \leqslant i \leqslant \mathcal{N}}$ as

$$\xi_j = \sum_{i=1}^{\mathcal{N}} \mathbf{P}_{ij} \phi_i^V. \tag{1.3.8}$$

The matrix $\mathbf{P} \in \mathbb{C}^{\mathcal{N} \times N}$ thus represents the reduced basis. This is a $\mathbf{B}_V$-orthogonal matrix (*i.e.*, $\mathbf{P}^* \mathbf{B}_V \mathbf{P} = \mathbf{I}$), where $\mathbf{B}_V \in \mathbb{C}^{\mathcal{N} \times \mathcal{N}}$ is the hermitian positive definite matrix with entries

$$[\mathbf{B}_V]_{ij} = \langle R_V \phi_j^V \phi_i^V \rangle, \quad 1 \leqslant i, j \leqslant \mathcal{N}. \tag{1.3.9}$$

We seek our reduced basis approximation $u_N(\mu) \in V_N$ as

$$u_N(\mu) = \sum_{j=1}^{N} \mathbf{x}_j(\mu) \xi_j, \tag{1.3.10}$$

where the coefficients $\mathbf{x}(\mu) = (\mathbf{x}_1(\mu), \ldots, \mathbf{x}_N(\mu)) \in \mathbb{C}^N$ are the coordinates of the reduced basis approximation $u_N(\mu) \in V_N$ expressed in the basis $\{\xi_1, \ldots, \xi_N\}$. The coordinates of the RB approximation $u_N(\mu) \in V_N$ in the $\{\phi_i^V\}_{1 \leqslant i \leqslant \mathcal{N}}$ basis are given by $\mathbf{u}_N(\mu) = \mathbf{P}\mathbf{x}(\mu) \in \mathbb{C}^{\mathcal{N}}$.

Galerkin reduced basis approximation

We begin with the case where $u_N(\mu) \in V_N$ is the Galerkin RB approximation. Recalling definition 1, that the Galerkin RB approximation $u_N(\mu) \in V_N$ satisfies the weak form:

$$\langle A(\mu) u_N(\mu), v_N \rangle = \langle f(\mu), v_N \rangle \quad \forall v_N \in V_N. \tag{1.3.11}$$

Equivalently, the coordinates $\mathbf{x}(\mu) \in \mathbb{C}^N$ of the Galerkin RB approximation satisfy the $N \times N$ linear system

$$(\mathbf{P}^* \mathbf{A}(\mu) \mathbf{P})\, \mathbf{x}(\mu) = \mathbf{P}^* \mathbf{f}(\mu) \tag{1.3.12}$$

Solving such a linear system can be achieved with $\mathcal{O}(N^3)$ complexity using a direct solver.

However, assembling this linear system for a given value of μ requires assembling the left-hand side matrix $\mathbf{P}^* \mathbf{A}(\mu) \mathbf{P}$ as well as the right-hand side vector $\mathbf{P}^* \mathbf{f}(\mu)$, which has a complexity dependent on $\mathcal{N}$. Fortunately, this can be circumvented using an efficient offline/online decoupling, as shown by the following proposition.

Proposition 1.3.1 (Galerkin efficient offline/online decoupling)**.** *If*

- *for all* $1 \leqslant q \leqslant Q^a$*, the* $N \times N$ *matrix* $\mathbf{P}^* \mathbf{A}_q \mathbf{P}$*; and*
- *for all* $1 \leqslant q \leqslant Q^f$*, the* N*-dimensional vector* $\mathbf{P}^* \mathbf{f}_q$

are pre-computed (during the so-called "offline phase"), then the linear system (1.3.12) *can be assembled for any value of* μ *with* $\mathcal{O}(Q^a N^2 + Q^f N)$ *complexity (during the so-called "online phase").*

Proof. Using the affine decompositions, the left-hand-side of (1.3.12) writes

$$\mathbf{P}^* \mathbf{A}(\mu) \mathbf{P} = \sum_{q=1}^{Q^a} \theta_q^a(\mu) \boxed{\mathbf{P}^* \mathbf{A}_q \mathbf{P}} \tag{1.3.13}$$

and the right-hand-side writes

$$\mathbf{P}^* \mathbf{f}(\mu) = \sum_{q=1}^{Q^f} \theta_q^f(\mu) \boxed{\mathbf{P}^* \mathbf{f}_q}. \tag{1.3.14}$$

If the boxed quantities are pre-computed (using operations with complexity dependent of $\mathcal{N}$ during the offline phase), the left-hand-side can be assembled in $\mathcal{O}(Q^a N^2)$ complexity and the right-hand-side in $\mathcal{O}(Q^f N)$ complexity. □

Least-squares reduced basis approximation

We now turn to the case where find $u_N(\mu) \in V_N$ is the least-squares RB approximation. Recalling proposition 1.2.3, $u_N(\mu) \in V_N$ satisfies the weak form:

$$\langle A(\mu)^* R_W^{-1} A(\mu) u_N(\mu), v_N \rangle = \langle A(\mu)^* R_W^{-1} f(\mu), v_N \rangle \quad \forall v_N \in V_N. \tag{1.3.15}$$

Equivalently, the coordinates $\mathbf{x}(\mu) \in \mathbb{C}^N$ of the least-squares RB approximation satisfy the $N \times N$ linear system

$$\left(\mathbf{P}^* \mathbf{A}(\mu)^* \mathbf{B}_W^{-1} \mathbf{A}(\mu) \mathbf{P}\right) \mathbf{x}(\mu) = \mathbf{P}^* \mathbf{A}(\mu)^* \mathbf{B}_W^{-1} \mathbf{f}(\mu) \tag{1.3.16}$$

As in the Galerkin case, (1.3.16) can be solved with $\mathcal{O}(N^3)$ complexity using a direct solver. The following proposition shows that there exists an efficient offline/online strategy for assembling this system.

Proposition 1.3.2 (Least-Squares efficient offline/online decoupling)**.** *If*
- *for all* $1 \leqslant p, q \leqslant Q^a$*, the* $N \times N$ *matrix* $\mathbf{P}^* \mathbf{A}_p^* \mathbf{B}_W^{-1} \mathbf{A}_q \mathbf{P}$ *; and*
- *for all* $1 \leqslant q \leqslant Q^f$ *and for all* $1 \leqslant p \leqslant Q^a$*, the* N*-dimensional vector* $\mathbf{P}^* \mathbf{A}_p^* \mathbf{B}_W^{-1} \mathbf{f}_q$

are pre-computed (during the so-called "offline phase"), then the linear system (1.3.16) *can be assembled for any value of* μ *with* $\mathcal{O}((Q^a)^2 N^2 + Q^a Q^f N)$ *complexity (during the so-called "online phase").*

Proof. Using the affine decompositions, the left-hand-side of (1.3.16) writes

$$\mathbf{P}^* \mathbf{A}(\mu)^* \mathbf{B}_W^{-1} \mathbf{A}(\mu) \mathbf{P} = \sum_{q=1}^{Q^a} \sum_{p=1}^{Q^a} \theta_q^a(\mu) \overline{\theta_p^a(\mu)} \boxed{\mathbf{P}^* \mathbf{A}_p^* \mathbf{B}_W^{-1} \mathbf{A}_q \mathbf{P}}. \tag{1.3.17}$$

the right-hand-side writes

$$\mathbf{P}^* \mathbf{A}(\mu)^* \mathbf{B}_W^{-1} \mathbf{f}(\mu) = \sum_{q=1}^{Q^f} \sum_{p=1}^{Q^a} \overline{\theta_p^a(\mu)} \theta_q^f(\mu) \boxed{\mathbf{P}^* \mathbf{A}_p^* \mathbf{B}_W^{-1} \mathbf{f}_q}. \tag{1.3.18}$$

if the boxed quantities are pre-computed (using operations with complexity dependent of $\mathcal{N}$ during the offline phase), the left-hand-side can be assembled in $\mathcal{O}((Q^a)^2 N^2)$ complexity and the right-hand-side in $\mathcal{O}(Q^a Q^f N)$ complexity. □

Remark (Redundant terms)**.** *Remark the hermitian property:*

$$\mathbf{P}^* \mathbf{A}_p^* \mathbf{B}_W^{-1} \mathbf{A}_q \mathbf{P} = \left(\mathbf{P}^* \mathbf{A}_q^* \mathbf{B}_W^{-1} \mathbf{A}_p \mathbf{P} \right)^*. \tag{1.3.19}$$

Using this property, the number of boxed quantities to be computed offline regarding the left-hand-side is not $(Q^a)^2$ *but is reduced to* $Q^a(Q^a + 1)/2$*.*

1.3.3 The non-affine case

When the operator $A(\mu)$ or right-hand side $f(\mu)$ is non-affine, the computationally efficient strategy described in section 1.3.2 cannot be used. For the RB method to be efficient, we need to replace the operator or the right-hand side by suitable affine approximation $\widetilde{A}(\mu)$ or $\widetilde{f}(\mu)$ with which the computationally efficient strategy described in section 1.3.2 can be used [84]. A useful tool in this context is the Empirical Interpolation Method (EIM), first introduced in Ref. [5].

The Empirical Interpolation Method

The EIM is an efficient method for approximating a given function $g : \Omega_{\text{space}} \times \mathcal{D}_{\text{param}} \to \mathbb{C}$ in separated form. The first variable $x \in \Omega_{\text{space}}$ must be thought of as a spatial variable, while $\mu \in \mathcal{D}_{\text{param}}$ must be thought of a parameter variable. The EIM constructs a set of M interpolation points $\{x_m\}_{1 \leqslant m \leqslant M}$ (also called the *magic points*), an interpolation matrix $\mathbf{B} \in \mathbb{C}^{M \times M}$ and EIM basis functions $h_m : \Omega_{\text{space}} \to \mathbb{C}$ for $1 \leqslant m \leqslant M$. A separated approximation can be defined as

$$\forall (x, \mu) \in \Omega_{\text{space}} \times \mathcal{D}_{\text{param}}, \quad \widetilde{g_M}(x; \mu) = \sum_{m=1}^{M} \varsigma_m(\mu) h_m(x). \tag{1.3.20}$$

with coefficients $\boldsymbol{\varsigma}(\mu) = (\varsigma_1(\mu), \ldots, \varsigma_M(\mu))^T$ solution to the $M \times M$ linear system

$$\mathbf{B}\boldsymbol{\varsigma}(\mu) = \boldsymbol{\phi}(\mu), \tag{1.3.21}$$

with right-hand side $\boldsymbol{\phi}(\mu) = (g(x^1; \mu), \ldots, g(x^M; \mu))^T$. The procedure is detailed in algorithm 1.1.

Algorithm 1.1: EIM algorithm

Input : $g : \Omega_{\text{space}} \times \mathcal{D}_{\text{param}} \to \mathbb{C}$; discrete surrogate sets $\Xi_{\text{space}} \subset \Omega_{\text{space}}$ and $\Xi_{\text{param}} \subset \mathcal{D}_{\text{param}}$, prescribed tolerance tol and number of iterations $M_{\max}$.

Output: Magic points $\{x_m\}_{1 \leqslant m \leqslant M}$, interpolation matrix $\mathbf{B} \in \mathbb{C}^{M \times M}$ and EIM basis functions $h_m : \Omega_{\text{space}} \to \mathbb{C}$ for $1 \leqslant m \leqslant M$

Compute $\mu^1 = \underset{\mu \in \Xi_{\text{param}}}{\operatorname{argmax}} \max_{x \in \Xi_{\text{space}}} |g(x; \mu)|$;

Compute $x^1 = \underset{x \in \Xi_{\text{space}}}{\operatorname{argmax}} |g(x; \mu^1)|$;

Set $h_1(\cdot) = \frac{g(\cdot; \mu^1)}{g(x^1; \mu^1)}$;

Set $\epsilon_1 = +\infty$, $\mathbf{B}_{11} = 1$ and $m = 1$, ;

while $m < M_{\max}$ *and* $\epsilon_m > tol$ **do**

- Compute $\mu^{m+1} = \underset{\mu \in \Xi_{\text{param}}}{\operatorname{argmax}} \max_{x \in \Xi_{\text{space}}} |g(x; \mu) - \widetilde{g_m}(x, \mu)|$;
- Compute $x^{m+1} = \underset{x \in \Xi_{\text{space}}}{\operatorname{argmax}} |g(x; \mu^{m+1}) - \widetilde{g_m}(x, \mu^{m+1})|$;
- Compute $\epsilon_{m+1} = |g(x^{m+1}; \mu^{m+1}) - \widetilde{g_m}(x^{m+1}, \mu^{m+1})|$;
- Set $h_{m+1}(\cdot) = \frac{g(\cdot; \mu^{m+1}) - \widetilde{g_m}(\cdot, \mu^{m+1})}{g(x^{m+1}; \mu^{m+1}) - \widetilde{g_m}(x^{m+1}, \mu^{m+1})}$;
- Set $\mathbf{B}_{i,m+1} = h_{m+1}(x^i)$ for all $1 \leqslant i \leqslant m+1$;
- $m \leftarrow m + 1$;

end

Let us recall the main properties of the EIM, which can be found in [5]:

- the interpolation matrix $\mathbf{B}$ is lower triangular with a unity diagonal, thus the linear system eq. (1.3.21) is always invertible;
- the EIM basis functions $\{h_m(\cdot)\}_{1\leqslant m\leqslant M}$ span the same subspace as the $\{g(\cdot;\mu^m)\}_{1\leqslant m\leqslant M}$, where the μ^m's are selected by algorithm 1.1;
- the following interpolation property holds: for all $1 \leqslant m \leqslant M$, $\widetilde{g_M}(\cdot;\mu^m) = g(\cdot;\mu^m)$.

Application to parametrized Laplace problem

Let us illustrate how the EIM can be used with our second model parametrized linear equation introduced in section 1.1. Namely, the operator stemming from the 2D Laplace equation with Dirichlet conditions writes, with $V = W = X_h^0(\Omega)$,

$$\forall (v,w) \in V \times W, \quad \langle A(\mu)v, w\rangle = \int_\Omega \kappa(x;\mu)\boldsymbol{\nabla} u(x)\cdot\boldsymbol{\nabla} v(x)\mathrm{d}x, \tag{1.3.22}$$

where we recall that $\kappa(\cdot;\mu) : \Omega \to \mathbb{R}$ defined by eq. (1.1.4) is a Gaussian conductivity centered on $\mu = (\mu_1,\mu_2)^T$. Applying the EIM to the conductivity yields a approximate separated form

$$\forall (x,\mu) \in \Omega\times\mathcal{D}, \quad \widetilde{\kappa_M}(x,\mu) = \sum_{m=1}^{M} \varsigma_m(\mu)h_m(x). \tag{1.3.23}$$

Having constructed an approximation for the conductivity, we may now introduce the approximate operator $\langle \widetilde{A}(\mu)v, w\rangle = \int_\Omega \widetilde{\kappa_M}(x;\mu)\boldsymbol{\nabla} v(x)\cdot\boldsymbol{\nabla} w(x)\mathrm{d}x$. The later is clearly affine with M terms, since

$$\widetilde{A}(\mu) = \sum_{m=1}^{M} \varsigma_m(\mu)A_m, \tag{1.3.24}$$

where for all $(v,w) \in V\times W$, $\langle A_m v, w\rangle = \int_\Omega h_m(x)\boldsymbol{\nabla} v(x)\cdot\boldsymbol{\nabla} w(x)\mathrm{d}x$. Intuitively, $\widetilde{A}(\mu)$ makes up a good approximation for $A(\mu)$ provided that $\widetilde{\kappa_M}(\cdot;\mu)$ makes up a good approximation for $\kappa(\cdot;\mu)$. More rigorous error estimates can be found in Refs. [37, 64]

1.4 Numerical illustration

1.4.1 Model problem 1: Helmholtz

Algebraic setting

For the parametrized Helmholtz problem, we recall that V $(= W)$ corresponds to the $\mathcal{N}$-dimensional FE approximation $X_h^0(]0,1[)$ defined by eq. (1.1.11). A basis for this FE

approximation space is $\{w_i\}_{1\leqslant i\leqslant \mathcal{N}}$, where the w_i's are the so-called Lagrange "nodal" (or "hat") basis functions, which are uniquely defined as the first order polynomials satisfying $w_i(x_j) = \delta_{ij}$, with compact support $[x_{i-1}, x_{i+1}]$.

The operator $A(\mu) \in \mathcal{L}(V, W')$ can be represented in the FE basis as $\mathbf{A}(\mu) = \mathbf{K} - \mu^2\mathbf{M}$, where $\mathbf{K}_{ij} = \int_{]0,1[} \frac{dw_j}{dx}\frac{dw_i}{dx}dx$ and $\mathbf{M}_{ij} = \int_{]0,1[} w_i w_j dxdx$, $1 \leqslant i, j \leqslant \mathcal{N}$. We can show by explicit calculations that $\mathbf{K}, \mathbf{M}$ are the $\mathcal{N} \times \mathcal{N}$ defined by

$$\mathbf{K} = \frac{1}{h}\begin{pmatrix} 2 & -1 & & \\ -1 & \ddots & \ddots & \\ & \ddots & \ddots & -1 \\ & & -1 & 2 \end{pmatrix}, \quad \mathbf{M} = h\begin{pmatrix} 2/3 & 1/6 & & \\ 1/6 & \ddots & \ddots & \\ & \ddots & \ddots & 1/6 \\ & & 1/6 & 2/3 \end{pmatrix}. \tag{1.4.1}$$

Similarly, the right-hand side can be represented in the FE basis by $\mathbf{f} \in \mathbb{C}^{\mathcal{N}}$ given by $\mathbf{f}_i = \int_{]0,1[} S w_i dx$ for all $1 \leqslant i \leqslant \mathcal{N}$.

This being set, we recall that the coordinates $\mathbf{u}(\mu) \in \mathbb{C}^{\mathcal{N}}$ of the FE solution $u(\mu) \in V$ in the FE basis can be obtained by solving the linear system $\mathbf{A}(\mu)\mathbf{u}(\mu) = \mathbf{f}$. In this example, we equip V and W with the same Riesz map (*i.e.*, $R_V = R_W$), represented in the FE basis by the matrix $\mathbf{B}_V = \mathbf{B}_W = \mathbf{K} + \mu^2_{\text{avg}}\mathbf{M}$, where $\mu_{\text{avg}} = 0.5(\mu_{\min} + \mu_{\max})$ the average wavenumber.

Reduced basis approximation in algebraic form

Given N points $\mu^{(1)}, \ldots, \mu^{(N)}$, we can build a Lagrange RB approximation subspace $V_N = \text{Span}\{u(\mu^{(1)}), \ldots, u(\mu^{(N)})\}$ as follows:

1. Compute the FE solutions $\mathbf{u}(\mu^{(1)}), \ldots, \mathbf{u}(\mu^{(N)})$;
2. Orthonormalize these N solutions, for instance using a Gram-Schmidt procedure [90, 13, 44] as

$$\begin{aligned} \mathbf{p}_1 &= \gamma_{11}\mathbf{u}(\mu^{(1)}), \\ \mathbf{p}_2 &= \gamma_{22}\mathbf{u}(\mu^{(2)}) + \gamma_{21}\mathbf{p}_1, \\ &\vdots \\ \mathbf{p}_N &= \gamma_{NN}\mathbf{u}(\mu^{(N)}) + \sum_{n=1}^{N-1}\gamma_{Ni}\mathbf{p}_i \end{aligned} \tag{1.4.2}$$

   where the coefficients γ_{ni} are chosen so that $\mathbf{P} = [\mathbf{p}_1|\cdots|\mathbf{p}_N] \in \mathbb{C}^{\mathcal{N}\times N}$ is $\mathbf{B}_V$-orthonormal (that is $\mathbf{P}^*\mathbf{B}_V\mathbf{P} = \mathbf{I}$) and $\text{Range}(\mathbf{P}) = \text{Span}\{\mathbf{u}(\mu^{(1)}), \ldots, \mathbf{u}(\mu^{(N)})\}$.

We recall that $\mathbf{P} \in \mathbb{C}^{\mathcal{N}\times N}$ represents the reduced basis in the FE basis and that the RB approximation $u_N(\mu) \in V_N$ is represented in the FE basis by $\mathbf{u}_N(\mu) = \mathbf{P}\mathbf{x}(\mu) \in \mathbb{C}^{\mathcal{N}}$. The coordinates $\mathbf{x}(\mu) \in \mathbb{C}^N$ of the RB approximation in the reduced basis, solves either the Galerkin linear system eq. (1.3.12) or the least-squares linear system eq. (1.3.16).

Results

We consider a Gaussian source term $S(x) = \exp(-\frac{(x-0.5)^2}{0.01})$ and a discretization with $\mathcal{N} = 1000$. For given value of N, we build the Lagrange RB approximation subspace based on points $\mu^{(1)}, \ldots, \mu^{(N)}$ uniformly distributed in $[\mu_{\min}, \mu_{\max}] = [1, 15]$, *i.e.*,

$$\mu^{(n)} = \mu_{\min} + \frac{n-1}{N-1}(\mu_{\max} - \mu_{\min}), \quad n = 1, \ldots, N. \tag{1.4.3}$$

Given this Lagrange RB approximation subspace, we can efficiently compute the Galerkin RB approximation $u_N(\mu)$ for all $\mu \in \Xi \subset [\mu_{\min}, \mu_{\max}]$, where Ξ is a given finite set with cardinality $\mathrm{Card}(\Xi) = 300$. This requires solving $\mathrm{Card}(\Xi)$ linear systems of size N. Since this is a toy problem, we can also compute the truth solution $u(\mu)$ for all $\mu \in \Xi$. This requires solving $\mathrm{Card}(\Xi)$ linear systems of size $\mathcal{N}$ (in a non-toy problem, $\mathcal{N}$ would be very large and so these computations would lead to prohibitive costs). Thus, we are able to compute the RB approximation error $\|u(\mu) - u_N(\mu)\|_V$ for all $\mu \in \Xi$.

We have plotted the maximum error $\max_{\mu\in\Xi} \|u(\mu) - u_N(\mu)\|_V$ and the mean error defined by

$$\operatorname*{mean}_{\mu\in\Xi} \|u(\mu) - u_N(\mu)\|_V = \frac{1}{\mathrm{Card}(\Xi)} \sum_{\mu\in\Xi} \|u(\mu) - u_N(\mu)\|_V \tag{1.4.4}$$

for different RB sizes $N = 2, \ldots, 9$. We have also plotted the maximum relative error $\max_{\mu\in\Xi} \|u(\mu) - u_N(\mu)\|_V / \|u(\mu)\|_V$ as well as the mean relative error. The results are shown on fig. 1.1.

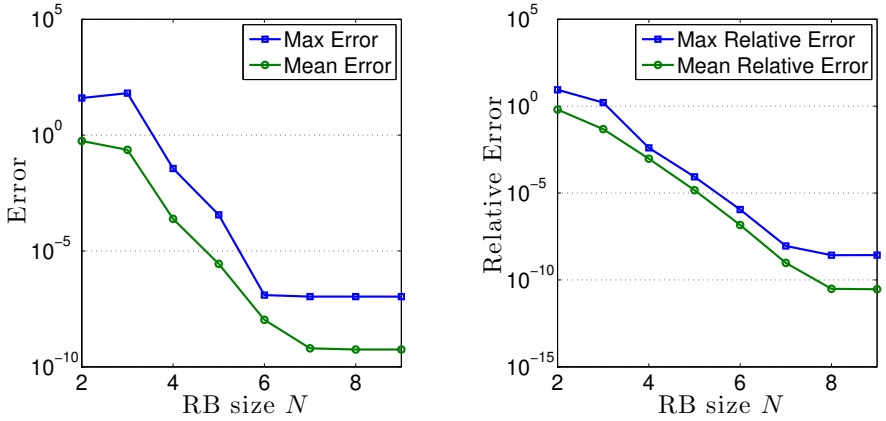


Figure 1.1: Maximum and mean error (left) and relative error (right) for Lagrange RB approximation subspaces of dimension $N = 2, \ldots, 9$.

We find that very few basis functions are sufficient for obtaining a very low RB approximation error. For example, $N = 5$ basis functions yield a mean relative error of $\mathcal{O}(10^{-5})$. In this situation, the truth and RB solutions are visually indistinguishable. We also find that the error stagnates after a threshold number of basis functions. Thus the RB of size

$N = 9$ does not yield to any better results than the RB of size $N = 8$. At this stage, it is worth recalling that the numerically unobtainable PDE solution $u^{\text{ex}}(\mu)$ satisfies

$$\|u_N(\mu) - u^{\text{ex}}(\mu)\|_V \leqslant \|u_N(\mu) - u(\mu)\|_V + \|u(\mu) - u^{\text{ex}}(\mu)\|_V. \tag{1.4.5}$$

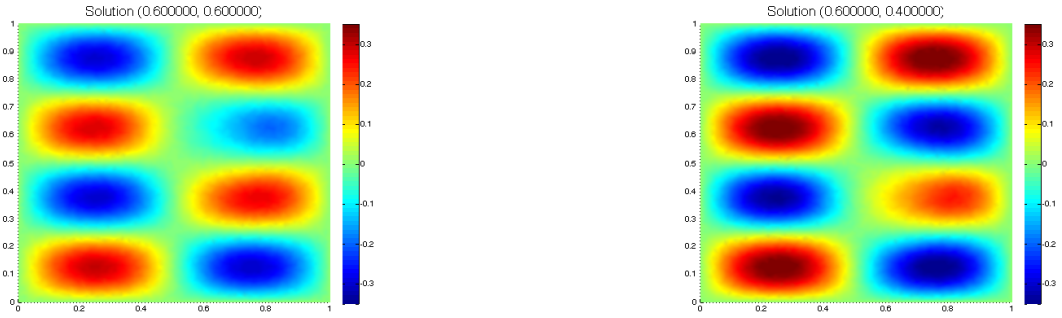
In the light of eq. (1.4.5), it is relevant to decrease the RB error $\|u_N(\mu) - u(\mu)\|_V$ to about the order of magnitude of the discretization error $\|u(\mu) - u^{\text{ex}}(\mu)\|_V$. However, there is no need to further decrease the RB error as this will not improve the overall error on the PDE solution $\|u_N(\mu) - u^{\text{ex}}(\mu)\|_V$, as nicely illustrated in [25].

1.4.2 Model problem 2: Laplace

For the parametrized Laplace problem, we recall that V $(= W)$ corresponds to the $\mathcal{N}$-dimensional FE approximation $X_h^0(\Omega)$ defined by eq. (1.1.14) with $\Omega =]0, 1[\times]0, 1[$. Given the imbedding $V \subset H^1(\Omega)$, the $\|\cdot\|_V$ norm is the usual $H^1(\Omega)$ norm. Based on a triangulation of the domain Ω, we obtain $\mathcal{N} = 3236$ degrees of freedom. We consider the source term

$$\forall x = (x_1, x_2) \in \Omega, \quad S(x) = 20\pi^2 \sin(2\pi x_1)\sin(4\pi x_2). \tag{1.4.6}$$

We consider the parameter space $\mathcal{D} = [0.4,\ 0.6] \times [0.4,\ 0.6]$. Figure 1.2 shows two truth solutions $u(\mu)$ for $\mu = (0.6,\ 0.6)$ and $\mu = (0.6,\ 0.4)$. We notice that the truth solutions exhibit a specific local behavior in the neighborhood of the conductivity peak at coordinates (μ_1, μ_2). We further find that a change in the parameter μ does not only have a local effect, but also affects the amplitude of the solution globally in Ω. Therefore this is an interesting parametrized problem to solve.



(a) Solution at $\mu = (0.6,\ 0.6)$.

(b) Solution at $\mu = (0.6,\ 0.4)$.

Figure 1.2: Truth solution $u(\mu)$ for two possible values of the parameter μ.

Affine approximation

As explained in section 1.3.3, the parametrized laplace operator $A(\mu)$ is non-affine. We build an affine approximation $\widetilde{A}(\mu)$ by applying the EIM to the parametrized conductivity

function $\kappa(\cdot;\mu) : \Omega \to \mathbb{R}$. We apply algorithm 1.1 with Ξ_{space} made of 3436 points in Ω and Ξ_{param} consisting of a 33×33 grid covering parameter space $\mathcal{D}$. The algorithm requires $M = 27$ iteration to converge to the prescribed tolerance $tol = 10^{-2}$.

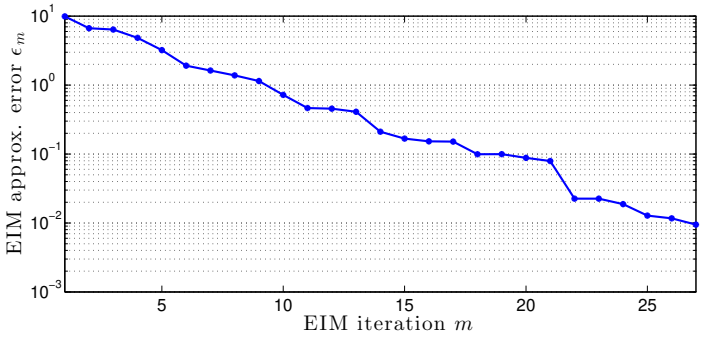


Figure 1.3: Convergence curve of the EIM (algorithm 1.1) applied to the parametrized conductivity defined by eq. (1.1.4).

For the sake of validation, we check that $\widetilde{A}(\mu)$ is indeed a good approximation for $A(\mu)$ by computing the error $\|u(\mu) - \widetilde{u}(\mu)\|_V$, where $u(\mu)$ is the truth solution (which solves $A(\mu)u(\mu) = f$) and where $\widetilde{u}(\mu)$ solves $\widetilde{A}(\mu)\widetilde{u}(\mu) = f$. Based on a (random) set $\Xi \subset \mathcal{D}$ of cardinality 50, we have been able to compute:

$$\max_{\mu\in\Xi} \frac{\|u(\mu) - \widetilde{u}(\mu)\|_V}{\|u(\mu)\|_V} \approx 3.25 \times 10^{-4}, \quad \operatorname*{mean}_{\mu\in\Xi} \frac{\|u(\mu) - \widetilde{u}(\mu)\|_V}{\|u(\mu)\|_V} \approx 1.62 \times 10^{-4}. \tag{1.4.7}$$

We conclude that the non-affine Laplace operator $A(\mu)$ can be safely replaced by its affine approximation $\widetilde{A}(\mu)$.

RB approximation

We build a Lagrange RB approximation subspace of dimension $N = 12$ based on points $\mu^{(1)}, \ldots, \mu^{(N)}$ randomly chosen in $\mathcal{D}$ (of course, such a random choice in $\mathcal{D}$ cannot be optimal, we shall soon explain in chapter 2 how these points can be intelligently chosen). These points are shown by the black stars on fig. 1.4. We consider the Galerkin RB approximation $u_N(\mu)$. Based on a (random) set $\Xi' \subset \mathcal{D}$ of cardinality 100, we have been able to compute:

$$\max_{\mu\in\Xi'} \frac{\|u(\mu) - u_N(\mu)\|_V}{\|u(\mu)\|_V} \approx 9.61 \times 10^{-3}, \quad \operatorname*{mean}_{\mu\in\Xi'} \frac{\|u(\mu) - u_N(\mu)\|_V}{\|u(\mu)\|_V} \approx 1.85 \times 10^{-3}. \tag{1.4.8}$$

Furthermore, we have plotted the relative error with respect to the parameter on fig. 1.4. We observe, quite intuitively, that the error is largest at the points μ that are far from the

$\mu^{(1)}, \ldots, \mu^{(N)}$ (see for example the top right corner). Where the error is smallest (near the $\mu^{(1)}, \ldots, \mu^{(N)}$, for instance in the middle), the relative error reaches at best around 1×10^{-4}. This is consistent with eq. (1.4.7); indeed, the RB approximation $u_N(\mu)$ cannot be a better approximation to $u(\mu)$ than $\widetilde{u}(\mu)$, since the RB solver stems from replacing $A(\mu)$ by its affine approximation $\widetilde{A}(\mu)$.

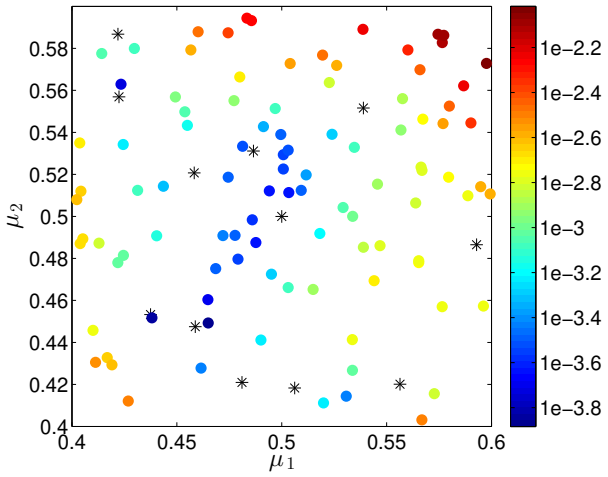


Figure 1.4: Relative error $\mu \in \Xi' \mapsto \frac{\|u(\mu)-u_N(\mu)\|_V}{\|u(\mu)\|_V}$ (circles, with color indicating the magnitude) and the $N = 12$ randomly chosen points $\mu^{(1)}, \ldots, \mu^{(N)}$ used for the Lagrange RB approximation (black stars).

1.5 Conclusions

In this chapter, we have presented a short review of the reduced basis method for parametrized linear PDEs. Because the solution of a parametrized linear PDE is unobtainable in practice, our reference is a high-fidelity approximation (the so-called truth approximation), typically obtained using the finite element method on a fine mesh. Thus, we have introduced a mathematical framework with finite-dimensional Hilbert spaces, which are typically finite element approximation spaces with very large dimension – say $\mathcal{N} = \mathcal{O}(10^{10})$ in the most extreme cases.

In short, the reduced basis method consists in projecting the high-fidelity problem onto a low-dimensional subspace and solving the projected problem, rather than the high-fidelity one, thus enabling significant computational savings. Two possible projections (Galerkin and Least-squares) have been presented and possible choices for the low-dimensional subspace have been reported. In this thesis, only Lagrange reduced basis approximation subspaces will be considered for ease of implementation and in view of obtaining the best computational performance.

In this chapter, we have presented the offline/online computational strategy in detail in the case of affine operators and right-hand sides. We have reviewed the Empirical Interpolation Method (EIM) for recovering affine approximations in the non-affine cases. Finally, the reduced basis method was put in action on two purely academic model problems with a relatively small number of degrees of freedom, *i.e.*, $\mathcal{N} = \mathcal{O}(10^3)$. The reduced basis method is able to reduce the problem to less than a dozen unknowns, while maintaining a very good level of accuracy. At this stage, the accuracy was evaluated by computing some high-fidelity solutions and comparing them to the reduced basis approximations. However, when the problem is large-scale, the cost of computing many high-fidelity solutions is intractable – which is precisely why one resorts to reduced basis methods! The next two chapters will present some strategies to reliably evaluate the accuracy of a reduced basis approximation in an *a posteriori* manner, that is, without having to compute the high-fidelity solutions. The notion of *a posteriori* error estimation will also prove very useful for optimally selecting the set of parameter samples at which the high-fidelity solutions have to be computed when constructing a Lagrange reduced basis.

Chapter 2

The classical *a posteriori* error estimation approach

Summary. In this second chapter, we address the question of *certification*. Indeed, the Reduced Basis Method is only relevant if one can assess the accuracy of the RB approximation. In order to maintain the best computational performance, it is key to be able to bound the RB approximation error without having to compute the high-fidelity solution. This is precisely the role of *a posteriori* error estimators. In this chapter, we recall the classical error estimator consisting of the residual norm divided by the inf-sup constant and show how it can be efficiently computed following an offline/online computational strategy. Readers already familiar with this method can skip to section 2.2.4, where we propose an original method for approximating the inf-sup constant without having to solve any generalized eigenvalue problems. The classical approach is illustrated on the parametrized Helmholtz model problem, while our approach is illustrated on the parametrized Laplace problem.

Contents

2.1 The need for *a posteriori* error estimation 32
2.1.1 The need for certification 32
2.1.2 Greedy RB approach 33
2.2 Classical inf-sup based error estimates 34
2.2.1 Error bound 34
2.2.2 Efficient computation of the residual norm 36
2.2.3 Efficient computation of the inf-sup constant 39
2.2.4 A heuristic approach 44
2.3 Numerical illustration 46

2.3.1 Model problem 1: Helmholtz 46
2.3.2 Model problem 2: Laplace 48
2.4 Conclusions 51

2.1 The need for *a posteriori* error estimation

2.1.1 The need for certification

A RB approximation $u_N(\mu) \in V_N \subset V$ is said to be *certified* with level of accuracy $\epsilon^{\text{rb}} > 0$ if it satisfies

$$\forall \mu \in \mathcal{D}, \quad \|u(\mu) - u_N(\mu)\|_V < \epsilon^{\text{rb}}. \tag{2.1.1}$$

We emphasize that, following the classical RB approach [105, 98] the RB approximation is certified with respect to a given truth approximation $u(\mu)$ in a finite approximation subspace V. Thus in this thesis, we forget about the numerically unobtainable PDE solution $u^{\text{ex}}(\mu)$. This presupposes that the discretization error $\|u(\mu) - u^{\text{ex}}(\mu)\|_V$ is "adequately small", in practice, this can be checked using classical finite element *a priori* or *a posteriori* estimates [122, 10, 20]. We mention that there exists completely different certification approaches, for instance in [124, 125, 126, 130, 25] where the RB approximation is certified with respect to the exact PDE solution.

The notion of *a posteriori* error estimator is key for determining the level of accuracy of a RB approximation.

Definition 5 (Error Estimator)**.** *For all $\mu \in \mathcal{D}$, let us denote $u(\mu) \in \mathcal{D}$ the truth solution and $u_N(\mu) \in V_N \subset V$ the RB approximation. An a posteriori error estimator is a function $\Delta_N : \mathcal{D} \to \mathbb{R}_+$ satisfying*

1. *the **reliability** property $\forall \mu \in \mathcal{D}$, $\|u(\mu) - u_N(\mu)\|_V \leqslant \Delta_N(\mu)$;*
2. *the **effectivity** property $\forall \mu \in \mathcal{D}$, $\exists c(\mu) > 0$, $\|u(\mu) - u_N(\mu)\|_V \geqslant c(\mu)\Delta_N(\mu)$;*
3. *the **a posteriori** property, i.e., the function $\Delta_N(\mu)$ can be evaluated without having to evaluate $u(\mu)$.*

Notice that when an *a posteriori* error estimator $\Delta_N : \mathcal{D} \to \mathbb{R}_+$ is available, the RB approximation is certified with a level of accuracy $\max_{\mu \in \mathcal{D}} \Delta_N(\mu)$.

By definition, an *a posteriori* error estimator bounds the absolute error. In the next proposition, we show how a bound on the absolute error can be translated into a bound on the relative error.

Proposition 2.1.1. *Let $\Delta_N : \mathcal{D} \to \mathbb{R}_+$ be an a posteriori error estimator satisfying*

$$\forall \mu \in \mathcal{D}, \quad \|u_N(\mu)\|_V > \Delta_N(\mu).$$

Then, there holds

$$\forall \mu \in \mathcal{D}, \quad \frac{\|u(\mu) - u_N(\mu)\|_V}{\|u(\mu)\|_V} \leqslant \frac{\Delta_N(\mu)}{\|u_N(\mu)\|_V} \left(1 - \frac{\Delta_N(\mu)}{\|u_N(\mu)\|_V}\right)^{-1}$$

Proof. By the triangular inequality and the reliability property, there holds for all $\mu \in \mathcal{D}$,

$$\|u_N(\mu)\|_V \leqslant \|u(\mu) - u_N(\mu)\|_V + \|u(\mu)\|_V \leqslant \Delta_N(\mu) + \|u(\mu)\|_V. \tag{2.1.2}$$

Thus $\|u(\mu)\|_V \geqslant \|u_N(\mu)\|_V - \Delta_N(\mu)$. Using $\|u_N(\mu)\|_V - \Delta_N(\mu) > 0$, we get

$$\frac{1}{\|u(\mu)\|_V} \leqslant \frac{1}{\|u_N(\mu)\|_V - \Delta_N(\mu)}. \tag{2.1.3}$$

We obtain the desired result by again applying the reliability property, *i.e.*,

$$\frac{\|u(\mu) - u_N(\mu)\|_V}{\|u(\mu)\|_V} \leqslant \frac{\Delta_N(\mu)}{\|u_N(\mu)\|_V - \Delta_N(\mu)} = \frac{\Delta_N(\mu)}{\|u_N(\mu)\|_V} \left(1 - \frac{\Delta_N(\mu)}{\|u_N(\mu)\|_V}\right)^{-1}. \tag{2.1.4}$$

□

Remark that the assumption $\|u_N(\mu)\|_V > \Delta_N(\mu)$ is easy to check in practice since the RB approximation $u_N(\mu)$ is efficiently computable and $\Delta_N(\mu)$ is an *a posteriori* quantity. Recalling the formula $(1 - X)^{-1} = 1 + X + X^2 + \ldots$, we remark that

$$\frac{\Delta_N(\mu)}{\|u_N(\mu)\|_V} \left(1 - \frac{\Delta_N(\mu)}{\|u_N(\mu)\|_V}\right)^{-1} = \frac{\Delta_N(\mu)}{\|u_N(\mu)\|_V} + \mathcal{O}\left(\frac{\Delta_N(\mu)^2}{\|u_N(\mu)\|_V^2}\right) \tag{2.1.5}$$

When the quantity $\frac{\Delta_N(\mu)}{\|u_N(\mu)\|_V}$ is small, then the high order terms become negligible relatively to the first order term. For instance, if the first order term is $\frac{\Delta_N(\mu)}{\|u_N(\mu)\|_V} = \mathcal{O}(10^{-3})$, then the second order term is $\frac{\Delta_N(\mu)^2}{\|u_N(\mu)\|_V^2} = \mathcal{O}(10^{-6})$; that is three orders of magnitude smaller than the first order term, therefore clearly negligible.

2.1.2 Greedy RB approach

An *a posteriori* error estimator $\Delta_N : \mathcal{D} \to \mathbb{R}_+$ is not only useful to assess the level of accuracy of a given RB approximation; it can also be used to construct a certified RB approximation with a prescribed level of accuracy $\epsilon^{\mathrm{rb}} > 0$ [105].

Namely, we consider a Lagrange RB approximation subspace of dimension $N = 1$, $V_1 = \mathrm{Span}\{u(\mu^{(1)})\}$ where $\mu^{(1)} \in \mathcal{D}$ is randomly picked in $\mathcal{D}$. The *a posteriori* error estimator $\Delta_N : \mathcal{D} \to \mathbb{R}_+$ is used in order to determine the current level of accuracy given by $\epsilon_N = \max_{\mu\in\mathcal{D}} \Delta_N(\mu)$. If the computed ϵ_N is below the prescribed tolerance ϵ^{rb}, then we are satisfied with current RB approximation. Otherwise, we identify the RB approximation with largest approximation error $u_N(\mu^\star) \in V_N$, where $\mu^\star$ is given by

$$\mu^\star = \operatorname*{argmax}_{\mu\in\mathcal{D}} \Delta_N(\mu). \tag{2.1.6}$$

We then enrich the RB approximation subspace as $V_{N+1} = V_N \oplus \mathrm{Span}\{u(\mu^\star)\}$ and re-iterate this procedure until the prescribed tolerance is reached. The overall greedy procedure is summarized by algorithm 2.1. The convergence of this algorithm is directly linked to the decay of the Kolmogorov N-width of the solution manifold $\mathscr{M} = \{u(\mu),\, \mu \in \mathcal{D}\}$, as shown in Refs. [12, 18]. As shown in [30], the choice of the cardinality of the discrete training set $\Xi \subset \mathcal{D}$ is intimately linked to the decay rate of the Kolmogorov width, in practice and for simplicity we shall always consider a fine set with about 1000 points.

Algorithm 2.1: Greedy RB construction.

Input : Discrete training set $\Xi \subset \mathcal{D}$, prescribed tolerance $\epsilon^{\mathrm{rb}} > 0$.
Output: Lagrange RB approximation space $V_N = \mathrm{Span}\{u(\mu^{(1)}), \ldots, u(\mu^{(N)})\}$.

Pick arbitrarily $\mu^{(1)}$ in Ξ;
Set $\epsilon_1 = +\infty$, $V_0 = \{0\}$ and initialize $N \leftarrow 1$,;
while $\epsilon_N < \epsilon^{\mathrm{rb}}$ **do**
  Compute $u(\mu^N)$;
  Update RB approximation subspace $V_N = V_{N-1} \oplus \mathrm{Span}\{u(\mu^{(N)})\}$;
  Find $\mu^{N+1} = \operatorname*{argmax}_{\mu\in\Xi} \Delta_N(\mu)$;
  $\epsilon_{N+1} = \max_{\mu\in\Xi} \Delta_N(\mu)$;
  Update $N \leftarrow N + 1$;
end

For completeness, the appendix section B.2 of this manuscript provides some implementation details on which operations should be performed per iteration of algorithm 2.1.

2.2 Classical inf-sup based error estimates

2.2.1 Error bound

We start with the following theorem, which states that the residual and error norms are equivalent norms.

Theorem 1 (Residual and error norm equivalence)**.** *Let $\tilde{v} \in V$. Then,*

$$\frac{1}{\gamma(\mu)}\|A(\mu)\tilde{v} - f(\mu)\|_{W'} \leqslant \|u(\mu) - \tilde{v}\|_V \leqslant \frac{1}{\alpha(\mu)}\|A(\mu)\tilde{v} - f(\mu)\|_{W'}$$

with $\alpha(\mu)$ and $\gamma(\mu)$ the inf-sup and continuity constants respectively, defined in (1.1.23)*.*

Proof. This is direct by the Banach-Necas-Babuska assumptions. The inf-sup condition yields,

$$\|A(\mu)\tilde{v} - f(\mu)\|_{W'} = \|A(\mu)(\tilde{v} - u(\mu))\|_{W'} \geqslant \alpha(\mu)\|\tilde{v} - u(\mu)\|_V;$$

while the continuity assumption yields,

$$\|A(\mu)\tilde{v} - f(\mu)\|_{W'} = \|A(\mu)(\tilde{v} - u(\mu))\|_{W'} \leqslant \gamma(\mu)\|\tilde{v} - u(\mu)\|_V.$$

□

By theorem 1, if the reduced basis residual $A(\mu)u_N(\mu) - f(\mu)$ converges to 0 in $\|\cdot\|_{W'}$ norm, then so does the reduced basis approximation error $u(\mu) - u_N(\mu)$ in $\|\cdot\|_V$ norm. Furthermore, the function $\mu \mapsto \frac{1}{\alpha(\mu)}\|A(\mu)u_N(\mu) - f(\mu)\|_{W'}$ is an *a posteriori* error estimator in the sense of definition 5. Indeed, it is reliable (upper bound for the error) and *a posteriori* (it can be computed without knowledge of the truth solution $u(\mu)$). It remains to prove that it is effective. This is shown by the following proposition.

Proposition 2.2.1 (Effectivity of classical inf-sup based estimator)**.** *There holds,*

$$1 \leqslant \frac{\frac{1}{\alpha(\mu)}\|A(\mu)u_N(\mu) - f(\mu)\|_{W'}}{\|u(\mu) - u_N(\mu)\|_V} \leqslant \frac{\gamma(\mu)}{\alpha(\mu)}.$$

Proof. The left inequality is a restatement of theorem 1. The right inequality is obtained from

$$\begin{aligned}\|A(\mu)u_N(\mu) - f(\mu)\|_{W'} &= \|A(\mu)(u_N(\mu) - u(\mu))\|_{W'} \\ &\leqslant \left(\sup_{v\in V}\frac{\|A(\mu)v\|_{W'}}{\|v\|_V}\right)\|u(\mu) - u_N(\mu)\|_V \\ &= \gamma(\mu)\|u(\mu) - u_N(\mu)\|_V.\end{aligned}$$

□

Typically, in resonant problems, the inf-sup constant may approach 0 near the resonant values of μ. In this situation, the upper bound $\gamma(\mu)/\alpha(\mu)$ may be very large, therefore the inf-sup based error bound provided by Theorem 1 tends to overestimate the error by possibly many orders of magnitude. This effect is well-known in the reduced basis community and effort is currently being made to circumvent this issue. Among the most

recent approaches we cite the hierarchical approach in Ref. [51] and the randomized approach in Ref. [113]. The natural-norm approach, originally introduced in the reduced basis context in Ref. [110], has been used successfully used to resolve effectivity issues [36]. This natural-norm approach will be explained in detail in chapter 3.

2.2.2 Efficient computation of the residual norm

We now explain how the reduced basis residual norm $\|A(\mu)u_N(\mu) - f(\mu)\|_{W'}$ can be efficiently computed using offline/online decoupling. We have the following result.

Proposition 2.2.2 (Residual norm efficient offline/online decoupling)**.** *If*
- *for all* $1 \leqslant p, q \leqslant Q^a$ *the* $N \times N$ *matrix* $\mathbf{P}^* \mathbf{A}_p \mathbf{B}_W^{-1} \mathbf{A}_q \mathbf{P}$*; and*
- *for all* $1 \leqslant q \leqslant Q^f$ *and for all* $1 \leqslant p \leqslant Q^a$ *the* N*-dimensional vector* $\mathbf{P}^* \mathbf{A}_p \mathbf{B}_W^{-1} \mathbf{f}_q$*;*
- *the* $Q^f \times Q^f$ *matrix with coefficients* $\mathbf{f}_q^* \mathbf{B}_W^{-1} \mathbf{f}_p$ *for* $1 \leqslant q, p \leqslant Q^f$*;*

are pre-computed (during the so-called "offline phase"), then the reduced basis residual norm can be computed for any value value of μ *with* $\mathcal{O}((Q^a)^2 N^2 + Q^a Q^f N + (Q^f)^2)$ *complexity (during the so-called "online phase").*

Proof. Using the notations from Chapter 1, we observe that

$$\|A(\mu)u_N(\mu) - f(\mu)\|_{W'} = \|\mathbf{A}(\mu)\mathbf{P}\mathbf{x}(\mu) - \mathbf{f}(\mu)\|_{\mathbf{B}_W^{-1}}. \tag{2.2.1}$$

Developing the square yields

$$\begin{aligned} \|A(\mu)u_N(\mu) - f(\mu)\|_{W'}^2 = {} & \mathbf{x}(\mu)^* \mathbf{P}^* \mathbf{A}(\mu)^* \mathbf{B}_W^{-1} \mathbf{A}(\mu) \mathbf{P} \mathbf{x}(\mu) \\ & - 2\Re\left\{\mathbf{x}(\mu)^* \mathbf{P}^* \mathbf{A}(\mu)^* \mathbf{B}_W^{-1} \mathbf{f}(\mu)\right\} \\ & + \mathbf{f}(\mu)^* \mathbf{B}_W^{-1} \mathbf{f}(\mu). \end{aligned} \tag{2.2.2}$$

An efficient offline/online decoupling for first two terms in (2.2.2) is provided by proposition 1.3.2; while the last term can be expressed as

$$\mathbf{f}(\mu)^* \mathbf{B}_W^{-1} \mathbf{f}(\mu) = \sum_{p=1}^{Q^f} \sum_{q=1}^{Q^f} \theta_q^f(\mu) \overline{\theta_p^f(\mu)} \boxed{\mathbf{f}_p^* \mathbf{B}_W^{-1} \mathbf{f}_q}. \tag{2.2.3}$$

We observe that the boxed scalar quantities do not depend on the parameter μ, thus they can be computed once and for all during the offline phase. □

Remark that, when the Least-Squares RB approximation is used, then $\mathbf{x}(\mu) \in \mathbb{C}^N$ solves the linear system eq. (1.3.16). In this situation, the expression of the RB residual norm is simplified to

$$\|A(\mu)u_N(\mu) - f(\mu)\|_{W'}^2 = \mathbf{x}(\mu)^* \mathbf{P}^* \mathbf{A}(\mu)^* \mathbf{B}_W^{-1} \mathbf{A}(\mu) \mathbf{P} \mathbf{x}(\mu) - \mathbf{f}(\mu)^* \mathbf{B}_W^{-1} \mathbf{f}(\mu). \tag{2.2.4}$$

Numerically robust formulas

Unfortunately, none of the formulas eq. (2.2.2) or eq. (2.2.4) are numerically robust for computing the residual norm [22, 19, 28]. Indeed, the squared residual norm is necessarily a positive quantity, yet it is obtained as the difference between two positive quantities. This situation corresponds to computing $r^2 = (a-b)^2$ by the formula $r^2 = a^2 - 2ab + b^2$. Yet this formula is well known to be numerically unstable, especially in the situation where a and b are very close. Namely, if $a = b + \mathcal{O}(\epsilon)$ where ϵ denotes machine precision, then the floating point representation of $a^2 - 2ab + b^2$ is $\mathcal{O}(\epsilon)$, whereas the floating point representation of $(a-b)^2$ is $\mathcal{O}(\epsilon^2)$. Thus the formula $r^2 = (a-b)^2$ is numerically more robust than the formula $r^2 = a^2 - 2ab + b^2$.

There exist many ways to compute the residual in a robust manner [22, 19, 28]. Following a QR algorithm or a Gram-Schmidt orthonormalization procedure [90, 13, 44], one can compute the matrix $\mathbf{Q} \in \mathbb{C}^{\mathcal{N} \times NQ^a}$ and the upper triangular matrix $\mathbf{R} \in \mathbb{C}^{NQ^a \times NQ^a}$ such that

$$\mathbf{QR} = [\mathbf{B}_W^{-1}\mathbf{A}_1\mathbf{p}_1| \cdots |\mathbf{B}_W^{-1}\mathbf{A}_{Q^a}\mathbf{p}_1| \cdots |\mathbf{B}_W^{-1}\mathbf{A}_1\mathbf{p}_N| \cdots |\mathbf{B}_W^{-1}\mathbf{A}_{Q^a}\mathbf{p}_N], \tag{2.2.5}$$

and such that $\mathbf{Q}$ is an $\mathbf{B}_W$-orthogonal matrix in the sense $\mathbf{Q}^*\mathbf{B}_W\mathbf{Q} = \mathbf{I}$. Let us further introduce the matrix

$$\mathbf{\Theta}(\mu) = \begin{pmatrix} \theta_1^a(\mu) & & \\ \vdots & & 0 \\ \theta_{Q^a}^a(\mu) & & \\ & \ddots & \\ & & \theta_1^a(\mu) \\ 0 & & \vdots \\ & & \theta_{Q^a}^a(\mu) \end{pmatrix} \in \mathbb{C}^{NQ^a \times N}. \tag{2.2.6}$$

This being set, we have a new efficient offline/online decoupling for computing the reduced basis residual norm.

Proposition 2.2.3 (Robust residual norm efficient offline/online decoupling)**.** *If*

- *for all* $1 \leqslant q \leqslant Q^f$ *the* NQ^a*-dimensional vector* $\mathbf{Q}^*\mathbf{f}_q$*; and*
- *for all* $1 \leqslant q, p \leqslant Q^f$ *the scalar quantity* $(\mathbf{B}_W^{-1}\mathbf{f}_k - \mathbf{Q}\mathbf{Q}^*\mathbf{f}_k)^*\mathbf{B}_W(\mathbf{B}_W^{-1}\mathbf{f}_q - \mathbf{Q}\mathbf{Q}^*\mathbf{f}_q)$*; and*
- *the* $NQ^a \times NQ^a$ *upper triangular matrix* $\mathbf{R}$*;*

are pre-computed (during the so-called "offline phase"), then the reduced basis residual norm can be computed for any value value of μ *with* $\mathcal{O}((Q^a)^2N^2 + Q^aQ^fN + (Q^f)^2)$ *complexity (during the so-called "online phase").*

Proof. It is clear that $\mathbf{QR\Theta}(\mu) = \mathbf{B}_W^{-1}\mathbf{A}(\mu)\mathbf{P}$. Using this information, the residual norm is expressed as

$$\begin{aligned} \|\mathbf{A}(\mu)\mathbf{P}\mathbf{x}(\mu) - \mathbf{f}(\mu)\|_{\mathbf{B}_W^{-1}} &= \|\mathbf{B}_W^{-1}\mathbf{A}(\mu)\mathbf{P}\mathbf{x}(\mu) - \mathbf{B}_W^{-1}\mathbf{f}(\mu)\|_{\mathbf{B}_W} \\ &= \|\mathbf{QR\Theta}(\mu)\mathbf{x}(\mu) - \mathbf{B}_W^{-1}\mathbf{f}(\mu)\|_{\mathbf{B}_W}. \end{aligned} \tag{2.2.7}$$

Recalling the Pythagorean theorem

$$\forall \mathbf{v} \in \mathbb{C}^{\mathcal{N}}, \quad \|\mathbf{v}\|^2_{\mathbf{B}_W} = \|\mathbf{Q}\mathbf{Q}^*\mathbf{B}_W\mathbf{v}\|^2_{\mathbf{B}_W} + \|\mathbf{v} - \mathbf{Q}\mathbf{Q}^*\mathbf{B}_W\mathbf{v}\|^2_{\mathbf{B}_W}, \tag{2.2.8}$$

and applying it to $\mathbf{v} = \mathbf{Q}\mathbf{R}\boldsymbol{\Theta}(\mu)\mathbf{x}(\mu) - \mathbf{B}_W^{-1}\mathbf{f}(\mu)$ recalling that $\mathbf{Q}$ is $\mathbf{R}_W$-orthogonal yields

$$\begin{aligned}\|\mathbf{A}(\mu)\mathbf{P}\mathbf{x}(\mu) - \mathbf{f}(\mu)\|^2_{\mathbf{B}_W^{-1}} = \|\mathbf{Q}\mathbf{R}\boldsymbol{\Theta}(\mu)\mathbf{x}(\mu) - \mathbf{Q}\mathbf{Q}^*\mathbf{f}(\mu)\|^2_{\mathbf{B}_W} \\ + \|\mathbf{B}_W^{-1}\mathbf{f}(\mu) - \mathbf{Q}\mathbf{Q}^*\mathbf{f}(\mu)\|^2_{\mathbf{B}_W}.\end{aligned} \tag{2.2.9}$$

Using the property $\forall \mathbf{c} \in \mathbb{C}^{NQ^a}$, $\|\mathbf{Q}\mathbf{c}\|_{\mathbf{B}_W} = \|\mathbf{c}\|_2$, where $\|\cdot\|_2$ denotes the euclidian norm on $\mathbb{C}^{NQ^a}$ we get

$$\begin{aligned}\|\mathbf{A}(\mu)\mathbf{P}\mathbf{x}(\mu) - \mathbf{f}(\mu)\|^2_{\mathbf{B}_W^{-1}} = \|\mathbf{R}\boldsymbol{\Theta}(\mu)\mathbf{x}(\mu) - \mathbf{Q}^*\mathbf{f}(\mu)\|^2_2 \\ + \|\mathbf{B}_W^{-1}\mathbf{f}(\mu) - \mathbf{Q}\mathbf{Q}^*\mathbf{f}(\mu)\|^2_{\mathbf{B}_W}.\end{aligned} \tag{2.2.10}$$

We now decompose each term using that $\mathbf{f}(\mu)$ is affine with Q^f terms. The first term to the right of Eq. eq. (2.2.10) is

$$\|\mathbf{R}\boldsymbol{\Theta}(\mu)\mathbf{x}(\mu) - \mathbf{Q}^*\mathbf{f}(\mu)\|^2_2 = \left\| \mathbf{R}\boldsymbol{\Theta}(\mu)\mathbf{x}(\mu) - \sum_{q=1}^{Q^f} \theta_q^f(\mu) \boxed{\mathbf{Q}^*\mathbf{f}_q} \right\|^2_2 \tag{2.2.11}$$

The boxed quantities in Eq. (2.2.11) (namely, Q^f vectors each of dimension NQ^a) do not depend on the parameter μ, thus they can be computed once and for all during the offline phase. Next, Eq. (2.2.11) can be evaluated online with $\mathcal{O}(N^2(Q^a)^2 + NQ^fQ^a)$ operations. The second term to the right of Eq. eq. (2.2.10) is

$$\begin{aligned}&\|\mathbf{B}_W^{-1}\mathbf{f}(\mu) - \mathbf{Q}\mathbf{Q}^*\mathbf{f}(\mu)\|^2_{\mathbf{B}_W} \\ &\qquad = \sum_{q=1}^{Q^f}\sum_{k=1}^{Q^f} \theta_q^f(\mu)\overline{\theta_k^f(\mu)} \boxed{(\mathbf{B}_W^{-1}\mathbf{f}_k - \mathbf{Q}\mathbf{Q}^*\mathbf{f}_k)^*\mathbf{B}_W(\mathbf{B}_W^{-1}\mathbf{f}_q - \mathbf{Q}\mathbf{Q}^*\mathbf{f}_q)}.\end{aligned} \tag{2.2.12}$$

Again, the boxed quantities do not depend on the parameter μ, thus they can be computed once and for all during the offline phase and so Eq. (2.2.12) can be evaluated online with $\mathcal{O}((Q^f)^2)$ operations. □

This alternative method for computing the residual norm is now robust. Indeed, looking at Eq. (2.2.10), we have decomposed the squared residual norm as the sum of two positive quantities. This situation corresponds to computing $r^2 = a^2 + b^2$, which is a robust formula, free from numerical instabilities.

Note that the online complexity is the same with the robust and the potentially numerically unstable formulas. However, the offline effort is not the same. Indeed, computing the matrices $\mathbf{Q}$ and $\mathbf{R}$ comes at the price of orthonormalizing NQ^a vectors, each of dimension

$\mathcal{N}$. In practice, a standard Gram-Schmidt orthonormalization procedure is not enough and some re-iterations must be considered in order to ensure numerical orthogonality [19]. Alternatively, the rank revealing QR algorithm can be employed [28].

Finally, let us remark that the framework for the robust computation of the residual norm can also be used to provide an alternative efficient offline/online decoupling for the Least-Squares RB solution. Namely, the following proposition is an alternative to proposition 1.3.2.

Proposition 2.2.4 (Robust Least-Squares efficient offline/online decoupling)**.** *If*
- *for all* $1 \leqslant q \leqslant Q^f$ *the* NQ^a*-dimensional vector* $\mathbf{Q}^*\mathbf{f}_q$*; and*
- *the* $NQ^a \times NQ^a$ *upper triangular matrix* $\mathbf{R}$*;*

are pre-computed (during the so-called "offline phase"), then the linear system (1.3.16) *can be assembled for any value of* μ *with* $\mathcal{O}((Q^a)^2N^2 + Q^aQ^fN)$ *complexity (during the so-called "online phase").*

Proof. The left-hand side of the linear system (1.3.16) can be assembled using the formula

$$\mathbf{P}^*\mathbf{A}(\mu)^*\mathbf{B}_W^{-1}\mathbf{A}(\mu)\mathbf{P} = \mathbf{\Theta}(\mu)^*\mathbf{R}^*\mathbf{R}\mathbf{\Theta}(\mu), \tag{2.2.13}$$

while the right-hand side can be assembled using

$$\begin{aligned}\mathbf{P}^*\mathbf{A}(\mu)^*\mathbf{B}_W^{-1}\mathbf{f}(\mu) &= \mathbf{\Theta}(\mu)^*\mathbf{R}^*\mathbf{Q}^*\mathbf{f}(\mu) \\ &= \sum_{q=1}^{Q^f} \theta_q^f(\mu)\mathbf{\Theta}(\mu)^*\mathbf{R}^*\boxed{\mathbf{Q}^*\mathbf{f}_q}.\end{aligned} \tag{2.2.14}$$

□

Compared to proposition 1.3.2, the strategy provided by proposition 2.2.4 is numerically more robust. Indeed, the formula eq. (2.2.13) assembles the left-hand side as $\mathbf{M}^*\mathbf{M}$ with $\mathbf{M} = \mathbf{R}\mathbf{\Theta}(\mu)$, thus the left-hand side is guaranteed to be hermitian. On the contrary, the strategy of proposition 1.3.2 tries to recover the theoretically hermitian left-hand side matrix using a sum of $(Q^a)^2$ non-hermitian matrices via the formula (1.3.13). Due to finite numerical precision, this formula may yield a non-hermitian left-hand side.

2.2.3 Efficient computation of the inf-sup constant

Using the notations of Chapter 1, the inf-sup constant $\alpha(\mu)$ defined by eq. (1.1.24a) is algebraically expressed as

$$\alpha(\mu) = \left(\inf_{\mathbf{v}\in\mathbb{C}^{\mathcal{N}}\setminus\{0\}} \frac{\mathbf{v}^*\mathbf{A}(\mu)^*\mathbf{B}_W^{-1}\mathbf{A}(\mu)\mathbf{v}}{\mathbf{v}^*\mathbf{B}_V\mathbf{v}}\right)^{1/2}. \tag{2.2.15}$$

Introducing the hermitian matrices $\mathbf{H}(\mu) = \mathbf{A}(\mu)^* \mathbf{B}_W^{-1} \mathbf{A}(\mu)$ and $\mathbf{X} = \mathbf{B}_V$, the inf-sup constant is equal to the square root of the smallest eigenvalue in the generalized eigenvalue problem

$$\begin{cases} \textit{Find } (\lambda, \mathbf{v}) \in \mathbb{R} \times \mathbb{C}^{\mathcal{N}} \setminus \{0\} \textit{ such that} \\ \mathbf{H}(\mu)\mathbf{v} = \lambda \mathbf{X}\mathbf{v}. \end{cases} \tag{G.E.P.}$$

Note that, given that $\mathbf{H}(\mu)$ is hermitian and positive-definite, the eigenvalues of (G.E.P.) are necessarily real and strictly positive. Thus, computing the discrete inf-sup constant $\alpha(\mu)$ for fixed parameter $\mu \in \mathcal{D}$ requires a large-scale generalized eigenvalue problem [97, Chapter 2, §2.4.6].

In practice, the generalized eigenvalue problem G.E.P. is preferably solved using iterative methods such as inverse iteration or the Lanczos method [90]. For completeness, these methods are recalled in appendix A. We observe that such eigensolvers are extremely expensive, since obtaining the smallest eigenvalue requires about a dozen $\mathbf{A}(\mu)$ solves and as many adjoint solves. Thus, computing the smallest eigenvalue for all possible values of μ in $\mathcal{D}$ is computationally unfeasible. In this context, we must turn to the *Successive Constraint Method* (SCM), which is well suited for efficiently constructing reliable lower bounds for the smallest eigenvalue without solving too many large-scale eigenvalue problems [61, 60, 112]. We now review the SCM to explain how this works.

Affine framework

The SCM however only applies in the context of an affinely parametrized matrix $\mathbf{H}(\mu)$, in the sense

$$\forall \mu \in \mathcal{D}, \quad \mathbf{H}(\mu) = \sum_{q=1}^{Q} \theta_q(\mu) \mathbf{H}_q, \tag{2.2.16}$$

where $\{\mathbf{H}_q\}_{q=1}^{Q}$ are $\mathcal{N} \times \mathcal{N}$ parameter-independent *hermitian* matrices and $\{\theta_q(\cdot)\}_{q=1}^{Q}$ are *real-valued* functions of $\mu \in \mathcal{D}$.

Note that, if our operator $A(\mu) \in \mathcal{L}(V, W')$ is affine in the sense of definition 3, then the parametrized matrix $\mathbf{H}(\mu)$ is automatically affine with hermitian terms and real-valued coefficients of $\mu \in \mathcal{D}$. Indeed, in this situation there holds

$$\forall \mu \in \mathcal{D}, \quad \mathbf{A}(\mu) = \sum_{q=1}^{Q^a} \theta_q^a(\mu) \mathbf{A}_q, \tag{2.2.17}$$

where $\{\mathbf{A}_q\}_{q=1}^{Q^a}$ are $\mathcal{N} \times \mathcal{N}$ parameter-independent complex matrices (not necessarily hermitian) and $\{\theta_q^a(\cdot)\}_{q=1}^{Q^a}$ are complex-valued functions of $\mu \in \mathcal{D}$. We can straightforwardly

show, in the fashion of [55, §3.3.3], the following formula

$$\mathbf{H}(\mu) = \sum_{q=1}^{Q^a} \left(|\theta_q^a(\mu)|^2 - \sum_{k=1}^{q-1} (\Re\{z_{kq}\} - \Im\{z_{kq}\}) - \sum_{k=q+1}^{Q^a} (\Re\{z_{kq}\} + \Im\{z_{kq}\}) \right) \mathbf{A}_q^* \mathbf{X}^{-1} \mathbf{A}_q$$
$$+ \sum_{q=1}^{Q^a} \sum_{k=q+1}^{Q^a} \left(\Re\{z_{kq}\} (\mathbf{A}_k + \mathbf{A}_q)^* \mathbf{X}^{-1} (\mathbf{A}_k + \mathbf{A}_q) + \Im\{z_{kq}\} (\mathbf{A}_k + i\mathbf{A}_q)^* \mathbf{X}^{-1} (\mathbf{A}_k + i\mathbf{A}_q) \right)$$

where $z_{qk} = \overline{\theta_k^a(\mu)} \theta_q^a(\mu)$. Thus, the desired affine decomposition (2.2.16) is achieved with at most $Q = (Q^a)^2$ terms. Remark that in the specific case where the functions $\{\theta_q^a(\cdot)\}_{q=1}^{Q^a}$ are real-valued, there holds $\Im\{z_{qk}\} = 0$ for all $q, k = 1, \ldots, Q^a$, therefore the affine decomposition (2.2.16) is achieved with only $Q = Q^a(Q^a + 1)/2$ terms.

