INSTITUT POLYTECHNIQUE DE PARIS

ENSTA IP PARIS

NNT : 2025IPPAE015

Thèse de doctorat

Discontinuous Galerkin finite element methods with transmission variables for time-harmonic wave propagation

Thèse de doctorat de l'Institut Polytechnique de Paris
préparée à l'École nationale supérieure de techniques avancées

École doctorale n°574 École doctorale de mathématiques Hadamard (EDMH)
Spécialité de doctorat: Mathématiques appliquées

Thèse présentée et soutenue à Palaiseau, le 20 novembre 2025, par

SIMONE PESCUMA

Composition du Jury :

Alexandre Ern Professeur, ENPC	Président
Lise-Marie Imbert-Gerard Associate Professor, The University of Arizona	Absente lors de la soutenance Rapporteuse
Sébastien Tordeux Professeur des universités, UPPA	Rapporteur
Théophile Chaumont-Frelet Chargé de recherche, Inria, Université de Lille	Examinateur
Emmanuel Perrey-Debain Professeur des universités, UTC	Examinateur
Ilaria Perugia Professeure, Universität Wien	Examinatrice
Axel Modave Chargé de recherche, CNRS, ENSTA	Directeur
Gwénaël Gabard Professeur des universités, Le Mans Université	Co-directeur

To Ambra and Giorgia,
I see the future in your eyes.

Acknowledgments

It feels very surreal writing these pages as if three years of my life have not just flashed in front of my eyes. That is not totally true because I have been taking some pauses from time to time to acknowledge the passing of time and to understand the evolution of events and people I have been revolving around. In these pages, which are the result of one of these pauses, I thank all the people that contributed, helped and supported me throughout these years of PhD.

First of all, I would like to thank my supervisors. Axel Modave really set up the best possible environment for my research. He has been extremely available and caring since the very first day while allowing me freedom to learn from my own mistakes. His rigorous way of working has been a clear path that I have constantly looked up to. Gwénaël Gabard has always pushed our conversations forward and shed the light onto unclear situations thanks to his deep knowledge. Moreover, Théophile Chaumont-Frelet, with whom I worked less regularly, impressed me with his wide expertise concerning many different topics and has been a people person that made the confrontation a relaxed and stimulating moment. I will never forget the whole-day-meetings we have often had in ENSTA to work together (that I refered to as "the meeting with the big 3"). We have always been capable of solving the problems we encountered and moving forward. In these occasions I learnt that different people with different backgrounds can enrich the conversation and collaborate to advance and progress: a very valuable lesson that enhances cooperation and diversity.

I am grateful for all the PhD students, post-docs and researchers of the UMA laboratory and, in particular, of the POEMS team. I loved working in a familiar, non-competitive and very welcoming environment that made me feel respected and heard. Special thanks to all the people with whom I shared the office throughout the years: Pierre A., Étienne, Rose-Cloé, Dylan and Pierre B., I admire each one of you and I hope you will capable to live the future you imagine for yourself.

I am thankful for my parents Rosa and Vincenzo and my sisters Chiara and Serena who have always supported and trusted me and all the choices they knew would benefit me and be best for my future even if they meant living far from them most of the time. In particular, my mum raised me teaching me that education is one of the most powerful tools that I could have because it allows freedom and that knowledge needs to be constantly sought. Every time I see my nieces Giorgia and Ambra's smiles and I hear them laugh I understand life's real meaning and my heart fills with love: I do see the future in their beautiful eyes.

I have always appreciated my dear friends Kevin and Camilla's support, admiration and friendship: feelings and emotions that are mutual, pure, unselfish and, most importantly, genuine. Growing up, maturing and metamorphosing together has always been and will always be an exciting and nourishing experience.

Words can not describe the affection that ties Caterina, Giulia and me (i.e. *La signora in giallo*) all together. In the past we supported each other to pave our way out of Polimi and now, with the same intensity and strength, we are always there to comfort, celebrate and listen to each other. When we are together I always feel completely free to be myself and they bring out my most genuine laughs: I owe them this powerful freedom and a very necessary light-heartedness and spontaneity.

Francesco and Lupo (i.e. *Boni*) remain important friends whose lives keep on being inspiring models for me. I love how we are capable of living at the same time similar and different experiences, sharing them and encouraging each other. Instead of weakening it, time is making our friendship stronger and stronger and it makes me feel loved and cared for.

Maddalena is for sure one of the most important people that I have in my life right now. She has always cared to listen to me when I needed to be comforted, even when words were not there, even when I could only express myself with a drawing or with some enigmatic sentences that I could not but write. Her special and unique sensitivity is a source of inspiration and admiration for me. I will always look forward to our trips together filled with laughs, art, good food (*alto en calorias*) and fun: Argentina is next.

Sasha, Magdalena, Elitsa, Gabriel and Masha have become a small, but important family here in Paris. It never stops amazing me how, coming from such different backgrounds, we have been able to bond and become such good friends. I am thankful for their daily presence and love. I apologize in advance to Magdalena if the content related to *porosity* in this thesis does not meet her expectation.

The friendship with Flavia and Evelina that blossomed here in Paris is one I care about the most. Since the very beginning (when we queued together for Mahmood's concert in 2022) it was characterized by lightness and laughs. With time we were able to know more, open up, listen, understand each other and be in the first row during many more concerts. I am so grateful for our shared love for Nutella: if it wasn't for it people would really have a hard time finding our waists.

My previous and current flatmates, Luca and Lorenzo respectively, coming from different backgrounds than mine, have been great examples that I admire of dedicated and professional workers who daily honor their jobs. It has been and it is a pleasure sharing our little, cold, imperfect, but beloved apartment in Pernety and spending some of our free time together along with beers and therapeutic laughs.

Thanks to my coach Olivier and teammates Gabin & Co at PAC Volley. Despite the fact that they got to know my shiest side, they have always provided a nice and stimulating environment at the gym where I could always step in and leave outside its walls bad thoughts and problems.

Zio, a Parigi c'è il sole?

In March 2024 my niece Giorgia sent me a vocal message through my sister Chiara's phone and asked me: "*Zio, a Parigi c'è il sole?*" ("*Uncle, is it sunny in Paris?*"). It was a difficult question to answer because the Sun was not shining outside in the sky, nor inside