Basic SCM

Denote $\lambda_{\min}(\mu)$ the smallest eigenvalue of (G.E.P.). The decomposition (2.2.16) is key to the SCM. Indeed, the method relies on the following statement

$$\lambda_{\min}(\mu) = \inf_{\mathbf{v} \in \mathbb{C}^{\mathcal{N}} \setminus \{0\}} \frac{\mathbf{v}^* \mathbf{H}(\mu) \mathbf{v}}{\mathbf{v}^* \mathbf{X} \mathbf{v}} = \inf_{\mathbf{v} \in \mathbb{C}^{\mathcal{N}} \setminus \{0\}} \sum_{q=1}^{Q} \theta_q(\mu) \frac{\mathbf{v}^* \mathbf{H}_q \mathbf{v}}{\mathbf{v}^* \mathbf{X} \mathbf{v}} = \inf_{\mathbf{v} \in \mathbb{C}^{\mathcal{N}} \setminus \{0\}} \theta(\mu)^T R(\mathbf{v}),$$

where we have introduced $\theta(\mu) \equiv [\theta_1(\mu) \ \cdots \ \theta_Q(\mu)]^T \in \mathbb{R}^Q$ and $R : \mathbb{C}^{\mathcal{N}} \setminus \{0\} \to \mathbb{R}^Q$ as

$$R(\mathbf{v}) \equiv \left[\frac{\mathbf{v}^* \mathbf{H}_1 \mathbf{v}}{\mathbf{v}^* \mathbf{X} \mathbf{v}} \ \cdots \frac{\mathbf{v}^* \mathbf{H}_Q \mathbf{v}}{\mathbf{v}^* \mathbf{X} \mathbf{v}} \right]^T.$$

Introducing the set $\mathcal{Y} \equiv \mathrm{Range}(R) \equiv \{y \in \mathbb{R}^Q : \exists \mathbf{v} \in \mathbb{C}^{\mathcal{N}} \setminus \{0\}, \ y = R(\mathbf{v})\}$, one has

$$\lambda_{\min}(\mu) = \inf_{y \in \mathcal{Y}} \theta(\mu)^T y.$$

Thus, the smallest eigenvalue problem is brought to an optimization problem over the set $\mathcal{Y}$. A first idea to characterize the set $\mathcal{Y}$ is to introduce the *bounding box*

$$\mathcal{B} \equiv \prod_{q=1}^{Q} \left[\inf_{\mathbf{v} \in \mathbb{C}^{\mathcal{N}} \setminus \{0\}} \frac{\mathbf{v}^* \mathbf{H}_q \mathbf{v}}{\mathbf{v}^* \mathbf{X} \mathbf{v}}, \sup_{\mathbf{v} \in \mathbb{C}^{\mathcal{N}} \setminus \{0\}} \frac{\mathbf{v}^* \mathbf{H}_q \mathbf{v}}{\mathbf{v}^* \mathbf{X} \mathbf{v}} \right] \subset \mathbb{R}^Q.$$

Clearly, $\mathcal{Y} \subset \mathcal{B}$ but this inclusion is usually not sharp.

The basic principle of the successive constraints method is to further characterize the set $\mathcal{Y}$ by using information from $J \geq 0$ eigensolves [61]. Let us assume that the smallest eigenvalues $\lambda_{\min}(\mu_1), \ldots, \lambda_{\min}(\mu_J)$ and associated generalized eigenvectors have been computed. These J eigensolves at parameter points μ belonging to the finite set $C_J = \{\mu_1, \ldots, \mu_J\} \subset \mathcal{D}$ can be performed using the algorithms presented in Appendix A. For

simplicity of notation, we shall now denote, for all $j \in \{1, \ldots, J\}$, $\lambda_j \equiv \lambda_{\min}(\mu_j)$ and $\mathbf{v}_j$ the associated eigenvector. This information is used to further characterize the set $\mathcal{Y}$.

Let $y \in \mathcal{Y}$. By definition, there exists $\mathbf{v}_y \in \mathbb{C}^{\mathcal{N}} \setminus \{0\}$ such that $y = R(\mathbf{v}_y)$. For all $j \in \{1, \ldots, J\}$ there holds

$$\theta(\mu_j)^T y = \frac{\mathbf{v}_y^* \mathbf{H}(\mu_j)\mathbf{v}_y}{\mathbf{v}_y^* \mathbf{X}\mathbf{v}_y} \geq \inf_{\mathbf{v}\in\mathbb{C}^{\mathcal{N}}\setminus\{0\}} \frac{\mathbf{v}^* \mathbf{H}(\mu_j)\mathbf{v}}{\mathbf{v}^* \mathbf{X}\mathbf{v}} = \lambda_j.$$

This establishes the inclusion $\mathcal{Y} \subset \mathcal{Y}_{LB}(C_J)$, where

$$\mathcal{Y}_{LB}(C_J) \equiv \{y \in \mathcal{B} : \theta(\mu_j)^T y \geq \lambda_j,\ j = 1, \ldots, J\}. \tag{2.2.18}$$

Thus, for all $\mu \in \mathcal{D}$, the quantity

$$\lambda_{LB}(\mu; C_J) \equiv \inf_{y\in\mathcal{Y}_{LB}(C_J)} \theta(\mu)^T y \tag{2.2.19}$$

is a lower bound for $\lambda_{\min}(\mu)$. It turns out that the optimization problem (2.2.19) can be put in the form of a linear program with Q variables, $2Q$ box constraints and J linear constraints [61].

In order to assess the accuracy of the lower bound $\lambda_{LB}(\mu, C_J)$, it is common to also compute an upper bound for $\lambda_{\min}(\mu)$. Clearly,

$$\mathcal{Y}_{UB}(C_J) \equiv \{R(\mathbf{v}_j),\ j = 1, \ldots, J\}, \tag{2.2.20}$$

is a subset of $\mathcal{Y}$. Thus, for all $\mu \in D$, the quantity

$$\lambda_{UB}(\mu; C_J) \equiv \min_{y\in\mathcal{Y}_{UB}(C_J)} \theta(\mu)^T y \tag{2.2.21}$$

will be an upper bound for $\lambda_{\min}(\mu)$. Solving this minimization problem is trivial, it can be done by enumeration because $\mathcal{Y}_{UB}$ is finite. When the relative difference between lower and upper bounds

$$\epsilon_{SCM}(\mu; C_J) \equiv \frac{\lambda_{UB}(\mu; C_J) - \lambda_{LB}(\mu; C_J)}{\lambda_{UB}(\mu; C_J)} \tag{2.2.22}$$

is below some prescribed tolerance, then the lower bound $\lambda_{LB}(\mu; C_J)$ is considered to be a good approximation of $\lambda_{\min}(\mu)$ from below. When this is not the case, the set C_J must be enriched, which means that more eigensolves are needed. Clearly, the parameter point $\mu \in \mathcal{D}$ where $\epsilon_{SCM}(\mu; C_J)$ is largest is a good candidate for the next eigensolve. This greedy strategy to compute the lower bounds is summarized in algorithm 2.2.

Computational complexity. Let us discuss the computational complexity of the basic SCM algorithm. The algorithm requires the solution of $2Q+J$ eigenvalue problems of size $\mathcal{N} \times \mathcal{N}$ for determining both the bounding box in the beginning and the smallest eigenpair in each iteration. In practice, parameter space $\mathcal{D}$ is discretized by a finite training set

Algorithm 2.2: Basic SCM algorithm

Compute for all $q = 1, \dots, Q$, $\sigma_q^- = \inf_{\mathbf{v} \in \mathbb{C}^{\mathcal{N}} \setminus \{0\}} \frac{\mathbf{v}^* \mathbf{H}_q \mathbf{v}}{\mathbf{v}^* \mathbf{X} \mathbf{v}}$ and $\sigma_q^+ = \sup_{\mathbf{v} \in \mathbb{C}^{\mathcal{N}} \setminus \{0\}} \frac{\mathbf{v}^* \mathbf{H}_q \mathbf{v}}{\mathbf{v}^* \mathbf{X} \mathbf{v}}$;

Initialize $J \leftarrow 0$, $\epsilon_{SCM} = \infty$, $C_0 = \emptyset$;

while $J \leq J_{\max}$ *and* $\epsilon_{SCM} > tol$ **do**

$\mu_{J+1} \leftarrow \operatorname{argmax}_{\mu \in \Xi \subset \mathcal{D}} \epsilon_{SCM}(\mu; C_J)$;

$C_{J+1} \leftarrow C_J \cup \{\mu_{J+1}\}$;

Eigensolve for eigenpair $(\lambda_{J+1}, \mathbf{v}_{J+1})$;

$\epsilon_{SCM} \leftarrow \max_{\mu \in D} \epsilon_{SCM}(\mu; C_{J+1})$;

$J \leftarrow J + 1$;

end

$\Xi \subset \mathcal{D}$, whose cardinality is usually $|\Xi| = \mathcal{O}(10^3)$. Computing lower bounds $\lambda_{LB}(\mu; C_J)$ for all $\mu \in \Xi$ requires $|\Xi|$ linear programs with Q variables and $2Q + J$ constraints. Since the upper bounds $\lambda_{UB}(\mu; C_J)$ for all $\mu \in \Xi$ are comparatively inexpensive to compute, the cost of the maximization of $\mu \mapsto \epsilon_{SCM}(\mu; C_J)$ over Ξ is essentially dominated by the costs of the linear programs.

Beyond the basic SCM. In the original paper [61], the set $\mathcal{Y}$ is further characterized by "positivity constraints". Indeed, knowing that $\lambda_{\min}(\mu) > 0$ (from the positive-definiteness of $\mathbf{H}(\mu)$), it is clear that

$$\forall \mu \in \mathcal{D},\ \forall y \in \mathcal{Y}, \quad \theta(\mu)^T y > 0. \tag{2.2.23}$$

Hence the idea is to replace the definition (2.2.18) of the set $\mathcal{Y}_{LB}(C_J)$ by

$$\begin{aligned} \mathcal{Y}_{LB}(C_J) = \{ & y \in \mathcal{B} : \theta(\mu_j)^T y \geq \lambda_j,\ j = 1, \dots, J \\ & \text{and } \theta(\mu')^T y \geq 0, \forall \mu' \in \mathcal{D} \}. \end{aligned} \tag{2.2.24}$$

When the set $\mathcal{D}$ is discretized by a finite set $\Xi \subset \mathcal{D}$, this amounts to $2Q+J+|\Xi|$ constraints in the linear program (2.2.19). This is of course unpractical when $|\Xi| = \mathcal{O}(10^3)$. To reduce costs, one keeps only $M_+ \geq 0$ positivity constraints among the $|\Xi|$. When solving the linear program (2.2.19) at parameter μ (that is, when computing $\lambda_{LB}(\mu; C_J)$) a good choice is to enforce positivity constraints locally around μ. Introducing the set $\mathcal{P}_{M_+}(\mu; \Xi)$ of the M_+ points in Ξ closest to μ (with respect to the euclidian norm), this amounts to replacing the $|\Xi|$ positivity constraints $\theta(\mu')^T y \geq 0, \forall \mu' \in \Xi$ by the M_+ local positivity constraints $\theta(\mu')^T y \geq 0, \forall \mu' \in P_{M_+}(\mu; \Xi)$. Thus, the number of constraints in the linear program is reduced to $2Q + J + M_+$.

The same idea can be used to replace the J constraints $\theta(\mu_j)^T y \geq \lambda_j,\ j = 1, \dots, J$ by the $M_\lambda \leq J$ constraints $\theta(\mu')^T y \geq \lambda_{\min}(\mu'),\ \forall \mu' \in \mathcal{P}_{M_\lambda}(\mu; C_J)$, where $\mathcal{P}_{M_\lambda}(\mu; C_J)$ is the

set of the M_λ points in C_J closest to μ (again, with respect to the euclidian norm). This reduces the number of constraints in the linear programs to $2Q + M_\lambda + M_+$.

Building on the original SCM paper [61], it is suggested in [27] to further replace the M_+ local positivity constraints by local "monotony constraints". At iteration $J \geq 1$, when solving the linear program (2.2.19) at parameter μ, one replaces the M_+ local positivity constraints $\theta(\mu')^T y \geq 0, \forall \mu' \in P_{M_+}(\mu; \Xi)$, by the M_+ local monotony constraints $\theta(\mu')^T y \geq \lambda_{LB}(\mu; C_{J-1}), \forall \mu' \in P_{M_+}(\mu; \Xi)$. This proves to significantly enhance the convergence properties of the original SCM algorithm. Further improvements of the SCM can be found in [112].

2.2.4 A heuristic approach

The function $\mu \mapsto \frac{1}{\alpha_{LB}(\mu)} \|A(\mu)u_N(\mu) - f(\mu)\|_{W'}$, where $\alpha_{LB}(\mu)$ is a lower bound of the inf-sup stability constant obtained using the SCM is a rigorous *a posteriori* error estimator. However, the SCM requires solving some large-scale eigenvalue problems of the form eq. (G.E.P.). These eigensolves can represent significant computational costs even when only a few of them are required.

In view of saving computational time and resources, we propose to replace the rigorous *a posteriori* error estimator by the heuristic indicator $\mu \mapsto \frac{1}{\hat{\alpha}(\mu)} \|A(\mu)u_N(\mu) - f(\mu)\|_{W'}$ where $\hat{\alpha}(\mu)$ is not a rigorous lower bound for the inf-sup constant. Our approach has the benefit of not requiring any large-scale eigensolve.

We propose to build the function $\mu \mapsto \hat{\alpha}(\mu)$ by sampling, at the iterations $N \geqslant 2$ of the Greedy algorithm, the quantity

$$\hat{\alpha}_N = \frac{\|A(\mu^N)u_{N-1}(\mu^N) - f(\mu^N)\|_{W'}}{\|u_{N-1}(\mu^N) - u(\mu^N)\|_V}. \tag{2.2.25}$$

This quantity can be computed efficiently, since the numerator is the residual norm (efficiently computed by the offline/online strategy) and the denominator consists in the RB approximation $u_{N-1}(\mu^N)$ (again, efficiently computed by the offline/online strategy) and $u(\mu^N)$ is readily available, since it is computed in the Nth iteration of the Greedy algorithm.

The quasi-constant case

Given the samples $\{\hat{\alpha}_N\}_{2 \leqslant N \leqslant N_{\max}}$, where $N_{\max}$ denotes the number of iterations performed by the Greedy algorithm, we can straightforwardly compute the mean $\hat{\alpha}$ and variance σ^2 as

$$\hat{\alpha} = \frac{1}{N^{\max} - 1} \sum_{N=2}^{N^{\max}} \hat{\alpha}_N, \quad \sigma^2 = \frac{1}{N^{\max} - 1} \sum_{N=2}^{N^{\max}} (\hat{\alpha}_N - \hat{\alpha})^2, \tag{2.2.26}$$

as well as the relative standard deviation $C_V = \sigma/\hat{\alpha}$ (expressed in %). We provide the following heuristic: *if* $C_V < 50\%$*, then* $\mu \mapsto \frac{1}{\hat{\alpha}}\|A(\mu)u_N(\mu) - f(\mu)\|_{W'}$ *is a relevant heuristic error indicator.*

As we shall see in the numerical examples, this heuristic performs very well for certain types of problems. However, if $C_V \geqslant 50\%$, one must resort to more advanced techniques.

Extension using interpolation

When $C_V \geqslant 50\%$, this means the distribution of the samples $\{\hat{\alpha}_N\}_{2\leqslant N\leqslant N_{\max}}$ is relatively stretched. In this situation, we propose to build the function $\mu \mapsto \hat{\alpha}(\mu)$ by interpolating the $N_{\max} - 1$ sampled points $\{\hat{\alpha}_N\}_{2\leqslant N\leqslant N_{\max}}$. As in [74, 97], we rely on a radial basis function (RBF) interpolation technique, interpolating the logarithm in order to be sure to a obtain a positive $\mu \mapsto \hat{\alpha}(\mu)$ interpolant. The RBF interpolant is given by

$$\log \hat{\alpha}(\mu) = \omega_0 + \omega^T \mu + \sum_{j=1}^{N^{\max}-1} c_j \phi(|\mu - \mu^{j+1}|), \tag{2.2.27}$$

where ϕ is a radial basis function, typically $\phi(r) = e^{-r^2}$ and $|\cdot|$ denotes the euclidian norm in $\mathbb{R}^p$ (recalling that $\mu \in \mathcal{D} \subset \mathbb{R}^p$, with p denoting the number of parameters). We require the unknown weighs $(\omega_0, \omega, c) \in \mathbb{R} \times \mathbb{R}^p \times \mathbb{R}^{N^{\max}-1}$ to satisfy the interpolation property

$$\log \hat{\alpha}(\mu^N) = \log \hat{\alpha}_N, \quad N = 2, \ldots, N^{\max}, \tag{2.2.28}$$

as well as the two conditions

$$\sum_{j=1}^{N^{\max}-1} c_j = 0, \qquad \sum_{j=1}^{N^{\max}-1} c_j \mu_\ell^{j+1} = 0, \ \ell = 1, \ldots, p. \tag{2.2.29}$$

Clearly, eq. (2.2.28) and eq. (2.2.29) lead to the following linear system

$$\begin{pmatrix} \mathbf{\Phi} & \mathbf{M}^T & \mathbf{1}^T \\ \mathbf{M} & 0 & 0 \\ \mathbf{1} & 0 & 0 \end{pmatrix} \begin{pmatrix} c \\ \omega \\ \omega_0 \end{pmatrix} = \begin{pmatrix} \hat{\boldsymbol{\alpha}} \\ 0 \\ 0 \end{pmatrix}, \tag{2.2.30}$$

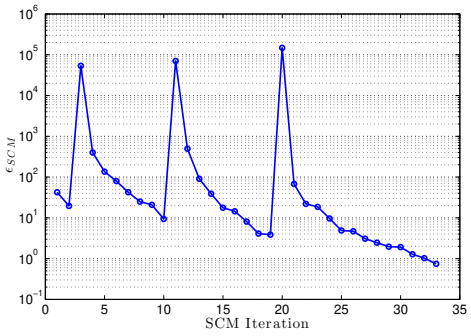
where $\mathbf{\Phi}_{ij} = \Phi(|\mu^{i+1} - \mu^{j+1}|)$ and $\mathbf{M}_{\ell,j} = \mu_\ell^{j+1}$ for $1 \leqslant i, j \leqslant N^{\max} - 1$ and $1 \leqslant \ell \leqslant p$. Here, we use the notations $\mathbf{1} = [1, \ldots, 1] \in \mathbb{R}^{N^{\max}-1}$ and $\hat{\boldsymbol{\alpha}} = [\log \hat{\alpha}_2, \ldots, \log \hat{\alpha}_{N^{\max}}]$.

Once the weighs $(\omega_0, \omega, c) \in \mathbb{R} \times \mathbb{R}^p \times \mathbb{R}^{N^{\max}-1}$ are obtained through solving the linear system eq. (2.2.30), the RBF interpolant $\hat{\alpha}(\mu)$ can be computed very efficiently for any value of $\mu \in \mathcal{D}$ using the formula eq. (2.2.27).

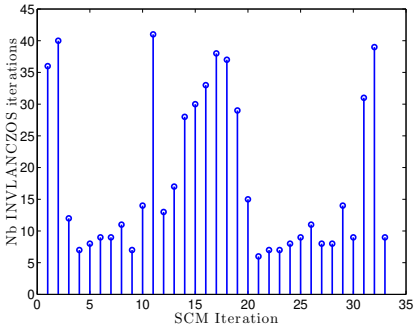
2.3 Numerical illustration

2.3.1 Model problem 1: Helmholtz

We come back to our Helmholtz model problem, introduced in chapter 1. We start by applying the SCM, in order to obtain a cheap lower bound for the inf-sup constant $\mu \mapsto \alpha(\mu)$. To this end, we apply algorithm 2.2 with a discrete set $\Xi \subset \mathcal{D} = [1, 15]$ made of 200 uniformly distributed points and with prescribed tolerance $tol = 0.9$. The large-scale generalized eigenvalue problems G.E.P. are solved via the inverse Lanczos algorithm with a prescribed tolerance of 10^{-7} (see Appendix A).



(a) Maximum relative difference between lower and upper bounds (see eq. (2.2.22)) per SCM iteration.



(b) Number of iterations in the inverse Lanczos algorithm (see Appendix A) per SCM iteration.

Figure 2.1: The SCM (see algorithm 2.2) applied to the Helmholtz model problem.

The algorithm requires $J = 33$ iterations to terminate. The convergence curve is shown on fig. 2.1a. We find that the SCM error does not decrease monotonically. Indeed, the iterations $3, 11$ and 20 are associated with an overshoot in the SCM error, this is due to the discovery of the resonant wavenumbers 2π, 3π and 4π. The SCM error recovers a monotonically decreasing behavior by iteration 20 and from there it decays at a constant rate until the prescribed tolerance is reached. Figure 2.1b shows the number of iterations in each call to the inverse Lanczos algorithm used to solve the large-scale generalized eigenvalue problem G.E.P.. In the worst cases (typically, in between two resonant wavenumbers where eigenvalue crossings occur), the number of iterations can reach 40, while in the more favorable cases the number of iterations is below 10. Figure 2.1b provides a measure of the computational costs associated to the SCM: namely, each inverse Lanczos iteration requires solving one direct problem *find* $\mathbf{x} \in \mathbb{C}^{\mathcal{N}}$ *such that* $\mathbf{A}(\mu)\mathbf{x} = \mathbf{y}$ and one adjoint problem *find* $\mathbf{x} \in \mathbb{C}^{\mathcal{N}}$ *such that* $\mathbf{A}(\mu)^*\mathbf{x} = \mathbf{y}$ (notice that the Helmholtz problem is self-adjoint therefore $\mathbf{A}(\mu) = \mathbf{A}(\mu)^*$, see section 1.4). The overall number of inverse Lanczos iterations performed throughout the $J = 33$ SCM iterations being 600, we conclude that the computational costs associated to the SCM is significant and may become prohibitive should $\mathcal{N}$ be very large (here $\mathcal{N} = 1000$ is far from any industrial

application).

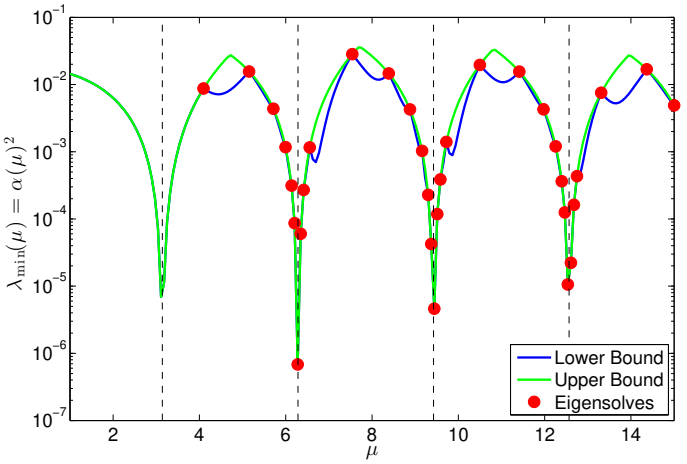
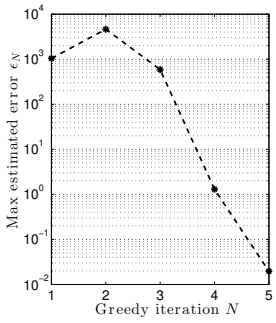


Figure 2.2: The SCM lower and upper bounds for the Helmholtz model problem. Dotted vertical lines indicate the resonant wavenumbers $\pi, 2\pi, 3\pi$ and 4π.

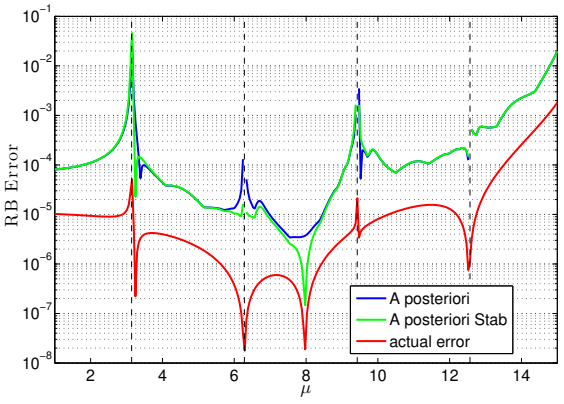
Fortunately, algorithm 2.2 only needs to be applied once. Then, we are able to compute, for any $\mu \in \mathcal{D}$, cheap lower and upper bounds $\alpha_{LB}(\mu)$, $\alpha_{UB}(\mu)$ such that the inf-sup constant $\alpha(\mu)$ satisfies $\alpha_{LB}(\mu) \leqslant \alpha(\mu) \leqslant \alpha_{UB}(\mu)$. Figure 2.2 shows the lower and upper bounds for 300 uniformly distributed points. As expected, we find that the inf-sup constant $\mu \mapsto \alpha(\mu)$ approaches 0 near the theoretical resonant wavenumbers $\pi, 2\pi, 3\pi$ and 4π represented by the dotted vertical lines.

Now that we can cheaply evaluate $\mu \mapsto \alpha_{LB}(\mu)$, we can apply the Greedy RB algorithm 2.1 with the error estimator $\Delta_N : \mu \mapsto \frac{1}{\alpha_{LB}(\mu)} \|A(\mu)u_N(\mu) - f(\mu)\|_{W'}$. Namely, we prescribe a tolerance $\epsilon^{\text{rb}} = 2 \times 10^{-2}$ and use a discrete surrogate set $\Xi \subset \mathcal{D}$ made up of 200 uniformly distributed points. The convergence curve of algorithm 2.1 shown on fig. 2.3a exhibits the exponential decrease of the maximum (over Ξ) of our error estimator with respect to the number of basis functions in the RB.

Now that a RB of size $N = 5$ is available, we can very efficiently evaluate the RB approximation $\mu \mapsto u_N(\mu)$ (in the numerical examples, we use the Galerkin RB approximation) as well as the associated *a posteriori* error estimator $\mu \mapsto \Delta_N(\mu)$. For the sake of validation, we have computed the finite element solution $u(\mu)$ for 300 uniformly distributed values of μ and have computed the actual error $\|u(\mu) - u_N(\mu)\|_V$. We compare the actual error to the *a posteriori* error estimator on fig. 2.3b. The *a posteriori* error estimator is computed using the two different formulas, without (blue curve) and with stabilization (green curve). The difference between the two formulas is clearly visible in the neighborhood of $\mu = 8$, where the stabilized formula is robust with respect to a very small residual norm, while the non-stabilized formula stagnates and provides a poorer estimate. Figure 2.3b illustrates theorem 1, since our *a posteriori* error estimator is indeed an upper bound for the actual error. Proposition 2.2.1 is also nicely illustrated, as we find our *a*



(a) Maximum of the error estimator over $\Xi \subset \mathcal{D}$ per Greedy iteration.



(b) Error and *a posteriori* error estimators.

Figure 2.3: A RB of size $N = 5$ for the Helmholtz model problem.

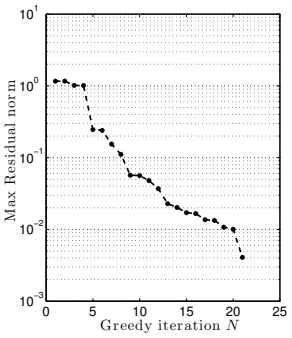
posteriori error estimator to be quite pessimistic. Indeed, the actual error is almost always overestimated by more than an order of magnitude. Furthermore, the amount of overestimation can be quite large, since we find it to reach 3 to 4 orders of magnitude near the resonant wavenumbers.

2.3.2 Model problem 2: Laplace

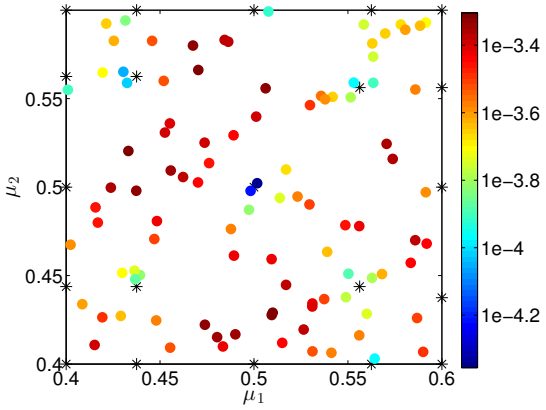
We come back to our Laplace model problem, introduced in chapter 1. We start by building a reduced basis using the Greedy algorithm 2.1. However, having not applied the SCM, we cannot cheaply evaluate any lower bound $\mu \mapsto \alpha_{LB}(\mu)$ for the inf-sup constant, thus the *a posteriori* error estimator $\mu \mapsto \frac{1}{\alpha_{LB}(\mu)} \|A(\mu)u_N(\mu) - f(\mu)\|_{W'}$ cannot be used. Furthermore, the residual norm $\|A(\mu)u_N(\mu) - f(\mu)\|_{W'}$ can not be efficiently evaluated for any value of μ, because the efficient offline/online computational strategy outlined in §2.2.2 cannot be set up, as $A(\mu)$ is non-affine. To circumvent this issue, we simply run the Greedy algorithm 2.1 setting $\Delta_N(\mu) = \|\widetilde{A}(\mu)u_N(\mu) - f(\mu)\|_{W'}$. Note that this quantity can be efficiently computed for any $\mu \in \mathcal{D}$ following the efficient offline/online computational strategy of §2.2.2 since $\widetilde{A}(\mu)$ is an affine approximation of $A(\mu)$. However, this quantity is not a rigorous *a posteriori* error estimator in the sense of definition 5. The convergence curve is shown fig. 2.4a and exhibits exponential decrease.

For the sake of validation, we compute the finite element solution $u(\mu)$ at 100 uniformly distributed points. We can thus obtain the relative error $\|u(\mu) - u_N(\mu)\|_V / \|u(\mu)\|_V$ at these 100 points, plotted on fig. 2.4b. We find the relative error to be at most around 10^{-3}. Figure 2.4b also shows the parameter points selected by the Greedy algorithm: namely, they are located mostly in the corners and on the edges of the compact parameter set $\mathcal{D} = [0.4, 0.6]^2$.

For large-scale problems, the brute-force strategy of computing those 100 finite element



(a) Maximum of the residual norm over $\Xi \subset \mathcal{D}$ per Greedy iteration.



(b) Relative error in the (μ_1, μ_2) plane, stars indicate the $N = 21$ parameters selected by the Greedy algorithm.

Figure 2.4: A RB of size $N = 21$ for the Laplace model problem.

solutions in order to obtain fig. 2.4b is time consumming and may even be computationally prohibitive. We now illustrate how the heuristic strategy presented in section 2.2.4 can be used to circumvent this. Namely, 20 residual/error ratio samples are available for free from the $N = 21$ Greedy iterations. Following our heuristic, $\mu \mapsto \frac{1}{\hat{\alpha}(\mu)} \|\widetilde{A}(\mu) u_N(\mu) - f(\mu)\|_{W'}$ should be a relevant error indicator. Combined with proposition 2.1.1, we find that the *a posteriori* quantity

$$\widetilde{\Delta_N^{\mathrm{rel}}}(\mu) = \frac{\|\widetilde{A}(\mu) u_N(\mu) - f(\mu)\|_{W'}}{\hat{\alpha}(\mu)\|u_N(\mu)\|} \left(1 - \frac{\|\widetilde{A}(\mu) u_N(\mu) - f(\mu)\|_{W'}}{\hat{\alpha}(\mu)\|u_N(\mu)\|}\right)^{-1} \tag{2.3.1}$$

should be a relevant indicator for the relative error. Note that eq. (2.3.1) is efficiently computable following the usual offline/online strategy. We provide evidence that our heuristic is good by studying the effectivity index

$$\mathrm{eff}(\mu) = \widetilde{\Delta_N^{\mathrm{rel}}}(\mu) \left(\frac{\|u(\mu) - u_N(\mu)\|_V}{\|u(\mu)\|_V}\right)^{-1}, \tag{2.3.2}$$

for the 100 values of μ uniformly distributed in $\mathcal{D}$ where we have computed the FE solution. The constant $\hat{\alpha}(\mu)$ is obtained by either: (i) the quasi-constant method $\hat{\alpha}(\mu) = \hat{\alpha}$ where $\hat{\alpha} \approx 3.56$ is the mean of the samples or (ii) by computing the radial basis function interpolant of the samples. We note that the quasi-constant applies in this case, because computing the relative standard deviation of the samples is well below 50% (we find $C_V = 13.95\%$).

We plot the effectivity distributions on fig. 2.5, in the quasi-constant case (top) and the RBF interpolant case (bottom). Note that an effectivity index of 1 means that the indicator $\widetilde{\Delta_N^{\mathrm{rel}}}(\mu)$ coincides with the relative error, an effectivity index < 1 (resp. > 1) means that the indicator underestimates (resp. overestimates) the relative error. In both cases, we

find the effectivities to be very close to 1, which confirms that our heuristic indicator is relevant. We notice a slightly less stretch distribution with a slightly better concentration of the effectivities near 1 using the RBF interpolant than using the quasi-constant method. This shows that there is a benefit of resorting to a μ-dependent constant $\hat{\alpha}(\mu)$.

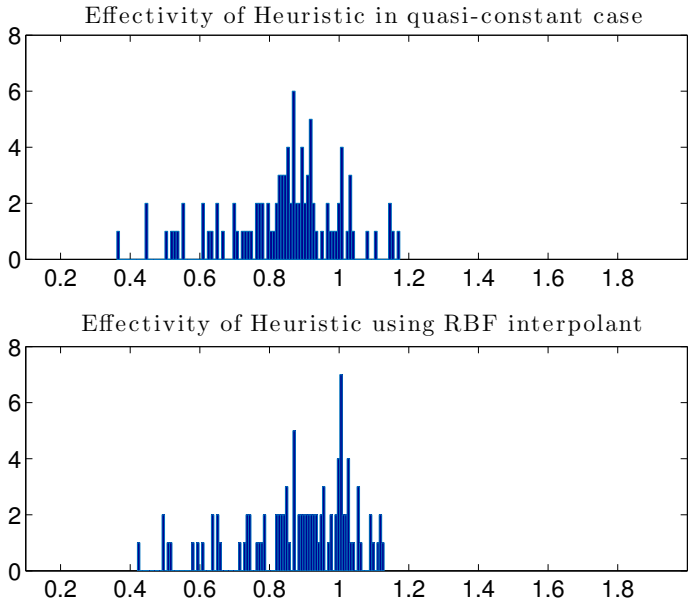


Figure 2.5: The distribution $\{\mathrm{eff}(\mu) \mid \mu \in \Xi\}$, where $\mathrm{eff}(\cdot)$ is defined by eq. (2.3.2) and $\Xi \subset \mathcal{D}$ denotes a uniformly distributed set of 100 parameter points. Top: in the quasi-constant case. Bottom: with the radial basis function interpolant $\mu \mapsto \hat{\alpha}(\mu)$.

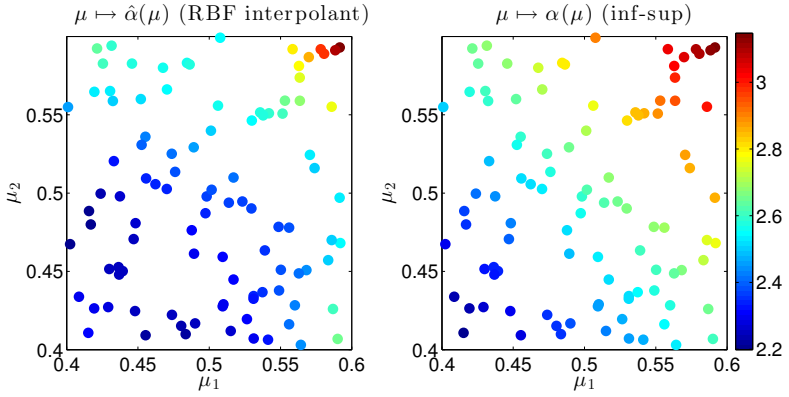


Figure 2.6: The RBF interpolant (left) and exact inf-sup constant (right) at 100 parameter points.

In order to check that our RBF interpolant $\mu \mapsto \hat{\alpha}(\mu)$ catches the true behavior of the inf-sup constant $\mu \mapsto \alpha(\mu)$, we have computed the inf-sup constant at the 100 considered parameter values using an inverse Lanczos algorithm. The comparison is plotted

on fig. 2.6. We find the RBF interpolant to be a decent approximation for the inf-sup constant, despite being extremely cheap to compute (contrary to the SCM, no large-scale generalized eigenvalue problems are solved).

2.4 Conclusions

In this chapter, we have recalled the notion of *a posteriori* error estimator. The latter can be used to assess the accuracy of a given reduced basis approximation with respect to the high-fidelity solution. Furthermore it can also be used to greedily select the parameter points at which the high-fidelity solutions should be computed in order to build an optimal Lagrange reduced basis subspace.

The classical error estimator consists in the residual norm divided by the parameter-dependent inf-sup constant. The offline/online computational strategy for efficiently computing this error estimator has been presented in detail. We have shown two methods for computing the residual norm: a default, potentially numerically unstable method and a numerically robust stabilized method. The Successive Constraints Method (SCM) constructing practical lower bounds for the parameter-dependent inf-sup constant has been reviewed. All these methods have been tested numerically on the parametrized Helmholtz model problem. Results highlight the benefits of the stabilized method for computing the residual norm. It is also found that the computational costs associated to SCM are very significant – although the number of eigensolves is limited, each eigensolve still requires solving numerous high-fidelity problems (*i.e.*, as many as the number of iterations in the eigensolver iterative process), the cost of which would be computationally prohibitive for large-scale problems.

We have proposed an original heuristic alternative to the SCM. In opposition to the SCM, the computational costs associated to our heuristic method are negligible, with no eigensolves required. The method performs well on the parametrized Laplace problem as we have been able to reconstruct a very cheap, but still relevant, approximation for the parameter-dependent inf-sup constant.

Chapter 3

The natural-norm *a posteriori* error estimation approach

Summary. In this chapter, we review the original concept of (primal) natural-norm for parametrized linear equations and introduce the concept of dual natural-norm. The natural-norms are used to derive residual-based *a posteriori* error bounds characterized by a $\mathcal{O}(1)$ stability constant. We translate these error bounds into very effective practical *a posteriori* error estimators for reduced basis approximations and show how they can be efficiently computed following an offline/online strategy. We prove that our practical dual natural-norm error estimator outperforms the classical inf-sup based error estimator in the self-adjoint case. Our findings are illustrated on anisotropic Helmholtz equations showing resonant behavior. Numerical results suggest that the proposed error estimator is able to successfully catch the correct order of magnitude of the reduced basis approximation error, thus outperforming the classical inf-sup based error estimator even for non self-adjoint problems.

Contents

3.1 Natural-norm error bounds 53
3.1.1 Reminders 53
3.1.2 Error bound using the primal natural-norm 53
3.1.3 Error bound using the dual natural-norm 55
3.1.4 Re-interpretation of the primal natural-norm approach as a right-preconditionning approach 58
3.2 Practical natural-norm *a posteriori* error estimators 59
3.2.1 Practical inf-sup based and primal natural-norm error estimators 59
3.2.2 Practical error estimator based on dual natural-norm 61
3.2.3 The self-adjoint case 63

3.3 Computational strategy 64
3.3.1 Offline/online strategy 64
3.3.2 Procedure for selecting anchor points 66
3.4 Numerical results 67
3.4.1 Problem setting 67
3.4.2 Self-adjoint case 68
3.4.3 Non self-adjoint case 72
3.5 Conclusions 75

3.1 Natural-norm error bounds

3.1.1 Reminders

This chapter is concerned with the parametrized linear equation (1.1.22), repeated here for convenience: *find* $u(\mu) \in V$ *such that*

$$A(\mu)u(\mu) = f(\mu) \text{ in } W', \tag{3.1.1}$$

where $A(\mu) \in \mathcal{L}(V, W')$ and $f(\mu) \in W'$ are the given μ-dependent linear operator and right-hand side. We recall (see section 1.1.4) that the complex Hilbert spaces V and W are finite dimensional with finite dimension $\mathcal{N} \gg 1$. We further recall the notation $R_V \in \mathcal{L}(V, V')$ for the inverse Riesz operator verifying $\|v\|^2 = \langle R_V v, v\rangle$ for all $v \in V$. The aim of this chapter is to bound the error $\|u(\mu) - u_N(\mu)\|_V$, where $u_N(\mu) \in V_N \subset V$ denotes a RB approximation of $u(\mu)$.

3.1.2 Error bound using the primal natural-norm

The original natural-norm concept has been introduced in the context of parametrized equations in Ref. [110]. In this work, we rename this original approach the *primal natural-norm* approach for reasons that will soon become obvious.

The primal natural-norm approach relies on the (primal) supremizer operator $T(\mu) \in \mathcal{L}(V, W)$, defined by

$$T(\mu) = R_W^{-1} A(\mu). \tag{3.1.2}$$

We can show that this operator satisfies the following supremizer property

$$\forall v \in V, \quad \frac{|\langle A(\mu)v, T(\mu)v\rangle|}{\|T(\mu)v\|_W} = \sup_{w \in W} \frac{|\langle A(\mu)v, w\rangle|}{\|w\|_W}. \tag{3.1.3}$$

We can use this supremizer operator to define the so-called *primal natural-norm*, which is a μ-dependent norm on V. We shall denote it $\|\|\|\cdot\|\|\|_{\mu,V}$. We define it as

$$\forall v \in V, \quad \|\|\|v\|\|\|_{\mu,V} = \|T(\mu)v\|_W, \tag{3.1.4}$$

or equivalently, using the property on Riesz maps $\|\cdot\|_{W'} = \|R_W^{-1}\cdot\|_W$,

$$\forall v \in V, \quad \|\|\|v\|\|\|_{\mu,V} = \|A(\mu)v\|_{W'} = (\langle A(\mu)v, R_W^{-1}A(\mu)v\rangle)^{1/2}. \tag{3.1.5}$$

Recalling that $A(\mu)$ is a weakly coercive operator satisfying the Banach-Nečas-Babuška assumptions, it is clear that the norm $\|\|\|\cdot\|\|\|_{\mu,V}$ is indeed a norm on V equivalent to the $\|\cdot\|_V$ norm, with equivalence constants independent from the dimension $\mathcal{N}$.

The original natural-norm concept consists in providing an error estimate not in the $\|\cdot\|_V$ norm, but rather in the $\|\|\|\cdot\|\|\|_{\mu,V}$ norm. To start with, note that for all $\mu \in \mathcal{D}$, the solution $u(\mu) \in V$ to $A(\mu)u(\mu) = f(\mu)$ satisfies, for all $\tilde{v} \in V$

$$\|\|\|u(\mu) - \tilde{v}\|\|\|_{\mu,V} = \|A(\mu)(u(\mu) - \tilde{v})\|_{W'} = \|A(\mu)\tilde{v} - f(\mu)\|_{W'}. \tag{3.1.6}$$

Thus, the primal natural-norm of the error coincides with the residual norm. In practice however, one is not satisfied with this result, because the natural-norm depends on μ. In order to circumvent this, we fix a value $\overline{\mu} \in \mathcal{D}$ and provide an error estimate in the $\|\|\|\cdot\|\|\|_{\overline{\mu},V}$ norm (which is no longer dependent on μ since $\overline{\mu}$ is fixed). In this situation, one can prove the following theorem.

Theorem 2. *Let $\overline{\mu} \in \mathcal{D}$. Define for all $\mu \in \mathcal{D}$ the primal natural-norm constant*

$$\alpha_{\overline{\mu}}(\mu) = \inf_{v \in V} \frac{\|\|\|v\|\|\|_{\mu,V}}{\|\|\|v\|\|\|_{\overline{\mu},V}}.$$

Then, for all $\mu \in \mathcal{D}$, the solution $u(\mu) \in V$ to (3.1.1) *satisfies*

$$\forall \tilde{v} \in V, \quad \|\|\|u(\mu) - \tilde{v}\|\|\|_{\overline{\mu},V} \leqslant \frac{1}{\alpha_{\overline{\mu}}(\mu)} \|A(\mu)\tilde{v} - f(\mu)\|_{W'},$$

furthermore, the inequality is an equality for $\mu = \overline{\mu}$ and $\alpha_{\overline{\mu}}(\overline{\mu}) = 1$.

Again, the question of the sharpness of this bound is raised. A significant improvement over the inf-sup-based error bound from Theorem 1 is that we now achieve equality at least when $\mu = \overline{\mu}$. The next proposition provides insight on the worst overestimation case scenario.

Proposition 3.1.1. *Let $\overline{\mu} \in \mathcal{D}$. Define for all $\mu \in \mathcal{D}$,*

$$\gamma_{\overline{\mu}}(\mu) = \sup_{v \in V} \frac{|||v|||_{\mu,V}}{|||v|||_{\overline{\mu},V}}.$$

Then, for all $\mu \in \mathcal{D}$, the solution $u(\mu) \in V$ to (3.1.1) *satisfies*

$$\forall \tilde{v} \in V, \quad 1 \leqslant \frac{\frac{1}{\alpha_{\overline{\mu}}(\mu)} \|A(\mu)\tilde{v} - f(\mu)\|_{W'}}{|||u(\mu) - \tilde{v}|||_{\overline{\mu},V}} \leqslant \frac{\gamma_{\overline{\mu}}(\mu)}{\alpha_{\overline{\mu}}(\mu)},$$

furthermore, $\gamma_{\overline{\mu}}(\overline{\mu}) = \alpha_{\overline{\mu}}(\overline{\mu}) = 1$.

This proposition reveals the potential benefits of the primal natural-norm approach. Indeed, under basic regularity assumptions, the quantity $\gamma_{\overline{\mu}}(\mu)/\alpha_{\overline{\mu}}(\mu)$, being equal to 1 for $\mu = \overline{\mu}$, will continue to be $\mathcal{O}(1)$ for values of μ in a neighborhood of $\overline{\mu}$. Thus, the amount of overestimation of the primal natural-norm error bound provided by Theorem 1 will be $\mathcal{O}(1)$ for values of μ adequately close to $\overline{\mu}$. Since $\overline{\mu}$ is a fixed value, chosen by the user, one can always consider a family of points $\{\overline{\mu}^k\}_{1 \leqslant k \leqslant K}$ and thus be able to estimate the correct order of magnitude of the error for any $\mu \in \mathcal{D}$. Notice however that the primal natural-norm approach still suffers from the problem of resonances.

Remark that the primal natural-norm error bound given by Theorem 2 does not bound the error in the norm of our choice. Indeed, the theorem bounds the error in the primal natural-norm $|||\cdot|||_{\overline{\mu},V}$ but not in the user-defined $\|\cdot\|_V$ norm. The dual natural-norm error bound (see next section) circumvents this; since it bounds the error in the $\|\cdot\|_V$ norm while maintaining a $\mathcal{O}(1)$ stability constant.

3.1.3 Error bound using the dual natural-norm

In order to go beyond the primal natural-norm approach, we need to introduce the adjoint operator $A(\mu)^* \in \mathcal{L}(W, V')$, defined by ${}_{W'}\langle A(\mu)v, w\rangle_W = {}_{V'}\langle A(\mu)^* w, v\rangle_V$ for all $v \in V$ and for all $w \in W$. We recall that the adjoint operator (or conjugate-transposed operator) is automatically continuous since it satisfies $\|A(\mu)^* w\|_{V'} \leqslant \gamma(\mu)\|w\|_W$ for all $w \in W$, where $\gamma(\mu)$ is the continuity constant of $A(\mu)$ defined by Eq. (1.1.24b). Furthermore, the adjoint is a weakly coercive operator, because $A(\mu)$ is a weakly coercive operator. Thus the following stability condition is satisfied

$$\beta(\mu) = \inf_{w \in W} \sup_{v \in V} \frac{|\langle A(\mu)^* w, v\rangle|}{\|v\|_V \|w\|_W} > 0. \tag{3.1.7}$$

Notice that $\beta(\mu) = \alpha(\mu)$. However, we shall distinguish the stability constant $\alpha(\mu)$ of $A(\mu)$ and the stability constant $\beta(\mu)$ of $A(\mu)^*$ for clarity.

The dual natural norm

As in the case of the primal, we now introduce the dual supremizer operator $\Upsilon(\mu) \in \mathcal{L}(W, V)$, defined by

$$\Upsilon(\mu) = R_V^{-1} A(\mu)^*. \tag{3.1.8}$$

We can show that this operator satisfies the following supremizer property

$$\forall w \in W, \quad \frac{|\langle A(\mu)^* w, \Upsilon(\mu) w\rangle|}{\|\Upsilon(\mu) w\|_W} = \sup_{v \in V} \frac{|\langle A(\mu)^* w, v\rangle|}{\|v\|_V}.$$

We now use the dual supremizer operator to define a natural-norm on W, which shall be denoted $|||\cdot|||_{\mu,W}$. We define it as

$$\forall w \in W, \quad |||w|||_{\mu,W} = \|\Upsilon(\mu) w\|_V, \tag{3.1.9}$$

or equivalently, using the property on Riesz maps $\|\cdot\|_{V'} = \|R_V^{-1} \cdot\|_V$,

$$\forall w \in W, \quad |||w|||_{\mu,W} = \|A(\mu)^* w\|_{V'} = \left(\langle A(\mu)^* w, R_V^{-1} A(\mu)^* w\rangle\right)^{1/2}. \tag{3.1.10}$$

It is clear form the weak coercivity of $A(\mu)^*$ that $|||\cdot|||_{\mu,W}$ is indeed a norm on W equivalent to the $\|\cdot\|_W$ norm, with equivalence constants independent from the dimension $\mathcal{N}$.

With this natural-norm on W, we can define a natural-norm on the dual W'. The latter will be called the *dual natural-norm* and be denoted $|||\cdot|||_{\mu,W'}$. Namely, it is quite classically defined as

$$\forall \ell \in W', \quad |||\ell|||_{\mu,W'} = \sup_{w \in W} \frac{|\langle \ell, w\rangle|}{|||w|||_{\mu,W}}. \tag{3.1.11}$$

Thus defined, it is clear that $|||\cdot|||_{\mu,W'}$ is an equivalent norm to $\|\cdot\|_{W'}$. The following proposition gives us a more convenient formula for expressing the dual natural-norm norm.

Proposition 3.1.2. *The $|||\cdot|||_{\mu,W'}$ norm defined by* (3.1.11) *is equivalently*

$$\forall \ell \in W', \quad |||\ell|||_{\mu,W'} = \|A(\mu)^{-1}\ell\|_V = \left(\langle R_V A(\mu)^{-1}\ell, A(\mu)^{-1}\ell\rangle\right)^{1/2}.$$

Proof. Let $\ell \in W'$ and $w \in W$. By the adjoint property, we have

$$\langle \ell, w\rangle = \langle \ell, A(\mu)^{-*} A(\mu)^* w\rangle = \langle A(\mu)^{-1}\ell, A(\mu)^* w\rangle.$$

Thus,

$$\begin{aligned} \forall \ell \in W', \quad |||\ell|||_{\mu,W'} = \sup_{w \in W} \frac{|\langle \ell, w\rangle|}{|||w|||_{\mu,W}} &= \sup_{w \in W} \frac{\langle A(\mu)^{-1}\ell, A(\mu)^* w\rangle}{\|A(\mu)^* w\|_{V'}} \\ &= \sup_{\phi \in V'} \frac{\langle A(\mu)^{-1}\ell, \phi\rangle}{\|\phi\|_{V'}} \\ &= \|A(\mu)^{-1}\ell\|_V. \end{aligned} \tag{3.1.12}$$

□

We can show the following norm equivalence, leaving the proof to the reader.

Proposition 3.1.3. *For all $\ell \in W'$, there holds,*

$$\beta(\mu)|||\ell|||_{\mu,W'} \leqslant \|\ell\|_{W'} \leqslant \gamma(\mu)|||\ell|||_{\mu,W'}.$$

Error bound

We now arrive to our ultimate goal of deriving error estimates using the dual natural norm. To start with, note that the error norm is exactly the dual natural-norm of the residual. Indeed, for all $\mu \in \mathcal{D}$, the solution $u(\mu) \in V$ to (3.1.1) satisfies, for all $\tilde{v} \in V$,

$$\|u(\mu) - \tilde{v}\|_V = |||A(\mu)\tilde{v} - f(\mu)|||_{\mu,W'}. \tag{3.1.13}$$

Notice the symmetry with Eq. (3.1.6), repeated here for convenience,

$$|||u(\mu) - \tilde{v}|||_{\mu,V} = \|A(\mu)\tilde{v} - f\|_{W'}. \tag{3.1.14}$$

All is now set to derive the error estimate using the dual natural-norm.

Theorem 3. *Let $\overline{\mu} \in \mathcal{D}$. For all $\mu \in \mathcal{D}$, define*

$$\sigma_{\overline{\mu}}(\mu) = \inf_{v\in V} \frac{|||A(\mu)v|||_{\overline{\mu},W'}}{\|v\|_V} \left(= \inf_{v\in V} \frac{\|A(\overline{\mu})^{-1}A(\mu)v\|_V}{\|v\|_V} \right).$$

Then, for all $\mu \in \mathcal{D}$ the solution $u(\mu) \in V$ to (3.1.1) *satisfies*

$$\forall \tilde{v} \in V, \quad \|u(\mu) - \tilde{v}\|_V \leqslant \frac{1}{\sigma_{\overline{\mu}}(\mu)} |||A(\mu)\tilde{v} - f(\mu)|||_{\overline{\mu},W'}.$$

Furthermore, the inequality is an equality for $\mu = \overline{\mu}$ and $\sigma_{\overline{\mu}}(\overline{\mu}) = 1$.

Proof. Start by

$$\begin{aligned} |||A(\mu)\tilde{v} - f(\mu)|||_{\overline{\mu}} &= \|A(\overline{\mu})^{-1}(A(\mu)\tilde{v} - f(\mu))\|_V \\ &= \|A(\overline{\mu})^{-1}A(\mu)(\tilde{v} - u(\mu))\|_V \\ &\geqslant \left(\inf_{v\in V} \frac{\|A(\overline{\mu})^{-1}A(\mu)v\|_V}{\|v\|_V} \right) \|u(\mu) - \tilde{v}\|_V. \end{aligned}$$

It remains to justify that $\sigma_{\overline{\mu}}(\mu) = \inf_{v\in V} \frac{\|A(\overline{\mu})^{-1}A(\mu)v\|_V}{\|v\|_V}$ is indeed > 0. For this, we bound from below by using the norm equivalence of Proposition 3.1.3,

$$\sigma_{\overline{\mu}}(\mu) = \inf_{v\in V} \frac{\|A(\overline{\mu})^{-1}A(\mu)v\|_V}{\|v\|_V} \geqslant \frac{1}{\gamma(\overline{\mu})} \inf_{v\in V} \frac{\|A(\mu)v\|_{W'}}{\|v\|_V} = \frac{\alpha(\mu)}{\gamma(\overline{\mu})}.$$

This lower bound is > 0 because the inf-sup constant $\alpha(\mu)$ is strictly positive.

To prove that the inequality is an equality when $\mu = \overline{\mu}$, we simply observe that $\sigma_{\overline{\mu}}(\overline{\mu}) = 1$ and come back to Eq. (3.1.13). □

Let us now give a result on the potential sharpness of this error bound.

Proposition 3.1.4. *Let $\overline{\mu} \in \mathcal{D}$. For all $\mu \in \mathcal{D}$, define*

$$\Sigma_{\overline{\mu}}(\mu) = \sup_{v \in V} \frac{|||A(\mu)v|||_{\overline{\mu},W'}}{\|v\|_V} \left(= \sup_{v \in V} \frac{\|A(\overline{\mu})^{-1}A(\mu)v\|_V}{\|v\|_V} \right).$$

Then, for all $\mu \in \mathcal{D}$ the solution $u(\mu) \in V$ to (3.1.1) *satisfies*

$$\forall \tilde{v} \in V, \quad 1 \leqslant \frac{\frac{1}{\sigma_{\overline{\mu}}(\mu)} |||A(\mu)\tilde{v} - f(\mu)|||_{\overline{\mu},W'}}{\|u(\mu) - \tilde{v}\|_V} \leqslant \frac{\Sigma_{\overline{\mu}}(\mu)}{\sigma_{\overline{\mu}}(\mu)},$$

furthermore $\Sigma_{\overline{\mu}}(\overline{\mu}) = \sigma_{\overline{\mu}}(\overline{\mu}) = 1$

Let us comment on the upper bound $\Sigma_{\overline{\mu}}(\mu)/\sigma_{\overline{\mu}}(\mu)$. This ratio can be interpreted as a *condition number*. It is $\mathcal{O}(1)$ for values of μ such that $A(\overline{\mu})^{-1}$ is a good left preconditioner for $A(\mu)$. With this understanding, the fact that both inequalities are equalities when $\mu = \overline{\mu}$ is due to the fact that $A(\overline{\mu})^{-1}$ is the ideal left preconditioner for $A(\overline{\mu})$. The sharpness of the bound provided by Theorem 3 is thus intimately linked to the properties of $A(\overline{\mu})^{-1}$ as left preconditioner for $A(\mu)$. In this sense, the dual natural-norm approach is a left preconditioning approach. Compared to existing left-preconditioning approaches in the reduced basis context[129, 4], here the preconditioner is parameter-independent.

3.1.4 Re-interpretation of the primal natural-norm approach as a right-preconditionning approach

We now re-interpret the primal natural-norm approach reviewed in Sec. 3.1.2 as a right preconditioning approach. This shows the symmetry between the primal and dual natural-norm approaches. Notice that the arguments used in this section strongly rely on the fact that the Hilbert space V, W (and topological duals V', W') are finite dimensional.

Proposition 3.1.5. *Let $\overline{\mu} \in \mathcal{D}$. For all $\mu \in \mathcal{D}$, the primal natural norm constant $\alpha_{\overline{\mu}}(\mu)$ (defined in Theorem 2) can be equivalently defined as*

$$\alpha_{\overline{\mu}}(\mu) = \inf_{\ell_w \in W'} \frac{\|A(\mu)A(\overline{\mu})^{-1}\ell_w\|_{W'}}{\|\ell_w\|_{W'}},$$

and the constant $\gamma_{\overline{\mu}}(\mu)$ (defined in Proposition 3.1.1) can be equivalently defined as

$$\gamma_{\overline{\mu}}(\mu) = \sup_{\ell_w \in W'} \frac{\|A(\mu)A(\overline{\mu})^{-1}\ell_w\|_{W'}}{\|\ell_w\|_{W'}}.$$

Proof. Let $\ell_w \in W'$. Then there exists a unique solution $v \in V$ to the problem $A(\overline{\mu})v = \ell_w$. Thus,

$$\inf_{\ell_w \in W'} \frac{\|A(\mu)A(\overline{\mu})^{-1}\ell_w\|_{W'}}{\|\ell_w\|_{W'}} = \inf_{v \in V} \frac{\|A(\mu)v\|_{W'}}{\|A(\overline{\mu})v\|_{W'}} = \inf_{v \in V} \frac{|||v|||_{\mu,V}}{|||v|||_{\overline{\mu},V}} = \alpha_{\overline{\mu}}(\mu).$$

We proceed analogously for $\gamma_{\overline{\mu}}(\mu)$, taking the supremum rather than infimum. □

In the light of this Proposition, we can now re-interpret the ratio $\gamma_{\overline{\mu}}(\mu)/\alpha_{\overline{\mu}}(\mu)$ of Proposition 3.1.1 as the condition number from preconditioning $A(\mu)$ to the right using $A(\overline{\mu})^{-1}$ as preconditioner.

3.2 Practical natural-norm *a posteriori* error estimators

We now explain how the error bounds can be translated into practical *a posteriori* error estimators. The first concern is the derivation of practical (*i.e.*, computable) lower bounds for the μ-dependent stability constants (inf-sup and natural-norm constants) which represent a computational bottleneck.

The second concern is that error bounds based on the concept of natural-norm are only expected to be sharp locally in the neighborhood of a so-called fixed *anchor point* $\overline{\mu} \in \mathcal{D}$. Therefore, in order to estimate the error globally over $\mathcal{D}$, one must consider K local natural-norms based on a discrete set of K anchor points $\mathcal{C}^K = \{\overline{\mu}^1, \ldots, \overline{\mu}^K\} \subset \mathcal{D}$ and an indicator function $\mathcal{I}^K : \mathcal{D} \to \mathcal{C}^K$ that maps each μ a unique "best" anchor point $\overline{\mu} \in \mathcal{C}^K$ in a sense that shall be defined shortly.

3.2.1 Practical inf-sup based and primal natural-norm error estimators

Practical lower bounds for the inf-sup constant $\mu \in \mathcal{D} \mapsto \alpha(\mu)$ can be directly obtained using the Successive Constraints Method (SCM); *e.g.* see Refs. [61, 27, 112]. A variant –

known as the natural-norm SCM; see Refs. [60, 26] – builds lower bounds for the inf-sup constant based on the following result, first shown in Ref. [110].

Proposition 3.2.1. *Let $\overline{\mu} \in \mathcal{D}$. Then for all $\mu \in \mathcal{D}$ the inf-sup constant $\alpha(\mu)$ (defined in* (1.1.24a)*) can be bounded from below as*

$$\alpha(\mu) \geqslant \alpha(\overline{\mu})\alpha_{\overline{\mu}}(\mu),$$

where $\alpha_{\overline{\mu}}(\mu)$ is the primal natural-norm constant (defined in Theorem 2).

Proof. This is clear from

$$\begin{aligned}\alpha(\mu) = \inf_{v\in V} \frac{|||v|||_{\mu,V}}{\|v\|_V} &= \inf_{v\in V} \frac{|||v|||_{\overline{\mu},V}}{\|v\|_V} \frac{|||v|||_{\mu,V}}{|||v|||_{\overline{\mu},V}} \\ &\geqslant \left(\inf_{v\in V} \frac{|||v|||_{\overline{\mu},V}}{\|v\|_V}\right)\left(\inf_{v\in V} \frac{|||v|||_{\mu,V}}{|||v|||_{\overline{\mu},V}}\right) = \alpha(\overline{\mu})\alpha_{\overline{\mu}}(\mu).\end{aligned}$$

□

In fact, the practical interest of this Proposition is very limited, because in practice the primal natural-norm constant is about as difficult to compute (or to approximate using SCM) as the inf-sup constant. The true interest of Proposition 3.2.1 is to replace the primal natural-norm constant by a more practical lower bound, provided by the following Proposition.

Proposition 3.2.2. *Let $\overline{\mu} \in \mathcal{D}$. For all $\mu \in \mathcal{D}$ define*

$$\overline{\alpha}_{\overline{\mu}}(\mu) = \inf_{v\in V} \frac{\Re\{\langle A(\mu)v, R_W^{-1}A(\overline{\mu})v\rangle\}}{|||v|||^2_{\overline{\mu},V}}.$$

Then, for all $\mu \in \mathcal{D}$, $\alpha_{\overline{\mu}}(\mu) \geqslant \overline{\alpha}_{\overline{\mu}}(\mu)$. Furthermore $\overline{\alpha}_{\overline{\mu}}(\overline{\mu}) = 1$.

Proof. Recalling that $|||v|||_{\mu,V} = \|A(\mu)v\|_{W'} = \sup_{w\in W} \frac{|\langle A(\mu)v,w\rangle|}{\|w\|_W}$ we get

$$\alpha_{\overline{\mu}}(\mu) = \inf_{v\in V} \frac{|||v|||_{\mu,V}}{|||v|||_{\overline{\mu},V}} = \inf_{v\in V} \sup_{w\in W} \frac{|\langle A(\mu)v, w\rangle|}{|||v|||_{\overline{\mu},V}\|w\|_W}.$$

We may choose the candidate supremizer $w = R_W^{-1}A(\overline{\mu})v$, yielding

$$\alpha_{\overline{\mu}}(\mu) \geqslant \inf_{v\in V} \frac{|\langle A(\mu)v, R_W^{-1}A(\overline{\mu})v\rangle|}{|||v|||_{\overline{\mu},V}\|R_W^{-1}A(\overline{\mu})v\|_W} = \inf_{v\in V} \frac{|\langle A(\mu)v, R_W^{-1}A(\overline{\mu})v\rangle|}{|||v|||^2_{\overline{\mu},V}} \geqslant \overline{\alpha}_{\overline{\mu}}(\mu),$$

where the last inequality simply stems from the fact that the modulus of a complex number is always greater than its real part. □

Remark. *Proposition 3.2.2 would provide a sharper lower bound had the real part been replaced by the modulus in the definition of the constant $\overline{\alpha}_{\overline{\mu}}(\mu)$. We choose to consider the real part and not the modulus – thus accepting a less sharp lower bound – for purely practical reasons. Namely, $\overline{\alpha}_{\overline{\mu}}(\mu)$ can be computed as the smallest eigenvalue in the generalized hermitian eigenvalue problem: find $(\lambda, w) \in \mathbb{R} \times W$ such that $\frac{1}{2}\left(A(\mu)A(\overline{\mu})^{-1}R_W + R_W A(\overline{\mu})^* A(\mu)^*\right) w = \lambda R_W w$ in W. Justification for the form of this generalized hermitian eigenvalue problem will be provided in the proof of Proposition 3.2.4.*

As shown in Ref. [110], under some regularity assumption on $\mu \mapsto A(\mu)$, the lower bound of Proposition 3.2.2 is second order accurate; in the sense

$$\alpha_{\overline{\mu}}(\mu) = \overline{\alpha}_{\overline{\mu}}(\mu) + \mathcal{O}(|\mu - \overline{\mu}|^2) \quad \text{as } \mu \to \overline{\mu}. \tag{3.2.1}$$

Combining the results from Proposition 3.2.2 and Proposition 3.2.1, we obtain the practical lower bound for the inf-sup constant: $\alpha(\mu) \geqslant \alpha(\overline{\mu})\overline{\alpha}_{\overline{\mu}}(\mu)$. Notice however that $\overline{\alpha}_{\overline{\mu}}(\mu)$ is not guaranteed to be positive (in opposition to $\alpha_{\overline{\mu}}(\mu)$, which is always > 0). It can typically turn negative when μ is "too distant" from $\overline{\mu}$ in some sense. When this is the case, our practical lower bound for the inf-sup constant becomes of no interest. For this reason, we define the *primal positivity coverage set* associated to any anchor point $\overline{\mu}$, as $\mathcal{D}^{\mathrm{pr}}_{\overline{\mu}} = \{\mu \in \mathcal{D}, \overline{\alpha}_{\overline{\mu}}(\mu) > 0\}$. In this context, a "good" set of anchor points $\mathcal{C}^K = \{\overline{\mu}^1, \ldots, \overline{\mu}^K\}$ is such that $\cup_{k=1}^K \mathcal{D}^{\mathrm{pr}}_{\overline{\mu}^k} = \mathcal{D}$ and a possible associated indicator function $\mathcal{I}^K$ maps each $\mu \in \mathcal{D}$ to the anchor point $\overline{\mu} \in \mathcal{C}^K$ such that the constant $\overline{\alpha}_{\overline{\mu}}(\mu)$ is largest, *i.e.*, $\mathcal{I}^K(\mu) = \underset{1 \leqslant k \leqslant K}{\operatorname{argmax}} \overline{\alpha}_{\overline{\mu}^k}(\mu)$.

All is now set to define our practical primal inf-sup based *a posteriori* error estimator

$$\forall \tilde{v} \in V, \quad \Delta^{\mathrm{pr}}_K(\tilde{v}; \mu) = \frac{1}{\alpha(\overline{\mu})\overline{\alpha}_{\overline{\mu}}(\mu)} \|A(\mu)\tilde{v} - f(\mu)\|_{W'}, \quad \text{with} \quad \overline{\mu} = \mathcal{I}^K(\mu). \tag{3.2.2}$$

3.2.2 Practical error estimator based on dual natural-norm

In the same fashion, we now construct a practical error estimator based on the the dual natural norm.

Proposition 3.2.3. *Let $\overline{\mu} \in \mathcal{D}$. For all $\mu \in \mathcal{D}$, define*

$$\overline{\sigma}_{\overline{\mu}}(\mu) = \inf_{v \in V} \frac{\Re\{\langle A(\mu)v, A(\overline{\mu})^{-*} R_V v \rangle\}}{\|v\|_V^2}.$$

Then for all $\mu \in \mathcal{D}$, $\overline{\sigma}_{\overline{\mu}}(\mu) \leqslant \sigma_{\overline{\mu}}(\mu)$. Furthermore $\overline{\sigma}_{\overline{\mu}}(\overline{\mu}) = 1$ and assuming $\mu \mapsto A(\mu)$ is differentiable in the neighborhood of $\overline{\mu}$ there holds

$$\sigma_{\overline{\mu}}(\mu) = \overline{\sigma}_{\overline{\mu}}(\mu) + \mathcal{O}(|\mu - \overline{\mu}|^2) \quad \text{as } \mu \to \overline{\mu}. \tag{3.2.3}$$

Proof. We start by the definition

$$\sigma_{\overline{\mu}}(\mu) = \inf_{v\in V} \frac{\|A(\overline{\mu})^{-1}A(\mu)v\|_V}{\|v\|_V} = \inf_{v\in V}\sup_{\ell\in V'} \frac{|\langle \ell, A(\overline{\mu})^{-1}A(\mu)v\rangle|}{\|v\|_V\|\ell\|_{V'}}. \tag{3.2.4}$$

Choose the candidate supremizer $\ell = R_V v$ and use the fact that the modulus of a complex number is always an upper bound for its real part.

In order to demontrate the second order accuracy, we can refer to same arguments in the case of the primal, see Ref. [110]. We repeat the essential steps for completeness. Let us start from the definition of $\sigma_{\overline{\mu}}(\mu)$ and proceed as follows

$$(\sigma_{\overline{\mu}}(\mu))^2 = \inf_{v\in V} \frac{\|A(\overline{\mu})^{-1}A(\mu)v\|_V^2}{\|v\|_V^2} = \inf_{v\in V} \frac{\|v + A(\overline{\mu})^{-1}(A(\mu)-A(\overline{\mu}))v\|_V^2}{\|v\|_V^2}. \tag{3.2.5}$$

If $\mu \mapsto A(\mu)$ is differentiable in the neighborhood of $\overline{\mu}$, then $\|A(\overline{\mu})^{-1}(A(\mu)-A(\overline{\mu}))v\|_V^2/\|v\|_V^2 = \mathcal{O}(|\mu-\overline{\mu}|^2)$. In this situation, developing the square in Eq. (3.2.5) yields

$$\begin{aligned}(\sigma_{\overline{\mu}}(\mu))^2 &= 1 + \inf_{v\in V} \frac{2\Re\{\langle R_V A(\overline{\mu})^{-1}(A(\mu)-A(\overline{\mu}))v, v\rangle\}}{\|v\|_V^2} + \mathcal{O}(|\mu-\overline{\mu}|^2)\\ &= 1 + 2\left(\inf_{v\in V} \frac{\Re\{\langle R_V A(\overline{\mu})^{-1}A(\mu)v, v\rangle\}}{\|v\|_V^2} - 1\right) + \mathcal{O}(|\mu-\overline{\mu}|^2).\end{aligned} \tag{3.2.6}$$

Thus, we have $\sigma_{\overline{\mu}}(\mu) = (1 + 2(\overline{\sigma}_{\overline{\mu}}(\mu) - 1) + \mathcal{O}(|\mu-\overline{\mu}|^2))^{1/2}$. We conclude by invoking the the formula $(1+t)^{1/2} = 1 + \frac{1}{2}t + \mathcal{O}(t^2)$ for $t = 2(\overline{\sigma}_{\overline{\mu}}(\mu) - 1)$, combined with the fact that $(\overline{\sigma}_{\overline{\mu}}(\mu) - 1) = \mathcal{O}(|\mu-\overline{\mu}|)$.

□

Notice that we have defined the lower bound $\overline{\sigma}_{\overline{\mu}}(\mu)$ of $\sigma_{\overline{\mu}}(\mu)$ using a real part and not a module, again for purely practical reasons (see Remark 3.2.1). Moreover, similar to its primal counterpart $\overline{\alpha}_{\overline{\mu}}(\mu)$, the constant $\overline{\sigma}_{\overline{\mu}}(\mu)$ is not guaranteed to be positive, so we must introduce the *dual positivity coverage set* associated to a given anchor point $\overline{\mu} \in \mathcal{C}^K$ as $\mathcal{D}_{\overline{\mu}}^{\mathrm{du}} = \{\mu \in \mathcal{D}, \overline{\sigma}_{\overline{\mu}}(\mu) > 0\}$. Again, a "good" set of anchor points $\mathcal{C}^K = \{\overline{\mu}^1, \ldots, \overline{\mu}^K\}$ is such that $\cup_{k=1}^K \mathcal{D}_{\overline{\mu}^k}^{\mathrm{du}} = \mathcal{D}$ and a possible associated indicator function $\mathcal{I}^K$ maps each $\mu \in \mathcal{D}$ to the anchor point $\overline{\mu} \in \mathcal{C}^K$ such that the constant $\overline{\sigma}_{\overline{\mu}}(\mu)$ is largest, *i.e.*, $\mathcal{I}^K(\mu) = \underset{1\leqslant k\leqslant K}{\operatorname{argmax}}\, \overline{\sigma}_{\overline{\mu}^k}(\mu)$.

All is now set to define our practical dual natural-norm based *a posteriori* error estimator

$$\forall \tilde{v} \in V, \quad \Delta_K^{\mathrm{du}}(\tilde{v};\mu) = \frac{1}{\overline{\sigma}_{\overline{\mu}}(\mu)} |||A(\mu)\tilde{v} - f(\mu)|||_{\mu,W'} \quad \text{with} \quad \overline{\mu} = \mathcal{I}^K(\mu). \tag{3.2.7}$$

3.2.3 The self-adjoint case

Proposition 3.2.4. *In the self-adjoint case, (i.e., $V = W$ and $A(\mu) = A(\mu)^*$), there holds,*

$$\forall \mu \in \mathcal{D}, \quad \overline{\alpha}_{\overline{\mu}}(\mu) = \overline{\sigma}_{\overline{\mu}}(\mu),$$

where $\overline{\alpha}_{\overline{\mu}}(\mu)$ is defined in Proposition 3.2.2 and $\overline{\sigma}_{\overline{\mu}}(\mu)$ is defined in Proposition 3.2.3.

Proof. This is straightforward observing that $\overline{\alpha}_{\overline{\mu}}(\mu)$ is equivalently given by

$$\overline{\alpha}_{\overline{\mu}}(\mu) = \inf_{w \in W} \frac{\Re\{\langle A(\mu)A(\overline{\mu})^{-1}R_W w, w\rangle\}}{\|w\|_W^2}.$$

□

Thanks to the preliminary result given by Proposition 3.2.4, we can show that the dual natural-norm error estimator necessarily outperforms the inf-sup based error estimator in the self-adjoint case.

Theorem 4. *Let $\mathcal{C}^K = \{\overline{\mu}^1, \dots, \overline{\mu}^K\} \subset \mathcal{D}$ and consider the self-adjoint case. Denote*

$$\mathcal{D}^+ = \bigcup_{k=1}^{K} \mathcal{D}_{\overline{\mu}^k}^{\mathrm{pr}} \left(= \bigcup_{k=1}^{K} \mathcal{D}_{\overline{\mu}^k}^{\mathrm{du}} \right).$$

Then, for all $\mu \in \mathcal{D}^+$, the solution $u(\mu) \in V$ to (3.1.1) *satisfies*

$$\forall \tilde{v} \in V, \quad \|u(\mu) - \tilde{v}\|_V \leqslant \Delta_K^{\mathrm{du}}(\tilde{v}; \mu) \leqslant \Delta_K^{\mathrm{pr}}(\tilde{v}; \mu),$$

with Δ_K^{du} the dual natural-norm error estimator defined by (3.2.7) *and Δ_K^{pr} the inf-sup based error estimator defined by* (3.2.2)*.*

Proof. Let $\mu \in \mathcal{D}$ and denote $\overline{\mu} = \mathcal{I}^K(\mu)$ the associated anchor point. Theorem 3 states that

$$\|u(\mu) - \tilde{v}\|_V \leqslant \frac{1}{\sigma_{\overline{\mu}}(\mu)} \|A(\overline{\mu})^{-1}(A(\mu)\tilde{v} - f(\mu))\|_V.$$

From the inequality $\frac{1}{\sigma_{\overline{\mu}}(\mu)} \leqslant \frac{1}{\overline{\sigma}_{\overline{\mu}}(\mu)}$, which stems from Proposition 3.2.3 combined to the fact that $\overline{\sigma}_{\overline{\mu}}(\mu) > 0$, we obtain the first inequality announced in the theorem.

The second inequality is a consequence of the equivalence of the $\|A(\overline{\mu})^{-1} \cdot\|_V$ norm and the $\|\cdot\|_{W'}$ norm, established in Proposition 3.1.3. Namely,

$$\|A(\overline{\mu})^{-1}(A(\mu)\tilde{v} - f(\mu))\|_V \leqslant \frac{1}{\beta(\overline{\mu})} \|A(\mu)\tilde{v} - f(\mu)\|_{W'}.$$

We have $\beta(\overline{\mu}) = \alpha(\overline{\mu})$ from the self-adjoint hypothesis. □

3.3 Computational strategy

3.3.1 Offline/online strategy

In this section, we consider a Reduced Basis (RB) approximation space $\widetilde{V_N} \subset V$, with small dimension $N \ll \mathcal{N}$ and a RB approximation $\widetilde{u_N}(\mu)$ in this N-dimensional RB subspace. The resulting RB approximation error can be bounded using our dual natural-norm *a posteriori* error estimator (3.2.7) as

$$\|u(\mu) - \widetilde{u_N}(\mu)\|_V \leqslant \Delta_K^{\mathrm{du}}(\widetilde{u_N}(\mu); \mu). \tag{3.3.1}$$

We now explain how $\Delta_K^{\mathrm{du}}(\widetilde{u_N}(\mu); \mu)$ can be efficiently computed. There are two components in our error estimator: a stability constant (namely, $\overline{\sigma}_{\overline{\mu}}(\mu)$) which we propose to replace by a cheap SCM lower bound, and the dual natural-norm of the residual which can be efficiently computed following an offline/online strategy.

The affine hypothesis

Following the Reduced Basis Method standard [54, 97] we require that the operator $A(\mu) \in \mathcal{L}(V, W')$ is affinely parametrized, that is, that there exist $Q \geqslant 1$ μ-independent operators $A_q \in \mathcal{L}(V, W')$, $1 \leqslant q \leqslant Q$ and complex valued functions $\varsigma_q : \mathcal{D} \to \mathbb{C}$, $1 \leqslant q \leqslant Q$ such that

$$\forall \mu \in \mathcal{D}, \quad A(\mu) = \sum_{q=1}^{Q} \varsigma_q(\mu) A_q. \tag{3.3.2}$$

Similarly, we require the right-hand-side $f(\mu) \in W'$ to be affinely parametrized, that is, that there exist $Q^f \geqslant 1$ μ-independent linear forms $f_q \in W'$, $1 \leqslant q \leqslant Q^f$ and complex valued functions $\varsigma_q^f : \mathcal{D} \to \mathbb{C}$, $1 \leqslant q \leqslant Q^f$ such that

$$\forall \mu \in \mathcal{D}, \quad f(\mu) = \sum_{q=1}^{Q^f} \varsigma_q^f(\mu) f_q. \tag{3.3.3}$$

Note that if the operator or right-hand side do not satisfy the affine assumption, the Empirical Interpolation Method (EIM) can be employed to recover affinely parametrized approximations [5, 84].

Computing a lower bound for the dual natural norm constant

Let $\overline{\mu} \in \mathcal{D}$ be a given anchor point. We now explain how a lower bound for $\overline{\sigma}_{\overline{\mu}}(\mu)$ can be efficiently computed for any μ query. This is key to the success of the proposed

method; otherwise our practical dual natural-norm error estimator would not be efficiently computable. By definition, $\overline{\sigma}_{\overline{\mu}}(\mu)$ is the smallest eigenvalue in the generalized hermitian eigenvalue problem: *find* $(\lambda, v) \in \mathbb{R} \times V$ *such that*

$$H_{\overline{\mu}}(\mu)v = \lambda R_V v \quad \text{in } V' \tag{3.3.4}$$

with $H_{\overline{\mu}}(\mu) \in \mathcal{L}(V, V')$ given by

$$\forall \mu \in \mathcal{D}, \quad H_{\overline{\mu}}(\mu) = \frac{1}{2}\left(A(\mu)^* A(\overline{\mu})^{-*} R_V + R_V A(\overline{\mu})^{-1} A(\mu)\right). \tag{3.3.5}$$

Note that $H_{\overline{\mu}}(\mu)$ is self-adjoint, but that it is not necessarily positive definite. An approach based on solving the generalized eigenvalue problem (3.3.4) for each μ query would lead to prohibitive computational costs. We propose to use the SCM in order to compute cheap lower and upper bounds for $\mu \mapsto \overline{\sigma}_{\overline{\mu}}(\mu)$, using a computationally efficient offline/online strategy. Clearly, using the affine representation (3.3.2) of $A(\mu)$, the $H_{\overline{\mu}}(\mu)$ operator admits the following affine representation

$$\begin{aligned} H_{\overline{\mu}}(\mu) = \sum_{q=1}^{Q} \Re\{\varsigma_q(\mu)\} \frac{1}{2}\left(A_q^* A(\overline{\mu})^{-*} R_V + R_V A(\overline{\mu})^{-1} A_q\right) \\ + \sum_{q=1}^{Q} \Im\{\varsigma_q(\mu)\} \frac{1}{2}\left(i A_q^* A(\overline{\mu})^{-*} R_V - i R_V A(\overline{\mu})^{-1} A_q\right). \end{aligned} \tag{3.3.6}$$

Thus, we have $H_{\overline{\mu}}(\mu) = \sum_{q=1}^{2Q} \theta_q(\mu) H_{\overline{\mu},q}$ where $H_{\overline{\mu},q} \in \mathcal{L}(V, V')$, $1 \leqslant q \leqslant 2Q$ are μ-independent *self-adjoint* operators and $\theta_q : \mathcal{D} \to \mathbb{R}$, $1 \leqslant q \leqslant 2Q$ are *real-valued* functions. In this context, the SCM can be readily applied; *e.g.* see Refs. [61, 60, 112].

The dual natural-norm of the RB residual

Given an anchor point $\overline{\mu} \in \mathcal{D}$, we explain how the dual natural-norm of the residual $|||A(\mu)\widetilde{u_N}(\mu) - f(\mu)|||_{\overline{\mu},W'}$ can be efficiently computed for all $\mu \in \mathcal{D}$. Let us assume the following decomposition for $\widetilde{u_N}(\mu)$ in the RB subspace $\widetilde{V_N} := \mathrm{Span}\{\xi_1, \ldots, \xi_N\} \subset V$,

$$\widetilde{u_N}(\mu) = \sum_{i=1}^{N} c_i(\mu)\xi_i. \tag{3.3.7}$$

With the RB method, the coefficients $c_i(\mu)$, $1 \leqslant i \leqslant N$ can be obtained very efficiently for any value of $\mu \in \mathcal{D}$ by solving a $N \times N$ linear system [97, 54]. For all $\mu \in \mathcal{D}$, there holds

$$|||A(\mu)\widetilde{u_N}(\mu) - f(\mu)|||^2_{\overline{\mu},W'} = \|A(\overline{\mu})^{-1}(A(\mu)\widetilde{u_N}(\mu) - f(\mu))\|^2_V \tag{3.3.8}$$

Using the affine representations Eqs. (3.3.2) and (3.3.3) and the expression (3.3.7) for the RB approximation, we get from developing the square

$$\begin{aligned}\|\!|A(\mu)\widetilde{u_N}(\mu) - f(\mu)|\!\|^2_{\overline{\mu},W'} = &\sum_{1\leqslant q,p\leqslant Q^f} \varsigma_q^f(\mu)\overline{\varsigma_p^f(\mu)}\langle R_V A(\overline{\mu})^{-1}f_q, A(\overline{\mu})^{-1}f_p\rangle \\ &\sum_{1\leqslant i,j\leqslant N}\sum_{1\leqslant q,p\leqslant Q} c_i(\mu)\overline{c_j(\mu)}\varsigma_q(\mu)\overline{\varsigma_p(\mu)}\langle R_V A(\overline{\mu})^{-1}A_q\xi_i, A(\overline{\mu})^{-1}A_p\xi_j\rangle \\ -2\sum_{1\leqslant q\leqslant Q}\sum_{1\leqslant p\leqslant Q^f}\sum_{1\leqslant i\leqslant N} &\Re\left\{\varsigma_q(\mu)\overline{\varsigma_p^f(\mu)}c_i(\mu)\langle R_V A(\overline{\mu})^{-1}A_q\xi_i, A(\overline{\mu})^{-1}f_p\rangle\right\}.\end{aligned} \tag{3.3.9}$$

Notice that none of the duality brackets depend on μ. Thus, these duality brackets can be computed once during the so-called *offline* phase. For each query $\mu \in \mathcal{D}$ (during the so-called *online* phase), the pre-computed duality brackets can be used to compute the dual natural-norm of the residual in $\mathcal{O}((Q^f)^2 + N^2Q^2 + NQQ^f)$ complexity using the formula (3.3.9).

Remark that during the offline phase, one must solve $NQ+Q^f$ problems of the form: *find* $y \in V$ *such that* $A(\overline{\mu})y = z$. Namely, for $z = A_q\xi_i$ ($1 \leqslant i \leqslant N$, $1 \leqslant q \leqslant Q$) and for $z = f_q$ ($1 \leqslant q \leqslant Q^f$). The number of problems to be solved is therefore $K(NQ + Q^f)$ when a set $\mathcal{C}^K$ of K anchor points is considered.

3.3.2 Procedure for selecting anchor points

We now present a strategy for constructing the set of anchor points $\mathcal{C}^K = \{\overline{\mu}^1, \ldots, \overline{\mu}^K\} \subset \mathcal{D}$. Let us adopt a discrete setting, by introducing an adequately fine discrete surrogate set $\Xi \subset \mathcal{D}$. Clearly, if we want to be able to estimate the error globally, the set of anchor points must be built in order that the following (discrete) *coverage property* holds

$$\forall \mu \in \Xi, \quad \exists \overline{\mu} \in \mathcal{C}^K, \quad \overline{\sigma}_{\overline{\mu}}(\mu) > \varrho, \tag{3.3.10}$$

or equivalently,

$$\forall \mu \in \Xi, \quad \max_{\overline{\mu}\in\mathcal{C}^K} \overline{\sigma}_{\overline{\mu}}(\mu) > \varrho, \tag{3.3.11}$$

where $\varrho \in [0, 1[$ is a prescribed threshold. Notice that with the strict inequalities, choosing $\varrho = 0$ will ensure that $\mu \mapsto \max_{\overline{\mu}\in\mathcal{C}^K} \overline{\sigma}_{\overline{\mu}}(\mu)$ remains strictly positive over Ξ, which is the minimum requirement to be able to estimate the error over Ξ. We leave the possibility of considering $\varrho > 0$ if one is interested in sharper error estimates.

The procedure that we propose consists in building a sequence $\{\Xi_k\}_{k\geqslant 1}$, with $\Xi_1 = \Xi$ that will ultimately converge to $\emptyset$ as follows:

1. Pick arbitrarily $\overline{\mu}^1$ in Ξ, set $\Xi_1 = \Xi$, $\mathcal{C}^0 = \emptyset$ and $k = 1$;
2. Update the set of anchor points $\mathcal{C}^k = \mathcal{C}^{k-1} \cup \{\overline{\mu}^k\}$;

3. Update training set $\Xi_{k+1} \leftarrow \Xi_k \setminus \{\mu \in \Xi_k,\ \overline{\sigma}_{\overline{\mu}^k}(\mu) > \varrho\}$;

4. If $\Xi_{k+1} \neq \emptyset$, then find
$$\overline{\mu}^{k+1} \leftarrow \underset{\mu\in\Xi_k}{\operatorname{argmin}} \max_{\overline{\mu}\in\mathcal{C}^k} \overline{\sigma}_{\overline{\mu}}(\mu), \tag{3.3.12}$$
set $k = k + 1$ and go back to (ii). Else, terminate.

At each iteration $k \geqslant 1$ such that $\Xi_k \neq \emptyset$, we consider a new anchor point $\overline{\mu}^k \in \Xi_k$, and construct the set $\Xi_k^+ = \{\mu \in \Xi_k,\ \overline{\sigma}_{\overline{\mu}^k}(\mu) > \varrho\}$ (this can be done efficiently using SCM). This set is guaranteed to be non-empty because it has at least one member: $\overline{\mu}^k$, using the fact that $\overline{\sigma}_{\overline{\mu}^k}(\overline{\mu}^k) = 1$ (see Proposition 3.2.3). Thus, the set $\Xi_{k+1} = \Xi_k \setminus \Xi_k^+$ is guaranteed to be a strict subset of Ξ_k. This demonstrates that the sequence $\{\Xi_k\}_{k\geqslant 1}$ converges to $\emptyset$ in at most $\mathrm{Card}(\Xi)$ iterations and so the procedure terminates. Note that, in practice, much less than $\mathrm{Card}(\Xi)$ iterations will be required for convergence as we shall see in the numerical examples.

Furthermore, there is no difficulty in showing that at iteration $k \geqslant 1$, we have
$$\Xi = \Xi_{k+1} \cup \left(\bigcup_{\kappa=1}^{k} \Xi_\kappa^+\right), \quad \Xi_{k+1} \cap \left(\bigcup_{\kappa=1}^{k} \Xi_\kappa^+\right) = \emptyset, \tag{3.3.13}$$
thus the procedure terminates at iteration K such that $\Xi = \bigcup_{\kappa=1}^{K} \Xi_\kappa^+$, which means that the discrete coverage property holds.

3.4 Numerical results

3.4.1 Problem setting

Let $\Omega =]0,1[\times]0,1[$. The domain boundary is divided into a Dirichlet boundary $\Gamma_D =]0,1[\times\{0\}$ and a Neumann boundary $\Gamma_N = \partial\Omega \setminus \Gamma_D$. Let $f \in L^2(\Omega)$ and $g \in H^{-1/2}(\Gamma_N)$. We consider the 2D Helmholtz equation, parametrized by $\mu = (\mu_1, \mu_2) \in \mathcal{D}$: *find* $u^{\mathrm{ex}}(\cdot;\mu) \in H^1(\Omega)$
$$\begin{cases} -\mathrm{div}\left(\begin{pmatrix} 1 & \nu \\ 0 & \mu_1 \end{pmatrix} \boldsymbol{\nabla} u^{\mathrm{ex}}(\mu)\right) - \mu_2 u^{\mathrm{ex}}(\mu) = f, \quad \text{in } \Omega, \\ u^{\mathrm{ex}}(\mu)|_{\Gamma_D} = 0, \quad \boldsymbol{\nabla} u^{\mathrm{ex}}(\mu)\cdot \boldsymbol{n}|_{\Gamma_N} = g. \end{cases} \tag{3.4.1}$$
The parameter μ_1 controls the anisotropy of the speed of sound, while the parameter μ_2 corresponds to the squared wavenumber. The constant ν (not a parameter) also controls the anisotropy of the speed of sound (isotropic speed of sound corresponds to $\nu = 0$, $\mu_1 = 1$). Notice that when $\nu = 0$ the problem is self-adjoint and corresponds to the benchmark proposed in Ref. [60] and addressed more recently in Ref. [113]. When $\nu > 0$, the problem is no longer self-adjoint.

We use the Finite Element (FE) method to discretize the weak form of (3.4.1). We define the Hilbert space V as the Lagrange P^1 approximation space, formed by globally continuous, piece-wise first-order polynomial functions that vanish on the boundary Γ_D. This FE space being $H^1(\Omega)$-conforming, the norm on V is simply the usual $\|\cdot\|_{H^1(\Omega)}$ norm. Using a triangulation of Ω, the dimension of this FE space is $\mathcal{N} = 3436$. The FE approximation $u(\mu) \in V$ is defined as the Galerkin projection of $u^{\text{ex}}(\mu)$ on V, which amounts to considering the test space $W = V$.

We further define a RB approximation $\widetilde{u_N}(\mu)$ as the Galerkin projection of $u(\mu)$ onto the RB space, meaning $\widetilde{u_N}(\mu) \in \widetilde{V_N} := \text{Span}\{u(\mu_1),\ldots, u(\mu_N)\}$, where the parameters $\mu_1, \ldots, \mu_N$ are selected in a greedy way based on the $\|\cdot\|_{W'}$ norm of the residual, following standard practice of the RB Method [54, 97]. Our goal will be able to estimate the RB error $\|u(\mu) - \widetilde{u_N}(\mu)\|_V$ for $\mu \in \mathcal{D}$.

Remark that, without any *a priori* knowledge on the possible location of resonant parameters, finding a "resonance-free" set $\mathcal{D}$ is not an easy task. In our numerical tests, we shall consider $\mathcal{D} \subset \widetilde{\mathcal{D}} = [0.8, 1.2] \times [10, 50]$, as in Ref. [60]. We consider the two possible values $\nu = 0$ (self-adjoint case) and $\nu = 0.5$ (non self-adjoint case). We can see the norm of the FE solution for 2000 random points in $\widetilde{\mathcal{D}}$ on Figs. 1 and 2. We can visually see 4 resonance lines where the norm of the FE solution is maximal. Notice that the location of resonance lines slightly differ between the self-adjoint and non self-adjoint cases.

Remark. *As can be seen Figs. 1 and 2, there are some resonant parameter values in the compact set $\widetilde{\mathcal{D}} = [0.8, 1.2] \times [10, 50]$. Denote $\mathcal{D}^{\text{res}}$ the set of resonant values in $\widetilde{\mathcal{D}}$ for which the Banach-Nečas-Babuška assumptions are not satisfied. In this work, we choose to address the problem as if we had no a priori knowledge of the existence of this set $\mathcal{D}^{\text{res}}$. As we shall see, our method is constructive of a discrete surrogate set for the "resonance-free" set $\mathcal{D} \subset \widetilde{\mathcal{D}}$ satisfying $\mathcal{D}^{\text{res}} \subset (\widetilde{\mathcal{D}} \setminus \mathcal{D})$.*

3.4.2 Self-adjoint case

The natural-norm constants

We test our anchor point selection procedure with $\varrho = 0$ and a surrogate set $\Xi \subset \widetilde{\mathcal{D}}$ made of 2000 random points (uniformly distributed). The algorithm terminates with $K = 6$ anchor points. On Fig. 3.3, we have plotted the obtained SCM lower bounds $\mu \mapsto \exp(\overline{\sigma}^{LB}_{\overline{\mu}}(\mu))$ for each of the $K = 6$ selected anchor points $\overline{\mu} \in \mathcal{C}^K$. The reason for taking the $\exp()$ is to obtain a better visualization, recalling that the lower bound for the dual natural-norm constant can become negative.

A close comparison with Fig. 1 reveals that the dual natural-norm constant $\overline{\sigma}_{\overline{\mu}}(\mu)$ is only positive for values of μ such that μ and $\overline{\mu}$ can be joined without crossing any resonance lines. Interestingly, we have found $K = 6$, when $K = 5$ could have been expected from Fig. 1. In fact, we have checked that the anchor point $\overline{\mu} = (0.80, 49.9)$ is indispensable

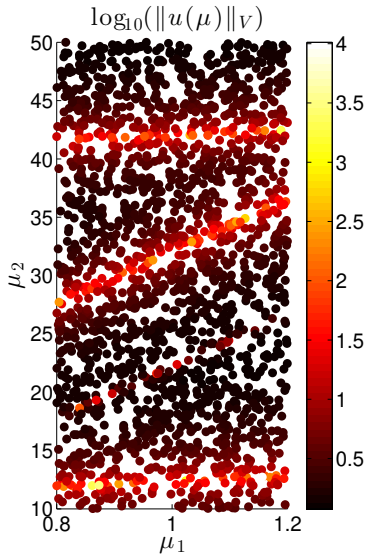


Figure 3.1: The norm of the FE solution, for 2000 random points in $\mathcal{D}$ in self-adjoint case.

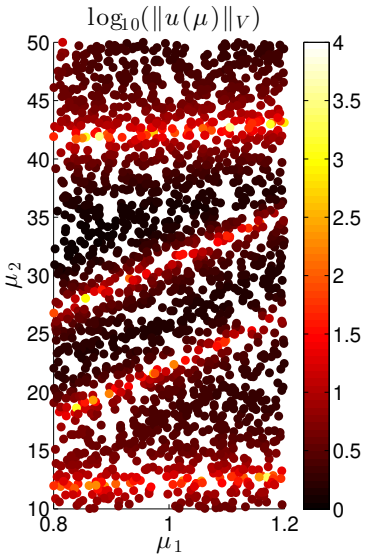


Figure 3.2: The norm of the FE solution, for 2000 random points in $\mathcal{D}$ in non self-adjoint case.

in order to achieve the positivity coverage, because there is indeed a resonance line to be crossed to reach this point starting from all previously selected anchor points. Let us now analyze the computational effort. At each iteration k of the anchor point selection procedure, a SCM algorithm is called in order to efficiently compute the lower bounds. In our numerical experiments, we have set the prescribed SCM tolerance to $tol = 0.9$. We have consigned in Table 1 the number of eigensolves of the generalized eigenvalue problem (3.3.4) performed during each call to the SCM.

k	1	2	3	4	5	6
Eigensolves	15	1	8	3	11	7

Table 3.1: Number of times that the generalized eigenvalue problem (3.3.4) must be solved at each iteration k of the anchor point selection procedure.

Comparing with Fig. 3.3, we find that the required number of eigensolves depends on the size of the positivity coverage. Typically, if the positivity coverage is vaster, then more eigensolves are needed.

Remark. *We notice a sensibility to the sampling of the surrogate set $\Xi \subset \widetilde{\mathcal{D}}$. Namely, a different sampling of the* 2000 *uniformly distributed random points in which the point* $(0.80, 49.9)$ *was absent led to* $K = 5$. *In this situation, we have not been able to find an index* $k = 1, \ldots, 5$ *such that* $\overline{\sigma}_{\overline{\mu}^k}(\mu)$ *for* $\mu = (0.80, 49.9)$ *was* > 0. *This illustrates the potential risks that the discrete positivity coverage property* (3.3.11) *is dependent on the choice of surrogate set* Ξ.

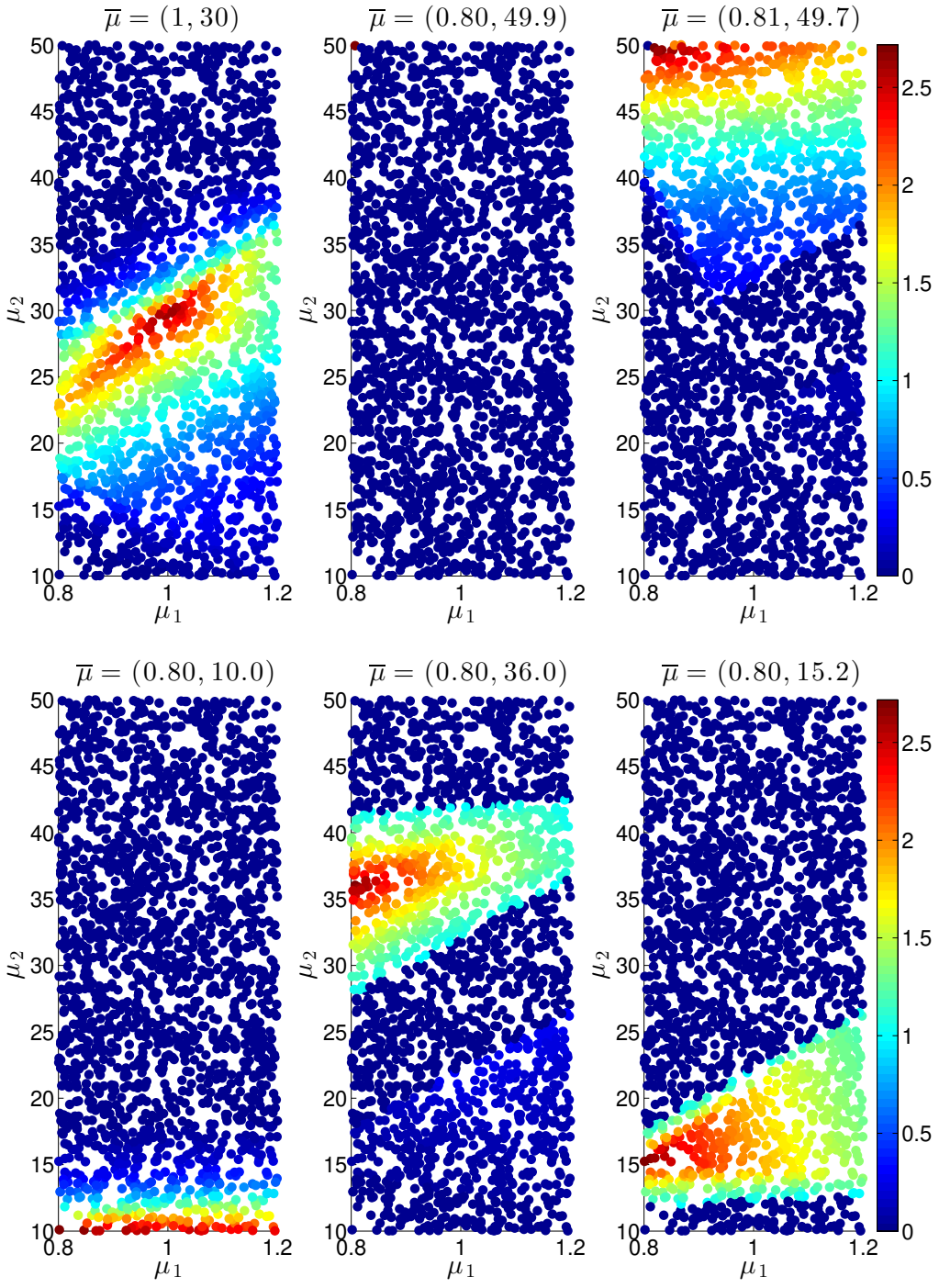


Figure 3.3: Self adjoint case: The SCM lower bound for the dual natural-norm constants $\mu \mapsto \exp(\overline{\sigma}_{\overline{\mu}}^{LB}(\mu))$ for the $K = 6$ successive values of $\overline{\mu}$ determined by the anchor point selection procedure.

Error estimates

We now consider a reduced basis approximation $\widetilde{V_N}$ of dimension $N = 15$. In order to assess the performance of our error estimators, we solve both FE and RB problems for all $\mu \in \Xi$, where $\Xi \subset \mathcal{D}$ is a random set of cardinality 2000. We have re-sampled the random points, thus this set Ξ is different from set the one used for selecting the anchor points. On Fig. 3.4, we have plotted two effectivity distributions; namely

- the effectivity distribution of the practical inf-sup based error estimator, that is $\{\Delta_K^{\mathrm{pr}}(\widetilde{u_N}(\mu);\mu)/\|u(\mu) - \widetilde{u_N}(\mu)\|_V,\ \mu \in \Xi\}$ (top);
- the effectivity distribution of the practical dual natural-norm based error estimator, that is $\{\Delta_K^{\mathrm{du}}(\widetilde{u_N}(\mu);\mu)/\|u(\mu) - \widetilde{u_N}(\mu)\|_V,\ \mu \in \Xi\}$ (bottom);

we have further consigned the essential statistics of these two effectivity distributions in Table 2.

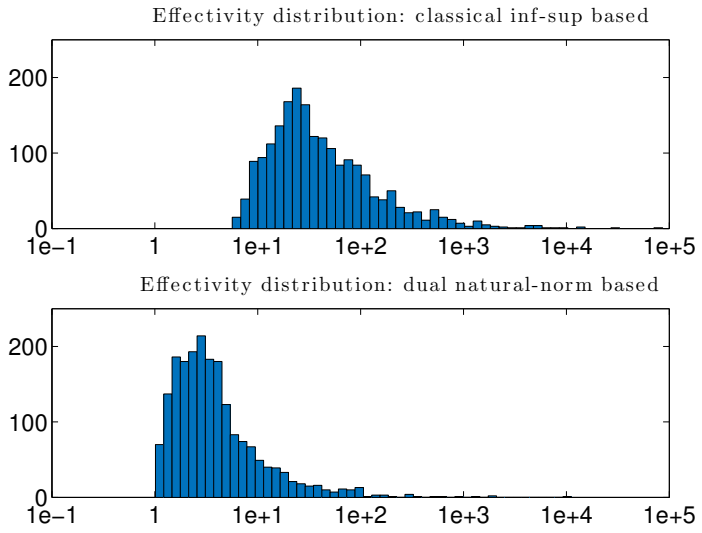


Figure 3.4: Self adjoint case: Effectivity distribution of the practical inf-sup based error estimator (top) and of the practical dual natural-norm based error estimator (bottom), obtained from 2000 random parameter samples in $\mathcal{D}$.

Error estimator	Max	85% quantile	Median	Mean
Inf-sup based	8.61×10^4	1.32×10^2	3.16×10^1	1.99×10^2
Dual natural-norm	1.11×10^4	9.51	3.17	1.74×10^1

Table 3.2: Effectivity statistics in self-adjoint case (based on the distribution shown on Fig. 3.4).

We find the inf-sup based error estimator to overestimate the error by at least one order of magnitude. On the contrary, the dual natural-norm based error estimator captures the

correct order of magnitude of the error for 85% of the considered values of $\mu \in \mathcal{D}$. For both error estimators, the amount of overestimation can become as large as 4 orders of magnitude. However, this phenomenon only occurs very locally; namely near the resonance lines. This is confirmed by Fig. 3.5, where we have plotted the effectivity in the (μ_1, μ_2) plane, and where we find all maximum values of effectivity to be located in the neighborhood of a resonant line. The tails of the distributions on Fig. 3.4 reflect the small probability of a random parameter value to be located very close to a resonant line.

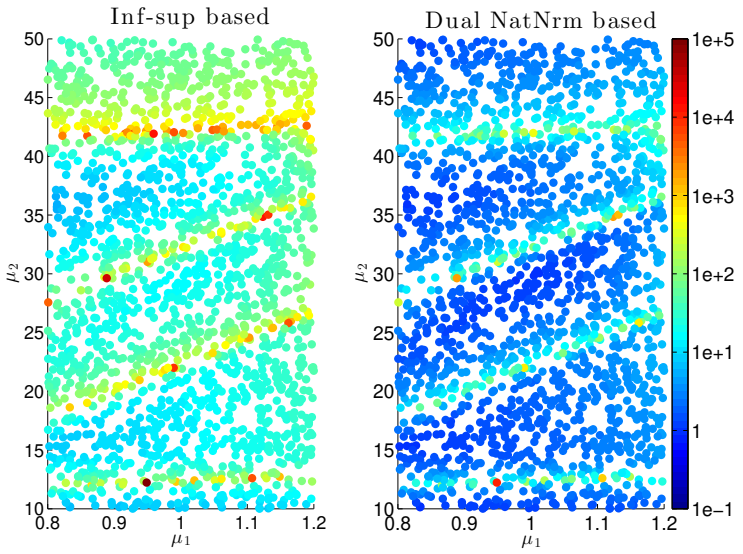


Figure 3.5: Self adjoint case: The effectivity of the inf-sup (left) and dual natural-norm (right) error estimators plotted as functions of $\mu = (\mu_1, \mu_2)$.

3.4.3 Non self-adjoint case

The natural-norm constants

We now address the non self-adjoint case. Recall that, in this case, there is a distinction between the primal natural-norm constant $\overline{\alpha}_{\overline{\mu}}(\mu)$ and the dual natural-norm constant $\overline{\sigma}_{\overline{\mu}}(\mu)$. We highlight the differences between these two constants in Fig. 3.6. It is worth noticing that the two coverage sets $\mathcal{D}^{\mathrm{pr}}_{\overline{\mu}} = \{\mu \in \mathcal{D},\, \overline{\alpha}_{\overline{\mu}}(\mu) > 0\}$ and $\mathcal{D}^{\mathrm{du}}_{\overline{\mu}} = \{\mu \in \mathcal{D},\, \overline{\sigma}_{\overline{\mu}}(\mu) > 0\}$ slightly differ.

We test our anchor point selection procedure with $\varrho = 0$ and a surrogate set $\Xi \subset \widetilde{\mathcal{D}}$ made of 2000 random points (uniformly distributed). In order for the algorithm to terminate, we had to slightly change the stopping criterion. Indeed, we found that stopping at the iteration k such that $\Xi_k = \emptyset$ was irrelevant in this situation due to the presence of resonances. We relaxed this criterion by stopping at the iteration k such that

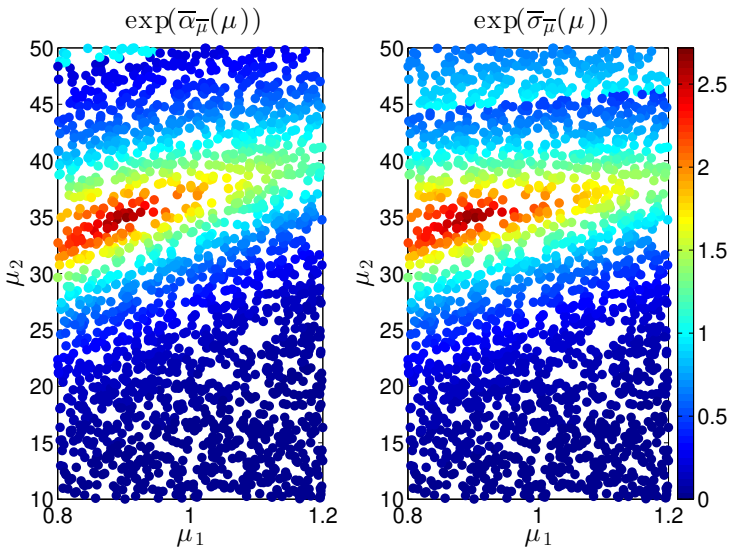


Figure 3.6: Non self-adjoint case: Comparison between the primal (left) and dual natural-norm (right) constants plotted as functions of $\mu = (\mu_1, \mu_2)$, with same anchor point $\overline{\mu} = (0.9, 35)$.

$\mathrm{Card}(\Xi_k) \leqslant \lceil (1-q)\mathrm{Card}(\Xi) \rceil$, where $q \in]0,1]$ is some fraction, $\mathrm{Card}(\Xi)$ the number of points in the initial surrogate parameter set and $\lceil \cdot \rceil$ denotes the ceiling operation. Under this new criterion with $q = 0.95$, the algorithm converged in $K = 15$ iterations. This means $K = 15$ anchor points are enough to obtain the discrete coverage property over $\Xi^+ = \Xi \setminus \Xi^-$, where Ξ^- denotes the set comprised of the 5% of parameter points over which the discrete coverage property is not satisfied; *i.e.*, for all $\mu \in \Xi^-$, for all $1 \leqslant k \leqslant K$, $\overline{\sigma}_{\overline{\mu}^k}(\mu) \leqslant 0$. We have checked that the points in set Ξ^- correspond to points near the resonance lines. Thus, our method is constructive of the set Ξ^+, which is a discrete surrogate set for the *a priori* unknown "resonance-free" set $\mathcal{D}$.

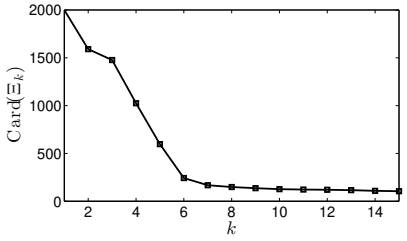


Figure 3.7: Convergence curve of anchor point selection procedure in the non self-adjoint case.

Looking at the converge curve on Fig. 3.7, we further find that the $K = 7$ first iterations already achieve 91.6% of the positivity coverage. The next iterations add small positivity coverage patches near the resonance lines to achieve the desired 95% positivity coverage property.

Error Estimates

We consider a reduced basis of size $N = 15$. In order to assess the performance of our dual natural-norm estimator, we solve both FE and RB problems for all $\mu \in \Xi$, where $\Xi \subset \mathcal{D}$ is a random set of cardinality 2000. We have re-sampled the random points, thus this set Ξ is different from set the one used for selecting the anchor points. On Fig. 3.8, we have plotted two effectivity distributions; namely

- the effectivity distribution of the exact inf-sup based error estimator, based on the computation of the exact inf-sup constant $\alpha(\mu)$, that is $\{\frac{1}{\alpha(\mu)}\|A(\mu)\widetilde{u_N}(\mu)-f(\mu)\|_{W'}/\|u(\mu)-\widetilde{u_N}(\mu)\|_V,\ \mu \in \Xi\}$ (top);

- the effectivity distribution of the practical dual natural-norm based error estimator, that is $\{\Delta_K^{\rm du}(\widetilde{u_N}(\mu);\mu)/\|u(\mu) - \widetilde{u_N}(\mu)\|_V,\ \mu \in \Xi\}$ (bottom);

We find 114 points in Ξ for which the discrete positivity coverage property does not hold. This corresponds to 5.7% of our points, slightly above the expected 5% and we have checked these points are all located near the resonance lines. We remove these 114 points from the initial Ξ set, and thus obtain a set of 1886 points for which our error estimator can be computed. Notice at this stage that, estimating the error at these 114 using our dual natural-norm error estimator is impossible with $K = 15$, but this would be possible by consider more anchor points $K > 15$. Inspection of the two distributions reveals that the exact inf-sup based estimator never provides the correct order of magnitude of the error, while the dual natural-norm error estimator does for most parameter values.

On Fig. 3.9, we have plotted the effectivity in the (μ_1, μ_2) plane. For the inf-sup based estimator, we find all maximum values of effectivity to be located in the neighborhood of a resonant line (as in the self-adjoint case). However, for the dual natural-norm based estimator, the maximum values are not always near a resonant line. Thus, the tail of the distribution on Fig. 3.8 (below) where the effectivity is large, does not necessarily correspond to parameter values located very near a resonance line.

In order to understand the origin of the tail of the distribution, we show on Fig. 3.10 the two stability constants at play: the inf-sup constant $\mu \mapsto \alpha(\mu)$ and the dual-natural norm constant $\mu \mapsto \max\limits_{1\leqslant k\leqslant K} \overline{\sigma}_{\overline{\mu}^k}(\mu)$. While the minimas of the inf-sup constant clearly mark the resonance lines, this is not the case for the dual natural-norm constant. In fact, we find that the values of μ for which the dual natural-norm constant is minimal correspond to the values of μ for which the effectivities are maximal on Fig. 3.9 (right). This confirms the relevance of effectivity bound from Proposition 3.1.4, which suggests that a small dual natural-norm constant will deteriorate the effectivity. Of course, it is always possible to obtain a $\mathcal{O}(1)$ dual natural-norm constant by adding more anchor points.

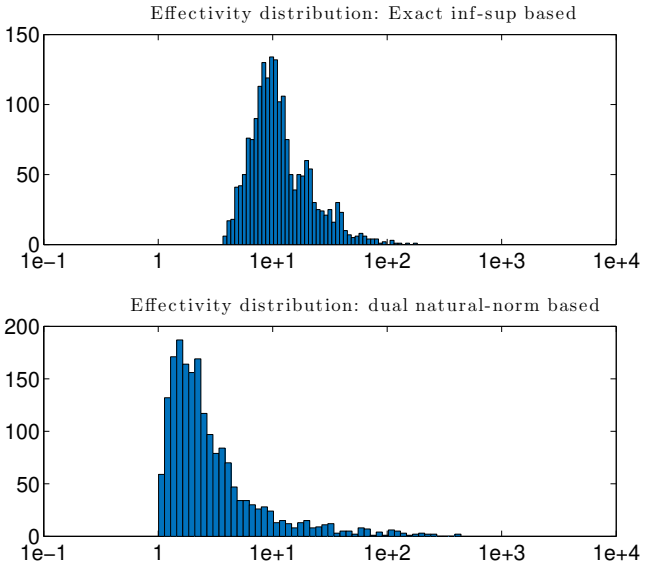


Figure 3.8: Non self-adjoint case: Effectivity distribution of the exact inf-sup based error estimator (top) and of the practical dual natural-norm based error estimator (bottom), obtained from 1886 random parameter samples in $\mathcal{D}$ satisfying the positivity coverage property.

3.5 Conclusions

In this chapter, we have developed both theoretical error bounds and practical *a posteriori* error estimators for reduced basis approximations to parametrized linear equations based on the concept of dual natural-norm. In comparison to the classical error bounds based on the inf-sup stability constant, the dual natural-norm error bounds are associated with a $\mathcal{O}(1)$ stability constant and are therefore very effective. Moreover, in opposition to the primal natural-norm approach, one is free to choose the norm $\|\cdot\|_V$ in which the error $u(\mu)-\widetilde{u_N}(\mu)$ should be measured, since the dual natural-norm is not a norm for measuring the error, but rather a norm for measuring the residual.

We have shown a computational strategy for efficiently computing the proposed dual natural-norm error estimator in the context of reduced basis approximations. This strategy was successfully applied to a Helmholtz equation parametrized by the wavenumber and anisotropy parameter. Numerical results show great potential, especially in the case of challenging problems with resonant parameters. In this context, the proposed method also provides a very practical way to determine a "resonance-free" set of parameters $\mathcal{D}$ out of a larger parameter set $\widetilde{\mathcal{D}}$ which contains resonant parameters at *a priori* unknown locations.

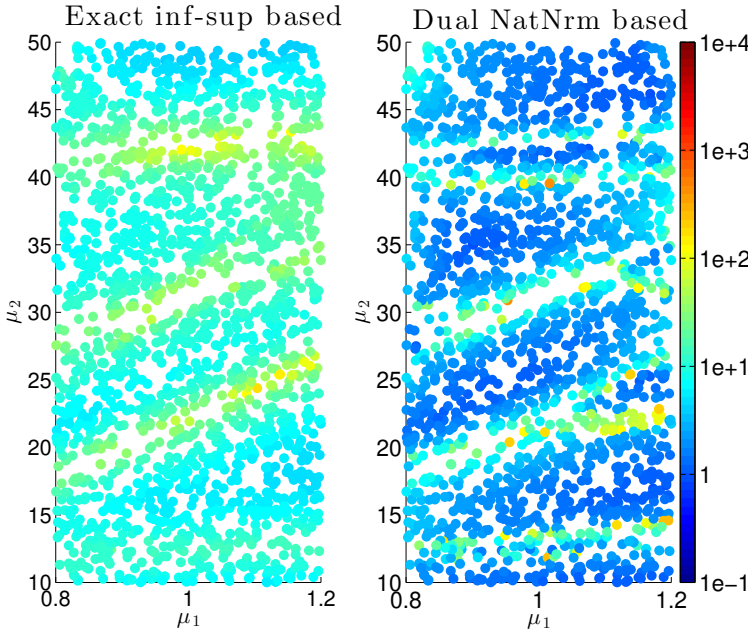


Figure 3.9: Non self adjoint case: The effectivity of the inf-sup (left) and dual natural-norm (right) error estimators plotted as functions of $\mu = (\mu_1, \mu_2)$. Notice the logarithm scale.

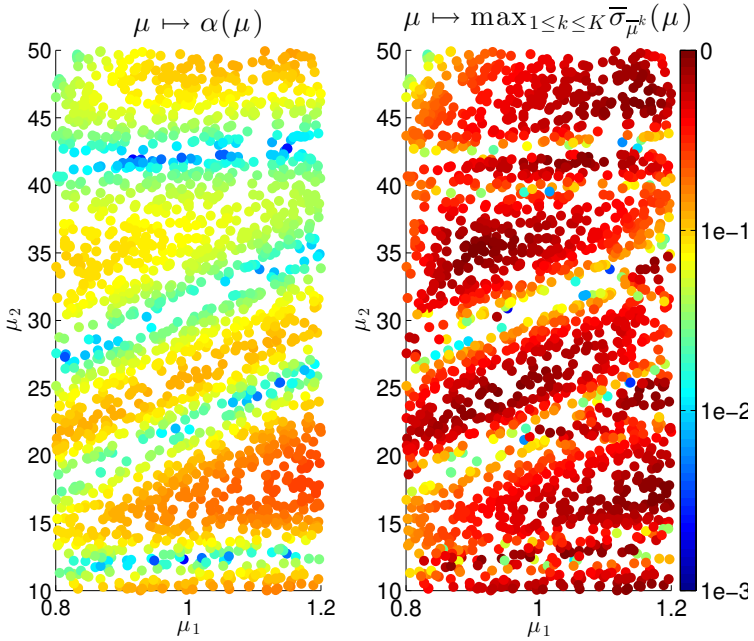


Figure 3.10: Non self adjoint case: The inf-sup constant (left) and dual natural-norm constant (right) as functions of $\mu = (\mu_1, \mu_2)$. Notice the logarithm scale.

Chapter 4

The reduced basis method in the context of multiple sources

Summary. This chapter is concerned with reduced basis approximations to parametrized problems featuring multiple sources. We propose two strategies: a *multiple RBs* strategy that consists in building distinct reduced basis approximation subspaces each one adapted to a specific source term and a *unique RB* strategy that consists in building a unique reduced basis approximation subspace for all the sources. In both cases, a block formulation is advantageously used in order to limit the overall number of problems to solve. We show that the block framework and associated reduced basis strategies are very well adapted to the class of parametrized problems with parametrized source, in which the traditionally μ-dependent source further depends on an additional parameter ν. In this situation, the EIM is used to incorporate the additional parameter ν in an approximate parametrized block formulation. We illustrate our reduced basis approach on an academic Laplace problem.

Contents

4.1 Introduction to parametrized problems with multiple sources 78
4.1.1 The parametrized problem with ℓ independent sources 78
4.1.2 Outline of two reduced basis strategies 80
4.2 Multiple RBs strategy 81
4.2.1 Motivation 81
4.2.2 Reduced basis approximation in a glance 82
4.2.3 Greedy construction 82
4.3 Unique RB strategy 83
4.3.1 A short review 83
4.3.2 The block formulation 84

4.3.3 Block reduced basis approximations 85
4.3.4 Greedy construction 87
4.3.5 Offline/Online strategy and complexity analysis 88
4.4 The parametrized problem with parametrized source 90
4.4.1 Problem formulation 90
4.4.2 Block approximation 92
4.4.3 Affine approximation of the block RHS 93
4.5 Numerical illustration 95
4.5.1 The parametrized problem with ℓ independent source terms . . 95
4.5.2 The parametrized problem with parametrized source 95
4.6 Conclusions 103

4.1 Introduction to parametrized problems with multiple sources

As usual in this thesis, V and W denote two Hilbert spaces with finite dimension $\mathcal{N}$, that must be thought of as finite element approximation spaces of infinite dimensional Sobolev spaces provided with the inherited norm respectively $\|\cdot\|_V$ and $\|\cdot\|_W$ together with the scalar product $(\cdot,\cdot)_V$ and $(\cdot,\cdot)_W$ respectively.

4.1.1 The parametrized problem with ℓ independent sources

Let $\mu \in \mathcal{D}_\mu$ denote the parameter, with $\mathcal{D}_\mu$ a compact set of $\mathbb{R}^{p_\mu}$, $p_\mu \geqslant 1$. We consider a μ-parametrized operator $A(\mu) \in \mathcal{L}(V, W')$ satisfying the Banach-Nečas-Babuška assumptions (we refer the reader to the mathematical framework introduced in chapter 1, section 1.1.4). We further consider a familly of $\ell \geqslant 1$ independent source terms $f_1(\mu), \ldots, f_\ell(\mu) \in W'$. The parametrized problem with ℓ distinct μ-parametrized source terms is the following: *find* $u_r(\mu) \in V$ *such that*

$$A(\mu)u_r(\mu) = f_r(\mu) \quad \text{in } W', \quad 1 \leqslant r \leqslant \ell. \tag{4.1.1}$$

Our objective is to build efficient subspace approximations to the ℓ solution manifolds $\mathscr{M}_r = \{u_r(\mu),\ \mu \in \mathcal{D}_\mu\}$ for $r = 1, \ldots, \ell$. At this stage, it is worth recalling that the problem (4.1.1) is linear. Consequently, the solution associated to any linear combination of the sources $\sum_{r=1}^{\ell} \beta_r f_r(\mu)$ with coefficients $\beta_1, \ldots, \beta_\ell \in \mathbb{C}$ can be straightforwardly obtained as the linear combination of the solutions $\sum_{r=1}^{\ell} \beta_r u_r(\mu)$ with the same coefficients. This property will be exploited in section 4.3.2 when a block formulation to the problem (4.1.1) will be introduced.

Model problem

As model problem, we consider the Laplace model problem (1.1.5) introduced in chapter 1, thus $V = W = X_h^0(\Omega)$, with $\Omega =]0,1[^2$. We recall that the parameter μ is two-dimensional (thus $p_\mu = 2$) and corresponds to the coordinates of the peak in the conductivity. We further recall that $\mathcal{D}_\mu = [0.4, 0.6]^2$. Rather than the right-hand side $f : w \in W \mapsto \int_\Omega S w \mathrm{d}\Omega$ with non-parametrized source term $S \in L^2(\Omega)$ given by eq. (1.4.6), we are now interested in the solutions $u_1(\mu), \ldots, u_4(\mu)$ associated to the four distinct μ-parametrized source terms $f_r(\mu) : w \in W \mapsto \int_\Omega S_r(x,\mu) w(x) \mathrm{d}x$, $1 \leqslant r \leqslant 4$, with

$$
\begin{aligned}
S_1(x;\mu) &= \mu_1 \exp\left(-\frac{(x_1-0.25)^2 + (x_2-0.25)^2)}{0.02}\right), \\
S_2(x;\mu) &= \mu_2 \exp\left(-\frac{(x_1-0.25)^2 + (x_2-0.75)^2)}{0.02}\right), \\
S_3(x;\mu) &= \mu_1 \exp\left(-\frac{(x_1-0.75)^2 + (x_2-0.25)^2)}{0.02}\right), \\
S_4(x;\mu) &= \mu_2 \exp\left(-\frac{(x_1-0.75)^2 + (x_2-0.75)^2)}{0.02}\right).
\end{aligned}
\tag{4.1.2}
$$

Figure 4.1 provides a visualization of the truth solutions associated to the sources $r = 1, 3, 4$ for two possible values of the parameter μ. As we can see, each source is clearly localized in its own corner of the domain.

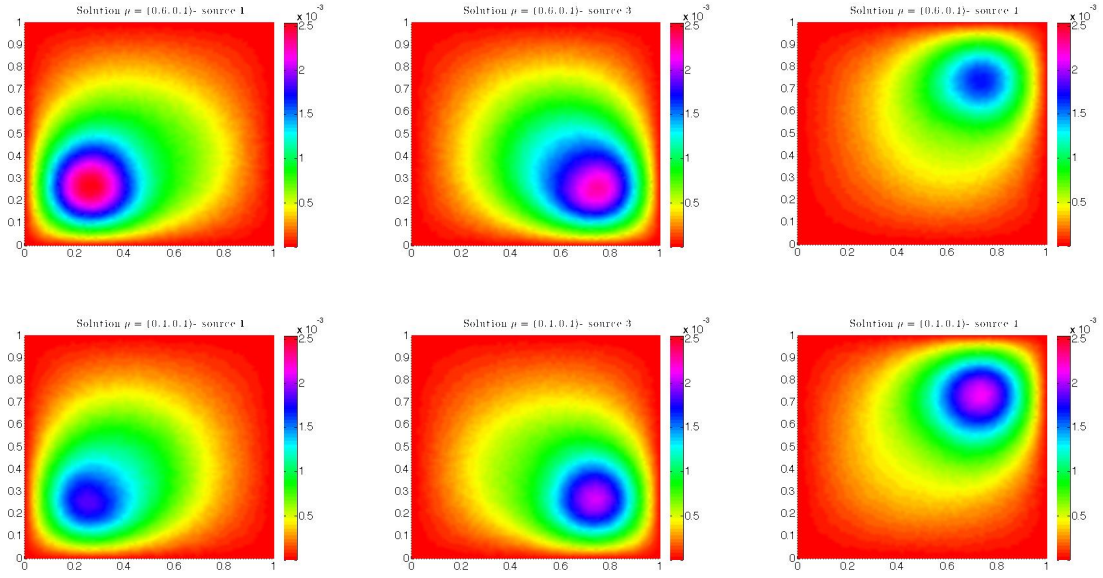


Figure 4.1: Truth solutions $u_r(\mu)$. First row: fixed $\mu = (0.6, 0.4)$. Second row: fixed $\mu = (0.4, 0.4)$. In columns: $r = 1, 3, 4$.

The truth solutions are affected by the parameter μ in two ways. First, the amplitude of the solutions depends on μ (in detail: μ_1 impacts the amplitude of the sources 1 and 3 while μ_2 impacts the sources 2 and 4). Second, the conductivity peak at coordinates $\mu = (\mu_1, \mu_2)$ provides a preferred direction of diffusion for the solution. This is particularly visible for the sources localized near the conductivity peak, namely for the source $r = 3$ at $\mu = (0.6, 0.4)$ (first row, second column on fig. 4.1) and for the source $r = 1$ at $\mu = (0.4, 0.4)$ (second row, first column on fig. 4.1).

Real-world problem

The next chapter will provide a real-world parametrized problem featuring multiple independent sources. Indeed, in section 5.3, we will study an antenna array: the varying parameter will be the frequency and we will be interested in computing the electric field generated by each individual antenna across the frequency range of interest.

4.1.2 Outline of two reduced basis strategies

Let us now discuss the possible reduced basis strategies for efficiently solving a parametrized problem featuring multiple sources. This chapter presents two strategies: the multiple RBs strategy or the unique RB strategy.

The multiple RBs strategy

The ℓ source terms being independent, the solution $u_r(\mu)$ associated to the source term $f_r(\mu)$ is potentially completely unrelated to the solution $u_p(\mu)$ associated to a different source term $f_p(\mu)$, $p \neq r$. In the extreme cases, these two solutions could even be orthogonal in the sense $(u_r(\mu), u_p(\mu))_V = 0$. In this situation, it seems preferable to build ℓ distinct reduced basis approximation subspaces, each dedicated to approximating the solutions associated to a given source term.

We propose to build the ℓ RB approximation subspaces relying on a unique family of parameter points $\{\mu_i\}_{1 \leqslant i \leqslant N}$ rather than on ℓ independently selected families $\{\mu_i^{(r)}\}_{1 \leqslant i \leqslant N^{(r)}}$, $r = 1, \ldots, \ell$. This choice is motivated by the fact that solving a problem with ℓ right-hand sides is more economical than solving ℓ problems with single right-hand sides. This strategy of multiple RBs is presented in section 4.2.

The unique RB strategy

Despite the ℓ source terms being independent, two truth solutions $u_r(\mu)$, $u_p(\mu)$ associated to two distinct sources $r \neq p$ might not be completely unrelated. In this situation, it

seems interesting to build a unique RB approximation subspace for approximating the truth solution $u_r(\mu)$ for any parameter value $\mu \in \mathcal{D}_\mu$ and any source index $r \in \{1, \dots, \ell\}$. This strategy of the unique RB is presented in section 4.3.

4.2 Multiple RBs strategy

4.2.1 Motivation

The multiple RBs strategy consists in constructing ℓ distinct reduced basis approximation subspaces, say $V^{(1)}, \dots, V^{(\ell)}$, respectively of dimension $N^{(1)}, \dots, N^{(\ell)}$ and each dedicated to approximating the solutions associated to a given source term.

A logical way to proceed is to build, for all $1 \leqslant r \leqslant \ell$, the RB approximation subspace $V^{(r)} \subset V$ by applying the greedy algorithm 2.1 to each parametrized problem $A(\mu)u_r(\mu) = f_r(\mu)$. These ℓ calls to the greedy algorithm are completely independent and could therefore be performed in parallel. In the end, one obtains ℓ distinct RB approximation subspaces under the form

$$V^{(r)} = \mathrm{Span}\{u_r(\mu_1^{(r)}), \dots, u_r(\mu_{N^{(r)}}^{(r)})\}, \quad 1 \leqslant r \leqslant \ell, \tag{4.2.1}$$

where $\mu_1^{(r)}, \dots, \mu_{N^{(r)}}^{(r)}$ denote the $N^{(r)}$ parameter points greedily selected during the rth greedy algorithm. Notice that for $r \neq p$, the set of parameter points selected during the rth greedy algorithm is *a priori* different from the set of parameter points selected during the pth greedy algorithm.

In the following, we propose an alternative strategy in which we choose a unique set of parameter points $\mu_1, \dots, \mu_N$ (we shall soon explain how they are chosen) and build ℓ distinct RB approximation subspaces as

$$V_N^{(r)} = \mathrm{Span}\{u_r(\mu_1), \dots, u_r(\mu_N)\}, \quad 1 \leqslant r \leqslant \ell. \tag{4.2.2}$$

The motivation for this choice is the following. It is computationally more advantageous to compute the ℓ truth solutions $u_r(\mu_n)$, $1 \leqslant r \leqslant \ell$ associated to a unique parameter point μ_n than to solve ℓ truth solutions $u_r(\mu_n^{(r)})$, $1 \leqslant r \leqslant \ell$ associated to ℓ distinct parameter points $\mu_n^{(r)}$, $1 \leqslant r \leqslant \ell$. Indeed, under the paradigm of direct solvers, the former relies on a single LU factorization of $A(\mu_n)$ which can be used to perform all ℓ forward-backward triangular processes, while the latter requires ℓ factorizations to be performed (one for each $A(\mu_n^r)$, $1 \leqslant r \leqslant \ell$). Knowing that a LU factorization is computed in $\mathcal{O}(\mathcal{N}^3)$ operations and that a forward-backward triangular process requires $\mathcal{O}(\mathcal{N}^2)$ operations (see for instance [75, §3.10]), the first strategy is clearly more efficient. This remains true under the paradigm of iterative solvers, where efficient block Krylov subspace recycling strategies can be set up to efficiently solve multiple right-hand sides with considerable speed-ups compared to successive calls to Krylov methods applied to single right-hand sides [115, 103].

4.2.2 Reduced basis approximation in a glance

Let ℓ distinct reduced basis approximation subspaces $V_N^{(1)}, \ldots, V_N^{(\ell)}$ given by eq. (4.2.2). We propose to build, for all $1 \leqslant r \leqslant \ell$ a reduced basis approximation $u_{r,N}(\mu) \in V_N^{(r)}$ of the solution $u_r(\mu) \in V$ to eq. (4.1.1), characterized by either the Galerkin projection (see definition 1) or the least-squares approximation (see definition 2).

In the affine case, we have the following error estimate provided by theorem 1

$$\|u_r(\mu) - u_{r,N}(\mu)\|_V \leqslant \frac{1}{\alpha(\mu)} \|A(\mu)u_{r,N}(\mu) - f_r(\mu)\|_{W'} \tag{4.2.3}$$

If $A(\mu)$ or $f_r(\mu)$ are non-affine, they are replaced by affine approximations $\widetilde{A}(\mu)$, $\widetilde{f}_r(\mu)$ in the sense of definitions 3 and 4. Thus, both the reduced basis approximation $u_{r,N}(\mu) \in V_N^{(r)}$ and the residual norm $\|\widetilde{A}(\mu)u_{r,N}(\mu) - \widetilde{f}_r(\mu)\|_{W'}$ can be efficiently computed following the traditional offline/online strategy [123].

4.2.3 Greedy construction

The ℓ distinct reduced basis approximation subspaces $V_N^{(1)}, \ldots, V_N^{(\ell)}$ are not built successively one after the other, rather, they are built simultaneously. To this end, let us introduce the reduced basis approximation error at some parameter value $\mu \in \mathcal{D}_\mu$ taking into account the ℓ independent sources given by $\max_{1 \leqslant r \leqslant \ell} \|u_r(\mu) - u_{r,N}(\mu)\|_V$.

In the light of the error estimate eq. (4.2.3), we define the error indicator

$$\delta_N(\mu) = \max_{1 \leqslant r \leqslant \ell} \|A(\mu)u_{r,N}(\mu) - f_r(\mu)\|_{W'}. \tag{4.2.4}$$

We propose to drive the greedy iterations using this error indicator. Thus, at the Nth iteration, the ℓ truth solutions $u_r(\mu_N)$, $1 \leqslant r \leqslant \ell$ are computed at the same parameter point μ_N and the parameter point μ_{N+1} to be considered in the next iteration is the maximizer of the error indicator $\delta_N(\mu)$ over all possible values of μ (in practice, over a surrogate set with finite cardinality). The overall strategy is summarized by algorithm 4.1.

Heuristic approach

We briefly explain how the heuristic approach introduced in section 2.2.4 can be adapted to the present situation.

At iteration $N \geqslant 1$ of algorithm 4.1, we propose to also retain the index r_{N+1} such that

$$r_{N+1} = \operatorname*{argmax}_{1 \leqslant r \leqslant \ell} \|A(\mu_{N+1})u_{r,N}(\mu_{N+1}) - f_r(\mu_{N+1})\|_{W'}. \tag{4.2.5}$$

Algorithm 4.1: Residual-driven RB generation

Input : Discrete training set $\Xi \subset \mathcal{D}_\mu$, target tolerance ϵ_{target} and maximum reduced basis size $N_{\max}$

Output: ℓ reduced basis approximation space $V_N^{(1)}, \ldots, V_N^{(\ell)}$

Pick arbitrarily μ_1 in Ξ;
Set $V_0^{(r)} = \{0\}$, $1 \leqslant r \leqslant \ell$;
Initialize $N \leftarrow 0$;
while $N \leq N_{\max}$ *and* $\epsilon > \epsilon_{\text{target}}$ **do**
  Solve $u_r(\mu_N)$ for all $r = 1, \ldots, \ell$ using a single factorization of $A(\mu_N)$ under the direct solver paradigm or an efficient block Krylov recycling strategy under the iterative solver paradigm;
  Update each reduced basis $V_N^{(r)} = V_{N-1}^{(r)} \oplus \text{Span}\{u_r(\mu_N)\}$, $1 \leqslant r \leqslant \ell$;
  $\mu_{N+1} \leftarrow \underset{\mu \in \Xi}{\text{argmax}}\ \delta_N(\mu)$;
  $\epsilon \leftarrow \underset{\mu \in \Xi}{\max}\ \delta_N(\mu)$;
  $N \leftarrow N + 1$;
end

Thus, at iteration $N \geqslant 2$, we can compute the following residual/error sample

$$\hat{\alpha}_N = \frac{\|A(\mu_N)u_{r^\star,N-1}(\mu_N) - f_{r^\star}(\mu_N)\|_{W'}}{\|u_{r^\star,N-1}(\mu_N) - u_{r^\star}(\mu_N)\|_V}, \quad r^\star = r_N. \tag{4.2.6}$$

These samples can be used in the heuristic approach explained in section 2.2.4.

Remark. *Note that the couple* (μ_{N+1}, r_{N+1}) *is given by*

$$(\mu_{N+1}, r_{N+1}) = \underset{(\mu,r) \in \Xi \times \{1,\ldots,\ell\}}{\text{argmax}} \|A(\mu)u_{r,N}(\mu) - f_r(\mu)\|_{W'}. \tag{4.2.7}$$

Thus, the couple $(\mu, r) \in \mathcal{D}_\mu \times \{1, \ldots, \ell\}$ *is viewed as an extended parameter upon which the residual norm is maximized.*

4.3 Unique RB strategy

4.3.1 A short review

Rather than ℓ distinct RB approximation subspaces each one adapted to a specific source term, it is possible to build a unique RB fit for all sources. To the best of our knowledge, the RB literature provides very few strategies for building a unique RB approximation subspace fit to approximate the solution $u_r(\mu)$ for any $\mu \in \mathcal{D}_\mu$ and any source index $r \in \{1, \ldots, \ell\}$.

In [114], the authors propose to view $(\mu, r) \in \mathcal{D}_\mu \times \{1, \ldots, \ell\}$ as an extended parameter. Under this approach, the greedy algorithm selects N extended parameter points (μ_1, r_1), $\ldots$, (μ_N, r_N) and the RB approximation subspace is built as

$$V_N = \mathrm{Span}\{u_{r_1}(\mu_1), \ldots, u_{r_N}(\mu_N)\}. \tag{4.3.1}$$

An alternative that is proposed in [128] and that is also used in [114] consists in selecting at each iteration $n \geqslant 1$ of the greedy algorithm a parameter point $\mu_n \in \mathcal{D}_\mu$ and a direction vector $z_n \in \mathbb{R}^\ell$ (or $\mathbb{C}^\ell$ if the problem is complex). Then, the RB approximation subspace is built as

$$V_N = \mathrm{Span}\left\{\sum_{r=1}^{\ell} z_{1r} u_r(\mu_1), \ldots, \sum_{r=1}^{\ell} z_{Nr} u_r(\mu_N)\right\}, \tag{4.3.2}$$

where for all $1 \leqslant n \leqslant N$ we have denoted $z_{n1}, \ldots, z_{n\ell}$ the components of the direction vector $z_n \in \mathbb{R}^\ell$. Compared to the reduced basis eq. (4.3.1), for which the nth reduced basis function is the truth solution associated to the source $f_{r_n}(\mu_n)$, here the nth reduced basis function is the truth solution associated to the linear combination of sources $\sum_{r=1}^{\ell} z_{nr} f_r(\mu_n)$.

4.3.2 The block formulation

Because linear combination of sources (and of solutions) are considered, it is useful to introduce $Z = \mathbb{R}^\ell$ (or $\mathbb{C}^\ell$ in the complex case) and to define the μ-parametrized continuous linear map $F(\mu) \in \mathcal{L}(Z, W')$ as

$$\forall z = (z_1, \ldots, z_\ell) \in Z, \quad F(\mu)z = \sum_{r=1}^{\ell} z_r f_r(\mu). \tag{4.3.3}$$

Observing that $F(\mu)\hat{e}_r = f_r(\mu)$ for $1 \leqslant r \leqslant \ell$, where $\hat{e}_r$ denotes the vector of $\mathbb{R}^\ell$ full of zeros, except the rth entry which is equal to 1; we clearly have $u_r(\mu) = U(\mu)\hat{e}_r$ where $U(\mu) \in \mathcal{L}(Z, V)$ is the solution to the following problem: *find* $U(\mu) \in \mathcal{L}(Z, V)$ *such that*

$$\forall z \in Z, \quad A(\mu)U(\mu)z = F(\mu)z \quad \text{in } W'. \tag{4.3.4}$$

In the following, the formulation eq. (4.3.4) will be called the *block* formulation.

Using the block notations, the RB approximation subspace defined by eq. (4.3.2) is equivalently given by

$$V_N = \mathrm{Span}\left\{U(\mu_1)z_1, \ldots, U(\mu_N)z_N\right\}. \tag{4.3.5}$$

We now propose to go beyond the state-of-the-art approach which can be found in [128, 114]. To this end, we introduce the richer RB approximation subspace defined by

$$V_N = \mathrm{Span}\{U(\mu_1)z_1^{(1)}, \ldots, U(\mu_1)z_1^{(d_1)}, \ldots, U(\mu_I)z_I^{(1)}, \ldots, U(\mu_I)z_I^{(d_I)}\}, \tag{4.3.6}$$

where $\mu_1, \ldots, \mu_I$ are greedily selected in $\mathcal{D}_\mu$ and for $1 \leqslant i \leqslant I$, $\{z_i^{(j)},\ 1 \leqslant j \leqslant d_i\}$ are d_i independent direction vectors in $Z = \mathbb{C}^\ell$. Notice that the dimension of V_N is at most (in practice, equal to) $N = \sum_{i=1}^I d_i$, therefore it is independent of ℓ.

In other words, we propose to use d_i directions vectors per parameter point μ_i, $i = 1 \ldots, I$, where the number of directions d_i will be adequately chosen. Indeed, since the number of sources ℓ may be quite large, it is not relevant to consider all possible directions $U(\mu_1)\hat{e}_1$,...,$U(\mu_1)\hat{e}_\ell$,..., $U(\mu_I)\hat{e}_1$, ..., $U(\mu_I)\hat{e}_\ell$ as this would lead to an approximation subspace with unacceptably high dimension, as anticipated in [128]. It is worth mentioning that the idea of retaining only the most useful directions to obtain an acceptable subspace dimension without compromising the quality approximation is widespread in the reduced basis literature, especially in the context of time-dependent problems [50, 87, 33].

In the previous works [128] and [114], only one direction vector is allowed per parameter point and thus the RB approximation subspace eq. (4.3.5) reaches a size N in N greedy iterations. Here, the RB approximation subspace eq. (4.3.6) can reach a size N in a number of greedy iterations $I < N$. Thus there is a hope to select fewer parameter points μ_i at which problem solves are required. Of course, the number of right-hand sides per required problem solve will be increased, but a computational advantage is still expected, keeping in mind that it is more advantageous to solve multiple right-hand sides with the same operator than to solve multiple distinct operators with single right-hand sides.

4.3.3 Block reduced basis approximations

Let us now define reduced basis approximations to the μ-parametrized block problem eq. (4.3.4).

Given a N-dimensional approximation subspace $V_N \subset V$, we propose to approximate the block solution $U(\mu) \in \mathcal{L}(Z, V)$ by a block reduced basis approximation $U_N(\mu) \in \mathcal{L}(Z, V_N)$ characterized by either

- the Galerkin problem (only in the situation where $V = W$)

$$\forall z \in Z,\ \forall v_N \in V_N,\ \langle \widetilde{A}(\mu)U_N(\mu)z, v_N\rangle = \langle \widetilde{F}(\mu)z, v_N\rangle, \tag{4.3.7}$$

- or the least-squares minimization problem

$$U_N(\mu) = \underset{\tilde{U}_N \in \mathcal{L}(Z,V_N)}{\operatorname{argmin}} \|\widetilde{A}(\mu)\tilde{U}_N - \widetilde{F}(\mu)\|^2_{Z\to W'}, \tag{4.3.8}$$

  where the $\|\cdot\|_{Z\to W'}$ norm classically denotes the norm on $\mathcal{L}(Z, W')$ given by

$$\forall F \in \mathcal{L}(Z, W'), \quad \|F\|_{Z\to W'} = \sup_{z\in Z} \frac{\|Fz\|_{W'}}{\|z\|_Z}. \tag{4.3.9}$$

In eqs. (4.3.7) and (4.3.8), $\widetilde{A}(\mu)$ (resp. $\widetilde{F}(\mu)$) denotes an affine approximation of $A(\mu)$ (resp. of $F(\mu)$). Recalling definitions 3 and 4, this means

$$\widetilde{A}(\mu) = \sum_{q=1}^{Q^a} \theta_q^a(\mu) A_q, \quad \widetilde{F}(\mu) = \sum_{q=1}^{Q^F} \theta_q^F(\mu) F_q, \tag{4.3.10}$$

with μ-independent $A_q \in \mathcal{L}(V, W')$, $1 \leqslant q \leqslant Q^a$ and μ-independent $F_q \in \mathcal{L}(Z, W')$, $1 \leqslant q \leqslant Q^F$.

Error estimates

Proposition 4.3.1 (Block RB error estimate)**.** *Let $V_N \subset V$ be a reduced basis approximation space and $U_N(\mu) \in \mathcal{L}(Z, V_N)$ be a reduced basis approximation of $U(\mu) \in \mathcal{L}(Z, V)$. Let*

$$\delta^F(\mu) = \|F(\mu) - \widetilde{F}(\mu)\|_{Z \to W'},$$
$$\delta_N^A(\mu) = \|(A(\mu) - \widetilde{A}(\mu))U_N(\mu)\|_{Z \to V}.$$

Then,

$$\|U(\mu) - U_N(\mu)\|_{Z \to V} \leqslant \frac{1}{\alpha(\mu)} \left(\|\widetilde{A}(\mu)U_N(\mu) - \widetilde{F}(\mu)\|_{Z \to W'} + \delta^F(\mu) + \delta_N^A(\mu) \right),$$

where $\alpha(\mu)$ denotes the inf-sup constant.

Proof. This is straightforward from

$$\begin{aligned} A(\mu)(U(\mu) - U_N(\mu)) &= F(\mu) - A(\mu)U_N(\mu) \\ &= \left(F(\mu) - \widetilde{F}(\mu)\right) + \left(\widetilde{F}(\mu) - \widetilde{A}(\mu)U_N(\mu)\right) \\ &+ \left(\widetilde{A}(\mu)U_N(\mu) - A(\mu)U_N(\mu)\right). \end{aligned} \tag{4.3.11}$$

One concludes by applying the triangular inequality and using the inf-sup condition. □

In practice, the affine approximations are chosen accurate enough so that $\delta^F(\mu)$ and $\delta_N^A(\mu)$ are very small compared to the target accuracy and can therefore be neglected in the error estimation. As already observed in [82], the affine approximation $\widetilde{A}(\mu)$ must match the operator $A(\mu)$ only on the RB approximation $U_N(\mu)$. Thus, in order to obtain a small $\delta_N^A(\mu)$, it is sufficient to guarantee that for all $v_N \in V_N$, $\widetilde{A}(\mu)v_N \approx A(\mu)v_N$ and there is no need to guarantee the stronger property $v \in V$, $\widetilde{A}(\mu)v \approx A(\mu)v$.

Let us define the residual operator $R(\mu) \in \mathcal{L}(Z, Z')$ as

$$R(\mu) = \left(\widetilde{A}(\mu)U_N(\mu) - \widetilde{F}(\mu)\right)^* R_W^{-1} \left(\widetilde{A}(\mu)U_N(\mu) - \widetilde{F}(\mu)\right), \tag{4.3.12}$$

where the asterisk denotes the adjoint. Recalling that $Z = Z' = \mathbb{C}^\ell$, we see that $R(\mu)$ is a hermitian positive-definite operator, whose spectrum is therefore a set of ℓ real positive numbers. Let us list the eigenvalues as $\lambda_1(\mu) \geqslant \lambda_2(\mu) \geqslant \cdots \geqslant \lambda_\ell(\mu)$ and associated eigenvectors as $z_1(\mu), \ldots, z_\ell(\mu)$, with entries repeated with multiplicity. From the min-max theorem, there holds

$$\lambda_1(\mu) = \sup_{z \in Z} \frac{\|\widetilde{A}(\mu)U_N(\mu)z - \widetilde{F}(\mu)z\|^2_{W'}}{\|z\|^2_Z} = \|\widetilde{A}(\mu)U_N(\mu) - \widetilde{F}(\mu)\|^2_{Z \to W'}. \qquad (4.3.13)$$

Note that the eigenvector $z_1(\mu)$ associated to the largest eigenvalue $\lambda_1(\mu)$ is the only possible direction considered in [128] and [114].

4.3.4 Greedy construction

Algorithm 4.2: Residual-driven RB generation

Input : Discrete training set $\Xi \subset \mathcal{D}_\mu$, target tolerance ϵ_{target} and maximum reduced basis size $N_{\max}$

Output: Reduced basis approximation space V_N

Pick arbitrarily μ_1 in Ξ;
Set $V_0 = \{0\}$, $d_1 = \ell$, $z_1^{(j)} = \hat{e}_j$, $1 \leqslant j \leqslant d_1$;
Initialize $I \leftarrow 1$, $N \leftarrow 0$;
while $N + d_I \leq N_{\max}$ *and* $\epsilon > \epsilon_{\text{target}}$ **do**
  Compute $U(\mu_I)z_I^{(j)}$ for all $1 \leqslant j \leqslant d_I$;
  Update reduced basis $V_{N+d_I} = V_N \oplus \text{Span}\{U(\mu_I)z_I^{(j)},\ 1 \leqslant j \leqslant d_I\}$;
  $N \leftarrow N + d_I$;
  Compute eigenvalues $\lambda_1(\mu) \geqslant \cdots \geqslant \lambda_\ell(\mu)$ and associated eigenvectors $z_1(\mu), \ldots, z_\ell(\mu)$ of the residual operator $R(\mu)$ for all $\mu \in \Xi$;
  $\mu_{I+1} \leftarrow \underset{\mu \in \Xi}{\text{argmax}}\ \sqrt{\lambda_1(\mu)}$;
  $\epsilon \leftarrow \sqrt{\lambda_1(\mu_{I+1})}$;
  Set d_{I+1} to be the largest integer i satisfying $\sqrt{\lambda_i(\mu_{I+1})} \geqslant \sqrt{\lambda_1(\mu_{I+1})}/100$;
  Set $z_{I+1}^{(j)} = z_j(\mu_{I+1})$ for $1 \leqslant j \leqslant d_{I+1}$;
  $I \leftarrow I + 1$;
end

To generate a reduced basis of the form eq. (4.3.6), we use the residual-driven greedy algorithm summarized by algorithm 4.2. During the first iteration $I = 1$, the reduced basis space V_N is initialized to a dimension $N = \ell$ as $V_N = \text{Span}\{U(\mu_1)\hat{e}_j,\ 1 \leqslant j \leqslant \ell\}$, where μ_1 is randomly selected in $\mathcal{D}_\mu$. The choice of the next parameter point μ_{I+1} to be considered is based on maximizing the residual $\mu \mapsto \|\widetilde{A}(\mu)U_N(\mu) - \widetilde{F}(\mu)\|_{Z \to W'}$ over a training set Ξ covering $\mathcal{D}_\mu$. The choice of the directions $z_{I+1}^{(j)}$, $1 \leqslant j \leqslant d_{I+1}$

to be considered is based on the eigenvectors associated to the largest eigenvalues of the residual operator $R(\mu_{I+1})$. These eigenvectors correspond to the directions $z \in Z$ in which the residual norm $\|\widetilde{A}(\mu_{I+1})U_N(\mu_{I+1})z - \widetilde{F}(\mu_{I+1})z\|_{W'}$ is maximal. We only choose to consider the leading directions, setting the number of eigenvectors d_{I+1} to be considered as the largest integer i satisfying the criterion

$$\sqrt{\lambda_i(\mu_{I+1})} \geqslant \frac{\sqrt{\lambda_1(\mu_{I+1})}}{100}. \tag{4.3.14}$$

We choose to formulate the criterion using the square roots of the eigenvalues, recalling from eq. (4.3.13) that the residual norm coincides with the square root of the largest eigenvalue.

Notice that the integer d_{I+1} can coincide with ℓ if the eigenvalues of the residual operator do not exhibit fast decay. This would mean that there are no preferred directions and the consequence would be that all directions are retained for the reduced basis. On the contrary, if this integer is small, then only a small number of directions contribute to the residual norm. In this situation, it is relevant to the require the next truth solves to be performed in these directions.

Heuristic approach

We briefly explain how the heuristic approach introduced in section 2.2.4 can be adapted to the present block situation.

At iteration $I \geqslant 2$, we compute

$$\hat{\alpha}_I = \frac{\|\widetilde{A}(\mu_I)U_N(\mu_I)z_I^{(1)} - \widetilde{F}(\mu_I)z_I^{(1)}\|_{W'}}{\|U(\mu_I)z_I^{(1)} - U_N(\mu_I)z_I^{(1)}\|_V}, \tag{4.3.15}$$

which corresponds to the ratio of the residual norm over the error norm at the point μ_I and in the direction $z_I^{(1)}$ (which is the leading direction of the residual). Computing $\hat{\alpha}_I$ is not expensive, as the residual norm is efficiently computable in $\mathcal{N}$-independent complexity and the solution $U(\mu_I)z_1^{(1)} \in V$ is computed anyway in this iteration of algorithm 4.2. We build the constant $\hat{\alpha}$ as the mean of samples $\{\hat{\alpha}_I\}_{2 \leqslant I \leqslant I_{\max}}$, where $I_{\max}$ denotes the number of iterations performed by algorithm 4.2. We propose to estimate the error using the heuristic

$$\|U(\mu) - U_N(\mu)\|_{Z \to V} \approx \frac{1}{\hat{\alpha}}\|\widetilde{A}(\mu)U_N(\mu) - \widetilde{F}(\mu)\|_{Z \to W'}. \tag{4.3.16}$$

4.3.5 Offline/Online strategy and complexity analysis

In order to analyze the computational complexity, we adopt an algebraic setting using the notations introduced in section 1.3.2. We have two bases at hand: $\{\phi_j^V\}_{1 \leqslant j \leqslant \mathcal{N}}$ for V and $\{\phi_i^W\}_{1 \leqslant i \leqslant \mathcal{N}}$ for W.

We recall that the operator $A(\mu) \in \mathcal{L}(V, W')$ is represented by the $\mathcal{N} \times \mathcal{N}$ matrix $\mathbf{A}(\mu)$ with entries $\langle A(\mu)\phi_j^V, \phi_i^W \rangle$, $1 \leqslant i, j \leqslant \mathcal{N}$. Similarly, the block right-hand side $F(\mu) \in \mathcal{L}(Z, W')$ is represented by the $\mathcal{N} \times \ell$ matrix $\mathbf{F}(\mu)$ with entries $\langle F(\mu)\hat{e}_j, \phi_i^W \rangle$, $1 \leqslant i \leqslant \mathcal{N}$, $1 \leqslant j \leqslant \ell$.

This being set, the solution $U(\mu) \in \mathcal{L}(Z, V)$ to the parametrized block problem eq. (4.3.4) is represented by the $\mathcal{N} \times \ell$ matrix $\mathbf{U}(\mu)$ satisfying

$$\mathbf{A}(\mu)\mathbf{U}(\mu) = \mathbf{F}(\mu). \tag{4.3.17}$$

Note that $\mathbf{U}(\mu) = [\mathbf{u}_1(\mu)|\cdots|\mathbf{u}_\ell(\mu)]$ where the rth column $\mathbf{u}_r(\mu) \in \mathbb{C}^{\mathcal{N}}$ holds the coordinates of $U(\mu)\hat{e}_r$ in the $\{\phi_j^V\}_{1 \leqslant j \leqslant \mathcal{N}}$ basis.

RB approximations

As usual in this thesis, the reduced basis is denoted $V_N = \text{Span}\{\xi_1, \ldots, \xi_N\}$, where the basis function $\xi_1, \ldots, \xi_N$ are orthonormal in the sense $(\xi_i, \xi_j)_V = \delta_{ij}$. Algebraically, the reduced basis is represented by the $\mathcal{N} \times N$ matrix $\mathbf{P} = [\mathbf{p}_1 \,|\, \cdots \,|\, \mathbf{p}_N]$, where the nth column $\mathbf{p}_n \in \mathbb{C}^{\mathcal{N}}$ corresponds to the coordinates of the basis function ξ_n in the $\{\phi_j^V\}_{1 \leqslant j \leqslant \mathcal{N}}$ basis.

In this context, the reduced basis approximation $U_N(\mu) \in \mathcal{L}(Z, V_N)$ defined by either the Galerkin problem eq. (4.3.7) or the least-squares minimization problem eq. (4.3.8) is represented by a $N \times \ell$ matrix $\mathbf{X}_N(\mu) = [\mathbf{x}_1(\mu) \,|\, \cdots \,|\, \mathbf{x}_\ell(\mu)]$, where the rth column $\mathbf{x}_r(\mu) \in \mathbb{C}^N$ holds the coordinates of $U_N(\mu)\hat{e}_r$ in the reduced basis $\{\xi_n\}_{1 \leqslant n \leqslant N}$.

In detail, the solution $\mathbf{X}_N(\mu) \in \mathbb{C}^{N \times \ell}$ to the Galerkin problem eq. (4.3.7) satisfies

$$\mathbf{P}^* \widetilde{\mathbf{A}}(\mu)\mathbf{P}\mathbf{X}_N(\mu) = \mathbf{P}^* \widetilde{\mathbf{F}}(\mu). \tag{4.3.18}$$

An efficient offline/online decoupling can be achieved analogously to proposition 1.3.1. Namely, if the matrices $\mathbf{P}^*\mathbf{A}_q\mathbf{P}$, $1 \leqslant q \leqslant Q^a$ (each of size $N \times N$) and the matrices $\mathbf{P}^*\mathbf{F}_q$, $1 \leqslant q \leqslant Q^F$ (each of size $N \times \ell$) are pre-computed offline, then the linear system (4.3.18) can be assembled for any value of μ with $\mathcal{O}(N^2 Q^a + N\ell Q^F)$ operations.

When the reduced basis approximation is defined by least-squares minimization problem eq. (4.3.8), the solution $\mathbf{X}_N(\mu) \in \mathbb{C}^{N \times \ell}$ satisfies

$$\mathbf{P}^* \widetilde{\mathbf{A}}(\mu)^* \mathbf{B}_W^{-1} \widetilde{\mathbf{A}}(\mu)\mathbf{P}\mathbf{X}_N(\mu) = \mathbf{P}^* \widetilde{\mathbf{A}}(\mu)^* \mathbf{B}_W^{-1} \widetilde{\mathbf{F}}(\mu). \tag{4.3.19}$$

Again, an efficient offline/online decoupling can be achieved analogously to proposition 1.3.2. If the matrices $\mathbf{P}^*{\mathbf{A}_q}^*\mathbf{B}_W^{-1}\mathbf{A}_p\mathbf{P}$, $1 \leqslant q, p \leqslant Q^a$ (each of size $N \times N$) and the matrices $\mathbf{P}^*{\mathbf{A}_p}^*\mathbf{B}_W^{-1}\mathbf{F}_q$, $1 \leqslant p \leqslant Q^a$, $1 \leqslant q \leqslant Q^F$ (each of size $N \times \ell$) are pre-computed offline, then the linear system (4.3.19) can be assembled for any value of μ with $\mathcal{O}(N^2(Q^a)^2 + N\ell Q^a Q^F)$ operations.

Solving either (4.3.18) or (4.3.19) requires $\mathcal{O}(N^3)$ to compute the LU factorization of the $N \times N$ system matrix and $\mathcal{O}(\ell N^2)$ for the ℓ forward-backward triangular processes applied to each of the ℓ right-hand sides (see [75, §3.10]).

Residual operator

The residual operator $R(\mu) \in \mathcal{L}(Z, Z')$ defined by eq. (4.3.12) is algebraically represented by the $\ell \times \ell$ residual matrix

$$\mathbf{R}(\mu) = \left(\widetilde{\mathbf{A}}(\mu)\mathbf{P}\mathbf{X}_N(\mu) - \widetilde{\mathbf{F}}(\mu)\right)^* \mathbf{B}_W^{-1} \left(\widetilde{\mathbf{A}}(\mu)\mathbf{P}\mathbf{X}_N(\mu) - \widetilde{\mathbf{F}}(\mu)\right). \tag{4.3.20}$$

An efficient offline/online decoupling can be achieved by adapting proposition 2.2.2. If the matrices $\mathbf{P}^*{\mathbf{A}_q}^*\mathbf{B}_W^{-1}\mathbf{A}_p\mathbf{P}$, $1 \leqslant q, p \leqslant Q^a$ (each of size $N \times N$), the matrices $\mathbf{P}^*{\mathbf{A}_p}^*\mathbf{B}_W^{-1}\mathbf{F}_q$, $1 \leqslant p \leqslant Q^a$, $1 \leqslant q \leqslant Q^F$ (each of size $N \times \ell$) and the matrices ${\mathbf{F}_p}^*\mathbf{B}_W^{-1}\mathbf{F}_q$, $1 \leqslant p, q \leqslant Q^F$ (each of size $\ell \times \ell$) are pre-computed offline, then the residual matrix $\mathbf{R}(\mu)$ can be assembled for any value value of μ with $\mathcal{O}(N^2(Q^a)^2 + N\ell Q^a Q^F + \ell^2(Q^F)^2)$ operations.

The eigenvalue decomposition of the residual matrix can be performed in $\mathcal{O}(\ell^3)$ using for instance the QR algorithm [90].

4.4 The parametrized problem with parametrized source

All the strategies presented in this chapter so far are concerned with the parametrized problem with ℓ independent source terms eq. (4.1.1), which can alternatively be expressed under the block form (4.3.4). In this section, we show that the proposed methods can also be applied to a different class of problems, namely, to the class of parametrized problems with parametrized source.

4.4.1 Problem formulation

In addition to the parameter $\mu \in \mathcal{D}_\mu$, let us introduce a new parameter $\nu \in \mathcal{D}_\nu$ where $\mathcal{D}_\nu$ denotes a compact set $\mathbb{R}^{p_\nu}$, $p_\nu \geqslant 1$. We now assume that the μ-parametrized right-hand side $f(\mu) \in W'$ further depends on this parameter ν. Thus, we have to deal with a (μ, ν)-parametrized right-hand side $f(\mu, \nu) \in W'$. However, this additional parameter ν does not affect the operator, which continues to depend only on μ. We are interested in the solutions to the problem: *find* $u(\mu, \nu) \in V$ *such that*

$$A(\mu)u(\mu, \nu) = f(\mu, \nu) \quad \text{in } W'. \tag{4.4.1}$$

We call this problem the *parametrized problem with parametrized source*.

Model problem

As model problem, let us again consider the Laplace model problem introduced in chapter 1. Rather than the right-hand side $f : w \in W \mapsto \int_\Omega S w \mathrm{d}\Omega$ with non-parametrized

source term $S \in L^2(\Omega)$ given by eq. (1.4.6), we now set $p_\nu = 2$, $\mathcal{D}_\nu = [10,\ 12] \times [0.2,\ 0.4]$ and consider the following (μ, ν)-parametrized source term

$$\forall x \in \Omega, \quad S(x; \mu, \nu) = S_{\mathcal{G}}(x; \mu, \nu_1) S_{\mathcal{C}}(x; \mu, \nu_2), \tag{4.4.2}$$

where $S_{\mathcal{G}}$ is the Gaussian centered on $\mu = (\mu_1, \mu_2)$ defined by

$$S_{\mathcal{G}}(x; \mu, \nu_1) = \exp\left(-\nu_1 \frac{(x_1 - \mu_1)^2 + (x_2 - \mu_2)^2}{0.01}\right), \tag{4.4.3}$$

and where $S_{\mathcal{C}}$ is the wave function defined by

$$S_{\mathcal{C}}(x; \mu, \nu_2) = \cos\left(\nu_2 \left(x_1 \cos\phi(\mu) + x_2 \sin\phi(\mu)\right)\right), \tag{4.4.4}$$

with angle $\phi(\mu)$ given by

$$\cos\phi(\mu) = \frac{\mu_1}{\sqrt{\mu_1^2 + \mu_2^2}}, \quad \sin\phi(\mu) = \frac{\mu_2}{\sqrt{\mu_1^2 + \mu_2^2}}. \tag{4.4.5}$$

Recalling that $\mu \in \mathcal{D}_\mu = [0.4, 0.6] \times [0.4, 0.6]$, one always has $\sqrt{\mu_1^2 + \mu_2^2} > 0$ thus the formulas eq. (4.4.5) hold true for all $\mu \in \mathcal{D}_\mu$. We draw attention to the fact that the wave function eq. (4.4.3) can be viewed as $S_{\mathcal{C}}(x; \mu, \nu_2) = \Re\{e^{i\nu_2 x \cdot \hat{\mu}}\}$ where i denotes the imaginary number, $\hat{\mu} = \mu / \sqrt{\mu_1^2 + \mu_2^2}$ is the normalized vector μ and $\cdot$ denotes the dot product.

It is worth noting that the parameter ν_1 affects the spreading of the Gaussian eq. (4.4.3), while the parameter ν_2 controls the frequency of the wave function eq. (4.4.4). The role of the parameter μ is twofold: not only does it control the location of the peak of the Gaussian, but it also defines the direction of the wave. Figure 4.2 gives an idea of the variety of truth solutions which can be obtained with such a parametrized source.

Real-world problem

A real-world application (not dealt with in this thesis) can be found in the context of acoustic or electromagnetic scattering applications. In this context, we want to solve the scattering of an incident plane wave across a frequency band (hence the parameter μ is the frequency as in chapters 5 and 6). The plane wave depends not only on the frequency μ, but also on its direction (even on its polarization in the case of an electromagnetic plane wave). Thus the additional parameter ν could represent the direction of the plane wave. The goal would be to resolve the scattering for all frequencies and for all directions of the incident plane waves, as is done in [39].

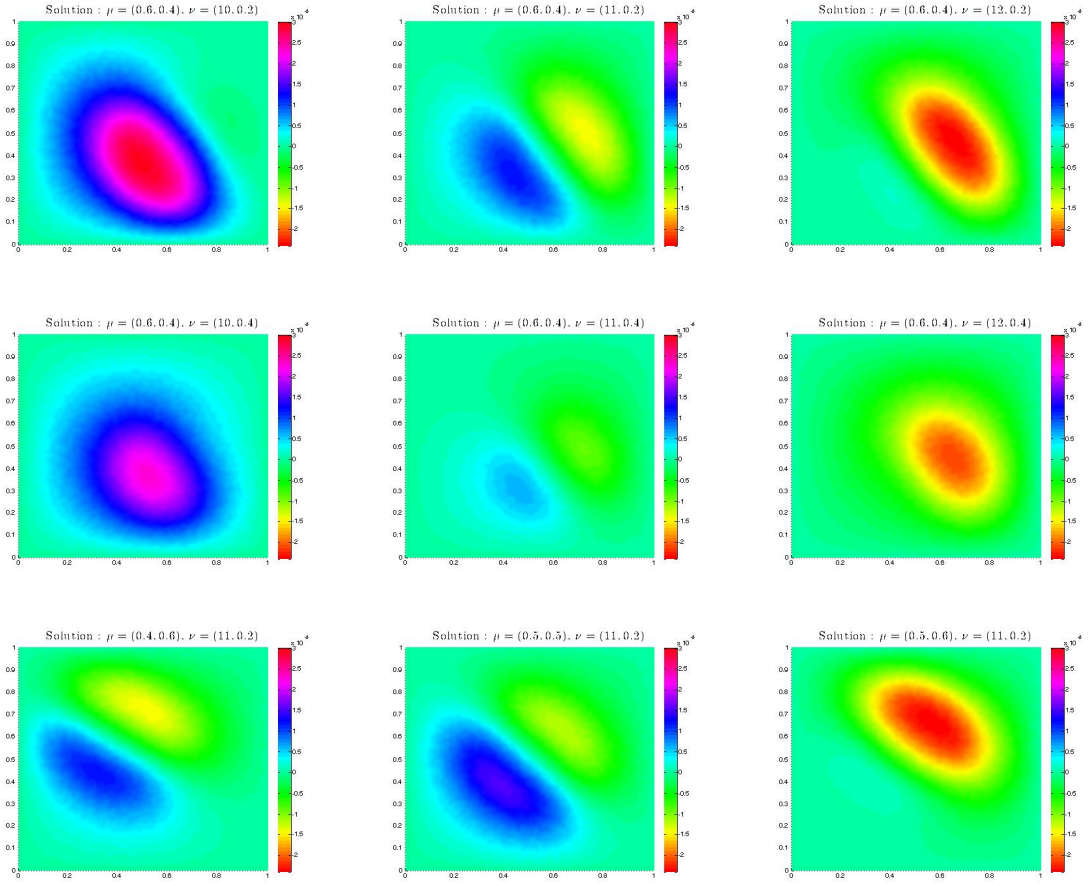


Figure 4.2: Truth solutions $u(\mu, \nu)$. First row: fixed $\mu = (0.6, 0.4)$, $\nu_2 = 0.1$ and three ν_1 configurations $\nu_1 = 10, 11, 12$. Second row: fixed $\mu = (0.6, 0.4)$, $\nu_2 = 0.4$ and three ν_1 configurations $\nu_1 = 10, 11, 12$. Third row: fixed $\nu = (11, 0.2)$ and three configurations $\mu = (0.4, 0.6), (0.5, 0.5), (0.5, 0.6)$.

4.4.2 Block approximation

In this work, we restrict ourselves to (μ, ν)-parametrized right-hand sides $f(\mu, \nu) \in W'$ given by

$$f(\mu, \nu) : w \in W \mapsto \int_\Omega S(x; \mu, \nu) w(x) \mathrm{d}x, \tag{4.4.6}$$

where $S(\cdot; \mu, \nu) \in L^2(\Omega)$ is a (μ, ν)-parametrized function (for example eq. (4.4.2)). In this situation, we claim that the parametrized problem with parametrized source eq. (4.4.1) can be approximated by a parametrized problem with ℓ independent source terms eq. (4.1.1), or equivalently by block problem (4.3.4). Thus, all the reduced basis strategies presented in this chapter can be applied.

To prove our claim, we apply the EIM [5] to the ν-parametrized function $S(\cdot; \cdot, \nu)$ of

the (x, μ) variable in $\Omega \times \mathcal{D}_\mu$ in order to represent it under an approximate linear way. Denoting ℓ the number of EIM steps (see algorithm 1.1), we obtain a set of ℓ parameter points $\{\nu^r\}_{1 \leqslant r \leqslant \ell}$, a set of ℓ interpolation points $\{(x^r, \mu^r)\}_{1 \leqslant r \leqslant \ell}$, a $\ell \times \ell$ interpolation matrix B and EIM basis functions defined on $\Omega \times \mathcal{D}_\mu$ denoted h_r, $1 \leqslant r \leqslant \ell$. The EIM approximant writes

$$\mathcal{I}_\ell S(x; \mu, \nu) = \sum_{r=1}^{\ell} z'_r(\nu) h_r(x, \mu), \tag{4.4.7}$$

with coefficients $z'(\nu) = (z'_1(\nu), \dots, z'_\ell(\nu))^T$ solution to the $\ell \times \ell$ linear system $Bz'(\nu) = \phi(\nu)$, with right-hand side $\phi(\nu) = (S(x^1; \mu^1, \nu), \dots, S(x^\ell; \mu^\ell, \nu))^T$. Using the $\ell \times \ell$ change-of-basis matrix C and the change-of-basis relationship $z'(\nu) = Cz(\nu)$, we can easily obtain the coefficients $z(\nu) = (z_1(\nu), \dots, z_\ell(\nu))^T$ such that the EIM approximant is expressed in the so-called *non-intrusive* way [23] as

$$\mathcal{I}_\ell S(x; \mu, \nu) = \sum_{r=1}^{\ell} z_r(\nu) S(x; \mu, \nu^r). \tag{4.4.8}$$

Next, we set $Z = \mathbb{R}^\ell$ (or $Z = \mathbb{C}^\ell$ in a complex setting) and define the μ-parametrized block right-hand side $F(\mu) \in \mathcal{L}(Z, W')$ as

$$\forall z = (z_1, \dots, z_\ell) \in Z, \quad F(\mu)z : w \in W \mapsto \sum_{r=1}^{\ell} z_r \int_\Omega S(x; \mu, \nu^r) w(x) \mathrm{d}x. \tag{4.4.9}$$

Intuitively, if $\mathcal{I}_\ell S(\cdot; \mu, \nu) \approx S(\cdot; \mu, \nu)$, then $F(\mu)z(\nu) \approx f(\mu, \nu)$. In this situation, we also have $U(\mu)z(\nu) \approx u(\mu, \nu)$ as a consequence of the Banach-Nečas-Babuška assumption. Indeed, we have the following proposition.

Proposition 4.4.1 (Block approximation of the (μ, ν)-parametrized solution)**.** *Let $u(\mu, \nu) \in V$ denote the solution to the parametrized problem with parametrized source* (4.4.1) *and $U(\mu) \in \mathcal{L}(Z, V)$ denote the solution to the parametrized block problem* (4.3.4) *with block right-hand side given by eq.* (4.4.9)*. Then, there holds*

$$\|u(\mu, \nu) - U(\mu)z(\nu)\|_V \leqslant \frac{1}{\alpha(\mu)} \|f(\mu, \nu) - F(\mu)z(\nu)\|_{W'},$$

where $\alpha(\mu)$ denotes the inf-sup constant.

This shows that, if the EIM approximation is adequately good, then the parametrized problem with parametrized source (4.4.1) can be well approximated by a parametrized block problem.

4.4.3 Affine approximation of the block RHS

As seen in section 4.3.5, the affine approximation $\widetilde{F}(\mu)$ of the block right-hand side $F(\mu)$ is key to the success of the offline/online computational strategy. We now give some

insight on how this approximation can be obtained.

We propose to again apply the EIM [5], this time to the μ-parametrized function $S(\cdot;\mu,\cdot)$ of the (x,ν) variable in $\Omega \times \mathcal{D}_\nu$. Let us denote M the number of EIM steps, $H_m(\cdot,\cdot)$, $1 \leqslant m \leqslant M$ the EIM basis functions defined on $\Omega \times \mathcal{D}_\nu$. The EIM constructs a set of points $\{\mu_m\}_{1\leqslant m\leqslant M}$, such that the subspace spanned by the M EIM basis functions coincides with the subspace spanned by the $S(\cdot;\mu_m,\cdot)$, $1 \leqslant m \leqslant M$. Thus the EIM approximant writes

$$\mathcal{I}_M S(x;\mu,\nu) = \sum_{m=1}^{M} t_m(\mu) S(x;\mu_m,\nu). \tag{4.4.10}$$

Recalling from section 4.4.2 that $S \approx \mathcal{I}_\ell S$ where $\mathcal{I}_\ell S$ is given by eq. (4.4.8), we propose to plug the approximation $S(\cdot;\mu_m,\cdot) \approx \mathcal{I}_\ell S(\cdot;\mu_m,\cdot)$ in eq. (4.4.10). This yields

$$S(x;\mu,\nu) \approx \sum_{r=1}^{\ell}\sum_{m=1}^{M} t_m(\mu) z_r(\nu) S(x;\mu_m,\nu^r). \tag{4.4.11}$$

This being set, we define our approximation $\widetilde{F}(\mu)$ of the μ-parametrized block right-hand side $F(\mu)$ defined by eq. (4.4.9) as

$$\forall z = (z_1,\ldots,z_\ell) \in Z,$$
$$\widetilde{F}(\mu)z : w \in W \mapsto \sum_{m=1}^{M} t_m(\mu) \sum_{r=1}^{\ell} z_r \int_\Omega S(x;\mu_m,\nu^r) w(x) \mathrm{d}x. \tag{4.4.12}$$

Observe that the quality of approximation in $\widetilde{F}(\mu)z(\nu) \approx f(\mu,\nu)$ is directly related to the combination of both EIM approximation errors (in ν and in μ) through eq. (4.4.11).

Clearly, $\widetilde{F}(\mu)$ is clearly affine in the sense of eq. (4.3.10) with $Q^F = M$ terms. Furthermore, the mth μ-independent term $F_m \in \mathcal{L}(Z,W')$ is given by

$$\forall z = (z_1,\ldots,z_\ell) \in Z, \quad F_m z : w \in W \mapsto \sum_{r=1}^{\ell} z_r \int_\Omega S(x;\mu_m,\nu^r) w(x) \mathrm{d}x. \tag{4.4.13}$$

Observing that $F_m z : w \in W \mapsto \sum_{r=1}^{\ell} z_r f(\mu_m,\nu^r)$, the μ-independent term $F_m \in \mathcal{L}(Z,W')$ can be easily assembled by assembling all the right-hand sides $f(\mu_m,\nu^r)$, for all $1 \leqslant r \leqslant \ell$. Hence the present approach is non-intrusive [23], as it does not involve any other assembly routine than the assembly routine for the right-hand side.

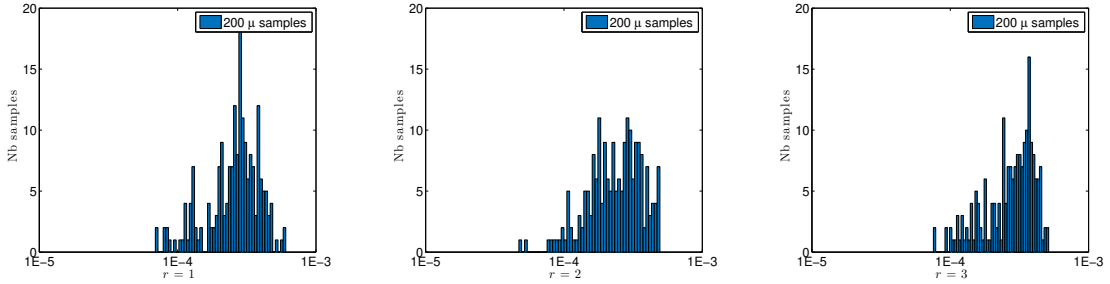
4.5 Numerical illustration

4.5.1 The parametrized problem with ℓ independent source terms

We consider the Laplace model problem with the $\ell = 4$ source terms given by eq. (4.1.2). The truth solutions plotted on fig. 4.1 confirm that $u_r(\mu)$ is completely different from $u_p(\mu)$, $p \neq r$. This motivates the choice of building ℓ distinct reduced basis approximation subspaces following the multiple RBs approach explained in section 4.2.

We run the greedy algorithm 4.1 setting the target to 1×10^{-5}. This takes $N = 21$ iterations. Overall, $N\ell = 84$ truth solutions are computed in the process, but thanks to the use of a unique set of parameter points $\{\mu_i\}_{1 \leqslant n \leqslant N}$ common to the ℓ RBs, the solver $A(\mu)^{-1}$ is called only $N = 21$ times, each time to solve $\ell = 4$ right-hand sides.

For validation, for each $r = 1, \ldots, \ell$ we compute 200 truth solutions $u_r(\mu)$ at random points in $\mathcal{D}_\mu$ and determine the relative approximation error $\|u_r(\mu) - u_{r,N}(\mu)\|_V / \|u_r(\mu)\|_V$. This means that overall, $200 \times 4 = 800$ truth solutions are computed. The distributions for $r = 1, 2, 3$ are shown on fig. 4.3.



(a) Distribution $r = 1$. (b) Distribution $r = 2$. (c) Distribution $r = 3$.

Figure 4.3: Distributions of the relative error $\|u_r(\mu) - u_{r,N}(\mu)\|_V / \|u_r(\mu)\|_V$, for 200 samples of μ.

The three distributions are much alike and confirm that the accuracy is relatively independent from the choice of the source term (indexed by r). In the present example, we are able to approximate the truth solution associated to any source with a relative error comprised between 0.1% and 0.01%. Furthermore, using our heuristic method based on the samples $\hat{\alpha}_N$ (see eq. (4.2.6)), we are able to obtain the correct order of magnitude of the relative error in a fully *a posteriori* manner, with effectivities very close to 1 (not shown here).

4.5.2 The parametrized problem with parametrized source

We now turn to the Laplace model problem with (μ, ν)-parametrized source given by eq. (4.4.2). Before reduced basis strategies can be applied, we must apply the EIM twice:

a first EIM that views ν as the varying parameter to obtain the block approximation and a second EIM that view μ as the varying parameter to obtain the affine approximation of the block right-hand side as explained above.

Block & affine approximation of the RHS

In order to obtain the block approximation, we apply the EIM to the ν-parametrized function $S(\cdot;\cdot,\nu)$ defined on $\Omega \times \mathcal{D}_\mu$ given by eq. (4.4.2). For the EIM, we discretize the set $\Omega \times \mathcal{D}_\mu$ for the (x, μ) variable as follows:

- $\Omega =]0,1[^2$ is discretized using an unstructured set of vertices $\Xi_x \subset \Omega$ with $n_x = 3436$ points;
- $\mathcal{D}_\mu = [0.4,\ 0.6]^2$ is discretized using a 8×8 uniform cartesian grid $\Xi_\mu \subset \mathcal{D}_\mu$, whose cardinality is $n_\mu = 64$.

Thus, $\Omega \times \mathcal{D}_\mu$ is discretized using $\Xi_x \times \Xi_\mu$ of cardinality $n_x n_\mu = 219,904$, each of these $n_x n_\mu$ point being a 4-dimensional vector of the form (x_1, x_2, μ_1, μ_2). Next, the set $\mathcal{D}_\nu = [10,\ 12] \times [0.2,\ 0.4]$ for the ν variable is discretized using a 8×8 uniform catesian grid $\Xi_\nu \subset \mathcal{D}_\nu$, whose cardinality is $n_\nu = 64$.

We build the EIM approximant $\mathcal{I}_\ell S(\cdot;\cdot,\nu)$ given by eq. (4.4.8). The EIM error ϵ_ℓ is computed as

$$\epsilon_\ell = \max_{\nu \in \Xi_\nu} \max_{(x,\mu) \in \Xi_x \times \Xi_\mu} \left| S(x;\mu,\nu) - \mathcal{I}_\ell S(x;\mu,\nu) \right|. \tag{4.5.1}$$

We show on fig. 4.4 the evolution of this quantity throughout the iterations of EIM. The algorithm is stopped as soon as the prescribed tolerance $tol = 5 \times 10^{-4}$ is reached on the EIM error. As can be seen on fig. 4.4, this takes $\ell = 17$ iterations.

Next, in order to obtain the affine approximation of the block right-hand side, we apply the EIM to the μ-parametrized function $S(\cdot;\mu,\cdot)$ defined on $\Omega \times \mathcal{D}_\nu$ given by eq. (4.4.2). Applying the EIM requires discretized surrogates of $\Omega \times \mathcal{D}_\nu$. We use the following:

- $\Omega =]0,1[^2$ is discretized using the same unstructured set of vertices $\Xi_x \subset \Omega$ as the one used to obtain the block approximation;
- $\mathcal{D}_\nu$ is discretized using the $\ell = 17$ points $\{\nu^r\}_{1 \leqslant r \leqslant \ell}$ selected in $\Xi_\nu \subset \mathcal{D}_\mu$ by the EIM applied to the ν-parameterized function $S(\cdot;\cdot,\nu)$.

Thus, $\Omega \times \mathcal{D}_\nu$ is discretized using $\Xi_x \times \{\nu^r\}_{1 \leqslant r \leqslant \ell}$ of cardinality $\ell n_x = 58,412$, each of these ℓn_x point being a 4-dimensional vector of the form (x_1, x_2, ν_1, ν_2). As for the μ variable, it is discretized using the 8×8 uniform cartesian grid $\Xi_\mu \subset \mathcal{D}_\mu$ already introduced. The EIM approximation error is defined as

$$\epsilon_M = \max_{\mu \in \Xi_\mu} \max_{(x,\mu) \in \Xi_x \times \{\nu^r\}_{1 \leqslant r \leqslant \ell}} \left| S(x;\mu,\nu) - \mathcal{I}_M S(x;\mu,\nu) \right|, \tag{4.5.2}$$

where $\mathcal{I}_M S$ is the EIM approximant defined by eq. (4.4.10). We show on fig. 4.4 the evolution of this quantity throughout the iterations. The algorithm is stopped as soon as

the prescribed tolerance $tol = 5 \times 10^{-4}$ is reached on the EIM error. As can be seen on fig. 4.4, this takes $M = 30$ iterations. Compared to the first EIM applied to $S(\cdot;\cdot,\nu)$, the convergence is much slower. In fact, the convergence curve resembles that obtained for the Gaussian conductivity alone (c.f. chapter 1, fig. 1.3).

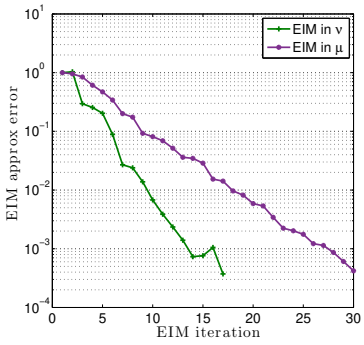


Figure 4.4: Convergence curves of the EIM applied to $S(\cdot;\cdot,\nu)$ (green curve) and to $S(\cdot;\mu,\cdot)$ (purple curve) to reach the prescribed tolerance $tol = 5 \times 10^{-4}$.

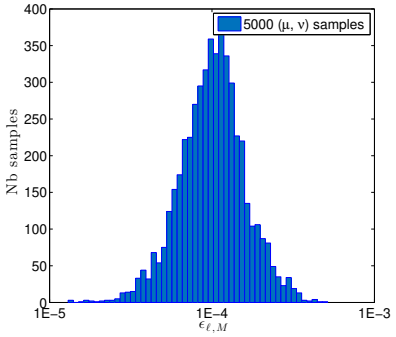


Figure 4.5: Distribution of the block+affine approximation error (4.5.3) for 5000 random (μ,ν) samples in $\mathcal{D}_\mu \times \mathcal{D}_\nu$.

In order to assess the overall quality of approximation in eq. (4.4.11); we take into account both the block approximation error (4.5.1) and the affine approximation error (4.5.2) by defining for all $(\mu,\nu) \in \mathcal{D}_\mu \times \mathcal{D}_\nu$ the block+affine approximation error as

$$\epsilon_{\ell,M}(\mu,\nu) = \max_{x\in\Xi_x} \left| S(x;\mu,\nu) - \mathcal{I}_{\ell,M} S(x;\mu,\nu) \right|, \tag{4.5.3}$$

where $\mathcal{I}_{\ell,M} S$ is given by

$$\mathcal{I}_{\ell,M} S(x;\mu,\nu) = \sum_{r=1}^{\ell} \sum_{m=1}^{M} t_m(\mu) z'_r(\nu) S(x;\mu_m,\nu_r). \tag{4.5.4}$$

We compute this block+affine approximation error for 5000 random (μ,ν) samples in $\mathcal{D}_\mu \times \mathcal{D}_\nu$ and plot the distribution on fig. 4.5. It is worth noting that the sample with the maximum error has an error 5.1518×10^{-4} which only slightly above the prescribed EIM tolerance $tol = 5 \times 10^{-4}$. This confirms that our choices of surrogate sets were adequate.

Construction of multiple RBs

To begin with, we apply the multiple RBs strategy by running algorithm 4.1 with a train set $\Xi \subset \mathcal{D}_\mu$ to a 33×33 uniform grid discretizing $\mathcal{D}_\mu = [0.4,\ 0.6]^2$ with target tolerance $\epsilon_{\text{target}} = 10^{-4}$. This constructs $\ell = 17$ distinct RBs all of the same size N, based on a unique set of selected parameter points $\{\mu_n\}_{1\leqslant n\leqslant N}$. We recall that the rth RB is given

by $V_N^{(r)} = \mathrm{Span}\{u_r(\mu_1), \ldots, u_r(\mu_N)\}$, c.f. eq. (4.2.2). In the present context, the truth solution $u_r(\mu)$ corresponds to the unique solution in V to $A(\mu)u_r(\mu) = f_r(\mu)$ where the right-hand side is given by $f_r(\mu) = f(\mu, \nu^r)$, where ν^r is the rth interpolation point selected by the EIM applied to $S(\cdot;\cdot,\nu)$.

For completeness, we recall that $f_r(\mu)$ is not affine, but that it can be adequately approximated by its affine approximation $f_r(\mu) \approx \widetilde{f}_r(\mu) = \sum_{m=1}^{M} t_m(\mu) f(\mu_m, \nu^r)$, with the coefficients $\{t_m(\mu)\}_{1\leqslant m\leqslant M}$ provided by the EIM applied to $S(\cdot;\mu,\cdot)$. Similarly, $A(\mu)$ is non-affine, but it is replaced by an efficient affine approximation with $Q^a = 27$ terms (we refer the reader to the first chapter of this thesis, section 1.4.2).

We find that $N = 19$ is the required basis size for algorithm 4.1 to converge to the prescribed tolerance, which means that the solver $A(\mu)^{-1}$ is called only 19 throughout algorithm 4.1, with $\ell = 17$ right-hand sides per solve. The convergence curve shown on fig. 4.6 exhibits exponential decay.

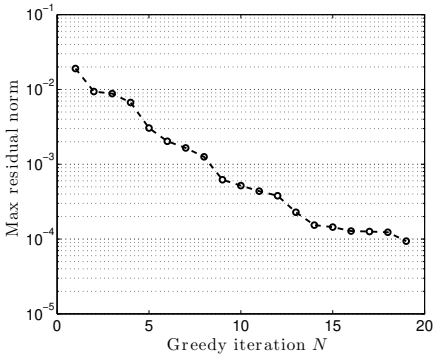


Figure 4.6: $\max_{(\mu,r)\in\Xi\times\{1,\ldots,\ell\}} \|\widetilde{A}(\mu)u_{r,N}(\mu) - \widetilde{f}_r(\mu)\|_{W'}$ throughout the greedy iterations.

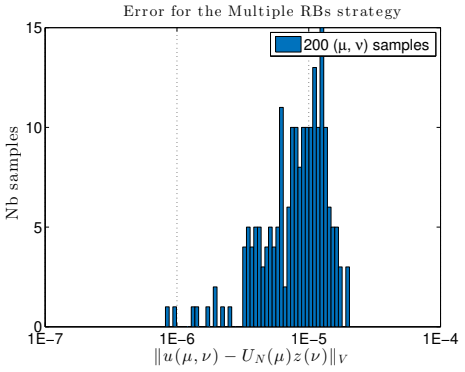


Figure 4.7: The actual RB approximation using the $\ell = 17$ RBs of size $N = 19$ error for 200 samples of $(\mu, \nu) \in \mathcal{D}_\mu \times \mathcal{D}_\nu$.

For validation purposes, we compute 200 truth solutions $u(\mu, \nu)$ at some parameters (μ, ν) randomly sampled in $\mathcal{D}_\mu \times \mathcal{D}_\nu$. Thus, we are able to compute 200 samples of the actual error $\|u(\mu,\nu) - U_N(\mu)z(\nu)\|_V$, where $U_N(\mu) \in \mathcal{L}(Z, V_N)$ is the RB approximation defined by

$$U_N(\mu) : z = (z_1, \ldots, z_\ell) \in Z \mapsto \sum_{r=1}^{\ell} z_r u_{r,N}(\mu). \tag{4.5.5}$$

To compute $U_N(\mu)z(\nu)$, one proceeds as follows:

(i) for all $1 \leqslant r \leqslant \ell$, compute the RB approximation $u_{r,N}(\mu) = \sum_{i=1}^{N} \mathbf{x}_{ri}(\mu)\xi_i^{(r)}$, where $V_N^{(r)} = \mathrm{Span}\{\xi_i^{(r)}, i = 1 \ldots, N\}$ denotes the rth reduced basis and where the RB coefficients $(\mathbf{x}_{r1}(\mu), \ldots, \mathbf{x}_{rN}(\mu)) \in \mathbb{C}^N$ are obtained by solving a $N \times N$ linear system,

(ii) perform the linear combination $U_N(\mu)z(\nu) = \sum_{r=1}^{\ell} z_r(\nu)u_{r,N}(\mu)$.

It is worth noting that (i) requires the resolution of ℓ linear systems of size $N \times N$, which amounts to $\mathcal{O}(\ell N^3)$ operations (*i.e.*, the dominant cost here comes from having to compute ℓ distinct LU factorizations). We further emphasize that because each $u_{r,N}(\mu)$ is a linear combination of N RB basis functions, the RB approximation $U_N(\mu)z(\nu)$ is in fact a linear combination of $N\ell = 323$ RB basis functions.

We plot the distribution of error on fig. 4.7. This confirms that the achieved level of absolute error is below 2×10^{-5}. We shall further comment on this distribution of the error when we will compare the *multiple RBs* strategy with the *unique RB* strategy.

Construction of a unique RB

We now test the unique RB stragy, by running algorithm 4.2 setting the train set $\Xi_\mu^{\mathtt{rb}}$ to a 33×33 uniform grid discretizing $\mathcal{D}_\mu = [0.4,\, 0.6]^2$ with target tolerance $\epsilon_{\text{target}} = 10^{-4}$. We name this run the *default init 100*. We run a variant of algorithm 4.2 called *default init 1000*, in which we change the usual criterion (4.3.14), setting the number of retained directions to the maximum integer i such that $\sqrt{\lambda_i(\mu^\star)} \geqslant \sqrt{\lambda_1(\mu^\star)}/1000$.

We ask whether it is relevant to consider all ℓ directions $z_1^{(j)} = \hat{e}_j$, $j = 1, \ldots, \ell$ in the first Greedy iteration. We propose a variant of algorithm 4.2 in which the initialization of the reduced basis is revisited:

- for all $\mu \in \Xi_\mu^{\mathtt{rb}}$, we consider the largest eigenvalue $\lambda_1(\mu)$ of the operator $R(\mu)$ with the zero reduced basis approximation $U_N(\mu) = 0$, this means $R(\mu) = \widetilde{F}(\mu)^* R_W^{-1} \widetilde{F}(\mu)$;
- we find $\mu^\star = \underset{\mu \in \Xi_\mu^{\mathrm{rb}}}{\operatorname{argmax}} \lambda_1(\mu)$ and consider the full spectrum of $R(\mu^\star)$, denoted $\lambda_1(\mu^\star) \geqslant \lambda_2(\mu^\star) \geqslant \cdots \geqslant \lambda_\ell(\mu^\star)$;
- we next determine d_1 using the criterion (4.3.14), *i.e.*, d_1 is the maximum integer i such that $\sqrt{\lambda_i(\mu^\star)} \geqslant \sqrt{\lambda_1(\mu^\star)}/100$. The d_1 initial directions $z_1^{(j)}$, $j = 1, \ldots, d_1$ are set to be the d_1 leading eigenvectors of $R(\mu^\star)$.

We name the run of this variant of algorithm 4.2 the *adaptive init 100*. We run another two variants called *adaptive init 10* and *adaptive init 1000* respectively, in which we change the usual criterion (4.3.14), setting the number of retained directions to the maximum integer i such that $\sqrt{\lambda_i(\mu^\star)} \geqslant \sqrt{\lambda_1(\mu^\star)}/10$ (resp. $\sqrt{\lambda_i(\mu^\star)} \geqslant \sqrt{\lambda_1(\mu^\star)}/1000$). We use this modified criterion for both the initialization phase and all the greedy iterations.

Finally, we run a last variant of algorithm 4.2, the *adaptive init 1*, which consists in setting the number of directions to be considered in the initialization and at each greedy iteration to 1, *i.e.* only the leading direction is considered. This corresponds to the original enrichment strategy proposed in [128] and [114].

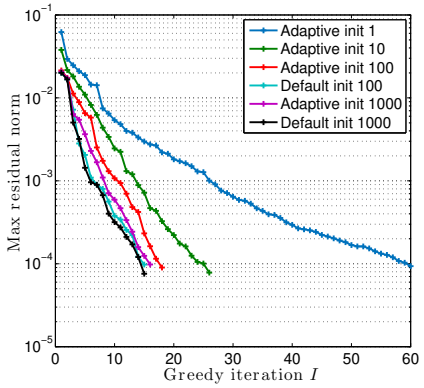


Figure 4.8: The quantity $\max_{\mu \in \Xi_\mu^{\mathrm{rb}}} \sqrt{\lambda_1(\mu)}$ throughout the greedy iterations.

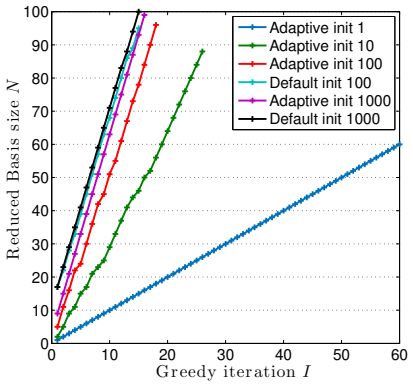


Figure 4.9: The reduced basis size throughout the greedy iterations.

Figure 4.8 shows the decrease of the maximum residual norm (which corresponds to $\max_{\mu \in \Xi_\mu^{\mathrm{rb}}} \sqrt{\lambda_1(\mu)}$, see section 4.3.3) throughout the greedy iterations and fig. 4.8 shows the RB size with respect to the number of greedy iterations. Section 4.5.2 summarizes the overall number of greedy iterations and final RB size for each of the runs.

For all runs, the maximum residual norm decays exponentially throughout the greedy iterations, as shown on fig. 4.8. The *default init 100* and *default init 1000* variants exhibit a much faster convergence than the *adaptive init 1* variant. However, the price to pay for this fast convergence is that the *default init 100* and *default init 1000* variants reach a much larger RB size ($N = 95$ and $N = 100$) than the *adaptive init 1* variant ($N = 60$). Clearly, one has to make a compromise between the number of greedy iterations and the final size of the RB. The *adaptive init 10* variant provides an example of such a compromise: the number of greedy iterations is $I = 26$ for a final RB size of $N = 88$.

	Greedy iterations I	Basis size N
Adaptive init 1	60	60
Adaptive init 10	26	88
Adaptive init 100	18	96
Default init 100	15	95
Adaptive init 1000	16	99
Default init 1000	15	100

Table 4.1: Number of greedy iterations and final basis size for different variants of the greedy algorithm 4.2.

At this stage, it is worth recalling that the number of greedy iterations I corresponds to the number of values of μ for which system solves with $A(\mu)$ are required (with potentially multiple right-hand sides). When the problem is large-scale, system solves with $A(\mu)$ represent very time and resource-consuming tasks. In this context, the use of multiple

directions as in the *Adaptive init 100*, *Default init 100*, *Adaptive init 1000* or *Default init 1000* variants leads to best computational performance. Admittedly, these variants do not yield RB approximation with smallest possible dimension. However, we argue that the extra computational costs from having to handle a RB subspace of size $N = 100$ rather than $N = 60$ are by far compensated by the much smaller number of problem solves. Here, there are exactly four times less large-scale problems solves ($I = 15$ versus $I = 60$). This suggests that the use of multiple directions has a great potential for significantly reducing the computational costs of building a RB with the greedy algorithm 4.2.

When it comes to choosing between the *adaptive* or the *default* initialization, the present results suggest better performance with the *default* initialization. Still, we believe that the *adaptive* initialization can be favorable when the number of sources ℓ is large (say 30 or 40), as this would prevent the RB size from increasing too fast during the first greedy iteration.

Validation

For the sake of validation, we compute 200 truth solutions $u(\mu, \nu)$ at some parameters (μ, ν) randomly samples in $\mathcal{D}_\mu \times \mathcal{D}_\nu$. Thus, we are able to compute 200 samples of the actual error $\|u(\mu, \nu) - U_N(\mu)z(\nu)\|_V$. Here, we choose to consider the RB approximations $U_N(\mu)$ from the two RBs generated during the *default init 100* and *adaptive init 1* runs respectively. These two RBs are completely different: the former consists of $N = 95$ linear combinations of truth solutions computed at only $I = 15$ distinct values of μ (with an average 6.333 linear combinations of truth solutions per value of μ), while the latter consists of $N = 60$ linear combination of truth solutions computed at $I = 60$ distinct values of μ (only one linear combination of truth solutions per value of μ).

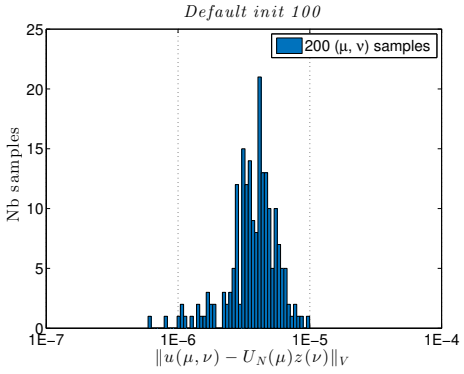


Figure 4.10: Distribution of the error for 200 samples, *default init 100*.

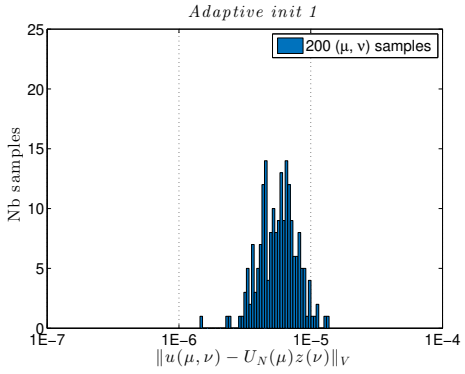


Figure 4.11: Distribution of the error for 200 samples, *adaptive init 1*.

Figures 4.10 and 4.11, show the distributions of the actual error for the two RBs generated during the *default init 100* and *adaptive init 1* runs.

We find that the *default init 100* yields slightly lower approximation errors. This is somewhat intuitive, since the RB constructed during this run is much larger than the RB constructed by the *adaptive init 1* variant (95 versus 60 basis functions). Nonetheless, the approximation properties offered by the RB of size $N = 60$ constructed by the *adaptive init 1* has comparable approximation properties, achieved with much fewer basis functions. We understand that this is due to the fact that each of the $N = 60$ basis functions are computed at a different value of the parameter μ.

Using the 200 truth solutions, we also compute the effectivity of our heuristic error indicator, defined as

$$\mathrm{eff}(\mu,\nu) = \frac{\|\widetilde{A}(\mu)U_N(\mu)z(\nu) - \widetilde{F}(\mu)z(\nu)\|_{W'}}{\hat{\alpha}\|u(\mu,\nu) - U_N(\mu)z(\nu)\|_V}. \tag{4.5.6}$$

The main statistics of the two effectivity distributions are consigned in table 4.2. Clearly, the effectivity is always very close to 1, meaning that the heuristic error indicator never under-, nor over-estimates the actual error too much. Note that for both runs, we have obtained similar values for the stability factor: $\hat{\alpha} \approx 3.834 \pm 6.05\%$ with the *Adaptive init 1* variant and $\hat{\alpha} \approx 3.43 \pm 1.35\%$ with the adaptive run. This confirms the robustness of the heuristic error estimation method on the present example.

	Min(eff)	Mean(eff)	Median(eff)	Max(eff)
Adaptive init 1	0.58	0.84	0.85	1.08
Default init 100	0.15	0.92	9.46	1.16

Table 4.2: Main statistics of the effectivity distribution for 200 samples with two different RBs generated by two variants of the greedy algorithm 4.2.

Comparison with the multiple RBs strategy

A close comparison between the error distributions figs. 4.10 and 4.11 obtained with the *unique RB* approximation strategy and the error distribution fig. 4.7 obtained with the *multiple RBs* approximation strategy reveals that the *unique RB* approximation strategy is more accurate. At first sight, this could look like an error, since we have prescribed $\epsilon^{\mathrm{target}} = 10^{-4}$ in both algorithm 4.1 and algorithm 4.2. However, this is not an error, since the criterion $\epsilon^{\mathrm{target}}$ in algorithm 4.1 is set on $\max_{(\mu,r)\in\Xi\times\{1,\ldots,\ell\}} \|\widetilde{A}(\mu)u_{r,N}(\mu) - \widetilde{f}_r(\mu)\|_{W'}$ while in algorithm 4.2 it is set on $\max_{\mu\in\Xi} \|\widetilde{A}(\mu)U_N(\mu) - \widetilde{F}(\mu)\|_{Z\to W'}$.

At this stage, it is worth recalling that

$$\|\widetilde{A}(\mu)U_N(\mu) - \widetilde{F}(\mu)\|_{Z\to W'} = \sup_{z\in Z} \frac{\|\widetilde{A}(\mu)U_N(\mu)z - \widetilde{F}(\mu)z\|_{W'}}{\|z\|_Z}. \tag{4.5.7}$$

Choosing $z = \hat{e}_r$ as candidate supremizer and recalling that $U_N(\mu)\hat{e}_r = u_{r,N}(\mu)$ and $\widetilde{F}(\mu)\hat{e}_r = \widetilde{f}_r(\mu)$, we find

$$\|\widetilde{A}(\mu)U_N(\mu) - \widetilde{F}(\mu)\|_{Z\to W'} \geqslant \max_{1\leqslant r\leqslant \ell} \|\widetilde{A}(\mu)u_{r,N}(\mu) - \widetilde{f}_r(\mu)\|_{W'}. \tag{4.5.8}$$

This shows that the stopping criterion is strongest in in algorithm 4.2 than in algorithm 4.1 and so one should not be surprised to generate a RB approximation with better accuracy than with algorithm 4.2.

In terms of the overall number of basis function computed for each RB strategy: algorithm 4.1 has computed $N\ell = 323$ basis functions in 19 solver calls while algorithm 4.2 has computed 95 basis functions in 15 solver calls (considering the original *Default init 100* variant). Clearly, the *unique RB* strategy is much more efficient than the *multiple RBs* on this particular example: it is able to achieve a better accuracy with less reduced basis functions computed in less solver calls.

4.6 Conclusions

In this chapter, we have presented the necessary adaptations of the reduced basis method to μ-parametrized problems featuring multiple sources. When there is a finite number ℓ of source terms, we proposed two distinct strategies.

The *multiple RBs* strategy simultaneously constructs ℓ distinct reduced basis approximation subspaces each of dimension N using only N calls to the solver $A(\mu)^{-1}$. In terms of computational performance, whatever the choice of paradigm for the solver (direct or iterative), this is more advantageous than successively constructing ℓ reduced basis approximation subspaces one after another. We have provided an illustration on an academic Laplace but refer to section 5.3 for a real-world application to antenna arrays and extensive analysis of computational costs.

The *unique RB* strategy relies on a single approximation subspace. Building on the original works [128] and [114], we have envisaged the possibility of enriching the reduced basis with more than one basis function per greedy iteration. To this end, we have introduced the residual operator which is a generalization of the notion of residual in the block context. We did not restrict ourselves to the leading eigenvector of the residual operator, but considered the d most dominant eigenvectors and have provided an empirical criterion for selecting the number d of eigenvectors to be considered. Numerical illustration of the proposed strategy has shown a reduction in the number of calls to the solver $A(\mu)^{-1}$ compared to the original strategy which uses only the leading eigenvector. For large-scale problems, this could represent a significant advantage, given that evaluation of the solver $A(\mu)^{-1}$ is usually the most time and resource-consuming task.

The methods presented in this chapter address the class of parametrized problems featuring numerous but still a finite number of sources. The scope of this work has been ex-

tended to the class of parametrized problem with an infinite number of sources parametrized by an additional parameter ν. This was done by applying the EIM to approximate the parametrized source using a finite number ℓ of sources. In this context, the number of EIM terms determines the finite number of sources ℓ to be considered and for which the RB strategies presented in this chapter can apply. A challenging academic example of this methodology was provided as illustration.

Chapter 5

Reduced basis method for frequency sweeps with edge finite elements

Summary. This chapter is devoted to the reduced basis method applied to frequency-parametrized Maxwell's equations solved using edge finite elements. First we present the governing equations and their high-fidelity discretization using Raviart-Thomas-Nédélec edge finite elements. We briefly review the FETI-2LM method for solving the resulting complex, non-hermitian, linear system on parallel architectures using domain-decomposition. Next, we apply the reduced basis method on two antenna problems of industrial interest. On a horn antenna, we compare the certified reduced basis approach (with SCM) to a heuristic strategy and show the superiority of the heuristic strategy. Finally, we show some results on large-scale antenna array problem. We successfully use the *multiple RBs* strategy from chapter 4 in order to efficiently solve the radiation patterns generated by each antenna in the antenna array.

Contents

5.1 Strong formulation and high-fidelity approximation 106
5.1.1 Governing equations 106
5.1.2 High-fidelity discretization 108
5.1.3 Numerical solver: FETI-2LM 110
5.2 The RBM for the frequency sweep problem 112
5.2.1 The frequency-parametrized problem 112
5.2.2 Results on the Horn Antenna test case 113
5.3 Application to antenna arrays 117
5.3.1 RBM with a single right-hand side 118
5.3.2 RBM with multiple right-hand sides 119

5.4 Conclusions 121

5.1 Strong formulation and high-fidelity approximation

5.1.1 Governing equations

Let $\Omega \subset \mathbb{R}^3$ be a bounded physical domain. For all time t in $\mathbb{R}$, electromagnetic phenomena are described using the four functions $\boldsymbol{D}$, $\boldsymbol{E}$, $\boldsymbol{B}$, $\boldsymbol{H}$ of the $(\boldsymbol{x}, t)$ variable in $\Omega \times \mathbb{R}$ and with values in $\mathbb{R}^3$ [24]. These functions are called *electric induction*, *electric field*, *magnetic induction* and *magnetic field* respectively. They are related to the given *charge density* $\rho : \Omega \times \mathbb{R} \to \mathbb{R}$ and the given *current density* $\boldsymbol{J} : \Omega \times \mathbb{R} \to \mathbb{R}^3$ by the Maxwell equations

$$-\frac{\partial \boldsymbol{D}}{\partial t} + \mathbf{curl}\, \boldsymbol{H} = \boldsymbol{J}, \tag{5.1.1a}$$

$$\frac{\partial \boldsymbol{B}}{\partial t} + \mathbf{curl}\, \boldsymbol{E} = 0, \tag{5.1.1b}$$

$$\operatorname{div} \boldsymbol{D} = \rho, \tag{5.1.1c}$$

$$\operatorname{div} \boldsymbol{B} = 0. \tag{5.1.1d}$$

called the *Maxwell-Ampère*, *Maxwell-Faraday*, *Gauss electrical* and *Gauss magnetic* laws, respectively. The charge and current densities further satisfy the following *conservation law*

$$\frac{\partial \rho}{\partial t} + \operatorname{div} \boldsymbol{J} = 0. \tag{5.1.2}$$

Engineering applications basically consist in acting on ρ and $\boldsymbol{J}$ in order to produce a desired electromagnetic field. Typically, in antenna applications, the current density is specifically tuned to generate a desired radiation distribution – the so-called antenna pattern [76].

In free space, the following *constitutive relations* hold

$$\boldsymbol{D} = \epsilon_0 \boldsymbol{E}, \qquad \boldsymbol{B} = \mu_0 \boldsymbol{H}, \tag{5.1.3}$$

with $\epsilon_0 = \frac{1}{36\pi} 10^{-9}$ (in $\mathrm{F/m}$) and $\mu_0 = 4\pi \cdot 10^{-7}$ (in $\mathrm{H/m}$) called respectively *permittivity* and *permeability* of free space. In the absence of charge and currents (*i.e.*, $\rho = 0$ and $\boldsymbol{J} = 0$), we deduce from the Maxwell equations and the constitutive relations that $\boldsymbol{B}$ and $\boldsymbol{E}$ both satisfy the *vectorial wave equation*

$$-\frac{1}{c_0^2}\frac{\partial^2 \boldsymbol{U}}{\partial t^2} + \boldsymbol{\Delta}\, \boldsymbol{U} = 0, \tag{5.1.4}$$

with $\boldsymbol{\Delta}$ denoting the vectorial Laplacian $\boldsymbol{\Delta} = \mathbf{grad} \operatorname{div} - \mathbf{curl}\, \mathbf{curl}$ and $c_0 = \frac{1}{\sqrt{\epsilon_0 \mu_0}} \approx 3 \cdot 10^8$ (in $\mathrm{m/s}$) denoting the velocity of electromagnetic waves in free space.

Time-harmonic Maxwell equations

Given that the Maxwell equations hold for all $t \in \mathbb{R}$, one can apply the Fourier transform. This amounts to considering electromagnetic fields of the form:

$$\begin{aligned} \boldsymbol{D}(\boldsymbol{x},t) &= \Re\{\boldsymbol{D}(\boldsymbol{x})e^{i\omega t}\}, \qquad \boldsymbol{E}(\boldsymbol{x},t) = \Re\{\boldsymbol{E}(\boldsymbol{x})e^{i\omega t}\}, \\ \boldsymbol{B}(\boldsymbol{x},t) &= \Re\{\boldsymbol{B}(\boldsymbol{x})e^{i\omega t}\}, \qquad \boldsymbol{H}(\boldsymbol{x},t) = \Re\{\boldsymbol{H}(\boldsymbol{x})e^{i\omega t}\}, \end{aligned} \tag{5.1.5}$$

where the Fourier variable $\omega > 0$ is called the *angular frequency*. The *frequency* is defined by $f = \frac{\omega}{2\pi}$. In this work, we prefer to work with the *wavenumber* $k = \frac{\omega}{c_0}$. In free space, we can express the stationary Maxwell equations in terms of the complex-valued fields $\boldsymbol{E}$, $\boldsymbol{H}$ only, by eliminating the $\boldsymbol{D}$, $\boldsymbol{B}$ fields using the constitutive relations (5.1.3). Thus, in the absence of charges this yields:

$$\begin{cases} -ik\boldsymbol{e} + \mathbf{curl}\,\boldsymbol{h} &= \boldsymbol{j}, \\ ik\boldsymbol{h} + \mathbf{curl}\,\boldsymbol{e} &= 0. \end{cases} \tag{5.1.6}$$

where $\boldsymbol{e} = \sqrt{\epsilon_0}\boldsymbol{E}$, $\boldsymbol{h} = \sqrt{\mu_0}\boldsymbol{H}$ and $\boldsymbol{j} = \sqrt{\mu_0}\boldsymbol{J}$ are the re-normalized fields. Let $L^2(\Omega)$ denote the space composed of all complex-valued, square integrable functions defined on Ω. Following [1, §2.2], we introduce the following Sobolev spaces

$$\boldsymbol{L}^2(\Omega) = \left(L^2(\Omega)\right)^3, \tag{5.1.7}$$

$$\boldsymbol{H}(\mathbf{curl},\Omega) = \left\{\boldsymbol{v} \in \boldsymbol{L}^2(\Omega),\ \mathbf{curl}\,\boldsymbol{v} \in \boldsymbol{L}^2(\Omega)\right\}. \tag{5.1.8}$$

Clearly, from the governing equations eq. (5.1.6), for a source field $\boldsymbol{j}$ in $\boldsymbol{L}^2(\Omega)$ both $\boldsymbol{e}$ and $\boldsymbol{h}$ are in $\boldsymbol{H}(\mathbf{curl},\Omega)$. Applying the curl operator to the second equation and plugging into the first, we eliminate the magnetic field and obtain the following equation for the electric field:

$$\mathbf{curl}\,\mathbf{curl}\,\boldsymbol{e} - k^2\boldsymbol{e} = -ik\boldsymbol{j}. \tag{5.1.9}$$

In this work, we consider more general constitutive relations than the free space relations eq. (5.1.3). To this end, we introduce two tensor fields $\underline{\underline{\epsilon}}, \underline{\underline{\nu}} : \Omega \to \mathbb{C}^{3\times 3}$ respectively corresponding to the *relative permittivity* and *inverse relative permeability* tensor fields. We assume that they are hermitian and positive definite for all $\boldsymbol{x} \in \Omega$. Under this more general setting, eq. (5.1.9) becomes

$$\mathbf{curl}\,\underline{\underline{\nu}}\mathbf{curl}\,\boldsymbol{e} - k^2\underline{\underline{\epsilon}}\boldsymbol{e} = -ik\boldsymbol{j}. \tag{5.1.10}$$

Boundary conditions

We consider the domain boundary to be split in two parts $\partial\Omega = \Gamma_A \cup \Gamma_D$ as shown on fig. 5.1. We impose that the part of the boundary Γ_D is a perfect electric conductor, which means that $\boldsymbol{e} \times \hat{\boldsymbol{n}} = 0$ on Γ_D where $\hat{\boldsymbol{n}}$ denotes the outgoing unitary normal. The part of the boundary Γ_A is purely artificial and only exists because the domain Ω must

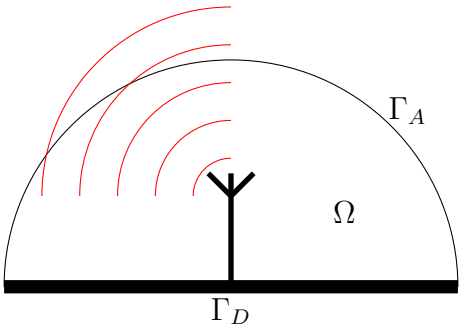


Figure 5.1: Schematic view of a computational domain with one radiating antenna.

be bounded for computational reasons. Indeed, the reality that we try to simulate is that of an unbounded physical domain. We enforce a so-called absorbing (or non-reflective) boundary condition on Γ_A, which lets the electromagnetic waves propagate out of the domain, as if the boundary did not exist. The development of such absorbing boundary conditions is an active area of research [8, 41]. Here, we consider the following first-order absorbing condition

$$ik\boldsymbol{e} \times \hat{\boldsymbol{n}} = \hat{\boldsymbol{n}} \times (\underline{\underline{\nu}}\mathbf{curl}\,\boldsymbol{e} \times \hat{\boldsymbol{n}}). \tag{5.1.11}$$

Thus the full set of equations that we wish to solve is given by

$$\begin{cases} \mathbf{curl}\,\underline{\underline{\nu}}\mathbf{curl}\,\boldsymbol{e} - k^2\underline{\underline{\epsilon}}\boldsymbol{e} = -ik\boldsymbol{j} & \text{in } \Omega, \\ \boldsymbol{e} \times \hat{\boldsymbol{n}} = 0 & \text{on } \Gamma_D, \\ ik\boldsymbol{e} \times \hat{\boldsymbol{n}} = \hat{\boldsymbol{n}} \times (\underline{\underline{\nu}}\mathbf{curl}\,\boldsymbol{e} \times \hat{\boldsymbol{n}}) & \text{on } \Gamma_A. \end{cases} \tag{5.1.12}$$

As shown in [1, §8.3.3], the natural function space of the electric field satisfying eq. (5.1.12) is

$$\begin{aligned} \boldsymbol{H}^+_{0,\Gamma_D}(\mathbf{curl},\Omega) = \{\boldsymbol{v} \in \boldsymbol{H}(\mathbf{curl},\Omega) \ &: \boldsymbol{v} \times \hat{\boldsymbol{n}}|_{\Gamma_D} = 0 \\ &\text{and } \boldsymbol{v} \times \hat{\boldsymbol{n}}|_{\Gamma_A} \in \boldsymbol{L}^2_t(\Gamma_A)\}, \end{aligned} \tag{5.1.13}$$

where $\boldsymbol{L}^2_t(\Gamma_A) = \{\boldsymbol{v} \in \boldsymbol{L}^2(\Gamma_A) \ : \ \boldsymbol{v} \cdot \hat{\boldsymbol{n}} = 0\}$ is the space of tangential square-integrable functions defined on Γ_A. It is endowed with the inner product

$$\begin{aligned} (\boldsymbol{v},\boldsymbol{w})_{\boldsymbol{H}^+_{0,\Gamma_D}(\mathbf{curl},\Omega)} = &\int_\Omega \underline{\underline{\nu}}\mathbf{curl}\,\boldsymbol{v} \cdot \mathbf{curl}\,\overline{\boldsymbol{w}}\mathrm{d}\Omega + c_1 \int_\Omega \underline{\underline{\epsilon}}\boldsymbol{v} \cdot \overline{\boldsymbol{w}}\mathrm{d}\Omega \\ &+ c_2 \int_{\Gamma_A} (\boldsymbol{v} \times \hat{\boldsymbol{n}}) \cdot (\overline{\boldsymbol{w}} \times \hat{\boldsymbol{n}})\mathrm{d}\Gamma, \end{aligned} \tag{5.1.14}$$

where $c_1, c_2 > 0$ are given constants. It is shown in Ref. [1, §8.3.3] that there exists a unique solution $\boldsymbol{e} \in \boldsymbol{H}^+_{0,\Gamma_D}(\mathbf{curl},\Omega)$ satisfying eq. (5.1.12).

5.1.2 High-fidelity discretization

The weak-form associated to eq. (5.1.12) is discretized using $\boldsymbol{H}(\mathbf{curl},\Omega)$-conforming finite elements known as the Raviart-Thomas-Nédélec finite elements [80, 78]. Let $\mathcal{T}_h$

be a partition of Ω into tetrahedral elements. For each tetrahedron $T \in \mathcal{T}_h$, define the zero-th local Raviart-Thomas-Nédélec space of complex-valued functions defined on T as $\boldsymbol{RTN}_0(T) = \{\boldsymbol{v} : \boldsymbol{x} \in T \mapsto \boldsymbol{\alpha} + \boldsymbol{\beta} \times \boldsymbol{x},\ \boldsymbol{\alpha}, \boldsymbol{\beta} \in \mathbb{C}^3\}$. The global approximation space is given by

$$\boldsymbol{RTN}_0 = \{\boldsymbol{v} \in \boldsymbol{H}(\mathbf{curl}, \Omega),\ \boldsymbol{v}|_T \in \boldsymbol{RTN}_0(T),\ \forall T \in \mathcal{T}_h\}. \tag{5.1.15}$$

The dimension of the $\boldsymbol{RTN}_0$ approximation space is finite and coincides with the number of edges in the mesh $\mathcal{T}_h$. Indeed, we recall that a degree of freedom is associated to each edge in the mesh [80]. Incorporation of the essential Dirichlet boundary condition yield the following finite element approximation space

$$\boldsymbol{V}_h^{\mathrm{RTN}} = \{\boldsymbol{v} \in \boldsymbol{RTN}_0,\ \boldsymbol{v} \times \hat{\boldsymbol{n}}|_{\Gamma_D} = 0\}, \tag{5.1.16}$$

whose dimension $\mathcal{N}$ coincides with the number of edges in the mesh $\mathcal{T}_h$ that do not lie on the part of the boundary Γ_D. In the Galerkin context, the discretized weak form writes: *find* $\boldsymbol{e_h} \in \boldsymbol{V}_h^{\mathrm{RTN}}$ *such that*

$$\forall \boldsymbol{v}_h \in \boldsymbol{V}_h^{\mathrm{RTN}}, \quad a(\boldsymbol{e}_h, \boldsymbol{v}_h) = f(\boldsymbol{v}_h), \tag{5.1.17}$$

where $a : \boldsymbol{V}_h^{\mathrm{RTN}} \times \boldsymbol{V}_h^{\mathrm{RTN}} \to \mathbb{C}$ is the continuous sesquilinear form defined by

$$\begin{aligned} a(\boldsymbol{e}_h, \boldsymbol{v}_h) = &\int_\Omega \underline{\underline{\nu}} \mathbf{curl}\, \boldsymbol{e}_h \cdot \mathbf{curl}\, \overline{\boldsymbol{v}_h} \mathrm{d}\Omega - k^2 \int_\Omega \underline{\underline{\epsilon}} \boldsymbol{e}_h \cdot \overline{\boldsymbol{v}_h} \mathrm{d}\Omega \\ &+ ik \int_{\Gamma_A} (\boldsymbol{e}_h \times \hat{\boldsymbol{n}}) \cdot (\overline{\boldsymbol{v}_h} \times \hat{\boldsymbol{n}}) \mathrm{d}\Gamma \end{aligned} \tag{5.1.18}$$

and $f : \boldsymbol{V}_h^{\mathrm{RTN}} \to \mathbb{C}$ is the continuous linear form defined by

$$f(\boldsymbol{v}_h) = -ik \int_\Omega \boldsymbol{j}_h \cdot \overline{\boldsymbol{v}_h} \mathrm{d}\Omega, \tag{5.1.19}$$

with $\boldsymbol{j}_h = \Pi_h \boldsymbol{j}$, where Π_h denotes the interpolation operator [29].

Denoting $\{\boldsymbol{w}_i\}_{1 \leqslant i \leqslant \mathcal{N}}$ a basis for the global Raviart-Thomas-Nédélec approximation space incorporating the Dirichlet essential boundary condition (see [45, Chapter 3, §5.3]), we introduce the matrix $\mathbf{A} \in \mathbb{C}^{\mathcal{N} \times \mathcal{N}}$ and vector $\mathbf{f} \in \mathbb{C}^{\mathcal{N}}$ with coefficients

$$\mathbf{A}_{ij} = a(\boldsymbol{w}_j, \boldsymbol{w}_i), \quad \mathbf{f}_j = f(\boldsymbol{w}_j), \quad 1 \leqslant i, j \leqslant \mathcal{N}. \tag{5.1.20}$$

Thus, the solution to eq. (5.1.17) is given by $\boldsymbol{e}_h = \sum_{i=1}^{\mathcal{N}} \mathbf{u}_i \boldsymbol{w}_i$ where $\mathbf{u} \in \mathbb{C}^{\mathcal{N}}$ solves the large-scale linear system $\mathbf{A}\mathbf{u} = \mathbf{f}$. The discrete inverse Riesz operator is given by the hermitian, positive-definite matrix $\mathbf{B} \in \mathbb{C}^{\mathcal{N} \times \mathcal{N}}$ given by

$$\mathbf{B}_{ij} = (\boldsymbol{w}_j, \boldsymbol{w}_i)_{\boldsymbol{H}^+_{0,\Gamma_D}(\mathbf{curl},\Omega)}, \quad 1 \leqslant i, j \leqslant \mathcal{N}. \tag{5.1.21}$$

5.1.3 Numerical solver: FETI-2LM

We now review the FETI-2LM domain decomposition method which is used to efficient solve the large-scale linear system $\mathbf{Au} = \mathbf{f}$ on multiple processors. The acronym FETI-2LM stands for Finite Element Tearing and Interconnecting with two Lagrange multipliers. A more detailed review of this method (and of FETI methods in general) can be found in [102].

Consider a splitting of the domain Ω in two non-overlapping subdomains $\overline{\Omega} = \overline{\Omega^{(1)}} \cup \overline{\Omega^{(1)}}$, as shown on fig. 5.2. Using the subscript i for the degrees of freedom located inside the

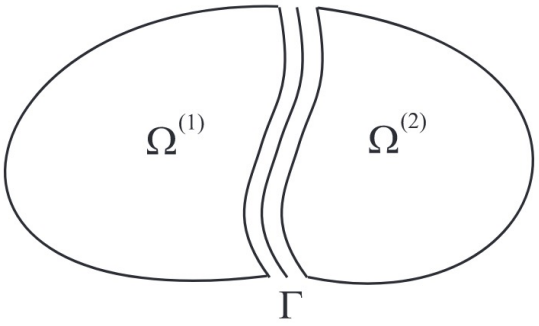


Figure 5.2: Splitting into two non-overlapping subdomains.

subdomains $\Omega^{(s)}, s = 1, 2$ and the subscript b for the degrees of freedom located on the interface $\partial\Omega^{(1)} \cap \partial\Omega^{(2)}$, the $\mathbf{Au} = \mathbf{f}$ linear system can be rewritten as

$$\begin{bmatrix} \mathbf{A}_{\mathrm{ii}}^{(1)} & 0 & \mathbf{A}_{\mathrm{ib}}^{(1)} \\ 0 & \mathbf{A}_{\mathrm{ii}}^{(2)} & \mathbf{A}_{\mathrm{ib}}^{(2)} \\ \mathbf{A}_{\mathrm{bi}}^{(1)} & \mathbf{A}_{\mathrm{bi}}^{(2)} & \mathbf{A}_{\mathrm{bb}}^{(1)} + \mathbf{A}_{\mathrm{bb}}^{(2)} \end{bmatrix} \begin{bmatrix} \mathbf{u}_{\mathrm{i}}^{(1)} \\ \mathbf{u}_{\mathrm{i}}^{(2)} \\ \mathbf{u}_{\mathrm{b}} \end{bmatrix} = \begin{bmatrix} \mathbf{f}_{\mathrm{i}}^{(1)} \\ \mathbf{f}_{\mathrm{i}}^{(2)} \\ \mathbf{f}_{\mathrm{b}}^{(1)} + \mathbf{f}_{\mathrm{b}}^{(2)} \end{bmatrix}. \tag{5.1.22}$$

Under the domain decomposition paradigm, each processor $s = 1, 2$ assembles their own local subdomain contributions:

$$\mathbf{A}^{(s)} = \begin{bmatrix} \mathbf{A}_{\mathrm{ii}}^{(s)} & \mathbf{A}_{\mathrm{ib}}^{(s)} \\ \mathbf{A}_{\mathrm{bi}}^{(s)} & \mathbf{A}_{\mathrm{bb}}^{(s)} \end{bmatrix}, \quad \mathbf{f}^{(s)} = \begin{bmatrix} \mathbf{f}_{\mathrm{i}}^{(s)} \\ \mathbf{f}_{\mathrm{b}}^{(s)} \end{bmatrix}. \tag{5.1.23}$$

In the FETI-2LM method, we consider given interface matrices $\mathbf{K}_{\mathrm{bb}}^{(s)}, s = 1, 2$ (for more detail on how these interface matrices should be chosen, see [104]) and search for two Lagrange multipliers $\boldsymbol{\lambda}_{\mathrm{b}}^{(s)}, s = 1, 2$ defined on the interface. Each processor $s = 1, 2$ can solve the following local problem

$$\begin{bmatrix} \mathbf{A}_{\mathrm{ii}}^{(s)} & \mathbf{A}_{\mathrm{ib}}^{(s)} \\ \mathbf{A}_{\mathrm{bi}}^{(s)} & \mathbf{A}_{\mathrm{bb}}^{(s)} + \mathbf{K}_{\mathrm{bb}}^{(s)} \end{bmatrix} \begin{bmatrix} \tilde{\mathbf{u}}_{\mathrm{i}}^{(s)} \\ \tilde{\mathbf{u}}_{\mathrm{b}}^{(s)} \end{bmatrix} = \begin{bmatrix} \mathbf{f}_{\mathrm{i}}^{(s)} \\ \mathbf{f}_{\mathrm{b}}^{(s)} + \boldsymbol{\lambda}_{\mathrm{b}}^{(s)} \end{bmatrix}. \tag{5.1.24}$$

The goal is to find proper Lagrange multipliers in such a way that the solution to the local problem eq. (5.1.24) is indeed the subdomain restriction of the solution of the global problem eq. (5.1.22). In other words, the goal is to satisfy

$$\begin{cases} \tilde{\mathbf{u}}_{\mathrm{i}}^{(s)} = \mathbf{u}_{\mathrm{i}}^{(s)}, \\ \tilde{\mathbf{u}}_{\mathrm{b}}^{(s)} = \mathbf{u}_{\mathrm{b}}, \end{cases} \qquad s = 1, 2. \tag{5.1.25}$$

We proceed in two steps to characterize the Lagrange multipliers.

1. Exploiting the last line of the global problem eq. (5.1.22) and the last line of the local problem eq. (5.1.24), we obtain

$$\begin{aligned} &\mathbf{A}_{\mathrm{bi}}^{(1)} \left(\mathbf{u}_{\mathrm{i}}^{(1)} - \tilde{\mathbf{u}}_{\mathrm{i}}^{(1)} \right) + \mathbf{A}_{\mathrm{bi}}^{(2)} \left(\mathbf{u}_{\mathrm{i}}^{(2)} - \tilde{\mathbf{u}}_{\mathrm{i}}^{(2)} \right) + \left(\mathbf{A}_{\mathrm{bb}}^{(1)} + \mathbf{A}_{\mathrm{bb}}^{(2)} \right) \mathbf{u}_{\mathrm{b}} \\ &- \left(\mathbf{A}_{\mathrm{bb}}^{(1)} + \mathbf{K}_{\mathrm{bb}}^{(1)} \right) \tilde{\mathbf{u}}_{\mathrm{b}}^{(1)} - \left(\mathbf{A}_{\mathrm{bb}}^{(2)} + \mathbf{K}_{\mathrm{bb}}^{(2)} \right) \tilde{\mathbf{u}}_{\mathrm{b}}^{(2)} = -\boldsymbol{\lambda}_{\mathrm{b}}^{(1)} - \boldsymbol{\lambda}_{\mathrm{b}}^{(2)}. \end{aligned} \tag{5.1.26}$$

   Thus, if eq. (5.1.25) are satisfied, then

$$\boldsymbol{\lambda}_{\mathrm{b}}^{(1)} + \boldsymbol{\lambda}_{\mathrm{b}}^{(2)} - \left(\mathbf{K}_{\mathrm{bb}}^{(1)} + \mathbf{K}_{\mathrm{bb}}^{(2)} \right) \tilde{\mathbf{u}}_{\mathrm{b}}^{(s)} = 0, \quad s = 1, 2. \tag{5.1.27}$$

2. Next, using the Schur complement matrix $\mathbf{S}_{\mathrm{bb}}^{(s)} = \mathbf{A}_{\mathrm{bb}}^{(s)} - \mathbf{A}_{\mathrm{bi}}^{(s)} \left(\mathbf{A}_{\mathrm{ii}}^{(s)} \right)^{-1} \mathbf{A}_{\mathrm{ib}}^{(s)}$, we can eliminate the interior degrees of freedom $\tilde{\mathbf{u}}_{\mathrm{i}}^{(s)}$ in eq. (5.1.24). Doing so, the interface degrees of freedom can be expressed as

$$\tilde{\mathbf{u}}_{\mathrm{b}}^{(s)} = \left(\mathbf{S}_{\mathrm{bb}}^{(s)} + \mathbf{K}_{\mathrm{bb}}^{(s)} \right)^{-1} \boldsymbol{\lambda}_{\mathrm{b}}^{(s)} + \left(\mathbf{S}_{\mathrm{bb}}^{(s)} + \mathbf{K}_{\mathrm{bb}}^{(s)} \right)^{-1} \mathbf{c}_{\mathrm{b}}^{(s)}, \tag{5.1.28}$$

   with $\mathbf{c}_{\mathrm{b}}^{(s)} = \mathbf{f}_{\mathrm{b}}^{(s)} - \mathbf{A}_{\mathrm{bi}}^{(s)} \left(\mathbf{A}_{\mathrm{ii}}^{(s)} \right)^{-1} \mathbf{f}_{\mathrm{i}}^{(s)}$.

Now all is set to fully characterize the Lagrange multipliers $\boldsymbol{\lambda}_{\mathrm{b}}^{(s)}, s = 1, 2$. Indeed, by plugging eq. (5.1.28) into eq. (5.1.27), we obtain the following interface equation

$$\mathbf{F} \begin{bmatrix} \boldsymbol{\lambda}_{\mathrm{b}}^{(1)} \\ \boldsymbol{\lambda}_{\mathrm{b}}^{(2)} \end{bmatrix} = \mathbf{d}, \tag{5.1.29}$$

with $\mathbf{F}$ and $\mathbf{d}$ the matrix and right-hand side defined as

$$\begin{aligned} \mathbf{F} &= \begin{bmatrix} \mathbf{I} & \mathbf{I} - \left(\mathbf{K}_{\mathrm{bb}}^{(1)} + \mathbf{K}_{\mathrm{bb}}^{(2)} \right) \left(\mathbf{S}_{\mathrm{bb}}^{(2)} + \mathbf{K}_{\mathrm{bb}}^{(2)} \right)^{-1} \\ \mathbf{I} - \left(\mathbf{K}_{\mathrm{bb}}^{(1)} + \mathbf{K}_{\mathrm{bb}}^{(2)} \right) \left(\mathbf{S}_{\mathrm{bb}}^{(1)} + \mathbf{K}_{\mathrm{bb}}^{(1)} \right)^{-1} & \mathbf{I} \end{bmatrix}, \\ \mathbf{d} &= \begin{bmatrix} \left(\mathbf{K}_{\mathrm{bb}}^{(1)} + \mathbf{K}_{\mathrm{bb}}^{(2)} \right) \left(\mathbf{S}_{\mathrm{bb}}^{(2)} + \mathbf{K}_{\mathrm{bb}}^{(2)} \right)^{-1} \mathbf{c}_{\mathrm{b}}^{(2)} \\ \left(\mathbf{K}_{\mathrm{bb}}^{(1)} + \mathbf{K}_{\mathrm{bb}}^{(2)} \right) \left(\mathbf{S}_{\mathrm{bb}}^{(1)} + \mathbf{K}_{\mathrm{bb}}^{(1)} \right)^{-1} \mathbf{c}_{\mathrm{b}}^{(1)} \end{bmatrix}. \end{aligned} \tag{5.1.30}$$

The interface equation (5.1.29) is solved using a Krylov method, typically the ORTHODIR method [127]. At each iteration of the Krylov method, a matrix-vector operator with $\mathbf{F}$ is required. Note that $\mathbf{F}$ is not assembled in practice, since the matrix-vector operation with $\mathbf{F}$ simply requires each processor to solve a local problem of the form eq. (5.1.24) followed by a communication phase [102].

Consequently, each processor successively solves numerous local problems eq. (5.1.24) throughout the Krylov method (*i.e.*, as many as the number of iterations required for the Krylov method to converge). In practice, a factorization of the local problem matrix is computed once and for all, so that these successive solves can be made very efficient. A sparse LU-factorization for the local problem matrix can be efficiently computed using for instance the PARDISO routines [108]. The cost of such factorization is reasonable, given that the size of the local problem matrix remains moderate.

5.2 The RBM for the frequency sweep problem

5.2.1 The frequency-parametrized problem

In this work, we wish to solve the discrete electric field $\boldsymbol{e}_h \in \boldsymbol{V}_h^{\mathrm{RTN}}$ solution to the finite element problem (5.1.17) not just for one value of the wavenumber k, but for a given range of values $k \in [k_{\min}, k_{\max}]$. We now introduce $\mu = k/C_{\mathrm{adim}}$ our varying parameter, where $C_{\mathrm{adim}} > 0$ is a constant used to adimensionalize the problem.

Under this parametrized setting, the sesquilinear form defined by eq. (5.1.18) is in fact μ-dependent, thus $a(\cdot,\cdot) = a(\cdot,\cdot;\mu)$. Moreover, the dependency in μ is trivially affine with $Q^a = 3$ and a possible parametrization is

$$a(\cdot,\cdot;\mu) = a_1(\cdot;\cdot) - \mu^2 a_2(\cdot,\cdot) + i\mu a_3(\cdot,\cdot), \tag{5.2.1}$$

where for all $\boldsymbol{u}_h, \boldsymbol{v}_h \in \boldsymbol{V}_h^{\mathrm{RTN}}$,

$$\begin{aligned} a_1(\boldsymbol{u}_h, \boldsymbol{v}_h) &= \int_\Omega \underline{\underline{\nu}} \mathbf{curl}\, \boldsymbol{u}_h \cdot \mathbf{curl}\, \overline{\boldsymbol{v}_h} \mathrm{d}\Omega, \\ a_2(\boldsymbol{u}_h, \boldsymbol{v}_h) &= C_{\mathrm{adim}}^2 \int_\Omega \underline{\underline{\epsilon}} \boldsymbol{u}_h \cdot \overline{\boldsymbol{v}_h} \mathrm{d}\Omega, \\ a_3(\boldsymbol{u}_h, \boldsymbol{v}_h) &= C_{\mathrm{adim}} \int_{\Gamma_A} (\boldsymbol{u}_h \times \hat{\boldsymbol{n}}) \cdot (\overline{\boldsymbol{v}_h} \times \hat{\boldsymbol{n}}) \mathrm{d}\Gamma. \end{aligned} \tag{5.2.2}$$

As for the right-hand side, the dependency in μ is also straightforwardly affine with $Q^f = 1$ and a possible parametrization is

$$f(\cdot;\mu) = i\mu f_1(\cdot), \tag{5.2.3}$$

where for all $\boldsymbol{v}_h \in \boldsymbol{V}_h^{\mathrm{RTN}}$,

$$f_1(\boldsymbol{v}_h) = C_{\mathrm{adim}} \int_\Omega \boldsymbol{j}_h \cdot \overline{\boldsymbol{v}_h} \mathrm{d}\Omega. \tag{5.2.4}$$

This being set, the frequency-parameterized discretized Maxwell equations read: *find* $\boldsymbol{e}_h(\mu) \in \boldsymbol{V}_h^{\mathrm{RTN}}$ *such that*

$$\forall \boldsymbol{v}_h \in \boldsymbol{V}_h^{\mathrm{RTN}}, \quad a(\boldsymbol{e}_h(\mu), \boldsymbol{v}_h; \mu) = f(\boldsymbol{v}_h). \tag{5.2.5}$$

Clearly, this formulation fits the framework proposed in section 1.1.4, with the correspondance given by table 5.1.

Abstract setting	Parametrized Maxwell
V	$\boldsymbol{V}_h^{\mathrm{RTN}}$
W	$\boldsymbol{V}_h^{\mathrm{RTN}}$
μ	$\mu = k/C_{\mathrm{adim}}$
$a(\cdot, \cdot, \mu)$	Equation (5.1.18)
$f(\mu)$	Equation (5.1.19)
$u(\mu)$	$\boldsymbol{e}_h(\mu)$

Table 5.1: Correspondance between the parametrized Maxwell equations and the general abstract setting of section 1.1.4.

5.2.2 Results on the Horn Antenna test case

Description of the the Horn Antenna test case

We consider a horn antenna and a substrate layer as shown on fig. 5.3 (we refer to [67] for a review on horn antenna concepts). The problem is distributed on 2 processors, with processor 1 handling $109,435$ local degrees of freedom (horn antenna subdomain) and processor 2 handling $30,265$ local degrees of freedom (substrate subdomain). Notice that the workload is not equally distributed on the 2 processors, which is non-optimal in terms of computational performance but allows more flexibility in the design processes.

The frequency range of interest is $18 - 22\mathrm{GHz}$. The coefficient C_{adim} is tuned in such a way that $\mu = k/C_{\mathrm{adim}}$ varies in $\mathcal{D} = [1.8,\ 2.2]$, thus the frequency f_{Hz} (in Hz) is easily recovered by the formula $f_{\mathrm{Hz}} = \mu \times 10^{10}$.

RB approximations

We want to build a RB approximation $u_N(\mu)$ of $u(\mu)$ with a target accuracy of 5%, which means

$$\forall \mu \in [\mu_{\min},\ \mu_{\max}], \quad \frac{\|u(\mu) - u_N(\mu)\|_V}{\|u(\mu)\|_V} < 5\%. \tag{5.2.6}$$

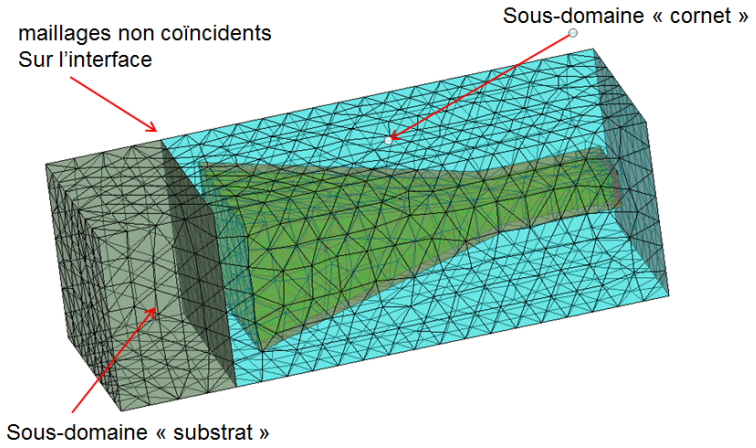


Figure 5.3: Meshes of the horn antenna test case.

To this end, we propose a certified and a heuristic approach.

- the certified approach is based on applying the greedy algorithm 2.1 driven by the indicator

$$\Delta_N^{\text{rel}}(\mu) = \frac{\Delta_N^{\text{abs}}(\mu)}{\|u_N(\mu)\|_V} \left(1 - \frac{\Delta_N^{\text{abs}}(\mu)}{\|u_N(\mu)\|_V}\right)^{-1}, \tag{5.2.7}$$

  where $\Delta_N^{\text{abs}}(\mu) = \frac{1}{\alpha_{LB}(\mu)} \|A(\mu)u_N(\mu) - f(\mu)\|_{W'}$ is a rigorous *a posteriori* error estimator, with $\mu \mapsto \alpha_{LB}(\mu)$ denoting a fast-to-evaluate lower bound for the inf-sup constant $\mu \mapsto \alpha(\mu)$ obtained from applying the SCM algorithm 2.2. Note that from proposition 2.1.1 eq. (5.2.7) is a rigorous upper bound for the RB relative error, hence the name "certified" for this approach.

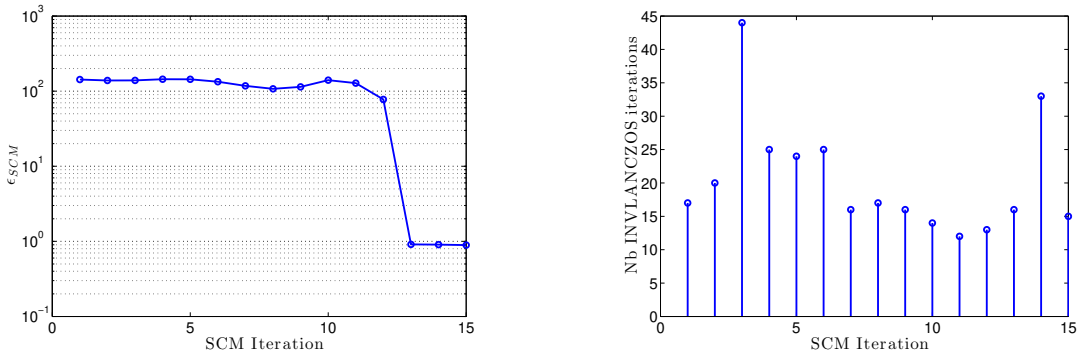
- the heuristic approach consists in driving the greedy algorithm 2.1 using the indicator

$$\widetilde{\Delta_N^{\text{rel}}}(\mu) = \frac{\|A(\mu)u_N(\mu) - f(\mu)\|_{W'}}{\hat{\alpha}\|u_N(\mu)\|_V}, \tag{5.2.8}$$

  where the constant $\hat{\alpha}$ is updated at each iteration of the greedy algorithm following the heuristic method presented in section 2.2.4 (only the quasi-constant case is considered).

For the SCM (only necessary for the certified approach), we use a discrete set $\Xi \subset \mathcal{D} = [1.8,\ 2.2]$ made of 200 uniformly distributed points and a prescribed tolerance $tol = 0.9$. The large-scale generalized eigenvalue problems G.E.P. are solved using the inverse Lanczos algorithm with a prescribed tolerance of 10^{-4} (see Appendix A).

The SCM takes $J = 15$ iterations to converge to the prescribed tolerance. The convergence curve shown on fig. 5.4a reveals that the error does not decrease until a threshold number of 12 eigensolves. The 13th eigensolves makes the error drop by two orders of magnitude. This is due to the fact the SCM error is defined as the maximum over all



(a) Maximum relative difference between lower and upper bounds per SCM iteration.

(b) Number of iterations in the inverse Lanczos algorithm per SCM iteration.

Figure 5.4: The SCM (see algorithm 2.2) applied to the Maxwell horn antenna problem.

possible values of μ of the relative difference between lower and upper bounds. The SCM error stagnates during the first 12 iterations because a new eigensolve at some value μ only improves the SCM bounds locally in the neighborhood of μ. It takes 13 iterations to cover the full parameter interval $[1.8,\ 2.2]$.

Figure 5.4b shows the number of iterations for each eigensolve. Most eigensolves require a number of iterations comprised between 10 and 20 which, in terms of computational costs, roughly corresponds to solving 20 to 40 high-fidelity problems.

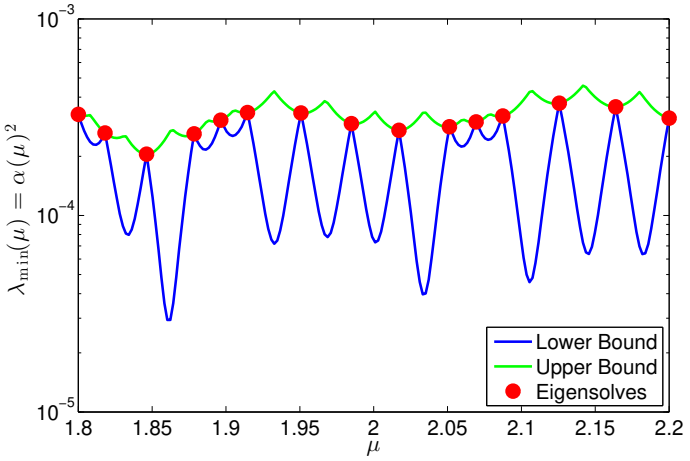
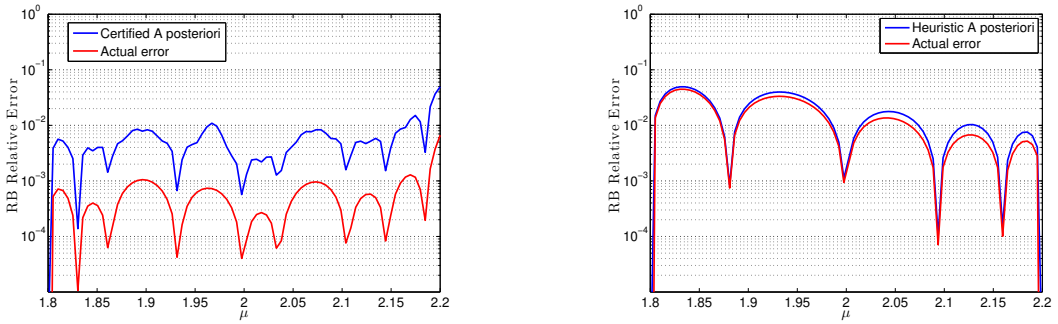


Figure 5.5: The SCM lower and upper bounds for the Maxwell horn antenna problem.

Figure 5.5 shows the lower and upper bounds obtained with the SCM. Looking at the exact eigensolves (red circles), the exact inf-sup constant $\mu \mapsto \alpha(\mu)$ seems to depend very little on μ. This is precisely the reason why the heuristic approach will perform well. We observe that the upper bounds better catch this "almost constant" behavior than the lower bounds. This is consistant with [112], in which the first order accuracy of the upper

bounds is demonstrated.

Applying the greedy algorithm 2.1 with a 5% prescribed tolerance, the certified approach yields a RB of size $N = 9$, while the heuristic yields a RB of size $N = 6$. In both approaches, the Galerkin RB approximation is considered. For the sake of validation, we have computed 80 truth solutions $u(\mu)$ at uniformly distributed values of μ. We use these truth solutions to compute the actual relative error $\|u(\mu) - u_N(\mu)\|_V / \|u(\mu)\|_V$ and compare the results with our *a posteriori* indicators $\Delta_N^{\text{rel}}(\mu)$ for the certified approach and $\widetilde{\Delta_N^{\text{rel}}}(\mu)$ for the heuristic approach. The comparisons are shown on fig. 5.6.



(a) Certified approach.

(b) Heuristic approach.

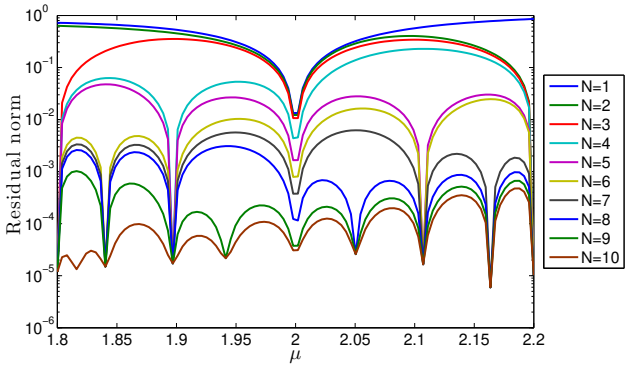
Figure 5.6: Two RB approximations for the Maxwell horn antenna problem.

The certified approach provides a RB approximation with an accuracy below 0.7% using $N = 9$ basis functions. It is therefore much more accurate than the prescribed 5%. The reason is that the certified indicator $\Delta_N^{\text{rel}}(\mu)$ overestimates the actual relative error roughly by a factor 10, as can be seen on fig. 5.6a. In opposition, the heuristic approach provides a RB approximation with an accuracy close to the prescribed 5% using just $N = 6$ basis functions. The heuristic indicator $\widetilde{\Delta_N^{\text{rel}}}(\mu)$ only slightly overestimates the actual relative error, as shown on fig. 5.6b. In terms of computational costs, the heuristic approach is much preferable than the certified approach: not only are the heavy computational costs associated to the SCM avoided, but also the final RB size is optimally small while still satisfying the prescribed tolerance, thus avoiding unnecessary computationally expensive truth solves.

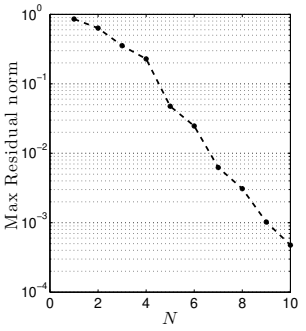
Convergence analysis

We now run the greedy algorithm driven by the residual norm for 10 iterations. Figure 5.7a shows the residual norm $\mu \mapsto \|A(\mu)u_N(\mu) - f(\mu)\|_{W'}$ for $N = 1, \dots, 10$. We observe that adding basis functions to the reduced basis does not only impact the residual norm locally in the neighborhood of the freshly computed frequency but globally over the frequency band of interest. This is particularly visible on the curve $N = 4$. Here, the freshly computed frequency is $\mu = 1.9$. Of course, the decrease is significant in the

neighborhood of $\mu = 1.9$, but we see the residual norm decrease for all other frequencies. This confirms the global approximation properties of the reduced basis. As shown on



(a) Residual norm with respect to μ for 10 successive RB sizes.



(b) Convergence curve of the maximum residual norm with respect to basis size.

Figure 5.7: Convergence of the RB method for the Maxwell horn antenna problem.

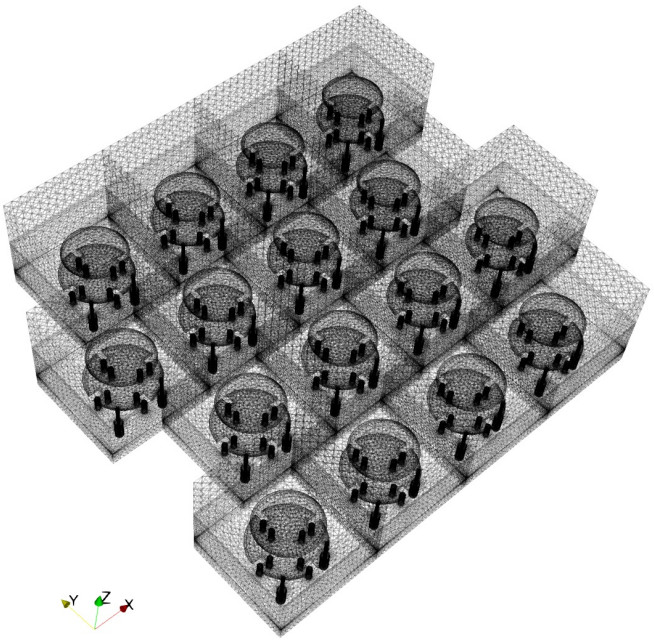
fig. 5.7, the maximum residual norm decreases exponentially with the reduced basis N. This corroborates the exponential decay of the truth solution manifold $\{u(\mu),\ \mu \in \mathcal{D}\}$.

We attract the attention of the reader that on fig. 5.7a the residual norm does not reach machine precision at the locations of the resolved frequencies. Indeed, the residual norm seems to stagnate somewhat around 10^{-5}. This is because the truth solutions $u(\mu)$ do not exactly satisfy the equation $A(\mu)u(\mu) = f(\mu)$. Indeed, the numerical solver FETI-2LM produces solutions with a relative residual $\|A(\mu)u(\mu) - f(\mu)\|_2/\|f(\mu)\|_2$ around 10^{-6}. The accuracy of the solver sets up a limit on the best possible accuracy that the reduced basis approximation can reach.

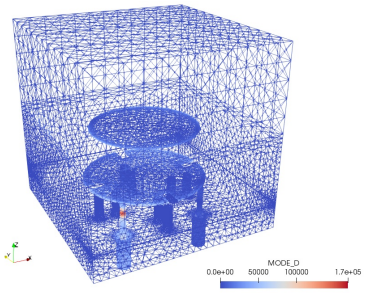
5.3 Application to antenna arrays

We now study a 4×4 antenna array shown on fig. 5.8a. Such antenna arrays are used for "beamforming", which includes for instance focusing the radiated signal in a given angular direction or towards a given point in space (see the review papers [121, 14]). or the book [52]). The problem consists of 16 antenna subdomains (see fig. 5.8b), each associated to a source term. The 16 antennas are independent from each other, therefore we are in the situation of a frequency-parametrized problem with $\ell = 16$ distinct source terms (see chapter 4). Notice that, although all 16 antennas are identical, no symmetry plane can be found on the antenna array and thus simulation of the full array is required.

In terms of mesh, each antenna subdomain is meshes with of $373,156$ degrees of freedom, thus the overall number of degrees of freedom for the full array is roughly $\mathcal{N} \sim 6,000,000$. The frequency range of interest is $8 - 12\text{GHz}$. As for the horn antenna test



(a) The 4×4 array comprised of 16 antenna subdomains (see right).



(b) One antenna subdomain (373, 156 degrees of freedom) .

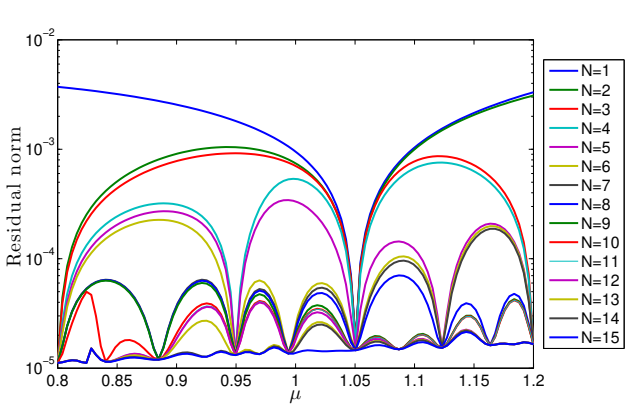
case, the adimensionalization coefficient is tuned in such a way that the set of parameters is $\mathcal{D} = [0.8,\ 1.2]$, therefore the frequency f_{Hz} (in Hz) can be easily recovered by the formula $f_{\text{Hz}} = \mu \times 10^{10}$.

5.3.1 RBM with a single right-hand side

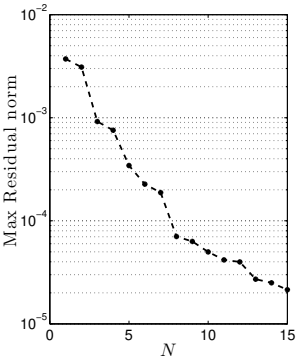
In this numerical experiment, all antennas are turned on. This means that there is one right-hand side with non-zero contributions from all subdomains (each sudomain being associated to a radiating antenna). We run the greedy algorithm driven by the residual norm in the euclidian norm $\mu \mapsto \|A(\mu)u_N(\mu) - f(\mu)\|_2$. The choice of the euclidian norm consists in setting the discrete inverse Riesz operator matrix $\mathbf{B}$ to the identity matrix. We have two motivations for this choice: (i) matrix-vector operations and system solves with the identity matrix are relatively inexpensive compared to similar operations performed with the "mathematically rigorous" discrete inverse Riesz operator matrix $\mathbf{B}$ given by eq. (5.1.21) and (ii) the FETI-2LM residual is also measured in the euclidian norm, which makes it easier to prescribe a tolerance on the RB residual that is consistent with prescribed tolerance on FETI-2LM iterations. Here, we guarantee that the FETI-2LM solver produces truth solutions $u(\mu)$ with a relative residual $\|A(\mu)u(\mu) - f(\mu)\|_2 / \|f(\mu)\|_2$ below 10^{-3}. Given that the euclidian norm of the right-hand side is roughly $\|f(\mu)\|_2 = \mathcal{O}(10^{-2})$, this means that $\|A(\mu)u(\mu) - f(\mu)\|_2 = \mathcal{O}(10^{-5})$. In order to reach about the same level

of accuracy with the RBM, we propose to set the prescribed tolerance in the RB greedy algorithm to 2.5×10^{-5}.

The algorithm terminates with a RB of size $N = 15$. Using 16 CPUs (one per MPI process associated to an antenna subdomain), the overall elapsed time was $1\text{h}45\text{min}$. We note that 92% of the elapsed time ($1\text{h}36\text{min}$) was spent on the FETI-2LM iterations and the remaining 8% on other operations including sparse matrix-vector operations, dot products, RB solves, etc.



(a) Residual norm with respect to μ for 10 successive RB sizes.



(b) Convergence curve of the maximum residual norm with respect to basis size.

Figure 5.9: Convergence of the RB method for the Maxwell array antenna problem, with only one activated antenna.

Figure 5.9a shows the residual norm throughout the iterations of the greedy algorithm and fig. 5.9b confirms the exponential decrease of the maximum residual norm. The stagnation of the RB residual around 10^{-5} is consistent with the prescribed tolerance set on the FETI-2LM iterations. We observe that the RB of size $N = 15$ is associated with residual norms across the full frequency band of interest of the same order magnitude than those obtained using the FETI-2LM numerical solver. This illustrates the ability of the RBM to recover very accurate solutions with minimal computational effort.

5.3.2 RBM with multiple right-hand sides

In such antenna application, we want to study the electric field generated by one illuminated antenna while the other antennas are turned off and we want to perform this analysis for each antenna. In other words, we are interested in the solutions associated to 16 distinct right-hand sides, the kth right-hand side being equal to 0 in all subdomains except in the kth subdomain which corresponds to the illuminated antenna.

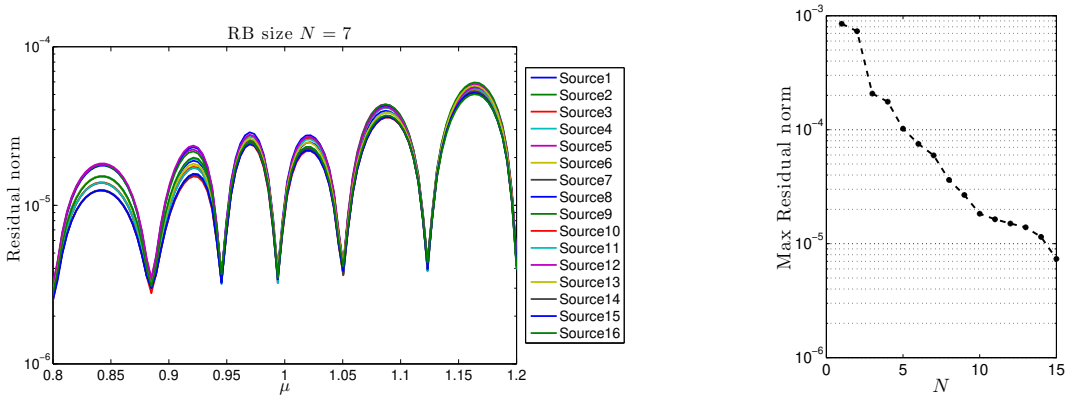
It takes about 390s ($6\text{min}30$) elapsed time to obtain the truth solution associated to one right-hand side using FETI-2LM on 16 CPUs (here, we use one CPU per MPI process).

This means that the computation of 16 truth solutions associated to 16 distinct sources could take:

- either the same time but using significantly more computational resources (namely, 256 CPUs would be required),
- or much longer (namely, about 1h44min elapsed time) with the same workstation.

An alternative, is to combine FETI-2LM with a block Krylov recycling strategy for efficiently solving multiple right-hand sides [40, 117, 103]. In this context, the number of arythmetic operations to be performed per FETI-2LM iteration increases with the number of right-hand sides, thus it becomes relevant to use multiple CPUs per MPI process in order to take the best advantage of the multi-threading possibilities offered by OpenMP. For example, on 48 CPUs (3 CPUs per MPI process) the 16 truth solutions associated to the 16 distinct radiating sources is obtained in about 950s (15min50) with the same prescribed tolerance 10^{-3} on the FETI-2LM iterations. This numerical experiment illustrates the use of FETI-2LM with a block Krylov recycling represents a clear advantage over successive calls to FETI-2LM with single right-hand side.

We simultaneously generate 16 RB approximation spaces by running the multi-source greedy algorithm 4.1 (which corresponds to the *multiple RBs* strategy presented in chapter 4). This algorithm takes the best advantage of the block Krylov recycling strategy. We briefly recall from chapter 4 that algorithm 4.1 consists in a greedy algorithm driven by the maximum residual norm over all 16 sources. Again, the euclidian norm is used for enhanced performance and improved readability of the prescribed tolerances. Knowing from previous numerical experiment in the single source configuration that $N = 15$ should be a sufficient RB size, we let the algorithm run for $N = 15$ iterations.



(a) Residuals associated to the 16 sources with a RB of size $N = 7$.

(b) Maximum residual norm (over all sources and all $\mu \in \Xi \subset \mathcal{D}$) with respect to RB size.

Figure 5.10: Convergence of the RB method for the Maxwell array antenna problem, with 16 distinct sources.

Figure 5.10a shows the residuals associated to each source at the iteration $N = 7$ of the multi-source greedy algorithm algorithm 4.1. Although they do not coincide, we find that all residuals have roughly the same behavior with respect to μ. Hence, it makes sense to select the next frequency to be computed by maximizing the residual over all sources and over a discrete training set $\Xi \subset \mathcal{D} = [0.8,\, 1.2]$. Here, based on the residuals plotted on fig. 5.10a, the algorithm selects the frequency $\mu = 1.1636$ to be computed in the 8th iteration.

Figure 5.10b shows the maximum residual norm over all sources and all $\mu \in \Xi$ throughout the greedy iterations. We find an exponential decrease which compares well with the single-source case.

The elapsed time for simultaneously generating the 16 RB models of size $N = 15$ on 48 CPUs is $4\text{h}16\text{min}$. Recalling that it took $1\text{h}45\text{min}$ on 16 CPUs to generate a RB model for one single source, a single right-hand side strategy would have required $9\text{h}20$ on 48 CPUs to construct these 16 RB models. Thus, we evaluate the multi-source strategy to be more than twice faster than the strategy that consists in repeated calls to single-source greedy algorithms.

5.4 Conclusions

In this chapter, we have successfully applied the reduced basis method to antenna applications in electromagnetism with frequency parameter. The underlying equations are the Maxwell equations solved using edge finite elements and a FETI-2LM domain-decomposition method. This problem fits the reduced basis framework with no particular difficulty, because both operator and right-hand sides are naturally affine in the frequency.

We have shown through a numerical example with a horn antenna that the inf-sup stability constant has minor dependency on the frequency. This is consistent with the fact that the underlying equations admit no real resonant frequencies. In this context, the heuristic method of approximating the inf-sup stability factor by a constant (*i.e.*, independent from the frequency) provides excellent results, with considerable speed-ups compared to the traditional certified reduced basis method based on the construction of rigorous lower bounds for the inf-sup stability constant using the SCM.

This chapter has also demonstrated the potential of the RBM for frequency sweeps with large-scale problems featuring multiple sources. The computational costs of the offline phase were limited thanks to the choice of the euclidian norm rather than the $\boldsymbol{H}^{+}_{0,\Gamma_D}(\mathbf{curl}, \Omega)$ norm and an efficient multiple right-hand side strategy using block Krylov recycling. Thanks to the RBM, we have been able to reduce an antenna array problem with $6,000,000$ unknowns and 16 sources to 16 RB models (one per source) each featuring only $N = 15$ unknowns, while maintaining a level of accuracy that is commensurate with that of the high-fidelity numerical solver.

Chapter 6

Reduced basis method for frequency sweeps with the boundary element method

Summary. In this chapter, we explain how the RBM can be used to address frequency-dependent problems in electromagnetic scattering. We consider electromagnetic scattering problems discretized with the boundary element method (BEM), which is well-known to provide complex, non-hermitian and fully populated linear systems. Our first contribution is the use of *non-intrusive local affine approximations* of the frequency-dependent BEM operator, which enables the user to better control the computational complexity of the offline phase. Our second contribution is the use of nested reduced basis approximation spaces rather than one global approximation space, in view of reducing the offline costs of the RBM. Our strategy is illustrated using various numerical examples of both academic and industrial interests.

Contents

6.1 Strong formulation and high-fidelity approximation 123
6.1.1 The Stratton-Chu integral representation formulas 123
6.1.2 Scattering by a perfect electric conductor 124
6.1.3 Discretization using the BEM 128
6.1.4 Numerical solvers 129
6.2 A non-intrusive RBM for frequency sweep analysis 130
6.2.1 Affine approximations for the frequency-parametrized problem 130
6.2.2 Non-intrusive local affine approximations 132
6.2.3 Non-intrusive RB approximation 135
6.2.4 Greedy construction 136
6.2.5 Monolithic construction 138

6.3 Numerical illustrations 139
6.3.1 EFIE vs CFIE: tests on the unit sphere 139
6.3.2 Approximation of the CFIE operator on the geometry of a fighter aircraft 142
6.3.3 RB approximations on the geometry of a fighter aircraft 143
6.3.4 Broadband frequency-sweeps 146
6.4 Conclusions 150

6.1 Strong formulation and high-fidelity approximation

6.1.1 The Stratton-Chu integral representation formulas

Recall from chapter 5 that, in the absence of charges, the (re-normalized) electric and magnetic fields $\boldsymbol{e}$, $\boldsymbol{h}$ satisfy the set of equations (5.1.6), repeated here for convenience

$$\begin{cases} -ik\boldsymbol{e} + \mathbf{curl}\,\boldsymbol{h} &= \boldsymbol{j}, \\ ik\boldsymbol{h} + \mathbf{curl}\,\boldsymbol{e} &= 0. \end{cases} \tag{6.1.1}$$

In the following, we consider no radiating sources, hence the source term $\boldsymbol{j}$ is set to zero. We consider regular fields $\boldsymbol{e}$, $\boldsymbol{h}$ satisfying (6.1.1) in some bounded interior domain $\Omega_i \subset \mathbb{R}^3$ as well as in the associated (unbounded) exterior domain $\Omega_e = \mathbb{R}^3 \setminus \overline{\Omega_i}$ and satisfying the following Silver-Muller radiation condition at infinity

$$\lim_{|\boldsymbol{x}|\to\infty} |\boldsymbol{x}| \left| \boldsymbol{e} - \boldsymbol{h} \times \frac{\boldsymbol{x}}{|\boldsymbol{x}|} \right| = 0, \tag{6.1.2}$$

where $|\cdot|$ denotes the usual euclidian norm of $\mathbb{R}^3$.

Denote $\Gamma = \partial\Omega_i$ the closed surface at the interface between the interior and exterior domains and $\hat{\boldsymbol{n}}$ the exterior unit normal to Γ, as shown on figure 6.1. We define the electric and magnetic surface currents on Γ as

$$\begin{aligned} \boldsymbol{j}_\Gamma &= \boldsymbol{h}_i \times \hat{\boldsymbol{n}} - \boldsymbol{h}_e \times \hat{\boldsymbol{n}}, \\ \boldsymbol{m}_\Gamma &= \boldsymbol{e}_i \times \hat{\boldsymbol{n}} - \boldsymbol{e}_e \times \hat{\boldsymbol{n}}, \end{aligned} \tag{6.1.3}$$

where $\boldsymbol{h}_i$ and $\boldsymbol{e}_i$ (resp. $\boldsymbol{h}_e$ and $\boldsymbol{e}_e$) denote the interior limits (resp. the exterior limits) of the magnetic field $\boldsymbol{h}$ and electric field $\boldsymbol{e}$.

We further introduce $\mathcal{G}(\cdot,\cdot;k)$, the ingoing fundamental solution of the 3-dimensional Helmholtz equation at wavenumber k,

$$\mathcal{G}(\boldsymbol{x},\boldsymbol{y};k) = \frac{e^{-ik|\boldsymbol{x}-\boldsymbol{y}|}}{4\pi|\boldsymbol{x}-\boldsymbol{y}|}, \quad \boldsymbol{x} \neq \boldsymbol{y}. \tag{6.1.4}$$

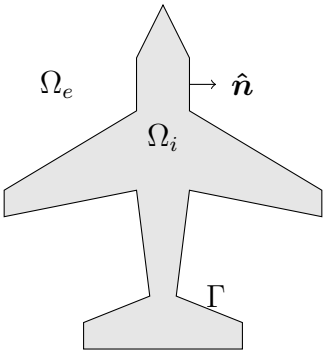


Figure 6.1: Schematic representation of the interior and exterior domains.

The stage is now set for the Stratton-Chu representation formulas [81, Theorem 5.5.1],

$$\forall \boldsymbol{x} \notin \Gamma, \quad \begin{cases} \boldsymbol{e}(\boldsymbol{x}) = -ik\mathscr{T}_k \boldsymbol{j}_\Gamma(\boldsymbol{x}) + \mathscr{K}_k \boldsymbol{m}_\Gamma(\boldsymbol{x}), \\ \boldsymbol{h}(\boldsymbol{x}) = ik\mathscr{T}_k \boldsymbol{m}_\Gamma(\boldsymbol{x}) + \mathscr{K}_k \boldsymbol{j}_\Gamma(\boldsymbol{x}), \end{cases} \tag{6.1.5}$$

with electric and magnetic potentials $\mathscr{T}_k$ and $\mathscr{K}_k$ respectively defined for all $\boldsymbol{x} \notin \Gamma$ by

$$\begin{aligned} \mathscr{T}_k \boldsymbol{j}_\Gamma(\boldsymbol{x}) = &\int_\Gamma \mathcal{G}(\boldsymbol{x}, \boldsymbol{y}; k) \boldsymbol{j}_\Gamma(\boldsymbol{y}) \mathrm{d}\Gamma_{\boldsymbol{y}} \\ &+ \frac{1}{k^2} \mathbf{grad}_{\boldsymbol{x}} \int_\Gamma \mathcal{G}(\boldsymbol{x}, \boldsymbol{y}; k) \mathrm{div}_{\Gamma, \boldsymbol{y}}\, \boldsymbol{j}_\Gamma(\boldsymbol{y}) \mathrm{d}\Gamma_{\boldsymbol{y}}, \end{aligned} \tag{6.1.6a}$$

$$\mathscr{K}_k \boldsymbol{m}_\Gamma(\boldsymbol{x}) = \mathbf{curl}_{\boldsymbol{x}} \int_\Gamma \mathcal{G}(\boldsymbol{x}, \boldsymbol{y}; k) \boldsymbol{m}_\Gamma(\boldsymbol{y}) \mathrm{d}\Gamma_{\boldsymbol{y}}. \tag{6.1.6b}$$

The Stratton-Chu representation formulas (6.1.5) have an exceptional consequence. Namely, that the electric and magnetic fields $\boldsymbol{e}$ and $\boldsymbol{h}$ can be computed anywhere in the interior domain Ω_i or in the exterior domain Ω_e, with just the knowledge of the electric and magnetic surface currents $\boldsymbol{j}_\Gamma$ and $\boldsymbol{m}_\Gamma$ defined on Γ.

6.1.2 Scattering by a perfect electric conductor

Let $\boldsymbol{e}^{\mathtt{inc}}, \boldsymbol{h}^{\mathtt{inc}}$ be given continuous incident fields. One can consider plane waves given by

$$\forall \boldsymbol{x} \in \mathbb{R}^3, \quad \begin{cases} \boldsymbol{e}^{\mathtt{inc}}(\boldsymbol{x}) &= \widehat{\boldsymbol{p}} e^{ik\widehat{\boldsymbol{d}}\cdot\boldsymbol{x}}, \\ \boldsymbol{h}^{\mathtt{inc}}(\boldsymbol{x}) &= (\widehat{\boldsymbol{d}} \times \widehat{\boldsymbol{p}}) e^{ik\widehat{\boldsymbol{d}}\cdot\boldsymbol{x}}, \end{cases} \tag{6.1.7}$$

where the unit vectors $\widehat{\boldsymbol{d}}, \widehat{\boldsymbol{p}}$ respectively denote the direction and polarization of the plane wave. We recall that the polarization $\widehat{\boldsymbol{p}}$ is always perpendicular to the direction $\widehat{\boldsymbol{d}}$.

We look for the total electric and magnetic fields $\boldsymbol{e}^{\mathtt{tot}}, \boldsymbol{h}^{\mathtt{tot}}$ in the exterior domain Ω_e under the following form $\boldsymbol{e}^{\mathtt{tot}} = \boldsymbol{e}^{\mathtt{inc}} + \boldsymbol{e}$ and $\boldsymbol{h}^{\mathtt{tot}} = \boldsymbol{h}^{\mathtt{inc}} + \boldsymbol{h}$ where $\boldsymbol{e}, \boldsymbol{h}$ denote the

scattered fields. We consider the interior domain Ω_i to be a perfect electric conductor (PEC), hence the following boundary condition is satisfied on Γ

$$\boldsymbol{e}_e^{\mathtt{tot}} \times \hat{\boldsymbol{n}} = 0, \tag{6.1.8a}$$

where we recall that $\boldsymbol{e}_e$ denotes the exterior limit on Γ of the field $\boldsymbol{e}$. Equivalently,

$$\boldsymbol{e}_e \times \hat{\boldsymbol{n}} = \boldsymbol{e}_e^{\mathtt{inc}} \times \hat{\boldsymbol{n}}. \tag{6.1.8b}$$

In the exterior domain Ω_e, the scattered fields $\boldsymbol{e}, \boldsymbol{h}$ must further satisfy the governing equations eq. (6.1.1) with no source term (*i.e.*, $\boldsymbol{j} = 0$) as well as the Silver-Muller radiation condition at infinity (6.1.2).

From the PEC nature of the interior domain, the total fields are zero inside the interior domain. This is recovered by extending the scattered fields $\boldsymbol{e}, \boldsymbol{h}$ to the interior domain Ω_i and imposing them to be equal to $-\boldsymbol{e}^{\mathtt{inc}}, -\boldsymbol{h}^{\mathtt{inc}}$. In other words,

$$\begin{cases} \boldsymbol{h} = -\boldsymbol{h}^{\mathtt{inc}}, \\ \boldsymbol{e} = -\boldsymbol{e}^{\mathtt{inc}}. \end{cases} \qquad \text{in } \Omega_i. \tag{6.1.9}$$

Exploiting the continuity of the incident field across Γ and the PEC boundary condition (6.1.8b), we deduce the continuity of the tangential trace of the electric field across Γ. We conclude that there are no magnetic surface currents on Γ, that is $\boldsymbol{m}_\Gamma = 0$. Thus, by virtue of the Stratton-Chu formulas (6.1.5), the scattered fields $\boldsymbol{e}, \boldsymbol{h}$ are represented by

$$\forall \boldsymbol{x} \notin \Gamma, \quad \begin{cases} \boldsymbol{e}(\boldsymbol{x}) = -ik\mathscr{T}_k \boldsymbol{j}_\Gamma(\boldsymbol{x}), \\ \boldsymbol{h}(\boldsymbol{x}) = \mathscr{K}_k \boldsymbol{j}_\Gamma(\boldsymbol{x}), \end{cases} \tag{6.1.10}$$

with

$$\boldsymbol{j}_\Gamma = -\boldsymbol{h}^{\mathtt{inc}} \times \hat{\boldsymbol{n}} - \boldsymbol{h}_e \times \hat{\boldsymbol{n}} = -\boldsymbol{h}_e^{\mathtt{tot}} \times \hat{\boldsymbol{n}}. \tag{6.1.11}$$

Physically, the electric surface current $\boldsymbol{j}_\Gamma$ therefore coincides with the exterior tangential trace of the total magnetic field. In order to build a formulation in terms of the only unknown $\boldsymbol{j}_\Gamma$, we study the limit of eq. (6.1.10) as $\boldsymbol{x}$ approaches the boundary Γ. In this work, we do not go into the details of how this limit is obtained (see [17, §5], [7, §4.1]), but we recall the main results. Let us introduce the *interior tangential components trace mapping*

$$\pi_T^i : \boldsymbol{v} \in \boldsymbol{C}^\infty(\overline{\Omega}_i) \mapsto \hat{\boldsymbol{n}} \times (\boldsymbol{v} \times \hat{\boldsymbol{n}})|_\Gamma. \tag{6.1.12}$$

Here, $\boldsymbol{C}^\infty(\overline{\Omega}_i) = \left(C^\infty(\overline{\Omega}_i)\right)^3$ is the space of infinitely differentiable vector fields defined on the closure of the interior domain. We similarly introduce the *exterior tangential components trace mapping* $\pi_T^e : \boldsymbol{v} \in \boldsymbol{C}^\infty(\overline{\Omega}_e) \mapsto \hat{\boldsymbol{n}} \times (\boldsymbol{v} \times \hat{\boldsymbol{n}})|_\Gamma$.

The Electric Field Integral Equation

The tangential components trace of the electric potential $\mathscr{T}_k$ is continuous across Γ. Indeed, given a surface current $\boldsymbol{j}_\Gamma$ defined on Γ, there holds

$$\boldsymbol{x} \in \Gamma, \quad \pi_T^i \mathscr{T}_k \boldsymbol{j}_\Gamma(\boldsymbol{x}) = \pi_T^e \mathscr{T}_k \boldsymbol{j}_\Gamma(\boldsymbol{x}) \equiv T\boldsymbol{j}_\Gamma(\boldsymbol{x}), \tag{6.1.13}$$

and the value of the tangential components trace is given by

$$\begin{aligned} \boldsymbol{x} \in \Gamma, \quad T\boldsymbol{j}_\Gamma(\boldsymbol{x}) = &\int_\Gamma \mathcal{G}(\boldsymbol{x},\boldsymbol{y};k)\boldsymbol{j}_\Gamma(\boldsymbol{y})\mathrm{d}\Gamma_{\boldsymbol{y}} \\ &+ \frac{1}{k^2}\mathbf{grad}_{\Gamma,\boldsymbol{x}} \int_\Gamma \mathcal{G}(\boldsymbol{x},\boldsymbol{y};k)\mathrm{div}_{\Gamma,\boldsymbol{y}}\, \boldsymbol{j}_\Gamma(\boldsymbol{y})\mathrm{d}\Gamma_{\boldsymbol{y}}. \end{aligned} \tag{6.1.14}$$

Applying either the interior or exterior tangential components trace (denoted π_T) to the Stratton-Chu representation formula (6.1.10) for the electric field yields,

$$\forall \boldsymbol{x} \in \Gamma, \quad -\pi_T \boldsymbol{e}^{\mathtt{inc}}(\boldsymbol{x}) = -ikT\boldsymbol{j}_\Gamma(\boldsymbol{x}). \tag{6.1.15}$$

We recognize an integral equation satisfied by the electric surface current $\boldsymbol{j}_\Gamma$. This equation is know as the *Electric Field Integral Equation* (EFIE).

The Magnetic Field Integral Equation

In opposition to the electric potential $\mathscr{T}_k$, the tangential components trace of the electric potential $\mathscr{K}_k$ is discontinuous across Γ. Indeed, given the surface current $\boldsymbol{m}_\Gamma$ defined on Γ, there holds

$$\boldsymbol{x} \in \Gamma, \quad \begin{cases} \pi_T^i \mathscr{K}_k \boldsymbol{m}_\Gamma(\boldsymbol{x}) = -\frac{1}{2}\hat{\boldsymbol{n}} \times \boldsymbol{m}_\Gamma(\boldsymbol{x}) + K\boldsymbol{m}_\Gamma(\boldsymbol{x}) \\ \pi_T^e \mathscr{K}_k \boldsymbol{m}_\Gamma(\boldsymbol{x}) = \frac{1}{2}\hat{\boldsymbol{n}} \times \boldsymbol{m}_\Gamma(\boldsymbol{x}) + K\boldsymbol{m}_\Gamma(\boldsymbol{x}), \end{cases} \tag{6.1.16}$$

where

$$\boldsymbol{x} \in \Gamma, \quad K\boldsymbol{m}_\Gamma(\boldsymbol{x}) = \int_\Gamma \frac{\partial}{\partial \hat{\boldsymbol{n}}_{\boldsymbol{y}}} \mathcal{G}(\boldsymbol{x},\boldsymbol{y};k) \times \boldsymbol{m}_\Gamma(\boldsymbol{y})\mathrm{d}\Gamma_{\boldsymbol{y}}. \tag{6.1.17}$$

Applying the interior tangential components trace to the Stratton-Chu representation formula (6.1.10) for the magnetic field yields another integral equation satisfied by the electric surface current $\boldsymbol{j}_\Gamma$, namely

$$\forall \boldsymbol{x} \in \Gamma, \quad -\pi_T^i \boldsymbol{h}(\boldsymbol{x}) = -\frac{1}{2}\hat{\boldsymbol{n}} \times \boldsymbol{j}_\Gamma(\boldsymbol{x}) + K\boldsymbol{j}_\Gamma(\boldsymbol{x}). \tag{6.1.18}$$

Observing that

$$\begin{aligned}\hat{\boldsymbol{n}} \times \pi_T^i \boldsymbol{h}(\boldsymbol{x}) &= \hat{\boldsymbol{n}} \times (\hat{\boldsymbol{n}} \times (\boldsymbol{h}_i(\boldsymbol{x}) \times \hat{\boldsymbol{n}})) \\ &= \hat{\boldsymbol{n}} \times (\boldsymbol{h}_i(\boldsymbol{x}) - (\boldsymbol{h}_i(\boldsymbol{x}) \cdot \hat{\boldsymbol{n}})\hat{\boldsymbol{n}}) \\ &= \hat{\boldsymbol{n}} \times \boldsymbol{h}_i(\boldsymbol{x}),\end{aligned} \tag{6.1.19}$$

and recalling that $\boldsymbol{h}_i$, the interior limit of $\boldsymbol{h}$, is equal to $-\boldsymbol{h}^{\mathtt{inc}}$, we obtain, from taking the left cross product by $\hat{\boldsymbol{n}}$ in (6.1.18), that,

$$\forall \boldsymbol{x} \in \Gamma, \quad \hat{\boldsymbol{n}} \times \boldsymbol{h}^{\mathtt{inc}}(\boldsymbol{x}) = \frac{1}{2}\boldsymbol{j}_\Gamma(\boldsymbol{x}) + \hat{\boldsymbol{n}} \times K\boldsymbol{j}_\Gamma(\boldsymbol{x}). \tag{6.1.20}$$

Note that we have used that $\hat{\boldsymbol{n}} \times (\hat{\boldsymbol{n}} \times \boldsymbol{j}_\Gamma) = \boldsymbol{j}_\Gamma$, which stems from the fact that $\boldsymbol{j}_\Gamma$ is a tangent field to Γ, recalling its definition (6.1.3). Equation (6.1.20) is known as the *Magnetic Field Integral Equation* (MFIE).

The Combined Field Integral Equation

At this point, we have derived two different integral equations satisfied by the electric surface current $\boldsymbol{j}_\Gamma$. A natural question that arises, which best solves the PEC scattering problem? This is a classical discussion [81, 32], out of which we retain the following points:

- **Robustness.** Both the EFIE and MFIE are ill-posed for some values of the wavenumber k, corresponding to the interior resonant wavenumbers. For such wavenumbers k, there exist spurious surface currents. This does not affect the electric and magnetic fields reconstructed using the Stratton-Chu representation formula when using the EFIE, but it does when using the MFIE.
- **Conditioning.** Being a first-kind integral equation, the EFIE is prone to ill-conditioning whereas the MFIE is a second-kind integral equation and has a better conditioning.
- **Flexibility.** The EFIE can be used to compute the electric surface currents open surfaces Γ, but this is not the case for the MFIE, that requires Γ to be a closed surface due to the explicit dependency in the outgoing normal.

A way to circumvent some of these issues, especially to obtain well-posed formulation for any wavenumber k, is to combine the EFIE and MFIE into one *Combined Field Integral Equation* (CFIE). That is,

$$\mathrm{CFIE} = (1-c)\mathrm{EFIE} + c\mathrm{MFIE}, \tag{6.1.21}$$

with a well-chosen value for $c \in]0, 1[$. In this work, we choose the value $c = 0.5$.

6.1.3 Discretization using the BEM

In order to numerically solve the electromagnetic scattering problem, we use the Boundary Element Method (BEM) to discretize the weak forms associated to the EFIE and MFIE. First, the surface Γ is meshed with triangles. We denote $\mathcal{T}_h$ the set of all the triangles in the mesh. Given a triangle $E \in \mathcal{T}_h$, the zero-th order local Raviart-Thomas space of complex-valued functions defined on E is given by $\boldsymbol{RT}_0(E) = \{\boldsymbol{v} : \boldsymbol{x} \in E \mapsto \boldsymbol{\alpha} + \beta\boldsymbol{x} \mid \boldsymbol{\alpha} \in \mathbb{C}^2,\ \beta \in \mathbb{C}\}$, see [99]. Following the standard BEM [56, 16], we choose as boundary element approximation space the global Raviart-Thomas space, given by

$$\boldsymbol{V}_h^{\mathrm{RT}} = \{\boldsymbol{v} \in \boldsymbol{H}^0_{\mathrm{div}}(\Gamma) \mid \boldsymbol{v}|_E \in \boldsymbol{RT}_0(E),\ \forall E \in \mathcal{T}_h\}, \tag{6.1.22}$$

where $\boldsymbol{H}^0_{\mathrm{div}}(\Gamma) = \{\boldsymbol{v} \in \boldsymbol{L}^2_t(\Gamma) \mid \mathrm{div}_\Gamma\, \boldsymbol{v} \in L^2(\Gamma)\}$ with $L^2(\Gamma)$ the Sobolev space of square-integrable complex-valued functions defined on Γ and $\boldsymbol{L}^2_t(\Gamma)$ the classical Sobolev space comprised of complex-valued functions $\boldsymbol{v} \in \boldsymbol{L}^2(\Gamma) = [L^2(\Gamma)]^3$ that are tangential to Γ, *i.e.*, satisfying $\boldsymbol{v} \cdot \widehat{\boldsymbol{n}} = 0$. Note that the boundary element approximation space $\boldsymbol{V}_h^{\mathrm{RT}}$ is finite dimensional; we denote $\mathcal{N}$ its dimension (equal to the number of triangles in the mesh). We now seek high-fidelity approximations $\boldsymbol{j}_{\Gamma,h}$ of $\boldsymbol{j}_\Gamma$ in the approximation space $\boldsymbol{V}_h^{\mathrm{RT}}$.

The EFIE

Formally (without going into more technical details), we can multiply the EFIE eq. (6.1.15) by any test function $\boldsymbol{w}_h \in \boldsymbol{V}_h^{\mathrm{RT}}$, integrate over Γ and integrate by parts. This yields the following discrete weak form: *find* $\boldsymbol{j}^{\mathrm{E}}_{\Gamma,h} \in \boldsymbol{V}_h^{\mathrm{RT}}$ *such that*

$$\forall \boldsymbol{w}_h \in \boldsymbol{V}_h^{\mathrm{RT}}, \quad \langle \boldsymbol{T}_h \boldsymbol{j}^{\mathrm{E}}_{\Gamma,h}, \boldsymbol{w}_h \rangle = \langle \boldsymbol{b}^{\mathrm{E}}_h, \boldsymbol{w}_h \rangle \tag{6.1.23}$$

where the discrete EFIE operator $\boldsymbol{T}_h$ is expressed as

$$\begin{aligned} \forall \boldsymbol{v}_h, \boldsymbol{w}_h \in \boldsymbol{V}_h^{\mathrm{RT}}, \quad \langle \boldsymbol{T}_h \boldsymbol{v}_h, \boldsymbol{w}_h \rangle = ik \int_\Gamma \overline{\boldsymbol{w}_h(\boldsymbol{x})} \cdot \int_\Gamma \mathcal{G}(\boldsymbol{x}, \boldsymbol{y}; k) \boldsymbol{v}_h(\boldsymbol{y}) \mathrm{d}\Gamma_{\boldsymbol{y}} \mathrm{d}\Gamma_{\boldsymbol{x}} \\ - \frac{i}{k} \int_\Gamma \mathrm{div}_{\Gamma,\boldsymbol{x}}\, \overline{\boldsymbol{w}_h(\boldsymbol{x})} \int_\Gamma \mathcal{G}(\boldsymbol{x}, \boldsymbol{y}; k) \mathrm{div}_{\Gamma,\boldsymbol{y}}\, \boldsymbol{v}_h(\boldsymbol{y}) \mathrm{d}\Gamma_{\boldsymbol{y}} \mathrm{d}\Gamma_{\boldsymbol{x}}, \end{aligned} \tag{6.1.24}$$

and the right-hand side (RHS) is given by

$$\forall \boldsymbol{w}_h \in \boldsymbol{V}_h^{\mathrm{RT}}, \quad \langle \boldsymbol{b}^{\mathrm{E}}_h, \boldsymbol{w}_h \rangle = \int_\Gamma \pi_T \boldsymbol{e}^{\mathtt{inc}}(\boldsymbol{x}) \cdot \overline{\boldsymbol{w}_h(\boldsymbol{x})} \mathrm{d}\Gamma_{\boldsymbol{x}}. \tag{6.1.25}$$

The MFIE

Similarly, we can multiply the MFIE eq. (6.1.20) by any test function $\boldsymbol{w}_h \in \boldsymbol{V}_h^{\mathrm{RT}}$ and integrate over Γ. This yields the following discrete weak form: *find* $\boldsymbol{j}^{\mathrm{M}}_{\Gamma,h} \in \boldsymbol{V}_h^{\mathrm{RT}}$ *such that*

$$\forall \boldsymbol{w}_h \in \boldsymbol{V}_h^{\mathrm{RT}}, \quad \langle \left(\frac{1}{2}\boldsymbol{I} + \boldsymbol{K}_h \right) \boldsymbol{j}^{\mathrm{M}}_{\Gamma,h}, \boldsymbol{w}_h \rangle = \langle \boldsymbol{b}^{\mathrm{M}}_h, \boldsymbol{w}_h \rangle \tag{6.1.26}$$

where the discrete the discrete MFIE operator $\frac{1}{2}\boldsymbol{I} + \boldsymbol{K}_h$ is given by

$$\begin{aligned}\forall \boldsymbol{v}_h, \boldsymbol{w}_h \in \boldsymbol{V}_h^{\mathrm{RT}}, \quad \langle \left(\frac{1}{2}\boldsymbol{I} + \boldsymbol{K}_h\right) \boldsymbol{v}_h, \boldsymbol{w}_h \rangle = \frac{1}{2}\int_\Gamma \overline{\boldsymbol{w}_h(\boldsymbol{x})} \cdot \boldsymbol{v}_h(\boldsymbol{x}) \mathrm{d}\Gamma_{\boldsymbol{x}} \\ + \int_\Gamma \overline{\boldsymbol{w}_h(\boldsymbol{x})} \cdot \left(\widehat{\boldsymbol{n}}(\boldsymbol{x}) \times \int_\Gamma \frac{\partial}{\partial \widehat{\boldsymbol{n}}(\boldsymbol{y})} \mathcal{G}(\boldsymbol{x}, \boldsymbol{y}; k) \times \boldsymbol{v}_h(\boldsymbol{y}) \mathrm{d}\Gamma_{\boldsymbol{y}}\right) \mathrm{d}\Gamma_{\boldsymbol{x}}.\end{aligned} \tag{6.1.27}$$

and RHS is given by

$$\forall \boldsymbol{w}_h \in \boldsymbol{V}_h^{\mathrm{RT}}, \quad \langle \boldsymbol{b}_h^{\mathrm{M}}, \boldsymbol{w}_h \rangle = \int_\Gamma \widehat{\boldsymbol{n}} \times \boldsymbol{h}^{\mathrm{inc}}(\boldsymbol{x}) \cdot \overline{\boldsymbol{w}_h(\boldsymbol{x})} \mathrm{d}\Gamma_{\boldsymbol{x}}. \tag{6.1.28}$$

The CFIE

This being set the discrete CFIE writes: *find* $\boldsymbol{j}_{\Gamma,h}^{\mathrm{C}} \in \boldsymbol{V}_h^{\mathrm{RT}}$ *such that*

$$\forall \boldsymbol{w}_h \in \boldsymbol{V}_h^{\mathrm{RT}}, \quad \langle \boldsymbol{C}_h \boldsymbol{j}_{\Gamma,h}^{\mathrm{C}}, \boldsymbol{w}_h \rangle = \langle \boldsymbol{b}_h^{\mathrm{C}}, \boldsymbol{w}_h \rangle \tag{6.1.29}$$

with operator

$$\boldsymbol{C}_h = (1-c)\boldsymbol{T}_h + c\left(\frac{1}{2}\boldsymbol{I} + \boldsymbol{K}_h\right) \tag{6.1.30}$$

and the associated right-hand side is given by $\boldsymbol{b}_h^{\mathrm{C}} = (1-c)\boldsymbol{b}_h^{\mathrm{E}} + c\boldsymbol{b}_h^{M}$.

6.1.4 Numerical solvers

In matrix form, the discrete problem writes as a linear system of equations $\mathbf{A}\mathbf{u} = \mathbf{f}$ where $\mathbf{A} \in \mathbb{C}^{\mathcal{N}\times\mathcal{N}}$ is a fully-populated and non-hermitian matrix (except for the EFIE where the matrix is actually hermitian). There exists essentially two kinds of numerical solvers for solving such a linear system of equations:

- Direct solvers: These solvers rely on a LU-factorization of $\mathbf{A}$, which takes $\mathcal{O}(\mathcal{N}^3)$ operations to compute. The computational costs increase rapidly with the problem size $\mathcal{N}$, thus the use of these solvers is prohibitive for large-scale problems. This issue can be circumvented by computing a LU-factorization not of the matrix $\mathbf{A}$, but rather of a hierarchical matrix ($\mathcal{H}$-matrix, see [6]) that approximates $\mathbf{A}$.
- Iterative solvers: These solvers rely on successive matrix-vector operations with $\mathbf{A}$ to build a Krylov subspace spanned by $\mathbf{b}, \mathbf{A}\mathbf{b}, \mathbf{A}^2\mathbf{b}, \mathbf{A}^3\mathbf{b}\ldots$ where $\mathbf{b} \in \mathbb{C}^{\mathcal{N}}$ is a starting vector. In traditional implementations, each matrix-vector operation with $\mathbf{A}$ takes $\mathcal{O}(\mathcal{N}^2)$ operations. Advanced techniques allow to reduce the complexity of each matrix-vector operation to $\mathcal{O}(\mathcal{N}\log\mathcal{N})$. Among these advanced techniques, we cite the fast multipole method (FMM), introduced in [47] in the context of many particle simulations, which progressively became a very popular method for solving

BEM systems in both electromagnetism and acoustics [68, 85, 49]. We also cite the adaptice cross-approximation (ACA) technique [65], which approximates whole blocks of the system matrix using low-rank submatrices.

In this work, we use the code MAXWELL3D developed by DEMR (Electromagnetism and radar department of ONERA). In this code, a parallel direct solver (which can run on parallel machines) is available as well as an iterative solver based the multi-level FMM (which can also run on parallel machines).

6.2 A non-intrusive RBM for frequency sweep analysis

6.2.1 Affine approximations for the frequency-parametrized problem

We consider an incident plane wave eq. (6.1.7) with fixed polarization $\widehat{\boldsymbol{p}}$ and fixed direction $\widehat{\boldsymbol{d}}$. We now view the wavenumber k as varying parameter, thus we set $\mu = k$ for following our usual notation. In this parametrized context, the EFIE, MFIE and CFIE discrete operators depend on μ, *i.e.*, $\boldsymbol{T}_h = \boldsymbol{T}_h(\mu)$, $\boldsymbol{K}_h = \boldsymbol{K}_h(\mu)$ and $\boldsymbol{C}_h = \boldsymbol{C}_h(\mu)$; as well as the associated right-hand sides, *i.e.,* $\boldsymbol{b}_h^{\mathrm{E}} = \boldsymbol{b}_h^{\mathrm{E}}(\mu)$, $\boldsymbol{b}_h^{\mathrm{M}} = \boldsymbol{b}_h^{\mathrm{M}}(\mu)$ and $\boldsymbol{b}_h^{\mathrm{C}} = \boldsymbol{b}_h^{\mathrm{C}}(\mu)$. For ease of presentation, we consider the CFIE, which is most general as it combines both EFIE and MFIE. We denote $\mathbf{A}(\mu)\mathbf{u}(\mu) = \mathbf{f}(\mu)$ the linear system to be solved for the CFIE.

Inspection of Eqs. (6.1.24) and (6.1.27) reveals that the discrete EFIE and MFIE operators are non-affine because the Green kernel couples the spatial variables to the wavenumber. A well-known strategy consists in recovering affine approximations by applying the EIM to the wavenumber-dependent kernel [39, 111]. One has to be cautious, because the EIM cannot be applied directly to the Green kernel because of its singularity when $|\boldsymbol{x}-\boldsymbol{y}| \to 0$. The classical way to circumvent this is to split the Green kernel in two terms as

$$\mathcal{G}(\boldsymbol{x},\boldsymbol{y};\mu) = \frac{e^{i\mu r}-1}{4\pi r} + \frac{1}{4\pi r}, \quad r = |\boldsymbol{x}-\boldsymbol{y}|. \tag{6.2.1}$$

The first term depends on μ and is non-singular (we shall use the superscript ^{ns} for "non-singular" in the following), while the second term does not depend on μ and is singular.

In order to obtain an affine approximation for the EFIE, we successively apply the EIM to the two non-singular functions $g_1^{\mathrm{ns}}(r;\mu) = i\mu\frac{e^{i\mu r}-1}{4\pi r}$ and $g_2^{\mathrm{ns}}(r;\mu) = \frac{-i}{\mu}\frac{e^{i\mu r}-1}{4\pi r}$. For ease of notation, we now use the notation $\star = 1, 2$. Recalling chapter 1, the EIM yields $M_\star \geq 1$ so-called *EIM basis functions* $h_1^{g\star}, \ldots, h_{M_\star}^{g\star}$ defined on $[0, r_{\max}]$, interpolation points $\{r_m^{g\star}\}_{1\leq m\leq M_\star}$ and a lower triangular *interpolation matrix* $\mathbf{B}^{g\star} \in \mathbb{C}^{M_\star\times M_\star}$ with unity diagonal [5]. The EIM interpolant is given by

$$g_\star^{\mathrm{ns}}(r;\mu) \approx \widetilde{g_\star^{\mathrm{ns}}}(r;\mu) = \sum_{m=1}^{M_\star} \varsigma_m^{g\star}(\mu) h_m^{g\star}(r), \tag{6.2.2}$$

with complex coefficients $\boldsymbol{\varsigma}^{g\star}(\mu) = (\varsigma_1^{g\star}(\mu), \ldots, \varsigma_{M_\star}^{g\star}(\mu))^T \in \mathbb{C}^{M_\star}$ solution to the $M_\star \times M_\star$ linear system

$$\mathbf{B}^{g\star}\boldsymbol{\varsigma}^{g\star}(\mu) = \mathbf{l}^{g\star}(\mu), \tag{6.2.3}$$

where $\mathbf{l}^{g\star}(\mu) = (g_\star^{\mathrm{ns}}(r_1^{g\star};\mu), \cdots, g_\star^{\mathrm{ns}}(r_{M_\star}^{g\star};\mu))^T \in \mathbb{C}^{M_\star}$. Replacing $g_\star^{\mathrm{ns}}$ by its EIM approximation $\widetilde{g_\star^{\mathrm{ns}}}$ for $\star = 1, 2$ in the expression of the EFIE operator Eq. (6.1.24) yields an affine approximation $\widetilde{\boldsymbol{T}}_h(\mu)$ for the EFIE operator given by

$$\begin{aligned}
\langle \widetilde{\boldsymbol{T}}_h(\mu)\boldsymbol{v}_h, \boldsymbol{w}_h\rangle = &\sum_{m=1}^{M_1} \varsigma_m^{g1}(\mu) \int_\Gamma \overline{\boldsymbol{w}_h(\boldsymbol{x})} \cdot \int_\Gamma h_m^{g1}(|\boldsymbol{x}-\boldsymbol{y}|)\boldsymbol{v}_h(\boldsymbol{y})\mathrm{d}\Gamma_{\boldsymbol{y}}\mathrm{d}\Gamma_{\boldsymbol{x}} \\
&+ i\mu \int_\Gamma \overline{\boldsymbol{w}_h(\boldsymbol{x})} \cdot \int_\Gamma \frac{1}{4\pi|\boldsymbol{x}-\boldsymbol{y}|}\boldsymbol{v}_h(\boldsymbol{y})\mathrm{d}\Gamma_{\boldsymbol{y}}\mathrm{d}\Gamma_{\boldsymbol{x}} \\
&+ \sum_{m=1}^{M_2} \varsigma_m^{g2}(\mu) \int_\Gamma \mathrm{div}_{\Gamma,\boldsymbol{x}}\, \overline{\boldsymbol{w}_h(\boldsymbol{x})} \int_\Gamma h_m^{g2}(|\boldsymbol{x}-\boldsymbol{y}|)\mathrm{div}_{\Gamma,\boldsymbol{y}}\, \boldsymbol{v}_h(\boldsymbol{y})\mathrm{d}\Gamma_{\boldsymbol{y}}\mathrm{d}\Gamma_{\boldsymbol{x}} \\
&- \frac{i}{\mu} \int_\Gamma \mathrm{div}_{\Gamma,\boldsymbol{x}}\, \overline{\boldsymbol{w}_h(\boldsymbol{x})} \int_\Gamma \frac{1}{4\pi|\boldsymbol{x}-\boldsymbol{y}|}\mathrm{div}_{\Gamma,\boldsymbol{y}}\, \boldsymbol{v}_h(\boldsymbol{y})\mathrm{d}\Gamma_{\boldsymbol{y}}\mathrm{d}\Gamma_{\boldsymbol{x}}.
\end{aligned} \tag{6.2.4}$$

Remarking that all the integrated terms are independent from the wavenumber; it is clear that $\widetilde{\boldsymbol{T}}_h(\mu)$ is affine with $(M_1 + M_2 + 2)$ terms. We can address the MFIE operator in a similar way, by observing that

$$\frac{\partial}{\partial \widehat{\boldsymbol{n}}(\boldsymbol{y})}\mathcal{G}(\boldsymbol{x},\boldsymbol{y};\mu) = (\psi^{\mathrm{ns}}(|\boldsymbol{x}-\boldsymbol{y}|;\mu) + \psi^{\mathrm{s}}(|\boldsymbol{x}-\boldsymbol{y}|))\,\frac{(\boldsymbol{y}-\boldsymbol{x})\cdot\boldsymbol{n}(\boldsymbol{y})}{|\boldsymbol{x}-\boldsymbol{y}|}, \tag{6.2.5}$$

with $\psi^{\mathrm{ns}}(\cdot;\mu)$ and ψ^{s} defined by

$$\psi^{\mathrm{ns}}(r;\mu) = i\mu\frac{e^{i\mu r}-1}{4\pi r} - \frac{e^{i\mu r}-1-i\mu r}{4\pi r^2}, \quad \psi^{\mathrm{s}}(r) = -\frac{1}{4\pi r^2}. \tag{6.2.6}$$

An affine approximation for the MFIE operator defined by Eq. (6.1.27) can be straightforwardly obtained from applying a third EIM to the function $g_3^{\mathrm{ns}} = \psi^{\mathrm{ns}}$. With this strategy, we obtain an affine approximation for the MFIE operator with $M_3 + 2$ terms. Thus, we obtain an affine approximation $\widetilde{\mathbf{A}}(\mu)$ for the μ-dependent CFIE system matrix $\mathbf{A}(\mu)$ with $Q^a = (M_1 + M_2 + M_3 + 4)$ terms given by

$$\widetilde{\mathbf{A}}(\mu) = \sum_{q=1}^{Q^a} \sigma_q(\mu)\mathbf{A}_q \tag{6.2.7}$$

with complex coefficients $\boldsymbol{\sigma}(\mu) = (\sigma_1(\mu), \ldots, \sigma_{Q^a}(\mu)) \in \mathbb{C}^{Q^a}$ solution to the linear

system $\mathbf{B}\boldsymbol{\sigma}(\mu) = \mathbf{l}(\mu)$, where

$$\mathbf{B} = \begin{pmatrix} \mathbf{B}^{g1} & & & & & & & \\ & 1 & & & & & & \\ & & \mathbf{B}^{g2} & & & & & \\ & & & 1 & & & & \\ & & & & \mathbf{B}^{g3} & & & \\ & & & & & 1 & & \\ & & & & & & 1 & \end{pmatrix}, \quad \mathbf{l}(\mu) = \begin{pmatrix} \mathbf{l}^{g1}(\mu) \\ i\mu \\ \mathbf{l}^{g2}(\mu) \\ \frac{-i}{\mu} \\ \mathbf{l}^{g3}(\mu) \\ 1 \\ \frac{1}{2} \end{pmatrix}. \tag{6.2.8}$$

Notice that $\mathbf{B} \in \mathbb{C}^{Q^a \times Q^a}$ is lower triangular with unity diagonal. Remark also that the last row of the system is associated to the mass term in the MFIE operator. At this stage, two main issues arise:

- the wavenumber-independent operator terms $\mathbf{A}_q$, $1 \leq q \leq Q^a$ in the affine decomposition Eq. (6.2.7) are non-standard integral operators (this is clear in Eq. (6.2.4)), which implies that new dedicated assembly/matrix-vector product routines have to be implemented in order to manipulate these non-standard integral operators. These new routines are non-trivial, because special treatment is required to integrate the singular kernel of some operator terms.
- the number of affine terms Q^a can be quite large, which can compromise the efficiency of the RB method, as reported in [39].

In the next section, we circumvent the first issue by proposing a non-intrusive approach. We circumvent the second issue by constructing local – rather than global – affine approximations.

6.2.2 Non-intrusive local affine approximations

In our method, a number of terms $Q \geq 2$ is prescribed by the user. A non-intrusive local affine approximation relies on $J \geq Q$ wavenumbers $\hat{\mu}_1 \leq \cdots \leq \hat{\mu}_J$ inside the range if interest $[\mu_{\min}, \mu_{\max}]$. Upon these J wavenumbers, we construct the $K = J - Q + 1$ sets

$$\mathcal{T}_k = \{\hat{\mu}_k, \ldots, \hat{\mu}_{k+Q-1}\}, \quad k \in \{1, \ldots, K\}, \tag{6.2.9}$$

each of cardinality Q. We define the *indicator function* $\mathcal{I} : [\mu_{\min}, \mu_{\max}] \to \{1, \ldots, K\}$ that maps each wavenumber μ to the index k such that $\mathcal{T}_k$ is the set of the Q points among the $\hat{\mu}_j$'s that are closest to μ.

This being set, we introduce for all $k \in \{1, \ldots, K\}$ the kth *non-intrusive local affine approximation* as

$$\forall \mu \in [\mu_{\min}, \mu_{\max}], \quad \mathbf{A}_k(\mu) = \sum_{q=1}^{Q} \theta_q^k(\mu) \mathbf{A}(\hat{\mu}_{k+q-1}), \tag{6.2.10}$$

where the θ_q^k's are wavenumber-dependent coefficients. As we shall see, $\mathbf{A}_k(\mu)$ will only be a good affine approximation for $\mathbf{A}(\mu)$ *locally* for values of μ in the subdomain $\mathcal{D}_k = \mathcal{I}^{-1}(k) = \{\mu \,|\, \mathcal{I}(\mu) = k\}$. In this section, we explain the construction process in detail.

The wavenumber-dependent coefficients

First, we explain how the wavenumber-dependent coefficients in Eq. (6.2.10) are defined given J available wavenumbers $\hat{\mu}_1 \leq \cdots \leq \hat{\mu}_J$. For this purpose, let $\mu \in [\mu_{\min}, \mu_{\max}]$ and denote $k = \mathcal{I}(\mu)$. Then $\boldsymbol{\theta}^k(\mu) = (\theta_1^k(\mu), \ldots, \theta_Q^k(\mu))^T \in \mathbb{C}^Q$ is defined by

$$\boldsymbol{\theta}^k(\mu) = \underset{\boldsymbol{\theta} \in \mathbb{C}^Q}{\operatorname{argmin}} \left\| \boldsymbol{\sigma}(\mu) - \sum_{q=1}^{Q} \theta_q \boldsymbol{\sigma}(\hat{\mu}_{k+q-1}) \right\|_2 \tag{6.2.11}$$

where $\|\cdot\|_2$ denotes the euclidian norm in $\mathbb{C}^{Q^a}$. Equivalently, the wavenumber-dependent coefficients satisfy

$$P_{\mathcal{T}_k}\left[\boldsymbol{\sigma}(\mu)\right] = \sum_{q=1}^{Q} \theta_q^k(\mu) \boldsymbol{\sigma}(\hat{\mu}_{k+q-1}), \tag{6.2.12}$$

where $P_{\mathcal{T}_k}[\cdot]$ denotes orthogonal projection from $\mathbb{C}^{Q^a}$ onto the Q-dimensional subspace $\mathrm{ColSpan}\{\boldsymbol{\sigma}(\hat{\mu}),\, \hat{\mu} \in \mathcal{T}_k\}$.

Construction using a localization procedure

We now provide an automatic procedure for selecting the wavenumbers $\hat{\mu}_1 \leq \cdots \leq \hat{\mu}_J$. There are two phases: phase 1 selects $Q+1$ wavenumbers using a classical greedy strategy and phase 2 selects more wavenumbers following a locally adaptive strategy until a prescribed tolerance is reached on the worst projection error.

Phase 1. The first phase consists in selecting a set of $Q+1$ wavenumbers using a classical greedy procedure driven by the projection error. Namely, at iteration $J \geq 1$ a set $\mathcal{C}_J$ of J wavenumbers is available. Thus, we can compute for all $\mu \in \Xi$ (where $\Xi \subset [\mu_{\min}, \mu_{\min}]$ is a discrete set) the vector $\boldsymbol{\sigma}(\mu)$ and its orthogonal projection $P_{\mathcal{C}_J}[\boldsymbol{\sigma}(\mu)]$ onto the J-dimensional subspace $\mathrm{ColSpan}\{\boldsymbol{\sigma}(\hat{\mu}),\, \hat{\mu} \in \mathcal{C}_J\}$. We then enrich the set $\mathcal{C}_J$ by adding the wavenumber $\hat{\mu}^\star \in \Xi$ for which the projection error is maximal. We continue this greedy selection procedure until $Q+1$ wavenumbers are selected. This procedure is summarized by Alg. 6.1.

Phase 2. At the start of the second phase, $Q+1$ wavenumbers are available from the first phase. Thus, Eq. (6.2.9) defines two sets $\mathcal{T}_k$, $k = 1, 2$. In this context, the indicator function maps each μ to the integer $k = \mathcal{I}(\mu)$ such that the projection of $\boldsymbol{\sigma}(\mu)$ must

Algorithm 6.1: Classical greedy (phase 1 of localization procedure)

Input : prescribed number of terms Q and a discrete set $\Xi \subset [\mu_{\min}, \mu_{\max}]$
Output: a set $\mathcal{C}_J = \{\hat{\mu}_j\}_{1 \leq j \leq J}$ with $J = Q + 1$.

Pick a random $\hat{\mu}^\star \in \Xi$;
Set $\mathcal{C}_1 = \{\hat{\mu}^\star\}$;
for $J = 1, \dots Q$ **do**
  Find $\hat{\mu}^\star = \underset{\mu \in \Xi}{\operatorname{argmax}} \, \| \boldsymbol{\sigma}(\mu) - P_{\mathcal{C}_J} \left[\boldsymbol{\sigma}(\mu) \right] \|_2$;
  Enrich $\mathcal{C}_{J+1} = \mathcal{C}_J \cup \{\hat{\mu}^\star\}$;
end

be performed onto the Q-dimensional subspace $\mathrm{ColSpan}\{\boldsymbol{\sigma}(\hat{\mu}),\, \hat{\mu} \in \mathcal{T}_k\}$. Thus, for any value of μ, the *local* projection error is given by $\| \boldsymbol{\sigma}(\mu) - P_{\mathcal{T}_k} \left[\boldsymbol{\sigma}(\mu) \right] \|_2$, with $k = \mathcal{I}(\mu)$. The locally adaptive strategy, summarized by Alg. 6.2, consists in selecting the wavenumbers that maximize the local projection error until a prescribed tolerance is reached on the maximal local projection error.

Algorithm 6.2: Locally adaptive strategy (phase 2 of localization procedure)

Input : prescribed tolerance $tol > 0$, a set $\mathcal{C}_J = \{\hat{\mu}_j\}_{1 \leq j \leq J}$ with $J = Q + 1$ obtained by Alg. 6.1
Output: enriched set $\mathcal{C}_J = \{\hat{\mu}_j\}_{1 \leq j \leq J}$

Set $K = 2$ and find $\mu^\star = \underset{\mu \in \Xi}{\operatorname{argmax}} \, \| \boldsymbol{\sigma}(\mu) - P_{\mathcal{T}_k} \left[\boldsymbol{\sigma}(\mu) \right] \|_2$, where $k = \mathcal{I}(\mu)$;
Compute $\epsilon = \| \boldsymbol{\sigma}(\mu^\star) - P_{\mathcal{T}_{k^\star}} \left[\boldsymbol{\sigma}(\mu^\star) \right] \|_2$, where $k^\star = \mathcal{I}(\mu^\star)$;
while $\epsilon > tol$ **do**
  Enrich $\mathcal{C}_{J+1} = \mathcal{C}_J \cup \{\hat{\mu}^\star\}$;
  Update $J \leftarrow J + 1$ and $K \leftarrow K + 1$;
  Find $\mu^\star = \underset{\mu \in \Xi}{\operatorname{argmax}} \, \| \boldsymbol{\sigma}(\mu) - P_{\mathcal{T}_k} \left[\boldsymbol{\sigma}(\mu) \right] \|_2$, where $k = \mathcal{I}(\mu)$;
  Compute $\epsilon = \| \boldsymbol{\sigma}(\mu^\star) - P_{\mathcal{T}_{k^\star}} \left[\boldsymbol{\sigma}(\mu^\star) \right] \|_2$, where $k^\star = \mathcal{I}(\mu^\star)$;
end

Discussion

We have explained how local affine approximations in the form of Eq. (6.2.10) could be constructed following an automatic procedure. At this point it is worth noticing that the proposed construction only requires the ability to evaluate $\mu \mapsto \boldsymbol{\sigma}(\mu)$, which, recalling Eq. (6.2.8), exclusively relies on the knowledge of the functions $g_\star^{\mathrm{ns}}$, $\star = 1, 2, 3$ and associated EIM interpolation matrices and interpolation points. Thus, the proposed construction is completely independent from the discretization with $\mathcal{N}$ degrees of freedom.

The rationale behind the proposed construction is the following [23]: at the end of Alg. 6.2, for any $\mu \in [\mu_{\min}, \mu_{\max}]$, the vector $\boldsymbol{\sigma}(\mu) \in \mathbb{C}^{Q^a}$ can be approximated by its orthogonal projection $P_{\mathcal{T}_k}[\boldsymbol{\sigma}(\mu)]$ with $k = \mathcal{I}(\mu)$ with an error smaller than the prescribed tolerance tol. Replacing $\boldsymbol{\sigma}(\mu)$ by $P_{\mathcal{T}_k}[\boldsymbol{\sigma}(\mu)]$ in the affine approximation for the CFIE operator $\widetilde{\mathbf{A}}(\mu)$ given by Eq. (6.2.7) and recalling the expression Eq. (6.2.12) for $P_{\mathcal{T}_k}[\boldsymbol{\sigma}(\mu)]$, we get

$$\widetilde{\mathbf{A}}(\mu) = \sum_{q=1}^{Q^a} \sigma_q(\mu)\mathbf{A}_q \approx \sum_{q=1}^{Q^a}\sum_{p=1}^{Q} \theta_p^k(\mu)\sigma_q(\hat{\mu}_{k+p-1})\mathbf{A}_q. \tag{6.2.13}$$

Swapping the summations we obtain

$$\widetilde{\mathbf{A}}(\mu) \approx \sum_{p=1}^{Q} \theta_p^k(\mu) \sum_{q=1}^{Q^a} \sigma_q(\hat{\mu}_{k+p-1})\mathbf{A}_q = \sum_{p=1}^{Q} \theta_p^k(\mu)\widetilde{\mathbf{A}}(\hat{\mu}_{k+p-1}). \tag{6.2.14}$$

Omitting the tilde in the RHS of Eq. (6.2.14) yields the non-intrusive local approximation proposed in Eq. (6.2.10). The tilde can indeed be omitted, since $\widetilde{\mathbf{A}}$ is – by design – a good approximation for $\mathbf{A}$. Ultimately, we obtain that $\mathbf{A}(\mu) \approx \sum_{p=1}^{Q} \theta_p^k(\mu)\mathbf{A}(\hat{\mu}_{k+p-1})$ which corresponds to our initial non-intrusive local affine approximation statement Eq. (6.2.10).

6.2.3 Non-intrusive RB approximation

We propose a revisited version of the RBM, specifically tailored for non-intrusive local affine approximations. As usual, a reduced basis is built using N solutions to the BEM linear system $\mathbf{A}(\mu)\mathbf{u}(\mu) = \mathbf{f}(\mu)$ at wavenumbers $\mu^{(1)}, \ldots, \mu^{(N)}$ inside the range of interest $[\mu_{\min}, \mu_{\max}]$. For convenience, let us index the wavenumbers in increasing order $\mu^{(1)} \leqslant \cdots \leqslant \mu^{(N)}$. Rather than one global RB $\mathbf{P} \in \mathbb{C}^{\mathcal{N}\times N}$ spanning the N solutions, we propose to build K nested RBs, in the sense that

$$\mathrm{Colspan}(\mathbf{P}_1) \subset \mathrm{Colspan}(\mathbf{P}_2) \subset \cdots \subset \mathrm{Colspan}(\mathbf{P}_K). \tag{6.2.15}$$

For $k \geqslant 1$, let $n_k \geqslant 0$ denote the number of wavenumbers among the N wavenumbers $\mu^{(1)}, \ldots, \mu^{(N)}$ which are in the kth subdomain $\mathcal{D}_k = \{\mu \,|\, \mathcal{I}(\mu) = k\}$. Thus, $N = n_1 + \cdots + n_K$. Notice that n_k may be 0 if none of the wavenumbers $\mu^{(1)}, \ldots, \mu^{(N)}$ is in the subdomain $\mathcal{D}_k$.

This being set, for $k \geqslant 1$ we propose to build $\mathbf{P}_k \in \mathbb{C}^{\mathcal{N}\times(n_1+\cdots+n_k)}$ as the RB spanning the $n_1 + \cdots + n_k$ solutions at the wavenumbers $\mu^{(1)}, \ldots, \mu^{(n_1+\cdots+n_k)}$. Notice that these wavenumbers are in the union of the k first subdomains $\mathcal{D}_1 \cup \cdots \cup \mathcal{D}_k$. In particular for $k = K$, the RB $\mathbf{P}_K \in \mathbb{C}^{\mathcal{N}\times N}$ spans the N solutions at wavenumbers $\mu^{(1)}, \ldots, \mu^{(N)}$.

Following standard practice of the RBM, the local RBs $\mathbf{P}_k$ are chosen $\mathbf{B}_V$-orthonormal (*i.e.*, $\mathbf{P}_k^*\mathbf{B}_V\mathbf{P}_k = \mathbf{I}$), where $\mathbf{B}_V \in \mathbb{C}^{\mathcal{N}\times\mathcal{N}}$ denotes the discretized inverse Riesz map. Here, we consider $\mathbf{B}_V$ to be the mass matrix with entries $\int_\Gamma \boldsymbol{\phi}_j \cdot \overline{\boldsymbol{\phi}_i}\mathrm{d}\Gamma$, $1 \leqslant i, j \leqslant \mathcal{N}$,

where $\{\boldsymbol{\phi}_i\}_{1\leqslant i\leqslant \mathcal{N}}$ denotes the basis of our discrete approximation space $\boldsymbol{V}_h^{\mathrm{RT}}$. This choice amounts to measuring the members of $\boldsymbol{V}_h^{\mathrm{RT}}$ using the $\boldsymbol{L}^2(\Gamma)$ norm.

For all $\mu \in [\mu_{\min}, \mu_{\max}]$ we define a local RB approximation $\mathbf{u}_N(\mu) \in \mathbb{C}^{\mathcal{N}}$ as the solution to the following least-squares optimization problem

$$\mathbf{u}_N(\mu) = \underset{\mathbf{u}_N \in \mathrm{Colspan}(\mathbf{P}_k)}{\operatorname{argmin}} \|\mathbf{A}_k(\mu)\mathbf{u}_N - \widetilde{\mathbf{f}}(\mu)\|^2_{\mathbf{B}_W^{-1}}, \quad k = \mathcal{I}(\mu), \tag{6.2.16}$$

where $\mathbf{A}_k(\mu)$ is our non-intrusive local affine approximation given by Eq. (6.2.10) and $\widetilde{\mathbf{f}}(\mu)$ is an affine approximation with Q^f terms for $\mathbf{f}(\mu)$, obtained by applying the EIM to the plane wave Eq. (6.1.7). Here $\mathbf{B}_W \in \mathbb{C}^{\mathcal{N}\times\mathcal{N}}$ denotes the discretized inverse Riesz operator on the test space (see chapter 1) and the $\|\cdot\|_{\mathbf{B}_W^{-1}}$ norm is defined by $\|\mathbf{f}\|_{\mathbf{B}_W^{-1}} = (\mathbf{f}^*\mathbf{B}_W^{-1}\mathbf{f})^{1/2} = \|\mathbf{B}_W^{-1}\mathbf{f}\|_{\mathbf{B}_W}$ for all $\mathbf{f} \in \mathbb{C}^{\mathcal{N}}$. We recall that the discrete operators stem from a Galerkin projection, consequently $\mathbf{B}_W = \mathbf{B}_V$ is taken to be the mass matrix.

Clearly, the local RB approximation defined by Eq. (6.2.16) can be efficiently assembled in no more[1] than $\mathcal{O}(N^2Q^2 + NQQ^f)$ and solved in $\mathcal{O}(N^3)$ operations using a direct solver following the usual offline/online strategy.

6.2.4 Greedy construction

Greedy RB generation frequency-sweep algorithm

The matrices $\mathbf{P}_k$, $1 \leq k \leq K$ are built following the frequency-sweep procedure summarized by Alg. 6.3.

At iteration $k \geq 1$, the wavenumbers at which the BEM linear system $\mathbf{A}(\mu)\mathbf{u}(\mu) = \mathbf{f}(\mu)$ must be solved to serve as new basis functions are selected by successively maximizing an error indicator $\mu \mapsto \delta_k(\mu)$ over a discrete surrogate set covering $\mathcal{D}_k$ until a prescribed tolerance is reached. Ideally, this error indicator should be a error estimator in the sense of definition 5. For the sake of simplicity, we choose the residual norm (6.2.16) scaled by a constant $\hat{\alpha}$, as explained in section 2.2.4. Thus our error indicator is given by

$$\delta_k(\mu) = \frac{1}{\hat{\alpha}} \frac{\|\mathbf{A}_k(\mu)\mathbf{u}_N(\mu) - \widetilde{\mathbf{f}}(\mu)\|_{\mathbf{B}_W^{-1}}}{\|\mathbf{u}_N(\mu)\|_{\mathbf{B}_V}} \tag{6.2.17}$$

The advantage of this error indicator is that it can be efficiently computed following the usual offline/online strategy, thus the maximum of $\mu \mapsto \delta_k(\mu)$ over $\mathcal{D}_k$ can be efficiently computed by evaluating $\delta_k(\mu)$ for all μ in some discrete surrogate set covering $\mathcal{D}_k$. Furthermore, recalling chapter 2 this indicator is expected to provide a good estimation (although not a rigorous bound) for the RB relative error $\|\mathbf{u}(\mu) - \mathbf{u}_N(\mu)\|_{\mathbf{B}_V} / \|\mathbf{u}(\mu)\|_{\mathbf{B}_V}$.

[1] the announced complexity corresponds to most computationally expensive scenario $k = K$, but the complexity is smaller for $k < K$.

Algorithm 6.3: Greedy RB generation frequency-sweep

Input : prescribed tolerance $\epsilon_{rel}^{\mathrm{rb}} > 0$ and a discrete set $\Xi \subset [\mu_{\min}, \mu_{\max}]$
Output: Nested RBs $\mathbf{P}_k$, $1 \leqslant k \leqslant K$.

Initialize $\mathbf{P}_0 = [\,]$ and $N = 0$;
for $k = 1, \ldots, K$ **do**
  Initialize $\mathbf{P}_k = \mathbf{P}_{k-1}$ and $n_k = 0$;
  Find $\mu^\star = \underset{\mu \in \Xi \cap \mathcal{D}_k}{\operatorname{argmax}} \delta_k(\mu)$;
  $\epsilon_{rel} \leftarrow \delta_k(\mu^\star)$;
  while $\epsilon_{rel} > tol$ **do**
    Compute $\mathbf{u}(\mu^\star)$ by solving the BEM;
    $\mathbf{B}_V$-orthonormalize $\mathbf{u}(\mu^\star)$ against the columns of $\mathbf{P}_k$ to obtain $\mathbf{p}_{n_k+1}$;
    $\mathbf{P}_k \leftarrow [\mathbf{P}_k \,|\, \mathbf{p}_{n_k+1}]$ and $n_k \leftarrow n_k + 1$;
    Find $\mu^\star = \underset{\mu \in \Xi \cap \mathcal{D}_k}{\operatorname{argmax}} \delta_k(\mu)$;
    $\epsilon_{rel} \leftarrow \delta_k(\mu^\star)$;
  end
  $N \leftarrow N + n_k$;
end

Complexity analysis

We analyze the computational costs associated to algorithm 6.3 in terms of the two most expensive tasks: system solves and matrix-vector products with the CFIE operator. The number of system solves is of course N. The number of matrix-vector product can be determined thanks to table 6.1. This table summarizes the matrix-vector product that need to be computed in the first three iterations of the algorithm.

In the first iteration $k = 1$, each time a reduced basis function $\mathbf{p}_i$ is computed (this happens n_1 times), the Q matrix-vector products $\mathbf{A}(\hat{\mu}_1)\mathbf{p}_i$, $\ldots$, $\mathbf{A}(\hat{\mu}_Q)\mathbf{p}_i$ need to be computed. This means that $n_1 Q$ matrix-vector products are required in the first iteration $k = 1$. Next, for $k \geqslant 2$, one starts to compute the block matrix product $\mathbf{A}(\hat{\mu}_{k+Q-1})\mathbf{P}_{k-1}$, where $\mathbf{P}_{k-1}$ is a block of $n_1 + \cdots + n_{k-1}$ vectors. Then, each time a reduced basis function $\mathbf{p}_i$ is computed (this happens n_k times), the Q matrix-vector products $\mathbf{A}(\hat{\mu}_k)\mathbf{p}_i$, $\ldots$, $\mathbf{A}(\hat{\mu}_{k+Q-1})\mathbf{p}_i$ need to be computed. Thus, the overall number of matrix operations is

- QN matrix-vector products;
- $K - 1$ block matrix-vector products with blocks of sizes n_1, $(n_1 + n_2)$, $\ldots$, $(n_1 + \ldots, n_{K-1})$.

	$\mathbf{A}(\hat{\mu}_1)$	$\mathbf{A}(\hat{\mu}_2)$	$\mathbf{A}(\hat{\mu}_3)$	...	$\mathbf{A}(\hat{\mu}_Q)$	$\mathbf{A}(\hat{\mu}_{Q+1})$	$\mathbf{A}(\hat{\mu}_{Q+2})$
$\mathbf{p}_1$	$\star$	$\star$	$\star$	$\cdots$	$\star$	$\diamond$	$\diamond$
$\vdots$	$\vdots$	$\vdots$	$\vdots$		$\vdots$	$\vert$	$\vert$
$\mathbf{p}_{n_1}$	$\star$	$\star$	$\star$	$\cdots$	$\star$	$\diamond$	$\vert$
$\mathbf{p}_{n_1+1}$		$\star$	$\star$	$\cdots$	$\star$	$\star$	$\vert$
$\vdots$		$\vdots$	$\vdots$		$\vdots$	$\vdots$	$\vert$
$\mathbf{p}_{n_1+n_2}$		$\star$	$\star$	$\cdots$	$\star$	$\star$	$\diamond$
$\mathbf{p}_{n_1+n_2+1}$			$\star$	$\cdots$	$\star$	$\star$	$\star$
$\vdots$			$\vdots$		$\vdots$	$\vdots$	$\vdots$
$\mathbf{p}_{n_1+n_2+n_3}$			$\star$	$\cdots$	$\star$	$\star$	$\star$

Table 6.1: Matrix vector products that need to be computed in the first three iterations $k = 1, 2, 3$ of algorithm 6.3. Rows represent reduced basis functions $\mathbf{p}_i$ and columns correspond to operators $\mathbf{A}(\hat{\mu}_j)$. The symbol $\star$ on row $\mathbf{p}_i$ and column $\mathbf{A}(\hat{\mu}_j)$ indicates that the matrix-vector product $\mathbf{A}(\hat{\mu}_j)\mathbf{p}_i$ is computed. The sequence $\diamond - \diamond$ on rows ranging from $\mathbf{p}_i$ to $\mathbf{p}_{i+\ell}$ and column $\mathbf{A}(\hat{\mu}_j)$ indicates that the block matrix- product $\mathbf{A}(\hat{\mu}_j)[\mathbf{p}_i| \dots |\mathbf{p}_{i+\ell}]$ is computed, where $[\mathbf{p}_i| \dots |\mathbf{p}_{i+\ell}]$ is a block of ℓ vectors. Each color represents an iteration of algorithm 6.3: $k = 1$ in red, $k = 2$ in blue and $k = 3$ in green.

	$\mathbf{A}(\hat{\mu}_1)$	$\mathbf{A}(\hat{\mu}_2)$	$\mathbf{A}(\hat{\mu}_3)$	...	$\mathbf{A}(\hat{\mu}_Q)$	$\mathbf{A}(\hat{\mu}_{Q+1})$	$\mathbf{A}(\hat{\mu}_{Q+2})$
$\mathbf{p}_1$	$\diamond$	$\diamond$	$\diamond$	$\cdots$	$\diamond$	$\diamond$	$\diamond$
$\vdots$	$\vert$	$\vert$	$\vert$		$\vert$	$\vert$	$\vert$
$\mathbf{p}_{n_1}$	$\diamond$	$\vert$	$\vert$	$\cdots$	$\vert$	$\vert$	$\vert$
$\mathbf{p}_{n_1+1}$		$\vert$	$\vert$	$\cdots$	$\vert$	$\vert$	$\vert$
$\vdots$		$\vert$	$\vert$		$\vert$	$\vert$	$\vert$
$\mathbf{p}_{n_1+n_2}$		$\diamond$	$\vert$	$\cdots$	$\vert$	$\vert$	$\vert$
$\mathbf{p}_{n_1+n_2+1}$			$\vert$	$\cdots$	$\vert$	$\vert$	$\vert$
$\vdots$			$\vert$		$\vert$	$\vert$	$\vert$
$\mathbf{p}_{n_1+n_2+n_3}$			$\diamond$	$\cdots$	$\diamond$	$\diamond$	$\diamond$

Table 6.2: Matrix vector products that need to be computed with the monolithic construction in the case of $K = 3$ subdomains (using the notations from table 6.1).

6.2.5 Monolithic construction

In practice, the computational costs associated to the greedy construction of the RB can be excessive. As we shall see in the numerical examples, the computational cost of the greedy construction is dominated by the Q matrix-vector products required per basis function. We analyze this situation as follows. At the kth iteration of the frequency sweep algorithm 6.3, for all $i \in \{1, \dots, n_k\}$, the ith iteration of the local greedy loop requires computing the Q matrix-vector products $\mathbf{A}(\hat{\mu}_k)\mathbf{p}_{N_{k-1}+i}, \dots, \mathbf{A}(\hat{\mu}_{k+Q-1})\mathbf{p}_{N_{k-1}+i}$ (here, we have denoted $N_{k-1} = n_1 + \cdots + n_{k-1}$ for conciseness). Thus, if we look at the CFIE operator $\mathbf{A}(\hat{\mu}_j)$ with fixed index $j \in \{k, \dots, k+Q-1\}$, the local greedy loop successively computes the n_k matrix vector products $\mathbf{A}(\hat{\mu}_j)\mathbf{p}_{N_{k-1}+1}, \dots, \mathbf{A}(\hat{\mu}_j)\mathbf{p}_{N_{k-1}+n_k}$.

Rather than successively computing these n_k matrix-vector products, a much better strategy in terms of computational performance would be to directly compute the block matrix-vector product $\mathbf{A}(\hat{\mu}_j)[\mathbf{p}_{N_{k-1}+1}|\cdots|\mathbf{p}_{N_{k-1}+n_k}]$. Unfortunately, this is not possible because the n_k basis functions are not known in advance, as they are successively built one after another following the greedy procedure.

This motivates an alternative construction which we choose to call the *monolithic construction* and which is also sketched in [23]. The paradigm of the monolithic construction is the following: the user pre-determines the number of CFIE solves and the values of the wavenumber for which the CFIE is to be solved. We emphasize that this paradigm is completely different from the paradigm of the greedy construction, where the user sets up a desired level of accuracy and lets the algorithm choose how many wavenumbers and the wavenumbers values at which to solve the CFIE. The advantage of the monolithic construction is that, the basis functions being known in advance, the overall computational performance can be enhanced with the use of block matrix-vector products.

Let N be the number of basis functions desired by the user and $\mu^{(1)}, \ldots, \mu^{(N)}$ denote the wavenumbers chosen by the user. Then, the monolithic construction consists in the following:

1. successively solve the N BEM linear systems and determine (through a Gram-Schmidt othonormalization procedure) the basis-functions $\mathbf{p}_1, \ldots, \mathbf{p}_N$;
2. for all $k = 1, \ldots, K-1$, compute the block matrix-vector product $\mathbf{A}(\hat{\mu}_k)\mathbf{P}_k$, with block size $n_1 + \cdots + n_k$;
3. for all $q = 1, \ldots, Q$, compute the block matrix-vector products $\mathbf{A}(\hat{\mu}_{K+q-1})\mathbf{P}_K$ with block size N.

This approach is summarized by table 6.2. Overall, the number of block matrix-vector products is $Q + K - 1$, therefore independent from the size of the reduced basis N. The block sizes in these block matrix-vector are however dependent on N, but as we shall see in the numerical examples, this does not significantly deteriorate the performance. Indeed, past an assembly phase (this is not necessarily a dense assembly of the matrix), it is relatively cheap to compute a matrix-vector product with a large number of vectors.

6.3 Numerical illustrations

6.3.1 EFIE vs CFIE: tests on the unit sphere

We consider the EFIE and CFIE on the unit sphere $\Gamma = \{\boldsymbol{x} \in \mathbb{R}^3,\ |\boldsymbol{x}| = 1\}$ and on the frequency window $400 - 600\mathrm{MHz}$.

To start with, we build non-intrusive local affine approximations for the EFIE and CFIE operators. This is done by successively applying the EIM to each μ-dependent ker-

nel. In this situation, the maximum distance between any two point on Γ is simply $r_{\max} = 2$. We use algorithm 1.1 with prescribed tolerance $tol^{\text{EIM}} = 10^{-5}$. The interval $[0, r_{\max}]$ is discretized using 2000 uniformly distributed points while the interval $[\mu_{\min}, \mu_{\max}] \approx [8.383,\ 12.575]$ is discretized using 1000 uniformly distributed points. We obtain $M_1 = 10$, $M_2 = 9$ and $M_3 = 11$. Thus the EFIE operator admits a traditional affine decomposition of the form eq. (6.2.7) with $Q^a = 21$ terms, while $Q^a = 34$ for the CFIE. We use our localization procedure algorithms 6.1 and 6.2 with prescribed number of terms $Q = 8$ and tolerance $tol = 10^{-2}$ and obtain a domain decomposition into $K = 2$ subdomains for both EFIE and CFIE operators.

For the right-hand side, we consider a direction $\widehat{\boldsymbol{d}} = \widehat{\boldsymbol{r}}(\theta, \phi)$ and a polarization $\widehat{\boldsymbol{d}} = \widehat{\boldsymbol{\phi}}(\theta, \phi)$ with fixed $\theta = \frac{\pi}{2}$ and $\phi = \frac{\pi}{4}$, using the notations provided by fig. 6.2. For the

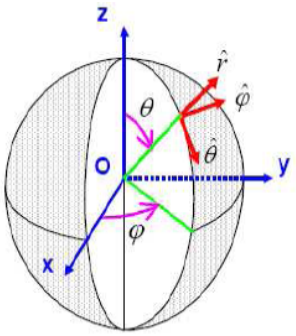


Figure 6.2: Conventions for the spherical coordinates.

plane wave eq. (6.1.7), we proceed as follows

$$e^{i\mu \boldsymbol{x}\cdot\widehat{\boldsymbol{d}}} = e^{i\mu|\boldsymbol{x}|\cos\left(\frac{\boldsymbol{x}}{|\boldsymbol{x}|},\widehat{\boldsymbol{d}}\right)}. \tag{6.3.1}$$

For general surface Γ, the function $\boldsymbol{x} \in \Gamma \mapsto |\boldsymbol{x}| \cos\left(\frac{\boldsymbol{x}}{|\boldsymbol{x}|}, \widehat{\boldsymbol{d}}\right)$ takes values in $[-R_{\max}, R_{\max}]$, where $R_{\max} = \max_{\boldsymbol{x}\in\Gamma} |\boldsymbol{x}|$. Thus, an affine approximation for the plane wave can be obtained by applying the EIM to the function $g^{\text{RHS}}(R, \mu) = e^{i\mu R}$ defined on $[-R_{\max}, R_{\max}] \times [\mu_{\min}, \mu_{\max}]$. Notice that $R_{\max} = 1$ in the case of the unit sphere. Using algorithm 1.1 with prescribed tolerance $tol^{\text{EIM}} = 10^{-5}$ and spatial and parameter intervals discretized with 1000 and 900 points respectively, we obtain an EIM approximant of g^{RHS} with $Q^f = 10$ terms.

After this preliminary work on the left and right-hand sides, we generate RB approximations. The truth BEM solutions are computed on a mesh with $\mathcal{N} = 6576$ triangle elements represented on fig. 6.3. Notice that the mesh did not come into play during the preliminary work on the left and right-hand sides. We use the greedy frequency sweep algorithm 6.3 setting the prescribed tolerance to $\epsilon^{\text{rb}}_{rel} = 5\%$. We obtain $n_1 = 3$, $n_2 = 2$ (thus $N = 5$) for both EFIE and CFIE.

We validate our RB approximation by computing the BEM solutions $\mathbf{u}(\mu)$ for 100 uniformly distributed points in $[\mu_{\min}, \mu_{\max}]$ and checking the level of relative error $\|\mathbf{u}(\mu) -$

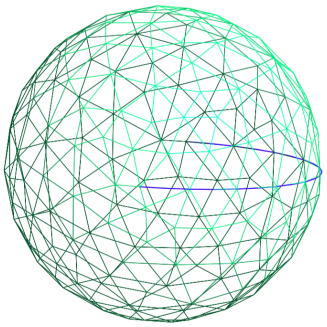


Figure 6.3: The unit sphere meshed with $\mathcal{N} = 6576$ triangle elements.

$\mathbf{u}_N(\mu)\|_{\mathbf{B}_V}/\|\mathbf{u}(\mu)\|_{\mathbf{B}_V}$. On figs. 6.4 and 6.5 we compare the exact relative error to our heuristic indicator $\mu \mapsto \delta_k(\mu)$ given by eq. (6.2.17).

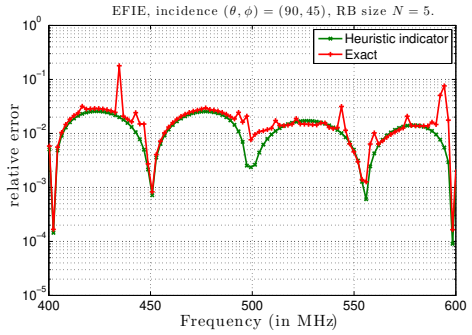


Figure 6.4: Relative error and indicator w.r.t. frequency using the EFIE.

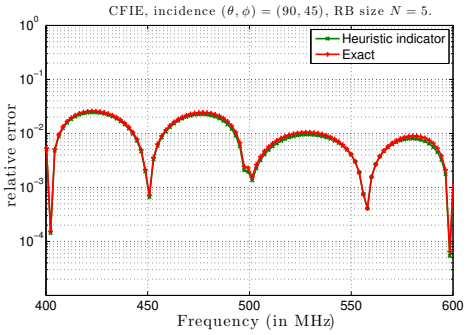


Figure 6.5: Relative error and indicator w.r.t. frequency using the CFIE.

Our heuristic indicator does not systematically coincide with the exact relative error in the case of the EFIE. Indeed, we find our heuristic indicator unable to catch the strong punctual discontinuities in the curve of the exact relative error (in red on fig. 6.4). The reason for the jagged curve of the exact relative error in the case of the EFIE is to be found in the presence of resonant frequencies where the inf-sup constant associated to the EFIE operator is very small [81, 55]. Punctually, in the neighborhood of these resonant frequencies, the exact relative error presents significant overshoots (reaching 17.9% at $434.7\mathrm{MHz}$, for instance). Unfortunately, our heuristic indicator fails to detect this and provides estimations that are a too optimistic.

However, in the case of the CFIE we find the curve of the exact relative error to be quite smooth and to coincide almost exactly with our heuristic indicator (see fig. 6.5). We believe that is because the CFIE, in opposition to the EFIE, does not suffer from the problem of resonances.

We conclude that, when it comes to construct RB approximations, the stable formulations such as the CFIE are to be preferred over the potentially resonant formulations such as the EFIE.

6.3.2 Approximation of the CFIE operator on the geometry of a fighter aircraft

We consider the CFIE on the geometry of a fighter aircraft shown on fig. 6.6. The size of this aircraft is roughly $15\mathrm{m}$ from nose to tail and about $8\mathrm{m}$ wingspan. The surface is meshed with $\mathcal{N} = 40,005$ triangular elements and the frequency window is $200 - 300\mathrm{MHz}$. Thus the rule of a thumb of 8 degrees of freedom per wavelength is satisfied.

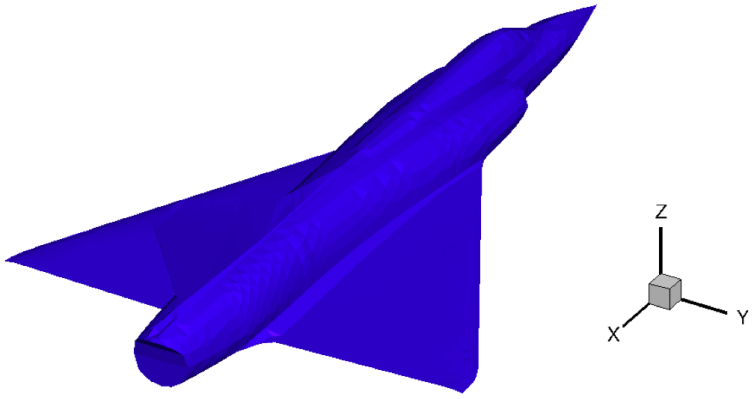


Figure 6.6: Geometry of the fighter aircraft.

We explore the use of algorithms 6.1 and 6.2 to construct non-intrusive local approximation of the frequency-dependent CFIE operator $\mathbf{A}(\mu)$. To this end, we first apply the EIM to each μ-parametrized kernel, using $r_{\max} = 15.2$ and prescribed tolerance $tol^{\mathrm{EIM}} = 10^{-5}$. We use 2000 uniformly distributed points for the spatial interval $[0, r_{\max}]$ and 1000 uniformly distributed points for the parameter interval $[\mu_{\min}, \mu_{\max}] \approx [4.191, 6.287]$. We obtain $M_1 = 16$, $M_2 = 14$ and $M_3 = 16$. Thus the CFIE operator admits a traditional affine decomposition of the form eq. (6.2.7) with $Q^a = 50$ terms. We build non-intrusive local affine approximations $\mathbf{A}_k(\mu)$ under the form eq. (6.2.10) using our localization procedure setting the prescribed number of terms to $Q = 6$ and use various tolerances tol ranging from 1.6 (which yields $K = 2$ subdomains) to 0.02 (which yields $K = 9$). We evaluate the quality of the approximation $\mathbf{A}(\mu) \approx \mathbf{A}_k(\mu)$ by computing for some samples of μ, the *LHS approximation error* given by

$$\varrho_k(\mu) = \frac{\|\mathbf{A}(\mu)\mathbf{v} - \mathbf{A}_k(\mu)\mathbf{v}\|_{\mathbf{B}_W^{-1}}}{\|\mathbf{v}\|_{\mathbf{B}_V}}, \tag{6.3.2}$$

where $\mathbf{v} = \mathbf{u}(\overline{\mu})$ is the BEM system solution at the average frequency $\overline{\mu} = 0.5(\mu_{\min} + \mu_{\max})$.

On fig. 6.7, we plot $\mu \mapsto \varrho_k(\mu)$ quantity for 50 uniform samples of μ for various prescribed tolerances in the localization procedure algorithm 6.2. As expected, a smaller prescribed tolerance yields a smaller LHS approximation error. Recalling that a new subdomain is created per iteration of algorithm 6.2, we find that each domain decomposition

step contributes to decreasing the LHS approximation error. We find that the error decreases not only locally in the neighborhood of the point $\hat{\mu}_j$ that is added, but globally for all values of μ in the interval. This is confirmed by fig. 6.8, where we plot the maximum and median LHS approximation error with respect to the number of subdomains in the domain decomposition. The convergence is exponential, with a constant rate for the median LHS approximation error. We notice that the convergence rate of the maximum LHS approximation error is reduced after $K = 5$ subdomains.

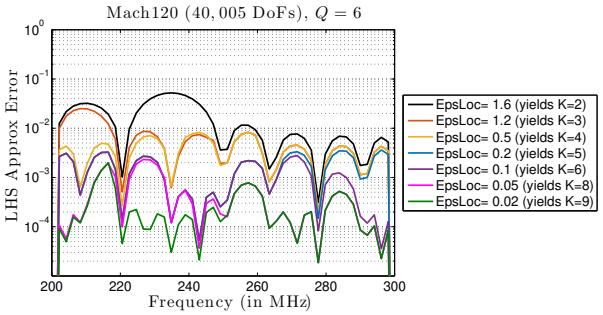


Figure 6.7: The LHS approximation error for various prescribed tolerances in algorithm 6.2.

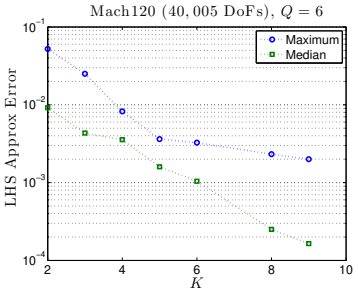


Figure 6.8: Maximum (over μ) of the *LHS approximation error* for various number of subdomains.

6.3.3 RB approximations on the geometry of a fighter aircraft

To start with, we build non-intrusive local affine approximations $\mathbf{A}_k(\mu)$ under the form eq. (6.2.10) using our localization procedure setting the prescribed number of terms to $Q = 6$ and tolerance $tol = 0.1$, which yields a domain decomposition of the frequency window into $K = 6$ subdomains (see section 6.3.2). As right-hand side, we consider an incident plane in the direction $\widehat{\boldsymbol{d}}(\theta, \phi)$ with $\theta = \pi/2$ (which corresponds to the plane of the wings) and $\phi = \pi/4$ (using the spherical coordinates defined on fig. 6.2). The incident plane wave is interpolated with $Q^f = 17$ terms using EIM.

Greedy VS monolithic construction

To start with, we build a reduced basis using the greedy construction algorithm 6.3 setting the prescribed tolerance to $\epsilon_{rel}^{\text{rb}} = 2.1\%$. On 8 CPUs, this takes $16\text{min}25$. We obtain $n_1 = 6$, $n_2 = 2$, $n_3 = 2$, $n_4 = 1$, $n_5 = 2$ and $n_6 = 2$, thus $N = 15$. Figure 6.9 shows the selected frequencies.

Next, we run the monolithic construction using the same $N = 15$ frequencies. On 8 CPUs, this takes $5\text{min}21$. Figure 6.10 summarizes the elapsed for the main operations: assembly of the right-hand side, matrix-vector operations with the CFIE operator, CFIE solves, mass matrix-vector products and mass solves. The other operations essentially consist in BLAS level 1 operations (*i.e.*, dot products).

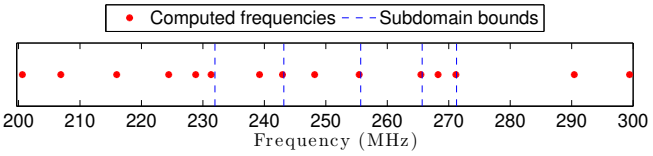


Figure 6.9: Frequencies selected by the greedy frequency sweep algorithm 6.3 with prescribed tolerance $\epsilon_{rel}^{\mathrm{rb}} = 2.1\%$.

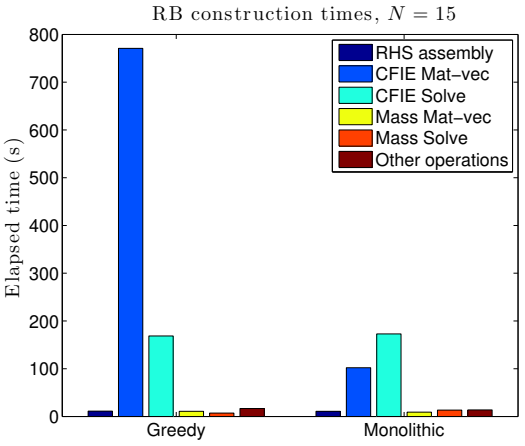


Figure 6.10: Elapsed times for the greedy and monolithic RB constructions for $N = 15$ reduced basis functions.

Clearly, the main difference between the greedy and monolithic construction is in the elapsed time on CFIE matrix-vector products: $12\mathrm{min}50$ for the greedy versus $1\mathrm{min}43$ for the monolithic construction. This is consistent with the number of CFIE matrix-vector products: about $QN + K - 1 = 95$ for the former and $Q + K - 1 = 11$ for the latter.

Quality assessment

We validate our RB approximation by computing 41 truth solutions at 41 uniformly distributed frequencies. This takes $10\mathrm{min}06$ on 8 CPUs. Notice that this is faster than the greedy construction ($16\mathrm{min}25$), but about twice longer than the monolithic construction ($5\mathrm{min}21$).

Figure 6.11 shows the relative error $\mu \mapsto \|\mathbf{u}(\mu) - \mathbf{u}_N(\mu)\|_{\mathbf{B}_V} / \|\mathbf{u}(\mu)\|_{\mathbf{B}_V}$ and heuristic indicator $\mu \mapsto \delta_k(\mu)$ given by eq. (6.2.17). We find the worst approximation error to be 1.5%, which is slightly below the prescribed 2.1%. This confirms that the heuristic indicator is successful. The frequencies with maximal error are located around $280\mathrm{MHz}$, where no frequencies have been selected (see fig. 6.9). For further validation, we compare on fig. 6.12 the monostatic Radar Cross Section (RCS) computed using the truth solution $\mathbf{u}(\mu)$ (black reference curve) or the RB approximation $\mathbf{u}_N(\mu)$ (magenta curve). Both

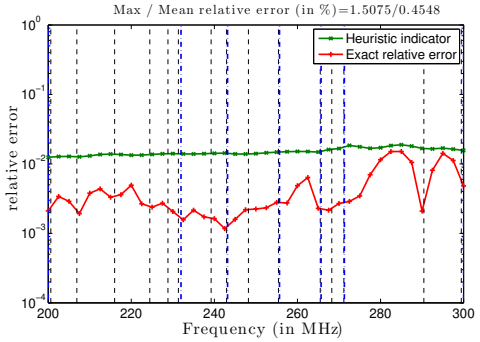


Figure 6.11: Relative error and heuristic error indicator for the greedy RB of size $N = 15$. Blue vertical lines indicate subdomain boundaries, black vertical lines indicate computed frequencies.

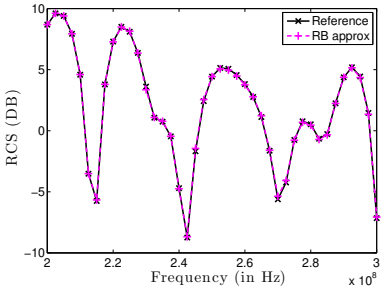


Figure 6.12: Reference and RB approximation of the Radar Cross Section (RCS) for the greedy RB of size $N = 15$.

curves are superimposed (even around the frequency 280MHz), which corroborates the good approximation properties of the greedy RB.

Which N frequencies for the monolithic construction ?

Clearly, for the best performance, we want to use the monolithic approach rather than the greedy approach. The logical question is which N frequencies should be used for the monolithic construction? A good choice is of course to consider the N frequencies selected by the greedy algorithm, but in practice, the greedy algorithm is too expensive due to the cost of matrix-vector products, as shown on fig. 6.10. We consider two possibilities:

- consider N uniformly distributed frequencies,
- consider N frequencies adaptively selected by algorithm 6.4. This algorithm proceeds in two steps: first, select $K - 1$ frequencies lying slightly to the left of each subdomain boundary. Then, select the next frequencies based on maximizing the distance to all previously selected frequencies. This aims at reproducing a distribution of selected frequencies similar to what the greedy algorithm can produce (see fig. 6.9).

We build two RBs of size $N = 15$ with uniformly distributed frequencies and with adaptively selected frequencies and assess the quality of each by computing the relative error. As can be seen on fig. 6.13, the uniform distribution of frequencies leads to large overshoots in the relative error in the neighborhood of the boundary between $\mathcal{D}_3$,$\mathcal{D}_4$ (around 255MHz) and $\mathcal{D}_5$, $\mathcal{D}_6$ (around 270MHz), where the error exceeds 5%. These errors are unacceptedly high. Indeed, we observe on fig. 6.15 that the RCS is deteriorated.

The RB using $N = 15$ frequencies by algorithm 6.4 does not suffer from this, as shown on fig. 6.14. The error stays beneath 1.6% and we find on fig. 6.15 that this level of error allows an accurate recovery of the RCS.

Algorithm 6.4: Algorithm to select N wavenumbers for the monolithic construction.

Input : number of wavenumbers N, domain decomposition $[\mu_{\min}, \mu_{\max}] = \cup_{k=1}^{K} \mathcal{D}_k$, with $\mathcal{D}_k = [\mu_{\min}^k, \mu_{\max}^k]$.

Output: a set $\mathcal{S}_N = \{\mu^{(n)}\}_{1 \leq n \leq N}$ of N selected wavenumbers.

Set $\mathcal{S}_0 = \emptyset$;
for $k = 1, \ldots, K$ **do**
  $\mu^\star = \mu_{\max}^k - \frac{1}{100}(\mu_{\max}^k - \mu_{\min}^k)$;
  Enrich $\mathcal{S}_k = \mathcal{S}_{k-1} \cup \{\hat{\mu}^\star\}$;
end
for $n = K+1, \ldots, N$ **do**
  $\mu^\star = \underset{\mu \in [\mu_{\min}, \mu_{\max}]}{\operatorname{argmax}} \min_{\mu' \in \mathcal{S}_n} |\mu - \mu'|$;
  Enrich $\mathcal{S}_n = \mathcal{S}_{n-1} \cup \{\hat{\mu}^\star\}$;
end

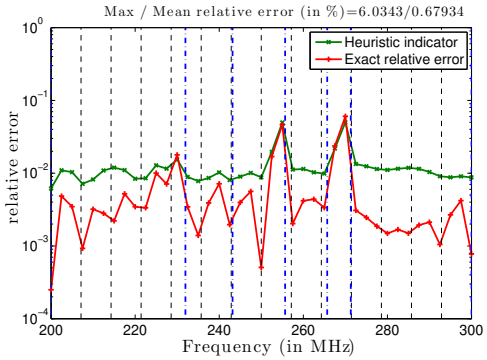


Figure 6.13: Relative error and heuristic error indicator for the RB with $N = 15$ uniform frequencies. Blue vertical lines indicate subdomain boundaries, black vertical lines indicate computed frequencies.

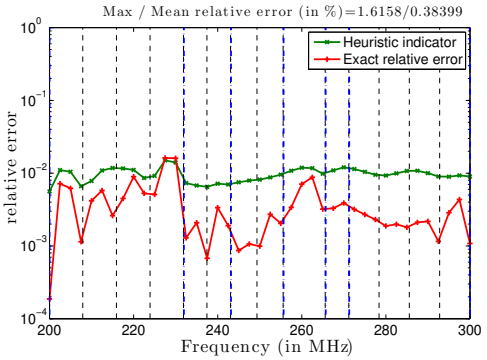


Figure 6.14: Relative error and heuristic error indicator using $N = 15$ frequencies by algorithm 6.4. Blue vertical lines indicate subdomain boundaries, black vertical lines indicate computed frequencies.

We conclude that algorithm 6.4 selects a better set of frequencies then the set of uniformly distributed frequencies.

6.3.4 Broadband frequency-sweeps

We again consider the CFIE on the geometry of the fighter aircraft fig. 6.6 with $40,005$ degrees of freedom. We want to compute the RCS on the band $100 - 400\text{MHz}$. We have seen in section 6.3.3 that a RB of size $N = 15$ was a good choice in order to recover a precise approximation of the RCS on the band $200 - 300\text{MHz}$. Which is the best strategy for broadband frequency-sweeps? We compare the following two strategies:

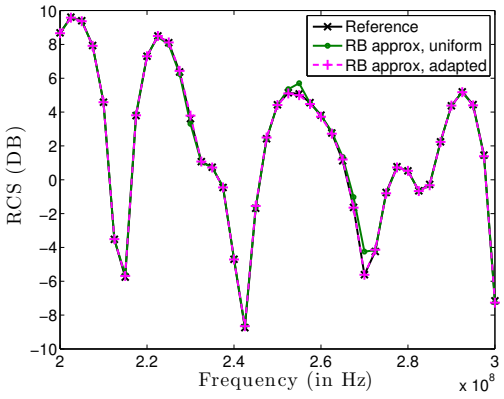


Figure 6.15: Reference RCS and RB approximations using $N = 15$ uniformly distributed or adaptively selected frequencies.

- *strategy with multiple RBs:* we build 3 distinct RBs: the first on $100 - 200$MHz, the second on $200 - 300$MHz and the third on $300 - 400$MHz;
- *strategy with a unique RB:* we build a unique RB for the full band $100 - 400$MHz.

Strategy with multiple RBs

For each band, we approximate the BEM operator using $Q = 6$ local terms and a prescribed tolerance 10^{-1} in the localization procedure (which yields $K = 6$ subdomains). We build RBs of size $N = 15$ using the monolithic construction with frequencies selected by algorithm 6.4. The elapsed times for the construction of each RB are consigned in table 6.3.

Band (Mhz)	$100 - 200$	$200 - 300$	$300 - 400$	Total (*i.e.*, $100 - 400$)
Elpased time (s)	680.63	321.79	275.26	1277.68

Table 6.3: Elapsed times for the construction of 3 RBs using monolithic construction with $N = 15$.

We find that the construction time is about twice longer on the band $100 - 200$MHz than on the other two bands. This is due to the fact that we are using the matrix-vector product accelerated with the FMM, which performs better when the mesh satisfies the "8 wavelengths per element" rule of the thumb. The matrix-vector with FMM is especially slow when the mesh is too refined, with occurs at the lower frequencies. Once the RBs are constructed, computing the RCS values on 121 uniformly distributed frequencies in the band $100 - 400$MHz is inexpensive. For validation, we have computed the reference RCS at these 121 frequency values in order to check that the RB approximations were accurate (see fig. 6.16). It is worth noting that these 121 high-fidelity computations took

2093 seconds. Thus, with no further optimization, the RB method provides a speed-up factor of about 2.

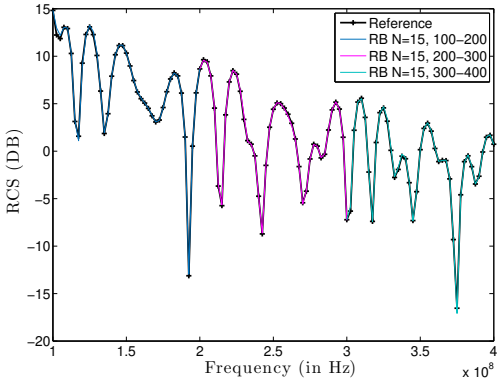


Figure 6.16: Reference RCS and RB approximations with the monolithic strategy with $N = 15$ on the three bands: $100 - 200$, $200 - 300$ and $300 - 400$MHz.

We have checked that $N = 15$ for each band was the optimal choice: indeed, a RB with $N < 15$ deteriorates the RCS and a RB with $N > 15$ means unnecessary computations.

Strategy with a unique RB

Since $N = 15$ reduced basis functions are needed per bandwidth of 100MHz, one would be tempted to think that $N = 45$ basis functions would be needed for a bandwidth of 300MHz. If that were true, then the strategy of building a unique RB valid over a broad band would bring no improvement over the strategy with multiple RBs. In fact, we are able to build a unique RB over the broad band with much less than $N = 45$ basis functions and thus the strategy with a unique RB is potentially competitive. Intuitively, this is linked to the fact that the basis functions from the band $100 - 200$MHz are useful to approximate the solutions on the rest of the band.

For the broadband $100 - 400$MHz, we approximate the BEM operator using $Q = 12$ local terms and a prescribed tolerance 10^{-1} in the localization procedure (which yields $K = 12$ subdomains). We start by building a RB of size $N = 45$. Since it provides an excellent approximation of the RCS over the broadband, we build smaller RBs of decreasing sizes $N = 44, 43, 42, \ldots$ until we find the optimally small RB, such that the RCS is captured with good accuracy. Figure 6.17 shows the RCS computed with RBs of sizes $N = 32, 30$ and 28 and the comparison with the reference RCS.

The RCS computed with the RB of size $N = 32$ coincides with the reference RCS. The RB of size $N = 30$ is in relatively good agreement with the reference, but the peak at 277MHz is not caught. With $N = 28$, the RB and reference RCS disagree not only at 277MHz but also at 240MHz. Clearly, RBs of size $N < 32$ are not able to properly resolve the RCS.

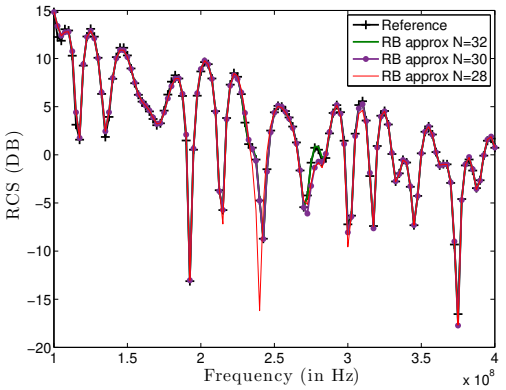


Figure 6.17: Reference RCS and RB approximations with the monolithic strategy with $N = 15$ on the full band $100 - 400$MHz.

In terms of computational time, building the RB of size $N = 32$ takes 1805 seconds. This is more than with the strategy with multiple RBs (1278 seconds), but still faster than performing 121 high-fidelity computations (2093 seconds). However, we believe that the strategy with a unique RB has great potential. Indeed, the number of CFIE solves and matrix-vector products is smaller with the unique RB strategy than with the multiple RB strategy. This is reflected on the elapsed times shown on fig. 6.18. Yet our current imple-

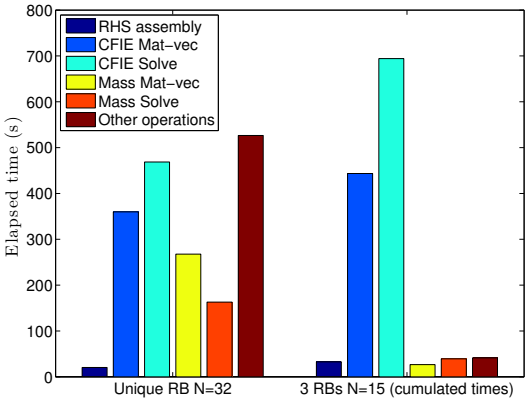


Figure 6.18: Elapsed times for the strategy with a unique RB $N = 32$ and the strategy with three RBs each of size $N = 15$.

mentation of the unique RB strategy spends about 30% of the elapsed time on non-critical operations, more specifically on data transfer. In double precision, the storage of all the necessary matrix-vector products $\mathbf{A}(\hat{\mu}_j)\mathbf{p}_i$ (see table 6.2) requires 221MBytes in the case $N = 15$ and 962MBytes in the case $N = 32$. Our code spends a significant amount of time on reading data on the disk. We believe that this situation could be circumvented by a more optimized implementation. Finally, the costs associated to operations with the mass matrix seems to scale badly with N. The number of mass matrix-vector products

explodes due to orthonormalization (here, we are using the stabilized method therefore we orthonormalize not only the RB of size N but also the matrix $\mathbf{Q}$ with NQ columns, see chapter 2), while the time elapsed on mass solves explodes because the block sizes become large. This seems to be an incompressible cost of the proposed strategy with a unique RB. A perspective to limit the block size would be to use localized, rather than nested, reduced basis spaces with a maximum size N^{loc} prescribed by the user [72].

6.4 Conclusions

In this chapter, we have proposed a non-intrusive reduced basis method for frequency-sweep analysis with the BEM. The notion of non-intrusiveness is understood in the sense of Casenave in [23], *i.e.*, the BEM code is only called to perform matrix-vector operations and system solves with standard BEM operators such as the EFIE, MFIE and CFIE that are introduced at the beginning of this chapter. In opposition to intrusive approaches [39, 42, 111, 77, 89], the proposed approach does not rely on any non-standard BEM operators and therefore it is not necessary to write new code for handling these, which saves precious engineering time. This first originality of our work resides in the use of *non-intrusive local affine approximations*, which rely on domain-decomposition of the frequency window into subdomains and the approximation of the BEM operator per subdomain. We have proposed a construction essentially based on the EIM [5]. We emphasize that our construction is completely mesh-free and therefore greatly differs (although similar in purpose) from the matrix-DEIM type approaches [82, 15].

The second originality of our work is the use of nested reduced basis approximation spaces. Thus, the size of the reduced basis increases with the frequency. In terms of computational strategy, we have proposed two approaches: the greedy and monolithic approaches. The first automatically selects the frequencies to be computed and detects how many frequencies have to be computed in order to reach a prescribed accuracy, the latter is more efficient in terms of computational performance, but it requires the user to provide the number and set of frequencies to be computed.

We conclude from numerical experiments that unconditionally stable formulations such as the CFIE are to be preferred over formulations with resonant frequencies, because the RB model is unable to detect the spurious modes. For the best computational performance, we advocate the monolithic approach rather than the greedy approach.

Future work will be to find some heuristics for guiding the choice of the number of reduced basis functions that are needed for a given bandwidth and given radio-electric size of the scattering object. This would enable the systematic use of the monolithic approach with guaranteed accuracy of the reduced basis approximation. We further envisage the construction of localized, rather than nested, reduced basis subspaces. This would circumvent the issue of the reduced basis size growing too rapidly with the frequency and therefore allow a better control over the computational complexity.

Chapter 7

Reduced basis method for aeroacoustic liner optimization in a discontinuous-Galerkin framework

Summary. This chapter is devoted to the application of the RBM for solving a parametrized problem in aeroacoustics. We want to solve the time-harmonic linearized Euler equations with impedance boundary conditions using a discontinuous Galerkin scheme. There are two varying parameters: the resistance and reactance, respectively the real and imaginary parts of impedance. The final goal is to be able to efficiently find the optimal impedance value, characterized by the best acoustic attenuation properties. The use of the RBM in this context is illustrated on a 2D axi-symmetrical model problem as well as on a 3D industrial aircraft engine nacelle problem. Significant speed-ups are obtained compared to traditional simulation campaigns.

Contents

7.1 Strong formulation and high-fidelity approximation 152
7.1.1 The time-harmonic linearized Euler equations 152
7.1.2 Discontinuous Galerkin scheme 155
7.2 The RBM for liner optimization 157
7.2.1 The impedance-parametrized problem 157
7.2.2 Results on the test case ALIAS 159
7.2.3 Validation campaign on the test case ALIAS 163
7.3 The RBM applied to a 3D nacelle engine 164
7.3.1 Problem description 164
7.3.2 Reduced basis approach 165

7.3.3 Validation campaign 168
7.4 Conclusions 169

7.1 Strong formulation and high-fidelity approximation

7.1.1 The time-harmonic linearized Euler equations

Let $\Omega \subset \mathbb{R}^3$ be a bounded spatial domain. For all time t in $\mathbb{R}$, fluid phenomena are described using four functions: three functions ρ, e and p of the $(\boldsymbol{x}, t)$ variables in $\Omega \times \mathbb{R}$ and with values in $\mathbb{R}$ and one function $\boldsymbol{v}$ of the $(\boldsymbol{x}, t)$ variables in $\Omega \times \mathbb{R}$ and with values in $\mathbb{R}^3$. These functions are called *density*, *total energy*, *pressure* and *velocity field* respectively. These four functions are related through the following *Euler equations*

$$\partial_t \begin{pmatrix} \rho \\ \rho u \\ \rho v \\ \rho w \\ e \end{pmatrix} + \partial_{x_1} \begin{pmatrix} \rho u \\ \rho u^2 + p \\ \rho uv \\ \rho uw \\ (e+p)u \end{pmatrix} + \partial_{x_2} \begin{pmatrix} \rho v \\ \rho uv \\ \rho v^2 + p \\ \rho vw \\ (e+p)v \end{pmatrix} + \partial_{x_3} \begin{pmatrix} \rho w \\ \rho uw \\ \rho vw \\ \rho w^2 + p \\ (e+p)w \end{pmatrix} = 0, \tag{7.1.1}$$

where we have denoted $\boldsymbol{v} = (u, v, w)^T$ the three velocity components and $\boldsymbol{x} = (x_1, x_2, x_3)^T$ the three components of the spatial variable. The pressure is further related to density and total energy by the *perfect gas law*

$$p = (\gamma - 1)\left(e - \frac{1}{2}\rho|\boldsymbol{v}|^2\right), \tag{7.1.2}$$

where $|\boldsymbol{v}| = \sqrt{u^2 + v^2 + w^2}$ denotes the euclidian norm of $\boldsymbol{v}$ and $\gamma > 1$ is the Laplace constant.

Linearized Euler equations

In order to obtain the *linearized Euler equations* (LEEs) [58, 73, 53], the Euler equations (7.1.1) are linearized around a given uniform steady flow. Only first-order perturbation terms are taken into account. More precisely, denote $(\rho_0, \boldsymbol{v}_0, p_0)$ the uniform steady flow and $(\rho_1, \boldsymbol{v}_1, p_1)$ the first-order dimensionless flow perturbation. Solution fields $(\rho, \boldsymbol{v}, p)$ are sought under the form

$$\begin{aligned} \rho &= \rho_0 + \epsilon\rho_0\rho_1 + \mathcal{O}(\epsilon^2), \\ \boldsymbol{v} &= \boldsymbol{v}_0 + \epsilon c_0 \boldsymbol{v}_1 + \mathcal{O}(\epsilon^2), \\ p &= p_0 + \epsilon\rho_0 c_0^2 p_1 + \mathcal{O}(\epsilon^2), \end{aligned} \tag{7.1.3}$$

with $c_0 = \sqrt{\gamma p_0/\rho_0}$ the speed of sound in the uniform steady flow, equal to around $340\mathrm{m/s}$ in air at rest. We further assume that the perturbation is isentropic, that is $p_1 = \rho_1$. Keeping only the first-order terms we find that the variable $\boldsymbol{\varphi} = (\boldsymbol{v}_1, p_1)^T$ satisfies [9]

$$\partial_t \boldsymbol{\varphi} + \sum_{i=1}^{3} A^i \partial_{x_i} \boldsymbol{\varphi} = 0, \tag{7.1.4}$$

with symmetric matrices

$$A^1 = \begin{pmatrix} u_0 & 0 & 0 & 1 \\ 0 & u_0 & 0 & 0 \\ 0 & 0 & u_0 & 0 \\ 1 & 0 & 0 & u_0 \end{pmatrix}, \; A^2 = \begin{pmatrix} v_0 & 0 & 0 & 0 \\ 0 & v_0 & 0 & 1 \\ 0 & 0 & v_0 & 0 \\ 0 & 1 & 0 & v_0 \end{pmatrix}, \; A^3 = \begin{pmatrix} w_0 & 0 & 0 & 0 \\ 0 & w_0 & 0 & 0 \\ 0 & 0 & w_0 & 1 \\ 0 & 0 & 1 & w_0 \end{pmatrix} \tag{7.1.5}$$

Boundary conditions

Denote $\partial\Omega$ the boundary of Ω and $\hat{\boldsymbol{n}} : \partial\Omega \to \mathbb{R}^3$ the outgoing unit normal. We denote $\hat{\boldsymbol{n}} = (n_1, n_2, n_3)^T$ the three components of the unit normal. We now specify the boundary conditions on $\partial\Omega$. The boundary is split as into four parts

$$\partial\Omega = \Gamma_{\text{in}} \cup \Gamma_{\text{out}} \cup \Gamma_{\text{wall}} \cup \Gamma_{\text{z}}, \tag{7.1.6}$$

with a different boundary condition imposed on each part of the boundary, as shown on fig. 7.1.

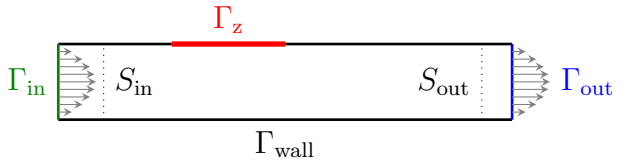


Figure 7.1: Model geometry for the LEEs.

- *Rigid wall boundary condition on* Γ_{wall}*:* the fluid is not allowed to penetrate through the boundary. Thus, the components of velocity tangential to Γ_{wall} must vanish [57], *i.e.*, $\boldsymbol{v}_1 \cdot \hat{\boldsymbol{n}} = 0$.
- *Impedance boundary condition on* Γ_{z}*:* this corresponds to an absorbing medium (the acoustic liner), fully characterized by an acoustic impedance Z [70, 2]. This boundary condition is expressed as $p_1 = Z\boldsymbol{v}_1 \cdot \hat{\boldsymbol{n}}$.

Remark. *On the boundary* Γ_{z}*, the mean flow is assumed to satisfy the rigid wall boundary condition* $\boldsymbol{v}_0 \cdot \hat{\boldsymbol{n}} = 0$.

In order to obtain proper in-flow and out-flow boundary conditions, the reflected waves should be prevented from propagating back into the domain. In this work, we rely the theory of characteristics [118, 43, 46, 119]. The flux on the boundary $\partial\Omega$ is first split into incoming and outgoing waves using the hyperbolicity of the system. Indeed, the flux matrix $F = \sum_{i=1}^{3} A^i n_i$ (defined on the boundary $\partial\Omega$) is real symmetric, thus there exists an orthonormal matrix P such that $F = P\Lambda P^T$, where Λ is the diagonal matrix holding the eigenvalues. We note that the sign of eigenvalues depends on the local mean flow normal velocity $\boldsymbol{v}_0 \cdot \boldsymbol{n}$. We can split $\Lambda = \Lambda^+ + \Lambda^-$, where Λ^+ (resp. Λ^-) holds the positive (resp. negative) eigenvalues. Similarly, the flux is split as $F = F^+ + F^-$, with $F^\pm = P\Lambda^\pm P^T$ the flux of incoming $(-)$ or outgoing $(+)$ waves. This being set, we can express the boundary conditions on Γ_{in} and Γ_{out}.

- *In-flow boundary condition on* Γ_{in}*:* on this boundary, a given fluid perturbation $\boldsymbol{\varphi}_{\text{in}}$ defined on Γ_{in} is given. The in-flow boundary condition consists in suppressing the outgoing waves and equaling the flux to the given incoming flux [57]. Mathematically, this corresponds to $F^-\boldsymbol{\varphi} = F^-\boldsymbol{\varphi}_{\text{in}}$.

- *Out-flow boundary condition on* Γ_{out}*:* this corresponds to an open boundary where the fluid is free to flow outside of the domain. Such artificial non-reflecting boundary condition is imposed in order to obtain a computationally bounded domain, when the physical domain is in fact unbounded [116, 59]. The concern of such a boundary condition is therefore purely numerical. Here, the outgoing waves are kept while the incoming waves are suppressed [118, 43], which consists in the boundary condition $F^-\boldsymbol{\varphi} = 0$ on Γ_{out}.

Time-harmonic regime

In this thesis, we further consider time-harmonic perturbation fields

$$p_1(\boldsymbol{x}, t) = \Re\left\{p_1(\boldsymbol{x})e^{j\omega t}\right\}, \quad \boldsymbol{v}_1(\boldsymbol{x}, t) = \Re\left\{\boldsymbol{v}_1(\boldsymbol{x})e^{j\omega t}\right\}, \tag{7.1.7}$$

thus the unknown functions are now the pressure perturbation amplitude $p_1 : \Omega \to \mathbb{C}$ and the velocity perturbation amplitude $\boldsymbol{v}_1 : \Omega \to \mathbb{C}^3$. We emphasize that these are complex-values fields. Thus, the aeroacoustic problem is expressed as follows: *find* $\boldsymbol{\varphi} = (\boldsymbol{v}_1, p_1)$ *solution to the time-harmonic LEEs*

$$\begin{cases} j\omega\boldsymbol{\varphi} + \sum_{i=1}^{3} A^i \partial_{x_i}\boldsymbol{\varphi} = 0 & \text{in } \Omega, \\ F^-\boldsymbol{\varphi} = F^-\boldsymbol{\varphi}_{\text{in}} & \text{on } \Gamma_{\text{in}}, \\ p_1 - Z\boldsymbol{v}_1 \cdot \hat{\boldsymbol{n}} = 0 & \text{on } \Gamma_{\text{z}}, \\ \boldsymbol{v}_1 \cdot \hat{\boldsymbol{n}} = 0 & \text{on } \Gamma_{\text{wall}}, \\ F^-\boldsymbol{\varphi} = 0 & \text{on } \Gamma_{\text{out}}. \end{cases} \tag{7.1.8}$$

7.1.2 Discontinuous Galerkin scheme

In this section, we review the discontinuous Galerkin scheme that will be used to solve the time-harmonic LEEs. The original method is presented in [35]. To start with, the physical domain Ω is approximated by a computational domain Ω_h, obtained by triangulation into $\mathcal{N}_{\text{elt}}$ non-overlapping elements $\mathcal{T}_h = \{T_e\}_{1 \leqslant e \leqslant \mathcal{N}_{\text{elt}}}$

$$\Omega \approx \Omega_h = \bigcup_{e=1}^{\mathcal{N}_{\text{elt}}} T_e. \tag{7.1.9}$$

We seek a high-fidelity approximation $\boldsymbol{\varphi}_h$ on the computational domain Ω_h in the first-order Lagrange approximation space

$$X_h^{\text{DG}} = \{\boldsymbol{\psi}_h \in (L^2(\Omega_h))^4 : \forall T_e \in \mathcal{T}_h,\, \boldsymbol{\psi}_h|_{T_e} \in (\mathbb{P}^1(T_e))^4\}. \tag{7.1.10}$$

Local formulation

In order to discretize eq. (7.1.8), we consider an element $T_e \in \mathcal{T}_h$ and denote $\boldsymbol{\varphi}_h^e = \boldsymbol{\varphi}_h|_{T_e}$. Multiplying by a test function and integrating over the element gives

$$\int_{T_e} j\omega \boldsymbol{\varphi}_h^e \cdot \overline{\boldsymbol{\psi}_h^e} \mathrm{d}\Omega + \sum_{i=1}^{3} \int_{T_e} A^i \partial_{x_i} \boldsymbol{\varphi}_h^e \cdot \overline{\boldsymbol{\psi}_h^e} \mathrm{d}\Omega = 0. \tag{7.1.11}$$

In order to ensure the connection between elements, we need to impose flux conservation across the interfaces between elements. This is done by adding a numerical flux term as follows [35]

$$\int_{T_e} j\omega \boldsymbol{\varphi}_h^e \cdot \overline{\boldsymbol{\psi}_h^e} \mathrm{d}\Omega + \sum_{i=1}^{3} \int_{T_e} A^i \partial_{x_i} \boldsymbol{\varphi}_h^e \cdot \overline{\boldsymbol{\psi}_h^e} \mathrm{d}\Omega + \int_{\partial T_e} \boldsymbol{F}_e(\boldsymbol{\varphi}_h^{e-}, \boldsymbol{\varphi}_h^{e+}) \cdot \overline{\boldsymbol{\psi}_h^{e-}} \mathrm{d}\Gamma = 0. \tag{7.1.12}$$

Following the Discontinuous-Galerkin (DG) paradigm, discontinuities are allowed across the interfaces between elements [101]; the numerical flux provides the link between the interior and exterior traces, defined for $\boldsymbol{x} \in \partial T_e$ as

$$\begin{cases} \boldsymbol{\varphi}_h^{e-}(\boldsymbol{x}) = \lim\limits_{\boldsymbol{y} \to \boldsymbol{x},\, \boldsymbol{y} \in T_e} \boldsymbol{\varphi}_h^e(\boldsymbol{y}) \\ \boldsymbol{\varphi}_h^{e+}(\boldsymbol{x}) = \lim\limits_{\boldsymbol{y} \to \boldsymbol{x},\, \boldsymbol{y} \notin T_e} \boldsymbol{\varphi}_h^e(\boldsymbol{y}). \end{cases} \tag{7.1.13}$$

The numerical flux depends on whether the element boundary ∂T_e intersects or not with the boundary $\partial\Omega_h$ of the computational domain. If $\partial T_e \cap \partial\Omega_h \neq \emptyset$, then the numerical flux enforces the boundary conditions. For the part of the boundary that does not intersect with the boundary, that is $\partial T_e \setminus (\partial T_e \cap \partial\Omega_h)$, the numerical flux simply ensures the connection between elements. Following [95], the fluxes are chosen as follows:

- *Interconnecting elements:* an upwind numerical flux is used to interconnect the elements, this corresponds to

$$\int_{\partial T_e \setminus (\partial T_e \cap \partial\Omega_h)} \boldsymbol{F}_e(\boldsymbol{\varphi}_h^{e-}, \boldsymbol{\varphi}_h^{e+}) \cdot \overline{\boldsymbol{\psi}_h^{e-}} \mathrm{d}\Gamma = \int_{\partial T_e \setminus (\partial T_e \cap \partial\Omega_h)} F^-(\boldsymbol{\varphi}_h^{e+} - \boldsymbol{\varphi}_h^{e-}) \cdot \overline{\boldsymbol{\psi}_h^{e-}} \mathrm{d}\Gamma, \tag{7.1.14a}$$

  where we recall that $F^- = P\Lambda^- P^T$ is the flux of incoming waves.

- *Rigid wall boundary condition:* the numerical flux for the rigid wall boundary condition on Γ_{wall} writes

$$\int_{\partial T_e \cap \Gamma_{\text{wall}}} \boldsymbol{F}_e(\boldsymbol{\varphi}_h^{e-}, \boldsymbol{\varphi}_h^{e+}) \cdot \overline{\boldsymbol{\psi}_h^{e-}} \mathrm{d}\Gamma = \int_{\partial T_e \cap \Gamma_{\text{wall}}} M_{\text{wall}} \boldsymbol{\varphi}_h^{e-} \cdot \overline{\boldsymbol{\psi}_h^{e-}} \mathrm{d}\Gamma \tag{7.1.14b}$$

  where $M_{\text{wall}} = \begin{pmatrix} \hat{\boldsymbol{n}} \otimes \hat{\boldsymbol{n}} & -\hat{\boldsymbol{n}} \\ -\hat{\boldsymbol{n}}^T & 1 \end{pmatrix}$ is a 4×4 real matrix.

- *Impedance boundary condition:* the numerical flux for the impedance boundary condition on Γ_{z} writes

$$\int_{\partial T_e \cap \Gamma_{\text{z}}} \boldsymbol{F}_e(\boldsymbol{\varphi}_h^{e-}, \boldsymbol{\varphi}_h^{e+}) \cdot \overline{\boldsymbol{\psi}_h^{e-}} \mathrm{d}\Gamma = \int_{\partial T_e \cap \Gamma_{\text{wall}}} \frac{Z-1}{Z+1} M_\beta \boldsymbol{\varphi}_h^{e-} \cdot \overline{\boldsymbol{\psi}_h^{e-}} \mathrm{d}\Gamma \tag{7.1.14c}$$

  where $M_\beta = \begin{pmatrix} \hat{\boldsymbol{n}} \otimes \hat{\boldsymbol{n}} & \hat{\boldsymbol{n}} \\ -\hat{\boldsymbol{n}}^T & -1 \end{pmatrix}$ is a 4×4 real matrix.

- *In-flow boundary condition:* the numerical flux for the in-flow boundary condition on Γ_{in} corresponds to an upwind numerical flux, interconnecting the interior trace with the given in flow-perturbation $\boldsymbol{\varphi}_{\text{in}}$, that is

$$\int_{\partial T_e \cap \Gamma_{\text{in}}} \boldsymbol{F}_e(\boldsymbol{\varphi}_h^{e-}, \boldsymbol{\varphi}_h^{e+}) \cdot \overline{\boldsymbol{\psi}_h^{e-}} \mathrm{d}\Gamma = \int_{\partial T_e \cap \Gamma_{\text{in}}} F^-(\boldsymbol{\varphi}_{in} - \boldsymbol{\varphi}_h^{e-}) \cdot \overline{\boldsymbol{\psi}_h^{e-}} \mathrm{d}\Gamma \tag{7.1.14d}$$

- *Out-flow boundary condition:* the numerical flux for the out-flow boundary condition on Γ_{out} writes

$$\int_{\partial T_e \cap \Gamma_{\text{out}}} \boldsymbol{F}_e(\boldsymbol{\varphi}_h^{e-}, \boldsymbol{\varphi}_h^{e+}) \cdot \overline{\boldsymbol{\psi}_h^{e-}} \mathrm{d}\Gamma = \int_{\partial T_e \cap \Gamma_{\text{out}}} -F^- \boldsymbol{\varphi}_h^{e-} \cdot \overline{\boldsymbol{\psi}_h^{e-}} \mathrm{d}\Gamma \tag{7.1.14e}$$

Global formulation

Using the local formulation eq. (7.1.12), valid over any element $T_e \in \mathcal{T}_h$, a global DG formulation can easily be obtained by summing over all elements in the triangulation $\mathcal{T}_h$. Namely, the global formulation reads: *find* $\boldsymbol{\varphi}_h \in X_h^{\text{DG}}$ *such that*

$$\forall \boldsymbol{\psi}_h \in X_h^{\text{DG}}, \quad a(\boldsymbol{\varphi}_h, \boldsymbol{\psi}_h) = f(\boldsymbol{\psi}_h), \tag{7.1.15}$$

where $a : X_h^{\mathrm{DG}} \times X_h^{\mathrm{DG}} \to \mathbb{C}$ is the continuous sesquilinear form defined by

$$\begin{aligned} a(\boldsymbol{\varphi}_h, \boldsymbol{\psi}_h) = \sum_{T_e \in \mathcal{T}_h} \Bigg(& \int_{T_e} j\omega \boldsymbol{\varphi}_h^e \cdot \overline{\boldsymbol{\psi}_h^e} \mathrm{d}\Omega + \sum_{i=1}^{3} \int_{T_e} A^i \partial_{x_i} \boldsymbol{\varphi}_h^e \cdot \overline{\boldsymbol{\psi}_h^e} \mathrm{d}\Omega \\ & + \int_{\partial T_e \setminus (\partial T_e \cap \partial \Omega_h)} F^-(\boldsymbol{\varphi}_h^{e+} - \boldsymbol{\varphi}_h^{e-}) \cdot \overline{\boldsymbol{\psi}_h^{e-}} \mathrm{d}\Gamma \\ & + \int_{\partial T_e \cap \Gamma_{\mathrm{wall}}} M_{\mathrm{wall}} \boldsymbol{\varphi}_h^{e-} \cdot \overline{\boldsymbol{\psi}_h^{e-}} \mathrm{d}\Gamma \\ & + \int_{\partial T_e \cap \Gamma_{\mathrm{z}}} \frac{Z-1}{Z+1} M_\beta \boldsymbol{\varphi}_h^{e-} \cdot \overline{\boldsymbol{\psi}_h^{e-}} \mathrm{d}\Gamma \\ & + \int_{\partial T_e \cap (\Gamma_{\mathrm{in}} \cup \Gamma_{\mathrm{out}})} -F^- \boldsymbol{\varphi}_h^{e-} \cdot \overline{\boldsymbol{\psi}_h^{e-}} \mathrm{d}\Gamma \Bigg), \end{aligned} \tag{7.1.16}$$

and $f : X_h^{\mathrm{DG}} \to \mathbb{C}$ is the continuous linear form defined by

$$f(\boldsymbol{\psi}_h) = \sum_{T_e \in \mathcal{T}_h} \int_{\partial T_e \cap \Gamma_{\mathrm{in}}} -F^- \boldsymbol{\varphi}_{in} \cdot \overline{\boldsymbol{\psi}_h^{e-}} \mathrm{d}\Gamma. \tag{7.1.17}$$

Denoting $\{\boldsymbol{w}_i\}_{1 \leqslant i \leqslant \mathcal{N}}$ a basis for the first-order Lagrange approximation space X_h^{DG}, we may introduce the matrix $\mathbf{A} \in \mathbb{C}^{\mathcal{N} \times \mathcal{N}}$ and vector $\mathbf{f} \in \mathbb{C}^{\mathcal{N}}$ with coefficients

$$\mathbf{A}_{ij} = a(\boldsymbol{w}_j, \boldsymbol{w}_i), \quad \mathbf{f}_j = f(\boldsymbol{w}_j), \quad 1 \leqslant i, j \leqslant \mathcal{N}. \tag{7.1.18}$$

Thus, the solution to eq. (7.1.15) is given by $\boldsymbol{\varphi}_h = \sum_{i=1}^{\mathcal{N}} \mathbf{u}_i \boldsymbol{w}_i$ where $\mathbf{u} \in \mathbb{C}^{\mathcal{N}}$ solves the large-scale linear system $\mathbf{A}\mathbf{u} = \mathbf{f}$. In our numerical simulations, we will use the FETI-2LM domain decomposition method introduced in section 5.1.3 in order to efficiently solve such large linear system on parallel architectures.

7.2 The RBM for liner optimization

7.2.1 The impedance-parametrized problem

In applied aeroacoustics, one solves the LEEs for pressure and velocity fields in order to compute the *energy density* and *acoustic intensity* fields given by the formulas of Cantrell & Hart [21], respectively

$$\begin{aligned} e &= \frac{p_1^2}{2\rho_0 c_0^2} + \frac{\rho_0}{2}|\boldsymbol{v}_1|^2 + (\boldsymbol{v}_0 \cdot \boldsymbol{v}_1)\frac{p_1}{c_0^2}, \\ \boldsymbol{i} &= p_1 \boldsymbol{v}_1 + \frac{p_1^2}{\rho_0 c_0^2}\boldsymbol{v}_0 + \rho_0(\boldsymbol{v}_0 \cdot \boldsymbol{v}_1)\boldsymbol{v}_1 + \frac{p_1}{c_0^2}(\boldsymbol{v}_0 \cdot \boldsymbol{v}_1)\boldsymbol{v}_0. \end{aligned} \tag{7.2.1}$$

In this work, we wish to solve the discretized time-harmonic LEEs eq. (7.1.15) not just for one impedance value, but for all possible values of impedance in a given set. The goal is to be able to find the optimal impedance value $Z^\star$ that maximizes the attenuation intensity given by

$$C = 10\log_{10}\left(\frac{\int_{S_{\text{out}}} \boldsymbol{i}\cdot \mathrm{d}\boldsymbol{S}}{\int_{S_{\text{in}}} \boldsymbol{i}\cdot \mathrm{d}\boldsymbol{S}}\right), \tag{7.2.2}$$

where $S_{\text{in}}, S_{\text{out}}$ are two control surfaces as shown on fig. 7.1. The optimal impedance value $Z^\star$ thus gives us the best absorbing properties [93].

In order to formulate the parametrized problem, let us introduce $\mu = (\Re(Z), \Im(Z))^T$ as our varying parameter, taking values in some compact set $\mathcal{D} \subset \mathbb{R}^2$. Note that, rather than one complex parameter, we treat two real parameters corresponding to the real and imaginary parts of impedance. Under this parametrized setting, the sesquilinear form defined by eq. (7.1.16) in fact depends on μ, thus $a(\cdot,\cdot) = a(\cdot,\cdot;\mu)$. Moreover, the dependency in μ is trivially affine, since

$$a(\cdot,\cdot;\mu) = \beta(\mu)a_1(\cdot;\cdot) + a_2(\cdot,\cdot), \tag{7.2.3}$$

with $\beta(\mu) = \frac{(\mu_1+j\mu_2)-1}{(\mu_1+j\mu_2)+1}$ the so-called *reflection coefficient* and $a_1 : X_h^{\text{DG}} \times X_h^{\text{DG}} \to \mathbb{C}$ defined by

$$a_1(\boldsymbol{\varphi}_h, \boldsymbol{\psi}_h) = \sum_{T_e\in\mathcal{T}_h} \int_{\partial T_e\cap\Gamma_{\text{wall}}} M_z\boldsymbol{\varphi}_h^{e-}\cdot\overline{\boldsymbol{\psi}_h^{e-}}\mathrm{d}\Gamma. \tag{7.2.4}$$

This being set, the impedance-parametrized discretized LEEs read: *find* $\boldsymbol{\varphi}_h(\mu) \in X_h^{\text{DG}}$ *such that*

$$\forall \boldsymbol{\psi}_h \in X_h^{\text{DG}}, \quad a(\boldsymbol{\varphi}_h(\mu), \boldsymbol{\psi}_h; \mu) = f(\boldsymbol{\psi}_h). \tag{7.2.5}$$

Clearly, this formulation fits the framework of section 1.1.4, with the correspondence given by table 7.1.

Abstract setting	Parametrized LEEs
V	X_h^{DG}
W	X_h^{DG}
μ	$\mu = (\Re(Z), \Im(Z))$
$a(\cdot,\cdot,\mu)$	Equation (7.1.16)
$f(\mu)$	Equation (7.1.17)
$u(\mu)$	$\boldsymbol{\varphi}_h(\mu)$

Table 7.1: Correspondance between the parametrized LEEs and the general abstract setting of section 1.1.4.

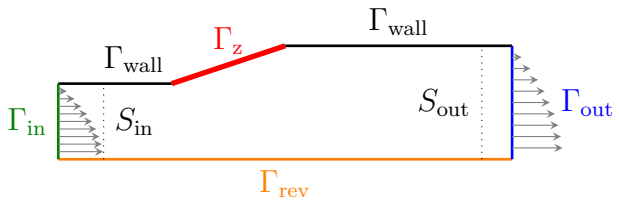


Figure 7.2: Geometry of the ALIAS test case

7.2.2 Results on the test case ALIAS

Description of the test case ALIAS

The test case ALIAS (Acoustic Liners for Air Conditioning Systems, see [93]) uses the axi-symmetric 2D geometry depicted on fig. 7.2. An axi-symmetric revolution boundary condition is enforced on the part of the border Γ_{rev} in order to simulate 3D flows on this geometry [34].

The global number of degrees of freedom is $\mathcal{N} = 118\,944$. The problem is distributed on 4 processors, with each processor handling a local number of degrees of freedom of roughly $30\,000$. We have reported the local number of degrees of freedom per processor in table 7.2. Notice that the sum of the local number of degrees of freedom is slightly larger

Process Number	1	2	3	4
Nb local DoFs	29 028	29 256	30 876	30 672

Table 7.2: Local number of degrees of freedom (DoFs) per processor for the test test ALIAS.

than $\mathcal{N}$. This is due to the redundancy of the degrees of freedom lying on the interface between the subdomains, as explained in section 5.1.3.

The frequency is set to $f = 1000\text{Hz}$. On the in-flow, we impose a non-radial and 1st order azimuthal acoustic duct mode $(m, n) = (0, 1)$ [79, 100, 91, 92]. A visualization of the truth pressure field at the impedance value $Z = 0.5 + 5j$ is presented on fig. 7.3. Visually, we see that the presence of the liner has little impact on the propagation of the acoustic mode.

Two reduced basis approximation spaces

We consider impedance values Z such that $\Re(Z) \in [-0.5, 5]$ ans $\Im(Z) \in [-5, 5]$. We build two reduced basis approximation spaces, each of dimension $N = 8$: a "Greedy RB" approximation space, built using the Greedy algorithm 2.1 which automatically selects the impedance values to be solved based on maximizing the residual-norm and a "By hand

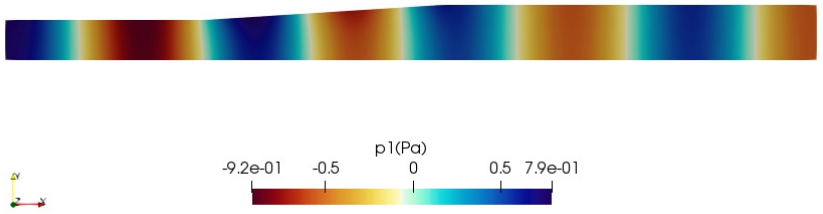


Figure 7.3: Truth pressure field, $Z = 0.5 + 5j$.

RB" approximation space, for which we have manually selected the impedance values to be solved. The solved impedance values for each RB are consigned in table 7.3. For simplicity, we choose to measure the residuals in the euclidian norm denoted $\|\cdot\|_2$.

Greedy RB	$\Re(Z)$	1.000	0.500	0.500	0.500	0.500	0.500	0.500	0.500
	$\Im(Z)$	0.250	-3.000	3.500	-0.833	1.062	-0.291	-1.645	1.875
By hand RB	$\Re(Z)$	1.500	4.000	2.750	2.750	2.750	2.750	2.750	2.750
	$\Im(Z)$	0.000	0.000	-4.000	-3.000	-2.000	2.000	3.000	4.000

Table 7.3: Two reduced basis approximation spaces for the test case ALIAS.

We find that the Greedy algorithm mostly selects impedance values with $\Re(Z) = 0.5$, which coincides with the smallest admissible $\Re(Z)$. We note that at the interface between the rigid wall and impedance boundary conditions, the impedance value drops from infinite to finite: thus the impedance values $\Re(Z) = 0.5$ are associated with the largest discontinuities.

Online, we compute the Least-Squares RB approximations for 4066 different values of Z on a 38×107 grid. For all these values of Z, we compute the attenuation intensity $C = C(Z)$ given by eq. (7.2.2). We obtain the two attenuation intensity maps shown on figs. 7.4 and 7.5. The two are essentially indistinguishable and allow quick identification of the optimal impedance value, namely $Z^\star = 0.5 + 0j$.

At the 4066 different values of $\mu = (\Re(Z), \Im(Z))$ computed online, we also compute our heuristic indicator $\mu \mapsto \|A(\mu)u_N(\mu) - f\|_2/(\hat{\alpha}\|u_N(\mu)\|_2)$, which is expected to provide a good approximation for the RB relative error $\|u(\mu) - u_N(\mu)\|_2/\|u(\mu)\|_2$ (see chapter 2). We obtain the two estimated relative error maps shown on figs. 7.6 and 7.7. These two estimated error maps are drastically different. The Greedy RB estimates a level of error of roughly 0.2% over the set of impedance values (maximum 0.32%), with local improvements in the neighborhood of the impedance values which have been solved for the re-

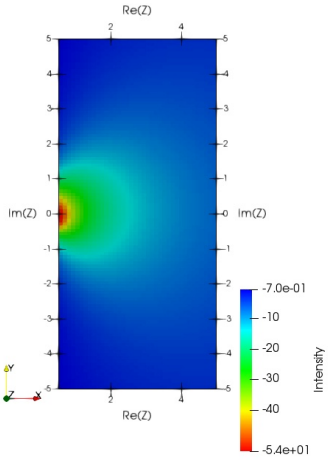


Figure 7.4: Attenuation intensity map using Greedy RB.

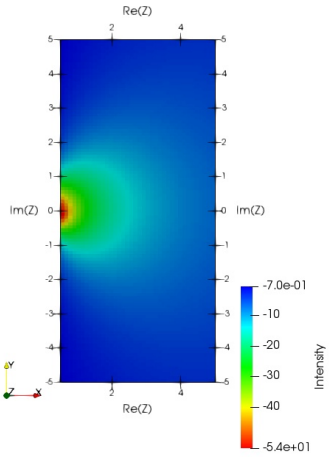


Figure 7.5: Attenuation intensity map using By hand RB.

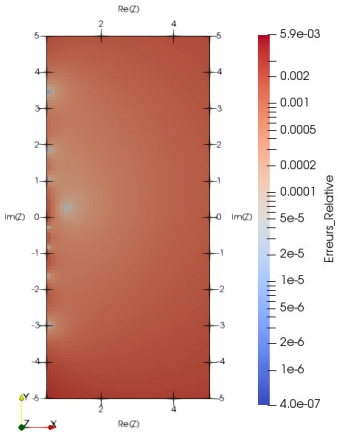


Figure 7.6: Estimated relative error map using Greedy RB.

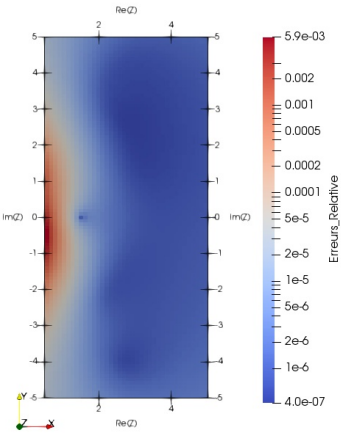


Figure 7.7: Estimated relative error map using By hand RB.

duced basis. For the By hand RB, we estimate very low errors across the set of impedance values, with maximum at the impedance values $0.5 \leqslant \Re(Z) \leqslant 1$ and $|\Im(Z)| < 3$ where the level of error is expected to be around 0.5% (maximum 0.59%). We insist that the two maps figs. 7.6 and 7.7 represent the estimated relative errors and not the actual relative errors. At this stage, we have no guarantee that the two coincide. The upcoming validation phase will give us more insight on the predictive power of these estimated relative error maps.

We show the truth pressure field at the optimal impedance value on fig. 7.8. In opposition to fig. 7.3, the presence of the acoustic liner strongly impacts the propagation of the acoustic mode, as the pressure perturbation is almost vanished.

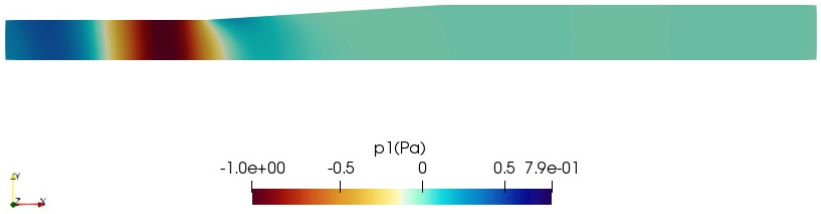


Figure 7.8: Truth pressure field at optimal impedance value $Z^\star = 0.5 + 0j$.

Visually, our two RB approximations of the pressure field at the optimal impedance value are essentially indistinguishable from the truth pressure field plotted on fig. 7.8. However, none of our two RB approximations recover the truth pressure field exactly. We define the RB approximation error in the pressure field as $|p_{DG} - p_{RB}|$, where p_{DG} denotes the truth pressure and p_{RB} the RB approximation of pressure. We plot this quantity on fig. 7.9 (Greedy case) and fig. 7.10 (By Hand case).

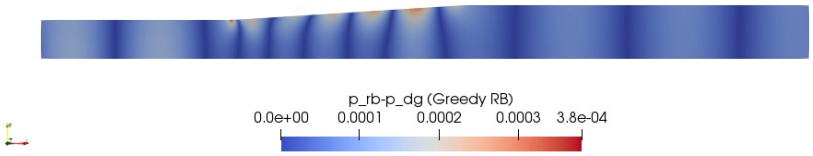


Figure 7.9: RB approximation error in the pressure field at optimal impedance value $Z^\star = 0.5 + 0j$ using the Greedy RB.

We find that the choice of the RB (Greedy or By hand) strongly impacts the RB approximation error in the pressure field. The Greedy RB approximation error is smallest upstream and downstream the liner, whereas it is largest in the vicinity of the liner. In opposition, the By hand RB approximation error is largest downstream the liner and is maximal at the junction between the outflow and axi-symmetrical boundary conditions. Moreover, we find that the By hand RB approximation error is not as smooth as the Greedy RB approximation error. We believe that the By Hand RB is less fit than the Greedy RB in approximating the field at the optimal impedance value $Z^\star = 0.5 + 0j$, because the truth solves performed by Greedy are at impedance values closer to $Z^\star$ than the ones manually selected for the By hand RB (see table 7.3).

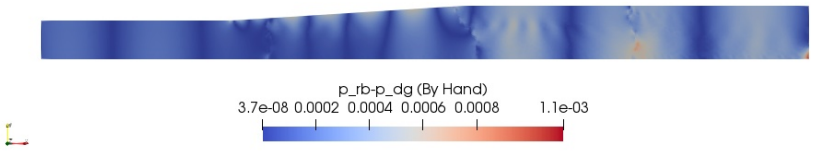


Figure 7.10: RB approximation error in the pressure field at optimal impedance value $Z^\star = 0.5 + 0j$ using By Hand RB.

7.2.3 Validation campaign on the test case ALIAS

For the sake of validation, we compute the truth solutions $u(\mu)$ at 500 randomly distributed impedance values $\mu = (\Re(Z), \Im(Z))$. This represents a significant computational effort (elapsed time $\approx$ 45min on 4 processors). We compute the Greedy RB approximation $u_N(\mu)$ for the same values of μ, at negligible computational costs (elapsed time $\approx 2\text{secs}$). In terms of computational performance, computing the RB solution is three orders of magnitude faster than computing the truth solution.

The effectivity index is defined as the ratio between the heuristic indicator and the relative error, *i.e.*,

$$\text{eff}(\mu) = \frac{\|A(\mu)u_N(\mu) - f\|_2/(\hat{\alpha}\|u_N(\mu)\|_2)}{\|u(\mu) - u_N(\mu)\|_2/\|u(\mu)\|_2}. \tag{7.2.6}$$

Figure 7.11 shows the effectivity index in the $(\Re(Z), \Im(Z))$ plane for the 500 values of μ. The distribution of the effectivity index is shown on fig. 7.12. We find the effectivity to be always greater than 1, which means that the heuristic indicator never underestimates the relative error. The relative error is overestimated by a factor comprised between 1.5 and 4, thus the heuristic indicator always catches the correct order of magnitude. We observe that the larger values of the effectivity index are located at impedance values Z with $\Im(Z) > 0$. This suggests that the inf-sup constant $\mu \mapsto \alpha(\mu)$ (of which $\hat{\alpha}$ is a rough estimation) has a dependency in μ.

In view to catching this effect, we have tried to incorporate a dependency in μ in the constant $\hat{\alpha}$ using radial basis function (RBF) interpolant as explained in section 2.2.4. Unfortunately, the matrix system to be solved for RBF interpolant was found to be singular. The reason for the non-invertibility of the matrix system to be solved for the RBF interpolant is the lack of information in the region $\Re(Z) > 0.5$. Indeed, the impedance values selected by the Greedy algorithm at the iterations $2, \ldots, N$ are all located on the axis $\Re(Z) = 0.5$ (see table 7.3), therefore the behavior in the region $\Re(Z) > 0.5$ can only be extrapolated and not interpolated.

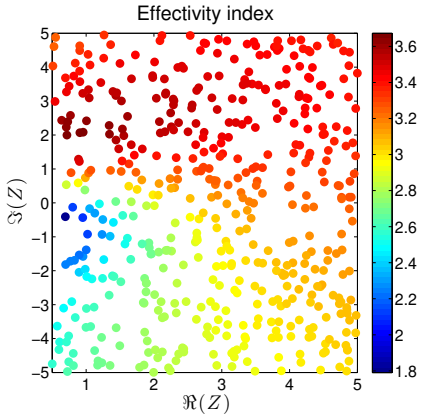


Figure 7.11: The effectivity index in the $(\Re(Z), \Im(Z))$ plane.

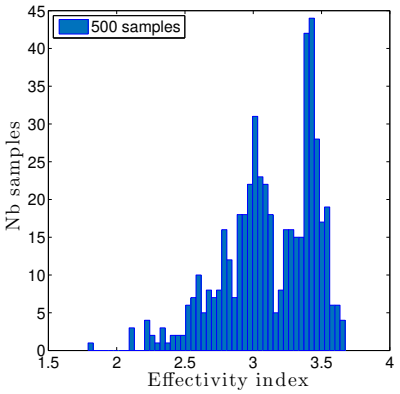


Figure 7.12: Distribution of the effectivity index.

7.3 The RBM applied to a 3D nacelle engine

7.3.1 Problem description

This application consists in optimizing the acoustic liners of an aircraft engine nacelle. We specifically consider the noise generated by the fan, thus our focus is on the acoustic liners mounted on the walls of the aircraft inlets (for a detailed description of liners mounted on aircraft nacelles, see [11]).

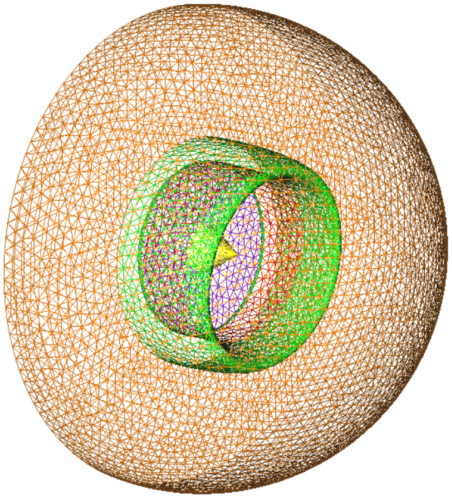


Figure 7.13: Mesh of the inlet of an aircraft engine nacelle.

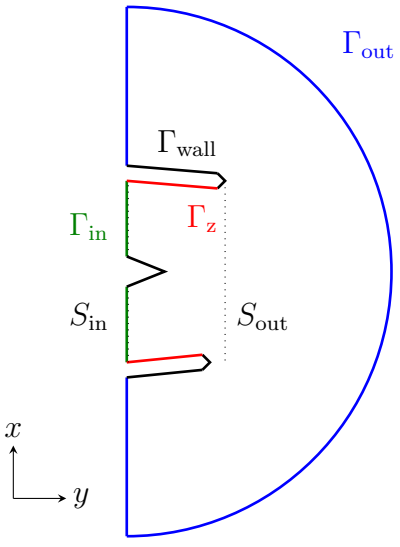


Figure 7.14: Schematic view showing boundary conditions and control surfaces.

We use the 3D geometry shown on figs. 7.13 and 7.14. Notice that there are no symmetry

planes in this geometry, because for design reasons the upper wall is slightly longer than the lower wall on which the liners are mounted (see fig. 7.14). This asymmetrical feature of aircraft inlets makes it necessary to perform the simulations in 3D. The frequency is set to $f = 587\text{Hz}$, which corresponds to a frequency at which peak noise attenuation is desired [63]. On the in-flow (inside the duct, where the fan is located see fig. 7.14), we impose a non-radial, 4th order azimuthal acoustic duct mode $(m, n) = (0, 4)$ [100, 92].

Taking into account the number of vertices in the mesh and a DG order $p = 2$, the overall number of degrees of freedom is $\mathcal{N} = 7,944,600$. For the best computational performance, the problem is distributed on 1024 processors, with each processor handling about $\sim 7,750$ degrees of freedom. The accuracy of the FETI-2LM solver is set so that the truth solution $u(\mu)$ satisfies the criterion $\|A(\mu)u(\mu) - f\|_2/\|f\|_2 < 10^{-5}$. In this situation, the computation of a truth solution for a given impedance value μ takes about $\sim 4\text{min}$, thus only about a dozen impedance values can be solved in an hour. In this context, the estimated cost of the attenuation intensity map for a 31×91 grid of impedance values is 188 hours – almost 8 days!

7.3.2 Reduced basis approach

We let the greedy algorithm run up to a RB of size $N = 8$, using a 25×25 grid in order to discretize the parameter set $\mathcal{D} = [-0.5,\ 5] \times [-5,\ 5]$. The selected impedance values are consigned in table 7.4. Similar to the ALIAS test case, we find that the greedy algorithm mostly selects impedance values with $\Re(Z) = 0.5$. The convergence curve showing the maximum residual norm throughout the greedy iterations is plotted on fig. 7.15. The decay of the maximum residual norm is clearly exponential, which confirms that the present reduced basis is relevant.

N	$\Re(Z)$	$\Im(Z)$
1	2.750	0.000
2	0.500	0.000
3	0.500	-2.916
4	0.500	-1.250
5	0.500	3.333
6	0.500	-0.833
7	0.500	0.833
8	0.500	-5.000

Table 7.4: Impedance values selected by the greedy algorithm applied to the inlet of the aircraft engine nacelle.

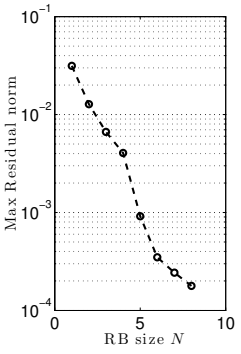


Figure 7.15: Maximum residual norm throughout the greedy algorithm applied to the inlet of the aircraft engine nacelle.

Figure 7.16 provides more details on the convergence history, showing the residual plotted

in the $(\Re(Z), \Im(Z))$ plane per greedy iteration. This confirms that the enrichment of the RB with a truth solution $u(\mu^\star)$ does not only improve the quality of approximation locally in the neighborhood of $\mu^\star$, but rather it improves the quality of approximation globally across parameter space. This is particularly spectacular upon adding the 5th truth solution at $\mu^\star = (0.5,\ 3.333)$ to the RB, where we find the residual norm to drop even in the two regions $\Re(Z) > 4$ and $\Im(Z) < -4$ which are very distant from $\mu^\star$. Note that this is also reflected in the convergence curve fig. 7.15, where the maximum residual norm drops from 4.05×10^{-3} at $N = 4$ to 9.14×10^{-4} at $N = 5$.

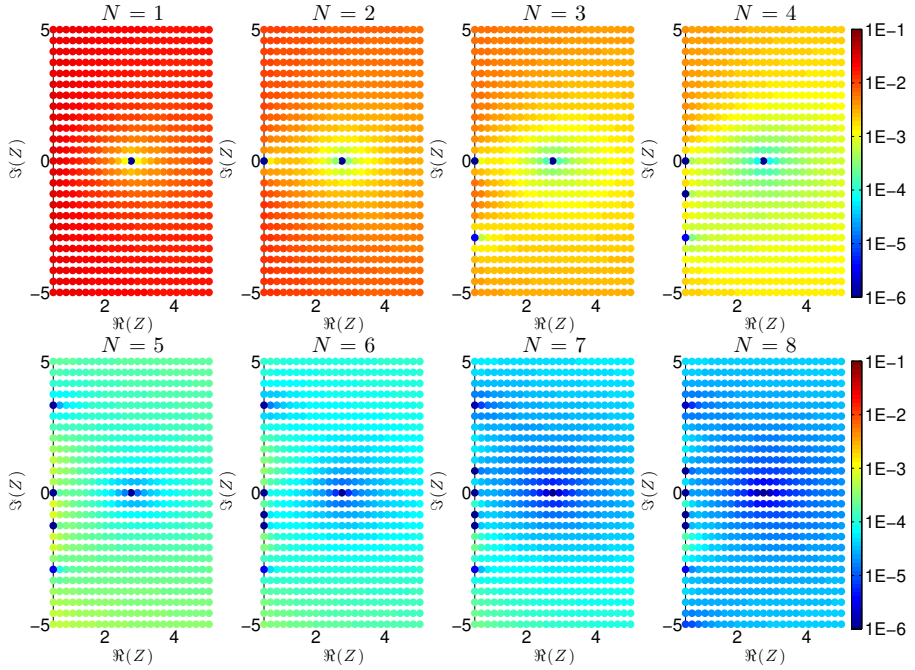


Figure 7.16: The RB residual norm in the $(\Re(Z), \Im(Z))$ plane for the aircraft engine nacelle with RBs of size $N = 1, \ldots, 8$.

Now that a RB of size $N = 8$ is at hand, we efficiently compute the attenuation intensity map for a 31×91 grid of impedance values (see fig. 7.17) and identify the optimal impedance $Z^\star = 1.1 + 0.111j$.

Figure 7.18 shows the pressure fields in the rigid case (*i.e.*, this corresponds to the absence of the liner, modeled by a rigid wall boundary condition) and in the presence of the optimal liner with impedance $Z^\star$. Clearly, the liner is able to drastically reduce the amplitude of the duct mode.

In terms of computational costs, the overall elapsed time is 38min on 1024 processors, taking into a account the offline phase during which the RB is constructed (with $N = 8$ high-fidelity solves) and the online phase during which the RB solver is evaluated $2,821$ times on a 31×91 impedance grid and associated attenuation intensities are computed. Note that, for the same budget (38min on 1024 processors), one would have been able to

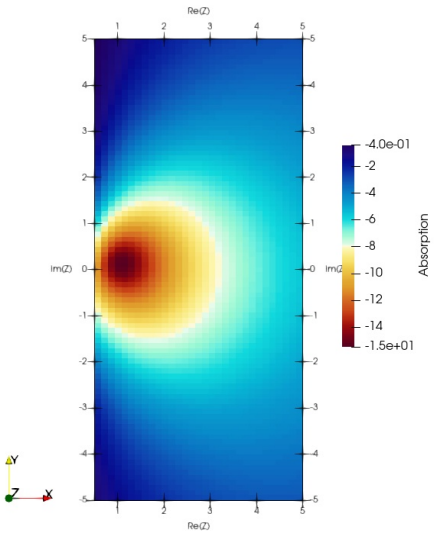


Figure 7.17: Attenuation intensity map for the aircraft engine nacelle, with greedy RB of size $N = 8$.

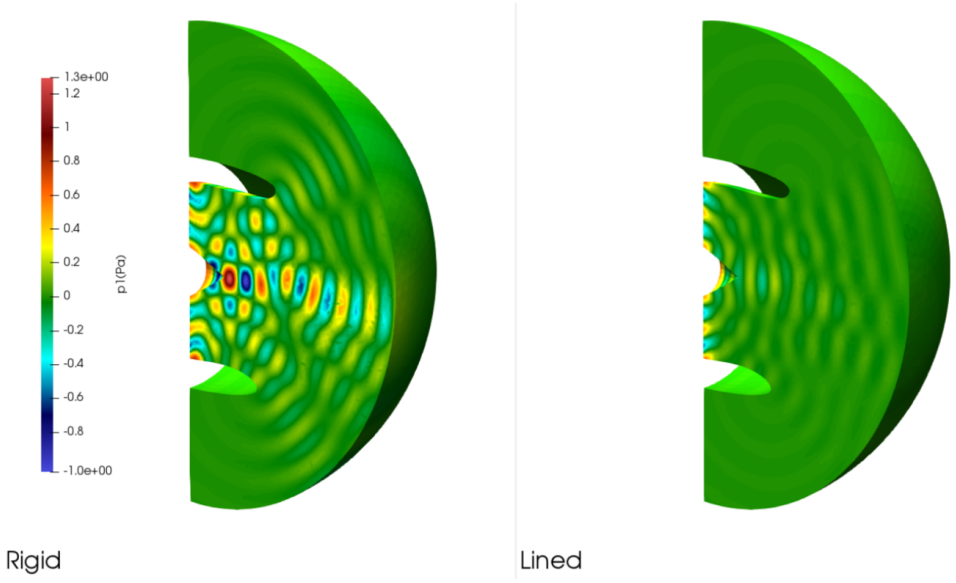


Figure 7.18: Pressure fields: rigid (wall BC) versus optimally lined (impedance BC with optimal impedance $Z^\star = 1.1 + 0.111j$)

evaluate only about ~ 10 truth solutions and associated attenuation intensities! Thus, the RBM provides a considerable speed-up.

7.3.3 Validation campaign

In opposition to the ALIAS test case, the computation of 500 truth solutions at 500 sampled impedance values is too computationally expensive (this would require about 2 days on a supercomputer with 1024 processors). For validation, we choose to compute only 5 truth solutions at chosen impedance values. Namely, we choose to compute the truth solutions at the optimal impedance value $Z^\star = 1.1 + 0.111j$ as well as 4 neighboring values. Table 7.5 shows the relative error between the truth and RB solution at these 5 impedance values, which is of the order of 0.06%. This takes into account the error on both velocity and pressure fields integrated over the whole computational domain.

Z	$\|u - u_N\|/\|u\|$	$\|Au_N - f\|/(\hat{\alpha}\|u_N\|)$	$\|C - C_N\|/\|C\|$
$1.1 + 0.111j$	6.94×10^{-4}	7.31×10^{-4}	1.13×10^{-4}
$0.8 + 0.333j$	6.89×10^{-4}	7.31×10^{-4}	2.61×10^{-4}
$0.8 - 0.111j$	7.82×10^{-4}	8.79×10^{-4}	1.70×10^{-4}
$1.4 + 0.333j$	5.42×10^{-4}	5.37×10^{-4}	1.21×10^{-4}
$1.4 - 0.111j$	5.91×10^{-4}	6.07×10^{-4}	7.81×10^{-5}

Table 7.5: Relative error on the reconstructed solution, heuristic error indicator and relative error on the attenuation intensity for 5 impedance values using a RB of size $N = 8$.

The difference between the pressure fields p_{DG} computed with the high-fidelity FETI-2LM solver and p_{RB} computed with the RB solver at the optimal impedance value is plotted on fig. 7.19. We have checked that the maximum amplitude of the error field $|p_{DG} - p_{RB}|$ does not exceed 0.08% of the maximum amplitude of the truth pressure field p_{DG} (compare the scales of fig. 7.19 and fig. 7.18). In opposition to the ALIAS test case, the error is not localized in the neighborhood of the liner as we find contributors to the error all across the flow.

The second column of table 7.5 gives the estimated relative error using the heuristic error indicator. During the greedy algorithm, 7 ratios between residual and error have been sampled. Computing the mean and variation coefficient, we find $\hat{\alpha} = 0.08022 \pm 24.9\%$. Here, the situation is similar to the ALIAS test case (see section 7.2.3), where the RBF interpolant cannot be computed because all the selected values of impedance have the same real part, consequently the only available information is the mean value $\hat{\alpha}$. Still, we find excellent agreement between the heuristic error indicator and the exact relative error with effectivities very close to 1.

Finally, let us compare the truth attenuation intensity $C(Z)$ (obtained by post-processing the truth solution) to the RB attenuation intensity $C_N(Z)$ (obtained by post-processing the RB approximation). The relative errors on the attenuation intensity at 5 impedance values are given in table 7.5. We find relative errors of about 0.02%. Thus, the RB attenuation intensity is very accurate, in fact, more accurate than the RB pressure and velocity fields. We suggest that this may be due to the fact that the attenuation intensity, as defined by eq. (7.2.2), is obtained by integrating the acoustic intensity field over the two

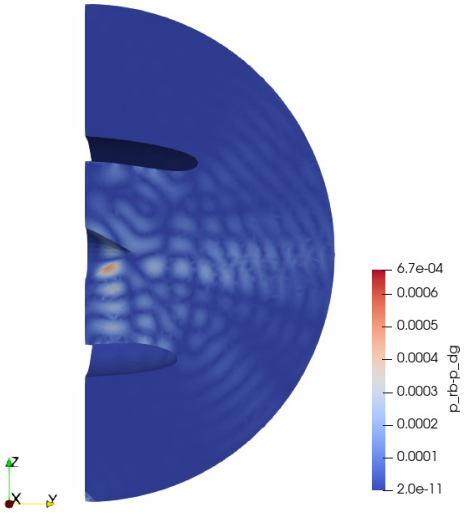


Figure 7.19: Error $|p_{DG} - p_{RB}|$ in the pressure field using a RB of size $N = 8$ at the optimal impedance $Z^\star = 1.1 + 0.111j$.

control surfaces S_{in} and S_{out} (see fig. 7.14). As can be seen on fig. 7.19, the maximum errors on the pressure field are not nearby these two control surfaces, thus the maximum errors are "invisible" (*i.e.*, they are not taken into account) in the process of calculating the attenuation intensity.

7.4 Conclusions

In this chapter, we have applied the RBM to the time-harmonic linearized Euler equations with parametrized impedance boundary condition. Thanks to the RBM, we have been able to obtain the acoustic pressure and velocity fields at a multitude of impedance values at very low costs and with good accuracy. Thus, the attenuation intensity map could be efficiently computed and the optimal impedance value could be rapidly and reliably identified. Furthermore, the accuracy of the attenuation intensity was even better than that of the reconstructed pressure and velocity fields due to integration over small parts of the computational domain.

This strategy has been illustrated on the ALIAS model problem with roughly $100,000$ degrees of freedom and on the aircraft engine nacelle with about $8,000,000$ degrees of freedom. Both applications take advantage of massively parallel supercomputers. In this context, the RBM proves to be very successful and enables significant speed-ups, yielding the results of a simulation campaign of 8 days in less than an hour!

The current implementation only deals with the time-harmonic linearized Euler equations, which is quite restrictive when it comes to industrial applications (this model is only relevant for the study of the noise generated by the fan, which is the application considered in this chapter). A perspective (far beyond the scope of this thesis) would be to generalize the RB approach to the non-linear, time-dependent Euler equations, in order to able to study and optimize the jet noise in aircraft engine nacelles. A shorter term perspective would be to incorporate the frequency as an additional parameter in the time-harmonic LEEs, thus the parameter would be a 3-dimensional parameter and we would be able to optimize the impedance for a frequency band, rather than for a fixed frequency. In this perspective, the acoustic duct mode, which is frequency-dependent, would have to be approximated using EIM in order to maintain the efficient offline/online decoupling of the reduced basis method.

Conclusions and perspectives

The purpose of this work

In computational electromagnetism, many applications require repeated numerical solutions of the time-harmonic Maxwell's equations over a frequency band. Similarly, in computational aeroacoustics, the optimization of a liner requires repeated numerical solutions of the time-harmonic linearized Euler equations for a vast set of impedance values. In both fields, there is a great interest in numerical methods for rapidly and reliably computing the solution of a PDE for multiple parameter queries.

Model order reduction techniques are particularly well suited for this purpose. In this thesis, we specifically considered the reduced basis method, which approximates the PDE solution at any parameter query using a linear combination of a small number of high-fidelity solutions computed for a small set of well-chosen parameter values. We aimed at defining the most efficient reduced basis strategies: (i) in terms of computational performance and (ii) with guaranteed bounds on the reduced basis approximation error.

Our contributions

A heuristic error estimation strategy. The traditional error estimation approach relies on lower bounds for the inf-sup stability constant. When the inf-sup stability constant has minor or very smooth dependency on the parameter, we proposed to simply replace the costly lower bounds for the inf-sup stability constant by an inexpensive quantity computed during the greedy iterations using the ratios between residual and error. We have shown throughout this thesis that this heuristic method provides excellent results for unconditionally stable time-harmonic problems.

Dual natural-norm error estimators. The traditional error estimator based on the inf-sup stability constant is well-known to significantly overestimate the error when the inf-

sup stability constant is very small, which typically occurs close to the resonant parameters in close-to-degenerate problems. We introduced a dual-natural norm and derived dual natural-norm error estimators characterized by a $\mathcal{O}(1)$ stability constant, thus very effective. We illustrated our approach on a Helmholtz problem exhibiting resonant behaviors and significantly improved the traditional inf-sup-based error estimator.

The reduced basis method with multiple sources. In chapter 4, we discussed the adaptation of the reduced basis method for applications featuring multiple sources. We proposed to enrich the reduced basis not with just one basis function, but rather with a block of basis functions associated to the first eigenvectors of the residual operator. This strategy could reduce the overall number of operator factorizations required for constructing the reduced basis.

A non-intrusive reduced basis method for frequency-sweeps with surface integral equations. In chapter 6, we presented an original reduced basis strategy for efficiently solving scattering problems in the context of multiple frequency queries. Our contribution includes the use of domain decomposition of the frequency window for mitigating the overall costs of the reduced basis method. The proposed method is non-intrusive and thus easily compatible with state-of-the-art methods such as the fast-multipole method. The potential of our method is demonstrated on numerous numerical examples.

Industrial applications. This thesis provides real-world industrial applications of the reduced basis method employed with various discretization techniques (edge finite elements, boundary elements and the discontinuous-Galerkin method). In our antenna and aeroacoustic applications, we tackled problems featuring millions of degrees of freedom using domain decomposition techniques in order to make the best use of parallel computational resources. In our electromagnetic scattering application, we went far beyond the academic example of the sphere, showing real-world examples on aircraft geometries and with matrix-vector products accelerated with the state-of-the-art fast-multiple method. Our results illustrate the benefit of the reduced basis method on large-scale configurations.

Perspectives of this work

We now discuss the remaining open questions and outline future work from both theoretical and algorithmic points of view.

Certifying the heuristic. We proposed to sample the ratios between residual and error throughout the greedy iterations in order to build a heuristic error indicator. In some

situations, this method proved to be relevant. So far, we have no *a priori* knowledge on the potential success of this heuristic. It would be nice to find easy-to-verify hypotheses under which the success of this method could be guaranteed.

Computational costs associated with stability constants. Rigorous certification of the reduced basis approximation traditionally relies on the computation of lower bounds for some parameter-dependent stability constants, *i.e.*, either the inf-sup constant or the primal/dual natural-norm constant. As shown in this thesis, the systematic use of the SCM for approximating these lower bounds is exceeding time and resource-consuming due to repeated solutions of large-scale generalized eigenvalue problems. The hierarchical or randomized error estimation approaches [51, 114], which do not rely on any stability constants, could be further investigated to avoid the heavy computational costs associated with the approximation of stability constants.

Localized reduced basis method for broadband applications. The use of localized reduced basis approximation spaces is a logical next step to be able to address large bandwidths in our frequency-parametrized applications. Rather than constantly increasing the basis size, which deteriorates both offline and online performances, the user would be able to prescribe a desired maximal basis size [72]. We expect this strategy to be particularly efficient for frequency-parametrized surface integral equations. Indeed, this would limit the number of right-hand sides per matrix-vector product, which is one of the main bottlenecks due to the fully-populated nature of integral operators even when efficient acceleration methods such as FMM are used.

More parameters and more sources. In real-world radar applications, the frequency is usually not the only parameter that varies. The development of stealth technologies for example typically requires to illuminate the scattering object with plane waves in all possible directions. Using the ideas of chapter 4, future work will be to incorporate the direction of the plane wave in a block formulation and to solve a frequency-parametrized block problem. This would enable an efficient computation of the radar cross section over a frequency band and for plane waves propagating in multiple directions. In our aeroacoustic applications, future work will consist in adding the frequency as an additional parameter in order to be able to optimize the fan noise over a frequency band, rather than at a fixed frequency.

Concluding remarks

Many applications in aeroacoustics and electromagnetism require the solutions of a linear PDE for a vast set of parameter values. We have shown in this work that the reduced basis

method is well suited for these complex industrial applications as it significantly reduces computational times while maintaining a very high level of accuracy.

The contributions presented in this thesis extend the reduced basis method for linear parametrized linear equations, with the potential to rapidly and reliably solve large-scale, time-harmonic problems for multiple sources and with varying parameters for various applications in the field of acoustic and electromagnetic wave propagation.

Appendix A

Methods for solving large-scale generalized eigenvalue problems

A.1 Context

In this appendix, we present some algorithms to solve the smallest eigenvalue in the generalized eigenvalue problem

$$\begin{cases} \textit{Find } (\lambda, \mathbf{v}) \in \mathbb{R} \times \mathbb{C}^{\mathcal{N}} \setminus \{0\} \textit{ such that} \\ \mathbf{H}\mathbf{v} = \lambda \mathbf{X}\mathbf{v}, \end{cases} \tag{G.E.P.}$$

where $\mathbf{H}$, $\mathbf{X}$ are hermitian matrices and $\mathbf{X}$ is positive definite. We focus on iterative methods, which are of particular interest in the context of large-scale eigenvalue problems, since they only require the ability to perform system solves and/or matrix-vector products with the matrices $\mathbf{H}$, $\mathbf{X}$.

Remark. *In the situation where* $\mathbf{H} = \mathbf{A}^*\mathbf{B}^{-1}\mathbf{A}$ *(see Chapter 2), the matrix* $\mathbf{H}$ *should of course never be assembled: each matrix-vector product* $\mathbf{u} \leftarrow \mathbf{H}\mathbf{v}$ *should consist in*

1. *computing the matrix-vector product* $\mathbf{u} \leftarrow \mathbf{A}\mathbf{v}$*;*
2. *solving the problem* $\mathbf{B}\mathbf{y} = \mathbf{u}$ *for* $\mathbf{y}$*;*
3. *and finally computing the matrix-vector product* $\mathbf{u} \leftarrow \mathbf{A}^*\mathbf{y}$.

Similarly, to solve $\mathbf{H}\mathbf{v} = \mathbf{u}$ *for* $\mathbf{v}$*, one should*

1. *solve the adjoint problem* $\mathbf{A}^*\mathbf{v} = \mathbf{u}$ *for* $\mathbf{v}$*;*
2. *compute the matrix-vector product* $\mathbf{y} \leftarrow \mathbf{B}\mathbf{v}$*;*
3. *and finally solve* $\mathbf{A}\mathbf{v} = \mathbf{y}$ *for* $\mathbf{v}$.

A.2 Inverse iteration

The first method consists in a power iteration method with the inverse operator $\mathbf{H}^{-1}$. The computational procedure is summarized by algorithm A.1. It converges to the smallest eigenpair $(\lambda^{(1)}, \mathbf{v}^{(1)})$ unless the starting vector $\mathbf{v}_0 \in \mathbb{C}^{\mathcal{N}}$ verifies $\mathbf{v}_0^* \mathbf{X} \mathbf{v}^{(1)} = 0$, which does not happen in practice when considering a random starting vector [90].

Let us rapidly discuss the computational complexity. At each iteration, there is one call to the solver $\mathbf{H}^{-1}$ and two calls to the matrix-vector product with $\mathbf{X}$ (the product $\mathbf{X}\mathbf{v}_{k-1}$ and the product $\mathbf{X}\mathbf{w}_k$ that serve to form $\|\mathbf{w}_k\|_{\mathbf{X}}$ and to form $\mathbf{X}\mathbf{v}_k$ required to compute ϵ_{k+1}).

Algorithm A.1: Generalized inverse iteration algorithm (adapted from [90], Chapter 4)

input : Starting vector $\mathbf{v}_0 \in \mathbb{C}^{\mathcal{N}}$, tolerance tol and maximum number of iterations $k_{\max}$
output: Approximate smallest eigenpair $(\lambda_k, \mathbf{v}_k)$

Initialize $k \leftarrow 1$, $\epsilon_1 \leftarrow \infty$;
while $k \leq k_{\max}$ *and* $\epsilon_k > tol$ **do**
  Solve $\mathbf{H}\mathbf{w}_k = \mathbf{X}\mathbf{v}_{k-1}$ for $\mathbf{w}_k$;
  $\mathbf{v}_k \leftarrow \mathbf{w}_k / \|\mathbf{w}_k\|_{\mathbf{X}}$;
  $\lambda_k \leftarrow \mathbf{v}_k^* \mathbf{X} \mathbf{v}_{k-1} / \|\mathbf{w}_k\|_{\mathbf{X}}$;
  $\epsilon_{k+1} \leftarrow \|\mathbf{v}_{k-1} - \mathbf{v}_k\|_{\mathbf{X}}$;
  $k \leftarrow k + 1$;
end

A.3 Lanczos method

The second method, the Lanczos method, relies on the construction of a Krylov subspace, either

$$\mathcal{K}^m(\mathbf{H}, \mathbf{v}_0) \equiv \mathrm{Span}\{\mathbf{v}_0, \mathbf{H}\mathbf{v}_0, \ldots, \mathbf{H}^{m-1}\mathbf{v}_0\}$$

for direct Lanczos, or $\mathcal{K}^m(\mathbf{H}^{-1}, \mathbf{v}_0)$ for inverse Lanczos, m being the number of Lanczos iterations and $\mathbf{v}_0 \in \mathbb{C}^{\mathcal{N}}$ some starting vector. This method converges to m eigenpairs $(\lambda^{(j)}, \mathbf{y}^{(j)})$, $j \in \{1, \ldots, \mathcal{N}\}$. In practice, the extreme eigenpairs (*i.e.*, smallest and largest) are first to converge [90].

The computational procedure of the direct Lanczos is summarized by algorithm A.2. The output is an approximation for the smallest eigenvalue. If one is interested in computing an approximation of the associated generalized eigenvector $\mathbf{v}^{(1)}$, with normalization $\|\mathbf{v}^{(1)}\|_{\mathbf{X}} = 1$, one can proceed by computing the Ritz vector $\mathbf{v}_k$ as

$$\mathbf{v}_k = \mathbf{Q}_k \mathbf{s}_k^{(1)},$$

where $\mathbf{Q}_k = [\mathbf{q}_1\, \mathbf{q}_2\, \cdots \mathbf{q}_k] \in \mathbb{C}^{\mathcal{N}\times k}$ and $\mathbf{s}_k^{(1)} \in \mathbb{C}^k$ denotes the 1st column of the matrix $\mathbf{S}_k$. The matrix $\mathbf{Q}_k$ holds the $\mathbf{X}$-orthogonal basis (*i.e.*, such that $\mathbf{Q}_k^*\mathbf{X}\mathbf{Q}_k = \mathbf{I}$) of the Krylov subspace $\mathcal{K}^k(\mathbf{H}, \mathbf{u}_1)$, while $\mathbf{s}_k^{(1)}$ is the eigenvector of $\mathbf{T}_k$ associated to the smallest Ritz value $\theta_k^{(1)}$, *i.e.*, $\mathbf{T}_k\mathbf{s}_k^{(1)} = \theta_k^{(1)}\mathbf{s}_k^{(1)}$, with normalization $\|\mathbf{s}_k^{(1)}\|_2 = 1$. Thus defined, $\mathbf{v}_k \to \mathbf{v}^{(1)}$ as $k \to \infty$. One can also compute an approximation for the second smallest eigenpair $(\lambda^{(2)}, \mathbf{v}^{(2)})$, by considering $(\theta_k^{(2)}, \mathbf{Q}_k\mathbf{s}_k^{(2)})$, and so on.

Algorithm A.2: Generalized Lanczos algorithm (adapted from [90], Chapter 15)

input : Starting vector $\mathbf{u}_1 \in \mathbb{C}^{\mathcal{N}}$, tolerance tol and maximum number of iterations $k_{\max}$
output: Approximate smallest eigenvalue λ_k

Initialize $\mathbf{p}_0 \leftarrow 0$, $k \leftarrow 1$, $\epsilon_1 \leftarrow \infty$;
$\mathbf{r}_1 \leftarrow \mathbf{X}\mathbf{u}_1$,;
$\beta_1 \leftarrow \sqrt{\mathbf{r}_1^*\mathbf{u}_1}$;
while $k \leq k_{\max}$ *and* $\epsilon_k > tol$ **do**
  $\mathbf{q}_k \leftarrow \mathbf{u}_k/\beta_k$;
  $\underline{\mathbf{u}}_k \leftarrow \mathbf{H}\mathbf{q}_k - \mathbf{p}_{k-1}\beta_k$;
  $\alpha_k \leftarrow \mathbf{q}_k^*\underline{\mathbf{u}}_k$;
  $\mathbf{p}_k \leftarrow \mathbf{r}_k/\beta_k$;
  $\mathbf{r}_{k+1} \leftarrow \underline{\mathbf{u}}_k - \mathbf{p}_k\alpha_k$;
  Solve $\mathbf{X}\mathbf{u}_{k+1} = \mathbf{r}_{k+1}$ for $\mathbf{u}_{k+1}$;
  $\beta_{k+1} \leftarrow \sqrt{\mathbf{u}_{k+1}^*\mathbf{r}_{k+1}}$;
  Diagonalize the tridiagonal matrix $\mathbf{T}_k = \begin{pmatrix} \alpha_1 & \beta_2 & & & \\ \beta_2 & \alpha_2 & \beta_3 & & \\ & \beta_3 & \ddots & \ddots & \\ & & \ddots & \ddots & \beta_k \\ & & & \beta_k & \alpha_k \end{pmatrix}$ in orthonormal basis as $\mathbf{T}_k = \mathbf{S}_k\mathbf{\Theta}_k\mathbf{S}_k^*$ with $\mathbf{\Theta}_k = \mathrm{diag}(\theta_k^{(1)}, \ldots, \theta_k^{(k)})$ ordered as $\theta_k^{(1)} \leq \theta_k^{(2)} \leq \cdots \leq \theta_k^{(k)}$;
  Update current smallest eigenvalue approximation $\lambda_k \leftarrow \theta_k^{(1)}$;
  Update current error $\epsilon_{k+1} \leftarrow \beta_{k+1}|s_{k1}|$, where s_{k1} corresponds to the coefficient on the 1st column, kth row of $\mathbf{S}_k$;
  $k \leftarrow k + 1$;
end

Let us now turn to the inverse Lanczos procedure, summarized by algorithm A.3. There are some minor, nonetheless significant, changes compared to the direct Lanczos. If one is interested in computing an approximation of the smallest generalized eigenvector $\mathbf{v}^{(1)}$, one can compute the Ritz vector $\mathbf{w}_k$, as $\mathbf{w}_k = \mathbf{Q}_k\mathbf{s}_k^{(k)}$, where $\mathbf{Q}_k = [\mathbf{q}_1\, \mathbf{q}_2\, \cdots \mathbf{q}_k] \in \mathbb{C}^{\mathcal{N}\times k}$ and $\mathbf{s}_k^{(k)} \in \mathbb{C}^k$ denotes the kth column of the matrix $\mathbf{S}_k$. At this point, one must be cautious because, being generated by an inverse Lanczos procedure, the matrix $\mathbf{Q}_k$ now holds a $\mathbf{X}^{-1}$-orthogonal basis (*i.e.*, such that $\mathbf{Q}_k^*\mathbf{X}^{-1}\mathbf{Q}_k = \mathbf{I}$) of the Krylov subspace

Algorithm A.3: Generalized inverse Lanczos algorithm (adapted from [90], Chapter 15)

input : Starting vector $\mathbf{r}_1 \in \mathbb{C}^{\mathcal{N}}$, tolerance tol and maximum number of iterations $k_{\max}$
output: Approximate smallest eigenvalue λ_k

Initialize $\mathbf{p}_0 \leftarrow 0$, $k \leftarrow 1$, $\epsilon_1 \leftarrow \infty$, $\beta_1 \leftarrow 1$;
$\mathbf{u}_1 \leftarrow \mathbf{X}\mathbf{r}_1$,;
$\tau \leftarrow \sqrt{\mathbf{r}_1^* \mathbf{u}_1}$;
$\mathbf{r}_1 \leftarrow \mathbf{r}_1/\tau$;
$\mathbf{u}_1 \leftarrow \mathbf{u}_1/\tau$;
while $k \leq k_{\max}$ *and* $\epsilon_k > tol$ **do**
  $\mathbf{q}_k \leftarrow \mathbf{u}_k/\beta_k$;
  Solve $\mathbf{H}\underline{\mathbf{u}}_k = \mathbf{q}_k$ for $\underline{\mathbf{u}}_k$;
  $\underline{\mathbf{u}}_k \leftarrow \underline{\mathbf{u}}_k - \mathbf{p}_{k-1}\beta_k$;
  $\alpha_k \leftarrow \mathbf{q}_k^* \underline{\mathbf{u}}_k$;
  $\mathbf{p}_k \leftarrow \mathbf{r}_k/\beta_k$;
  $\mathbf{r}_{k+1} \leftarrow \underline{\mathbf{u}}_k - \mathbf{p}_k\alpha_k$;
  $\mathbf{u}_{k+1} \leftarrow \mathbf{X}\mathbf{r}_{k+1}$;
  $\beta_{k+1} \leftarrow \sqrt{\mathbf{u}_{k+1}^* \mathbf{r}_{k+1}}$;
  Diagonalize the tridiagonal matrix $\mathbf{T}_k = \begin{pmatrix} \alpha_1 & \beta_2 & & & \\ \beta_2 & \alpha_2 & \beta_3 & & \\ & \beta_3 & \ddots & \ddots & \\ & & \ddots & \ddots & \beta_k \\ & & & \beta_k & \alpha_k \end{pmatrix}$ in orthonormal basis as $\mathbf{T}_k = \mathbf{S}_k \mathbf{\Theta}_k \mathbf{S}_k^*$ with $\mathbf{\Theta}_k = \mathrm{diag}(\theta_k^{(1)}, \ldots, \theta_k^{(k)})$ ordered as $\theta_k^{(1)} \leq \theta_k^{(2)} \leq \cdots \leq \theta_k^{(k)}$;
  Update current smallest eigenvalue approximation $\lambda_k \leftarrow 1/\theta_k^{(k)}$;
  Update current error $\epsilon_{k+1} \leftarrow \beta_{k+1}|s_{kk}|$, where s_{kk} corresponds to the coefficient on the kth column, kth row of $\mathbf{S}_k$;
  $k \leftarrow k + 1$;
end

$\mathcal{K}^k(\mathbf{H}^{-1}, \mathbf{u}_1)$. Thus, the Ritz vector $\mathbf{w}_k$ is in fact an approximation of the generalized eigenvector $\mathbf{w}$, with normalization $\|\mathbf{w}\|_{\mathbf{X}^{-1}} = 1$, associated to the *largest* eigenvalue in

$$\mathbf{H}^{-1}\mathbf{w} = \lambda \mathbf{X}^{-1}\mathbf{w}.$$

Let us introduce the new variable $\mathbf{v} = \mathbf{X}^{-1}\mathbf{w}$. The normalization $\|\mathbf{v}\|_{\mathbf{X}} = 1$ is straightforward from the normalization of $\mathbf{w}$. Furthermore we can easily show that $\mathbf{v}$ is the generalized eigenvector associated to the *smallest* eigenvalue in

$$\mathbf{H}\mathbf{v} = \lambda \mathbf{X}\mathbf{v},$$

thus $\mathbf{v} = \mathbf{v}^{(1)}$. To conclude, one can approximate the generalized eigenvector $\mathbf{v}^{(1)}$ by the quantity $\mathbf{v}_k = \mathbf{X}^{-1}\mathbf{w}_k = \mathbf{X}^{-1}\mathbf{Q}_k\mathbf{s}_k^{(k)}$. This strategy can be extended to obtain an

approximation for the second smallest eigenpair $(\lambda^{(2)}, \mathbf{v}^{(2)})$, by $(1/\theta_k^{(k-1)}, \mathbf{X}^{-1}\mathbf{Q}_k\mathbf{s}_k^{(k-1)})$, where $\mathbf{s}_k^{(k-1)} \in \mathbb{C}^k$ denotes the $(k-1)$th column of the matrix $\mathbf{S}_k$, and so on.

Let us briefly discuss the computational complexity of both direct and inverse Lanczos procedures. In the case of direct Lanczos, there is one call to the matrix-vector product with $\mathbf{H}$ and one call to the solver $\mathbf{X}^{-1}$ per iteration, while in the case of inverse Lanczos there is one call to the matrix-vector product with $\mathbf{X}$ and one call to the solver $\mathbf{H}^{-1}$ per iteration. For both, there is one call to a symmetric tridiagonal eigensolver per iteration. The latter operation can be made very efficient with dedicated algorithms, such as the QR (or QL) algorithm [90] and is usually not a large-scale operation, as the symmetric tridiagonal matrix involved is only of size $k \times k$ at iteration k and k remains relatively small.

Appendix B

Implementation details

B.1 Four possible offline phases

The offline phase of the RBM depends on the choice of approximation (either Galerkin or least-squares approximation, see chapter 1) and on the choice of numerical method for computing the residual norm (either default or stabilized, see chapter 2).

	Galerkin	Least-squares
Default	for all $1 \leqslant q \leqslant Q^a$, the matrix $\mathbf{P}^* \mathbf{A}_q \mathbf{P} \in \mathbb{C}^{N \times N}$ for all $1 \leqslant q \leqslant Q^f$, the vector $\mathbf{P}^* \mathbf{f}_q \in \mathbb{C}^N$ for all $1 \leqslant p, q \leqslant Q^a$, the matrix $\mathbf{P}^* \mathbf{A}_p^* \mathbf{B}_W^{-1} \mathbf{A}_q \mathbf{P} \in \mathbb{C}^{N \times N}$ for all $1 \leqslant q \leqslant Q^f$, $1 \leqslant p \leqslant Q^a$, the vector $\mathbf{P}^* \mathbf{A}_p^* \mathbf{B}_W^{-1} \mathbf{f}_q \in \mathbb{C}^N$ for $1 \leqslant q, p \leqslant Q^f$, the scalars $\mathbf{f}_q^* \mathbf{B}_W^{-1} \mathbf{f}_p \in \mathbb{C}$	for all $1 \leqslant p, q \leqslant Q^a$, the matrix $\mathbf{P}^* \mathbf{A}_p^* \mathbf{B}_W^{-1} \mathbf{A}_q \mathbf{P} \in \mathbb{C}^{N \times N}$ for all $1 \leqslant q \leqslant Q^f$, $1 \leqslant p \leqslant Q^a$, the vector $\mathbf{P}^* \mathbf{A}_p^* \mathbf{B}_W^{-1} \mathbf{f}_q \in \mathbb{C}^N$ for $1 \leqslant q, p \leqslant Q^f$, the scalars $\mathbf{f}_q^* \mathbf{B}_W^{-1} \mathbf{f}_p \in \mathbb{C}$
Stabilized	for all $1 \leqslant q \leqslant Q^a$, the matrix $\mathbf{P}^* \mathbf{A}_q \mathbf{P} \in \mathbb{C}^{N \times N}$ for all $1 \leqslant q \leqslant Q^f$, the vector $\mathbf{P}^* \mathbf{f}_q \in \mathbb{C}^N$ the upper triangular matrix $\mathbf{R} \in \mathbb{C}^{NQ^a \times NQ^a}$ for all $1 \leqslant q \leqslant Q^f$ the vector $\mathbf{Q}^* \mathbf{f}_q \in \mathbb{C}^{NQ^a}$ for all $1 \leqslant q, p \leqslant Q^f$ the scalar $(\mathbf{B}_W^{-1} \mathbf{f}_k - \mathbf{Q}\mathbf{Q}^* \mathbf{f}_k)^* \mathbf{B}_W (\mathbf{B}_W^{-1} \mathbf{f}_q - \mathbf{Q}\mathbf{Q}^* \mathbf{f}_q) \in \mathbb{C}$	the upper triangular matrix $\mathbf{R} \in \mathbb{C}^{NQ^a \times NQ^a}$ for all $1 \leqslant q \leqslant Q^f$ the vector $\mathbf{Q}^* \mathbf{f}_q \in \mathbb{C}^{NQ^a}$ for all $1 \leqslant q, p \leqslant Q^f$ the scalar $(\mathbf{B}_W^{-1} \mathbf{f}_k - \mathbf{Q}\mathbf{Q}^* \mathbf{f}_k)^* \mathbf{B}_W (\mathbf{B}_W^{-1} \mathbf{f}_q - \mathbf{Q}\mathbf{Q}^* \mathbf{f}_q) \in \mathbb{C}$

Table B.1: Quantities to be computed during the offline phase in four case scenarios.

Using the notations of chapters 1 and 2, table B.1 summarizes the quantities to be computed during the offline phase with different choices of approximation and numerical method for computing the residual norm. We observe that the least-squares approximation is associated with less offline pre-computed quantities than the Galerkin approximation.

B.2 Reduced basis updates throughout the greedy iterations

During the Nth iteration of algorithm 2.1, a reduced basis with $N-1$ basis functions $\mathbf{p}_1, \ldots, \mathbf{p}_{N-1}$ is available and a Nth basis function $\mathbf{p}_N \in \mathbb{C}^{\mathcal{N}}$ is to be added.

If a Galerkin projection is used, the matrices $\mathbf{P}^* \mathbf{A}_q \mathbf{P} \in \mathbb{C}^{N \times N}$, $1 \leqslant q \leqslant Q^a$ have to be computed, where $\mathbf{P} = [\mathbf{p}_1 | \cdots | \mathbf{p}_N] \in \mathbb{C}^{\mathcal{N} \times N}$ denotes the reduced basis. In order to compute these matrices, one observes that for $1 \leqslant q \leqslant Q^a$,

$$\mathbf{P}^* \mathbf{A}_q \mathbf{P} = \left(\begin{array}{ccc|c} \mathbf{p}_1^* \mathbf{A}_q \mathbf{p}_1 & \cdots & \mathbf{p}_1^* \mathbf{A}_q \mathbf{p}_{N-1} & \mathbf{p}_1^* \mathbf{A}_q \mathbf{p}_N \\ \vdots & & \vdots & \vdots \\ \mathbf{p}_{N-1}^* \mathbf{A}_q \mathbf{p}_1 & \cdots & \mathbf{p}_{N-1}^* \mathbf{A}_q \mathbf{p}_{N-1} & \mathbf{p}_{N-1}^* \mathbf{A}_q \mathbf{p}_N \\ \hline \mathbf{p}_N^* \mathbf{A}_q \mathbf{p}_1 & \cdots & \mathbf{p}_N^* \mathbf{A}_q \mathbf{p}_{N-1} & \mathbf{p}_N^* \mathbf{A}_q \mathbf{p}_N \end{array} \right). \tag{B.2.1}$$

The upper left block of size $(N-1) \times (N-1)$ is already available from the previous greedy iteration, thus only the last row and last column have to be computed.

In order to compute the last column, one needs to perform the matrix-vector product $\mathbf{A}_q \mathbf{p}_N$, followed by N dots products. We propose to store this matrix-vector product in memory for future use. Thus, in order to compute the last row of $\mathbf{P}^* \mathbf{A}_q \mathbf{P}$, one does not need to compute matrix-vector products $\mathbf{A}_q \mathbf{p}_1, \ldots, \mathbf{A}_q \mathbf{p}_{N-1}$ because these are already available in memory from the previous greedy iterations. In this way, the last row of $\mathbf{P}^* \mathbf{A}_q \mathbf{P}$ can be computed in just $N-1$ dot products.

This strategy for computing the matrices $\mathbf{P}^* \mathbf{A}_q \mathbf{P} \in \mathbb{C}^{N \times N}$, $1 \leqslant q \leqslant Q^a$ can be readily extended to other quantities to be computed during the offline phase. For instance, in order to compute the matrices $\mathbf{P}^* \mathbf{A}_p^* \mathbf{B}_W^{-1} \mathbf{A}_q \mathbf{P} \in \mathbb{C}^{N \times N}$ for $1 \leqslant q, p \leqslant Q^a$, $p \geqslant q$, one observes that

$$\mathbf{P}^* \mathbf{A}_p^* \mathbf{B}_W^{-1} \mathbf{A}_q \mathbf{P} = \left(\begin{array}{ccc|c} \mathbf{p}_1^* \mathbf{A}_p^* \mathbf{B}_W^{-1} \mathbf{A}_q \mathbf{p}_1 & \cdots & \mathbf{p}_1^* \mathbf{A}_p^* \mathbf{B}_W^{-1} \mathbf{A}_q \mathbf{p}_{N-1} & \mathbf{p}_1^* \mathbf{A}_p^* \mathbf{B}_W^{-1} \mathbf{A}_q \mathbf{p}_N \\ \vdots & & \vdots & \vdots \\ \mathbf{p}_{N-1}^* \mathbf{A}_p^* \mathbf{B}_W^{-1} \mathbf{A}_q \mathbf{p}_1 & \cdots & \mathbf{p}_{N-1}^* \mathbf{A}_p^* \mathbf{B}_W^{-1} \mathbf{A}_q \mathbf{p}_{N-1} & \mathbf{p}_{N-1}^* \mathbf{A}_p^* \mathbf{B}_W^{-1} \mathbf{A}_q \mathbf{p}_N \\ \hline \mathbf{p}_N^* \mathbf{A}_p^* \mathbf{B}_W^{-1} \mathbf{A}_q \mathbf{p}_1 & \cdots & \mathbf{p}_N^* \mathbf{A}_p^* \mathbf{B}_W^{-1} \mathbf{A}_q \mathbf{p}_{N-1} & \mathbf{p}_N^* \mathbf{A}_p^* \mathbf{B}_W^{-1} \mathbf{A}_q \mathbf{p}_N \end{array} \right). \tag{B.2.2}$$

Again, the upper left block of size $(N-1) \times (N-1)$ is already available from the previous greedy iteration and does not need to be re-computed. Computing the last column

requires the vector $\mathbf{B}_W^{-1}\mathbf{A}_q\mathbf{p}_N$ (*i.e.*, Riesz representer of the matrix vector product $\mathbf{A}_q\mathbf{p}_N$) followed by N dot products with the vectors $\mathbf{A}_p\mathbf{p}_1, \ldots, \mathbf{A}_p\mathbf{p}_N$ (stored in memory and therefore readily available). Computing the last row requires the vector $\mathbf{B}_W^{-1}\mathbf{A}_p\mathbf{p}_N$ (*i.e.*, Riesz representer of the matrix vector product $\mathbf{A}_p\mathbf{p}_N$) followed by $N-1$ dot products with the vectors $\mathbf{A}_q\mathbf{p}_1, \ldots, \mathbf{A}_q\mathbf{p}_{N-1}$ (stored in memory and therefore readily available).

Bibliography

[1] F. Assous, P. Ciarlet, and S. Labrunie. *Mathematical foundations of computational electromagnetism*. Springer, 2018.

[2] N. Atalla and F. Sgard. Modeling of perforated plates and screens using rigid frame porous models. *Journal of sound and vibration*, 303(1-2):195–208, 2007.

[3] M. Bachmayr and A. Cohen. Kolmogorov widths and low-rank approximations of parametric elliptic pdes. *Mathematics of Computation*, 86(304):701–724, 2017.

[4] O. Balabanov and A. Nouy. Randomized linear algebra for model reduction. part i: Galerkin methods and error estimation. *Advances in Computational Mathematics*, 45(5):2969–3019, 2019.

[5] M. Barrault, Y. Maday, N. C. Nguyen, and A. T. Patera. An 'empirical interpolation'method: application to efficient reduced-basis discretization of partial differential equations. *Comptes Rendus Mathematique*, 339(9):667–672, 2004.

[6] M. Bebendorf. *Hierarchical matrices*. Springer, 2008.

[7] A. Bendali and K. Lemrabet. Asymptotic analysis of the scattering of a time-harmonic electromagnetic wave by a perfectly conducting metal coated with a thin dielectric shell. *Asymptotic Analysis*, 57(3-4):199–227, 2008.

[8] J.-P. Berenger. A perfectly matched layer for the absorption of electromagnetic waves. *Journal of computational physics*, 114(2):185–200, 1994.

[9] M. Bernacki and S. Piperno. A dissipation-free time-domain discontinuous galerkin method applied to three-dimensional linearized euler equations around a steady-state non-uniform inviscid flow. *Journal of Computational Acoustics*, 14(04):445–467, 2006.

[10] C. Bernardi and R. Verfürth. Adaptive finite element methods for elliptic equations with non-smooth coefficients. *Numerische Mathematik*, 85(4):579–608, 2000.

[11] G. Bielak, J. Gallman, R. Kunze, P. Murray, J. Premo, M. Kosanchick, A. Hersh, J. Celano, B. Walker, J. Yu, et al. Advanced nacelle acoustic lining concepts development. Technical report, 2002.

[12] P. Binev, A. Cohen, W. Dahmen, R. DeVore, G. Petrova, and P. Wojtaszczyk. Convergence rates for greedy algorithms in reduced basis methods. *SIAM journal on mathematical analysis*, 43(3):1457–1472, 2011.

[13] Å. Björck and C. C. Paige. Loss and recapture of orthogonality in the modified gram–schmidt algorithm. *SIAM journal on matrix analysis and applications*, 13(1):176–190, 1992.

[14] E. Björnson, L. Sanguinetti, H. Wymeersch, J. Hoydis, and T. L. Marzetta. Massive mimo is a reality—what is next?: Five promising research directions for antenna arrays. *Digital Signal Processing*, 94:3–20, 2019.

[15] D. Bonomi, A. Manzoni, and A. Quarteroni. A matrix deim technique for model reduction of nonlinear parametrized problems in cardiac mechanics. *Computer Methods in Applied Mechanics and Engineering*, 324:300–326, 2017.

[16] A. Buffa and S. H. Christiansen. The electric field integral equation on lipschitz screens: definitions and numerical approximation. *Numerische Mathematik*, 94(2):229–267, 2003.

[17] A. Buffa and R. Hiptmair. Galerkin boundary element methods for electromagnetic scattering. In *Topics in computational wave propagation*, pages 83–124. Springer, 2003.

[18] A. Buffa, Y. Maday, A. T. Patera, C. Prud'homme, and G. Turinici. A priori convergence of the greedy algorithm for the parametrized reduced basis method. *ESAIM: Mathematical modelling and numerical analysis*, 46(3):595–603, 2012.

[19] A. Buhr, C. Engwer, M. Ohlberger, and S. Rave. A numerically stable a posteriori error estimator for reduced basis approximations of elliptic equations. *arXiv preprint arXiv:1407.8005*, 2014.

[20] Z. Cai and S. Zhang. Robust equilibrated residual error estimator for diffusion problems: Conforming elements. *SIAM Journal on Numerical Analysis*, 50(1):151–170, 2012.

[21] R. Cantrell and R. Hart. Interaction between sound and flow in acoustic cavities: Mass, momentum, and energy considerations. *The Journal of the Acoustical Society of America*, 36(4):697–706, 1964.

[22] F. Casenave, A. Ern, and T. Lelièvre. Accurate and online-efficient evaluation of the a posteriori error bound in the reduced basis method. *ESAIM: Mathematical Modelling and Numerical Analysis*, 48(1):207–229, 2014.

[23] F. Casenave, A. Ern, and T. Lelièvre. A nonintrusive reduced basis method applied to aeroacoustic simulations. *Advances in Computational Mathematics*, 41(5):961–986, 2015.

[24] M. Cessenat. *Mathematical methods in electromagnetism: linear theory and applications*, volume 41. World scientific, 1996.

[25] J. H. Chaudhry, L. N. Olson, and P. Sentz. A least-squares finite element reduced basis method. *SIAM Journal on Scientific Computing*, 43(2):A1081–A1107, 2021.

[26] Y. Chen. A certified natural-norm successive constraint method for parametric inf–sup lower bounds. *Applied Numerical Mathematics*, 99:98–108, 2016.

[27] Y. Chen, J. S. Hesthaven, Y. Maday, and J. Rodríguez. A monotonic evaluation of lower bounds for inf-sup stability constants in the frame of reduced basis approximations. *Comptes Rendus Mathematique*, 346(23-24):1295–1300, 2008.

[28] Y. Chen, J. Jiang, and A. Narayan. A robust error estimator and a residual-free error indicator for reduced basis methods. *Computers & Mathematics with Applications*, 77(7):1963–1979, 2019.

[29] P. Ciarlet Jr and J. Zou. Fully discrete finite element approaches for time-dependent maxwell's equations. *Numerische Mathematik*, 82(2):193–219, 1999.

[30] A. Cohen, W. Dahmen, R. DeVore, and J. Nichols. Reduced basis greedy selection using random training sets. *ESAIM: Mathematical Modelling and Numerical Analysis*, 54(5):1509–1524, 2020.

[31] A. Cohen and R. DeVore. Kolmogorov widths under holomorphic mappings. *IMA Journal of Numerical Analysis*, 36(1):1–12, 2016.

[32] D. Colton and R. Kress. *Integral equation methods in scattering theory*. SIAM, 2013.

[33] W. Dahmen, C. Plesken, and G. Welper. Double greedy algorithms: Reduced basis methods for transport dominated problems∗. *ESAIM: Mathematical Modelling and Numerical Analysis*, 48(3):623–663, 2014.

[34] R. Della Ratta Rinaldi, A. Iob, and R. Arina. An efficient discontinuous galerkin method for aeroacoustic propagation. *International Journal for Numerical Methods in Fluids*, 69(9):1473–1495, 2012.

[35] P. Delorme, P. Mazet, C. Peyret, and Y. Ventribout. Computational aeroacoustics applications based on a discontinuous galerkin method. *Comptes Rendus Mécanique*, 333(9):676–682, 2005.

[36] S. Deparis. Reduced basis error bound computation of parameter-dependent Navier-stokes equations by the natural norm approach. *SIAM Journal on Numerical Analysis*, 46(4):2039–2067, 2008.

[37] J. L. Eftang, M. A. Grepl, and A. T. Patera. A posteriori error bounds for the empirical interpolation method. *Comptes Rendus Mathematique*, 348(9-10):575–579, 2010.

[38] A. Ern and J.-L. Guermon. *Theory and practice of finite elements*, volume 159. Springer, 2004.

[39] M. Fares, J. S. Hesthaven, Y. Maday, and B. Stamm. The reduced basis method for the electric field integral equation. *Journal of Computational Physics*, 230(14):5532–5555, 2011.

[40] C. Farhat, L. Crivelli, and F. X. Roux. Extending substructure based iterative solvers to multiple load and repeated analyses. *Computer methods in applied mechanics and engineering*, 117(1-2):195–209, 1994.

[41] X. Feng. Absorbing boundary conditions for electromagnetic wave propagation. *Mathematics of computation*, 68(225):145–168, 1999.

[42] M. Ganesh, J. S. Hesthaven, and B. Stamm. A reduced basis method for electromagnetic scattering by multiple particles in three dimensions. *Journal of Computational Physics*, 231(23):7756–7779, 2012.

[43] M. B. Giles. Nonreflecting boundary conditions for euler equation calculations. *AIAA journal*, 28(12):2050–2058, 1990.

[44] L. Giraud, J. Langou, and M. Rozloznik. The loss of orthogonality in the gram-schmidt orthogonalization process. *Computers & Mathematics with Applications*, 50(7):1069–1075, 2005.

[45] V. Girault and P.-A. Raviart. *Finite element methods for Navier-Stokes equations: theory and algorithms*, volume 5. Springer Science & Business Media, 2012.

[46] D. Givoli. Non-reflecting boundary conditions. *Journal of computational physics*, 94(1):1–29, 1991.

[47] L. Greengard and V. Rokhlin. A fast algorithm for particle simulations. *Journal of computational physics*, 73(2):325–348, 1987.

[48] C. Greif and K. Urban. Decay of the kolmogorov n-width for wave problems. *Applied Mathematics Letters*, 96:216–222, 2019.

[49] N. A. Gumerov and R. Duraiswami. *Fast multipole methods for the Helmholtz equation in three dimensions*. Elsevier, 2005.

[50] B. Haasdonk and M. Ohlberger. Reduced basis method for finite volume approximations of parametrized linear evolution equations. *ESAIM: Mathematical Modelling and Numerical Analysis*, 42(2):277–302, 2008.

[51] S. Hain, M. Ohlberger, M. Radic, and K. Urban. A hierarchical a posteriori error estimator for the Reduced Basis Method. *Advances in Computational Mathematics*, 45(5-6):2191–2214, 2019.

[52] R. L. Haupt. *Antenna arrays: a computational approach*. John Wiley & Sons, 2010.

[53] J. S. Hesthaven. On the analysis and construction of perfectly matched layers for the linearized euler equations. *Journal of computational Physics*, 142(1):129–147, 1998.

[54] J. S. Hesthaven, G. Rozza, B. Stamm, et al. *Certified reduced basis methods for parametrized partial differential equations*, volume 590. Springer, 2016.

[55] J. S. Hesthaven, B. Stamm, and S. Zhang. Certified reduced basis method for the electric field integral equation. *SIAM Journal on Scientific Computing*, 34(3):A1777–A1799, 2012.

[56] R. Hiptmair and C. Schwab. Natural boundary element methods for the electric field integral equation on polyhedra. *SIAM Journal on Numerical Analysis*, 40(1):66–86, 2002.

[57] R. Hixon, S.-H. Shih, and R. R. Mankabadi. Evaluation of boundary conditions for computational aeroacoustics. *AIAA journal*, 33(11):2006–2012, 1995.

[58] F. Q. Hu. On absorbing boundary conditions for linearized euler equations by a perfectly matched layer. *Journal of computational physics*, 129(1):201–219, 1996.

[59] F. Q. Hu. A stable, perfectly matched layer for linearized euler equations in unsplit physical variables. *Journal of Computational Physics*, 173(2):455–480, 2001.

[60] D. Huynh, D. Knezevic, Y. Chen, J. S. Hesthaven, and A. Patera. A natural-norm successive constraint method for inf-sup lower bounds. *Computer Methods in Applied Mechanics and Engineering*, 199(29-32):1963–1975, 2010.

[61] D. B. P. Huynh, G. Rozza, S. Sen, and A. T. Patera. A successive constraint linear optimization method for lower bounds of parametric coercivity and inf–sup stability constants. *Comptes Rendus Mathematique*, 345(8):473–478, 2007.

[62] K. Ito and S. S. Ravindran. A reduced-order method for simulation and control of fluid flows. *Journal of computational physics*, 143(2):403–425, 1998.

[63] M. G. Jones, F. Simon, and R. Roncen. Broadband and low-frequency acoustic liner investigations at nasa and onera. *AIAA Journal*, 60(4):2481–2500, 2022.

[64] D. Klindworth, M. A. Grepl, and G. Vossen. Certified reduced basis methods for parametrized parabolic partial differential equations with non-affine source terms. *Computer methods in applied mechanics and engineering*, 209:144–155, 2012.

[65] S. Kurz, O. Rain, and S. Rjasanow. The adaptive cross-approximation technique for the 3d boundary-element method. *IEEE transactions on Magnetics*, 38(2):421–424, 2002.

[66] T. Lassila, A. Manzoni, and G. Rozza. On the approximation of stability factors for general parametrized partial differential equations with a two-level affine decomposition. *ESAIM: Mathematical Modelling and Numerical Analysis*, 46(6):1555–1576, 2012.

[67] E. Lier. Review of soft and hard horn antennas, including metamaterial-based hybrid-mode horns. *IEEE Antennas and Propagation Magazine*, 52(2):31–39, 2010.

[68] C.-C. Lu and W. C. Chew. A multilevel algorithm for solving a boundary integral equation of wave scattering. *Microwave and Optical Technology Letters*, 7(10):466–470, 1994.

[69] Z. Luo and G. Chen. *Proper orthogonal decomposition methods for partial differential equations*. Academic Press, 2018.

[70] D. MAA. Theory and design of microperforated panel sound-absorbing constructions. 1975.

[71] Y. Maday. Reduced basis method for the rapid and reliable solution of partial differential equations. 2006.

[72] Y. Maday and B. Stamm. Locally adaptive greedy approximations for anisotropic parameter reduced basis spaces. *SIAM Journal on Scientific Computing*, 35(6):A2417–A2441, 2013.

[73] R. Mankbadi, R. Hixon, S.-H. Shih, and L. Povinelli. Use of linearized euler equations for supersonic jet noise prediction. *AIAA journal*, 36(2):140–147, 1998.

[74] A. Manzoni and F. Negri. Heuristic strategies for the approximation of stability factors in quadratically nonlinear parametrized pdes. *Advances in Computational Mathematics*, 41(5):1255–1288, 2015.

[75] C. D. Meyer. *Matrix analysis and applied linear algebra*, volume 71. Siam, 2000.

[76] T. A. Milligan. *Modern antenna design*. John Wiley & Sons, 2005.

[77] A. Monje-Real and V. de la Rubia. Electric field integral equation fast frequency sweep for scattering of nonpenetrable objects via the reduced-basis method. *IEEE Transactions on Antennas and Propagation*, 68(8):6232–6244, 2020.

[78] P. Monk et al. *Finite element methods for Maxwell's equations*. Oxford University Press, 2003.

[79] P. Mungur and G. Gladwell. Acoustic wave propagation in a sheared fluid contained in a duct. *Journal of Sound and Vibration*, 9(1):28–48, 1969.

[80] J.-C. Nédélec. Mixed finite elements in $\mathbb{R}3$. *Numerische Mathematik*, 35(3):315–341, 1980.

[81] J.-C. Nédélec. *Acoustic and electromagnetic equations: integral representations for harmonic problems*, volume 144. Springer Science & Business Media, 2013.

[82] F. Negri, A. Manzoni, and D. Amsallem. Efficient model reduction of parametrized systems by matrix discrete empirical interpolation. *Journal of Computational Physics*, 303:431–454, 2015.

[83] N. Ngoc Cuong, K. Veroy, and A. T. Patera. Certified real-time solution of parametrized partial differential equations. In *Handbook of materials modeling*, pages 1529–1564. Springer, 2005.

[84] N. C. Nguyen. A posteriori error estimation and basis adaptivity for reduced-basis approximation of nonaffine-parametrized linear elliptic partial differential equations. *Journal of Computational Physics*, 227(2):983–1006, 2007.

[85] N. Nishimura. Fast multipole accelerated boundary integral equation methods. *Appl. Mech. Rev.*, 55(4):299–324, 2002.

[86] M. Nonino, F. Ballarin, G. Rozza, and Y. Maday. Overcoming slowly decaying kolmogorov n-width by transport maps: application to model order reduction of fluid dynamics and fluid–structure interaction problems. *arXiv preprint arXiv:1911.06598*, 2019.

[87] A. Nouy. A priori model reduction through proper generalized decomposition for solving time-dependent partial differential equations. *Computer Methods in Applied Mechanics and Engineering*, 199(23-24):1603–1626, 2010.

[88] M. Ohlberger and S. Rave. Reduced basis methods: Success, limitations and future challenges. *arXiv preprint arXiv:1511.02021*, 2015.

[89] D. Panagiotopoulos, E. Deckers, and W. Desmet. Krylov subspaces recycling based model order reduction for acoustic bem systems and an error estimator. *Computer Methods in Applied Mechanics and Engineering*, 359:112755, 2020.

[90] B. N. Parlett. *The symmetric eigenvalue problem*. SIAM, 1998.

[91] L. Pascal, E. Piot, and G. Casalis. Discontinuous galerkin method for acoustic modes computation in lined ducts. In *18th AIAA/CEAS Aeroacoustics Conference (33rd AIAA Aeroacoustics Conference)*, page 2153, 2012.

[92] L. Pascal, E. Piot, and G. Casalis. Discontinuous galerkin method for the computation of acoustic modes in lined flow ducts with rigid splices. *Journal of Sound and Vibration*, 332(13):3270–3288, 2013.

[93] E. Piot, J.-P. Brazier, F. Simon, V. Fascio, C. Peyret, and J. Ingenito. Design, manufacturing and demonstration of acoustic liners for air conditioning systems. In *22nd AIAA/CEAS Aeroacoustics Conference*, page 2788, 2016.

[94] T. A. Porsching. Estimation of the error in the reduced basis method solution of nonlinear equations. *Mathematics of Computation*, 45(172):487–496, 1985.

[95] J. Primus, E. Piot, and F. Simon. An adjoint-based method for liner impedance eduction: Validation and numerical investigation. *Journal of Sound and Vibration*, 332(1):58–75, 2013.

[96] C. Prud'Homme, D. V. Rovas, K. Veroy, L. Machiels, Y. Maday, A. T. Patera, and G. Turinici. Reliable real-time solution of parametrized partial differential equations: Reduced-basis output bound methods. *J. Fluids Eng.*, 124(1):70–80, 2002.

[97] A. Quarteroni, A. Manzoni, and F. Negri. *Reduced basis methods for partial differential equations: an introduction*, volume 92. Springer, 2015.

[98] A. Quarteroni, G. Rozza, and A. Manzoni. Certified reduced basis approximation for parametrized partial differential equations and applications. *Journal of Mathematics in Industry*, 1(1):1–49, 2011.

[99] P.-A. Raviart and J.-M. Thomas. A mixed finite element method for 2-nd order elliptic problems. In *Mathematical aspects of finite element methods*, pages 292–315. Springer, 1977.

[100] S. W. Rienstra. A classification of duct modes based on surface waves. *Wave motion*, 37(2):119–135, 2003.

[101] B. Rivière. *Discontituous Galerkin Methods for solving elliptic and parabolic equations: theory and implementation*. SIAM, 2008.

[102] F.-X. Roux and A. Barka. Feti methods. In *Computational Electromagnetics*, pages 651–685. Springer, 2014.

[103] F.-X. Roux and A. Barka. Block krylov recycling algorithms for feti-2lm applied to 3-d electromagnetic wave scattering and radiation. *IEEE Transactions on Antennas and Propagation*, 65(4):1886–1895, 2017.

[104] F.-X. Roux, F. Magoules, L. Series, and Y. Boubendir. Approximation of optimal interface boundary conditions for two-lagrange multiplier feti method. In *Domain Decomposition Methods in Science and Engineering*, pages 283–290. Springer, 2005.

[105] G. Rozza, D. B. P. Huynh, and A. T. Patera. Reduced basis approximation and a posteriori error estimation for affinely parametrized elliptic coercive partial differential equations. *Archives of Computational Methods in Engineering*, 15(3):229–275, 2008.

[106] Y. Saad. *Iterative methods for sparse linear systems*. SIAM, 2003.

[107] T. J. Santner, B. J. Williams, W. I. Notz, and B. J. Williams. *The design and analysis of computer experiments*, volume 1. Springer, 2003.

[108] O. Schenk and K. Gärtner. Two-level dynamic scheduling in pardiso: Improved scalability on shared memory multiprocessing systems. *Parallel Computing*, 28(2):187–197, 2002.

[109] W. Schilders. Introduction to model order reduction. In *Model order reduction: theory, research aspects and applications*, pages 3–32. Springer, 2008.

[110] S. Sen, K. Veroy, D. B. P. Huynh, S. Deparis, N. C. Nguyen, and A. T. Patera. "natural norm" a posteriori error estimators for reduced basis approximations. *Journal of Computational Physics*, 217(1):37–62, 2006.

[111] Y. Shi, X. Chen, Y. Tan, H. Jiang, and S. Liu. Reduced-basis boundary element method for fast electromagnetic field computation. *JOSA A*, 34(12):2231–2242, 2017.

[112] P. Sirkovic and D. Kressner. Subspace acceleration for large-scale parameter-dependent hermitian eigenproblems. *SIAM Journal on Matrix Analysis and Applications*, 37(2):695–718, 2016.

[113] K. Smetana, O. Zahm, and A. T. Patera. Randomized residual-based error estimators for parametrized equations. *SIAM Journal on Scientific Computing*, 41(2):A900–A926, 2019.

[114] K. Smetana, O. Zahm, and A. T. Patera. Randomized residual-based error estimators for parametrized equations. *SIAM journal on scientific computing*, 41(2):A900–A926, 2019.

[115] K. M. Soodhalter. Block krylov subspace recycling for shifted systems with unrelated right-hand sides. *SIAM Journal on Scientific Computing*, 38(1):A302–A324, 2016.

[116] C. K. Tam and Z. Dong. Radiation and outflow boundary conditions for direct computation of acoustic and flow disturbances in a nonuniform mean flow. *Journal of computational acoustics*, 4(02):175–201, 1996.

[117] R. Tezaur, A. Macedo, and C. Farhat. Iterative solution of large-scale acoustic scattering problems with multiple right hand-sides by a domain decomposition method

with lagrange multipliers. *International Journal for Numerical Methods in Engineering*, 51(10):1175–1193, 2001.

[118] K. W. Thompson. Time dependent boundary conditions for hyperbolic systems. *Journal of computational physics*, 68(1):1–24, 1987.

[119] I. Toulopoulos and J. A. Ekaterinaris. Artificial boundary conditions for the numerical solution of the euler equations by the discontinuous galerkin method. *Journal of Computational Physics*, 230(15):5974–5995, 2011.

[120] G. Turinici, A. T. Patera, and Y. Maday. A priori convergence theory for reduced-basis approximations of single-parametric elliptic partial differential equations. 2002.

[121] B. D. Van Veen and K. M. Buckley. Beamforming: A versatile approach to spatial filtering. *IEEE assp magazine*, 5(2):4–24, 1988.

[122] R. Verfürth. Robust a posteriori error estimators for a singularly perturbed reaction-diffusion equation. *Numerische Mathematik*, 78(3):479–493, 1998.

[123] K. Veroy, C. Prud'Homme, D. Rovas, and A. Patera. A posteriori error bounds for reduced-basis approximation of parametrized noncoercive and nonlinear elliptic partial differential equations. In *16th AIAA Computational Fluid Dynamics Conference*, page 3847, 2003.

[124] M. Yano. A reduced basis method with exact-solution certificates for steady symmetric coercive equations. *Computer Methods in Applied Mechanics and Engineering*, 287:290–309, 2015.

[125] M. Yano. A minimum-residual mixed reduced basis method: Exact residual certification and simultaneous finite-element reduced-basis refinement. *ESAIM: Mathematical Modelling and Numerical Analysis*, 50(1):163–185, 2016.

[126] M. Yano. A reduced basis method for coercive equations with an exact solution certificate and spatio-parameter adaptivity: energy-norm and output error bounds. *SIAM Journal on Scientific Computing*, 40(1):A388–A420, 2018.

[127] D. M. Young and K. C. Jea. Generalized conjugate-gradient acceleration of non-symmetrizable iterative methods. *Linear Algebra and its applications*, 34:159–194, 1980.

[128] O. Zahm, M. Billaud-Friess, and A. Nouy. Projection-based model order reduction methods for the estimation of vector-valued variables of interest. *SIAM Journal on Scientific Computing*, 39(4):A1647–A1674, 2017.

[129] O. Zahm and A. Nouy. Interpolation of inverse operators for preconditioning parameter-dependent equations. *SIAM Journal on Scientific Computing*, 38(2):A1044–A1074, 2016.

[130] S. Zhang. Primal-dual reduced basis methods for convex minimization variational problems: Robust true solution a posteriori error certification and adaptive greedy algorithms. *SIAM Journal on Scientific Computing*, 42(6):A3638–A3676, 2020.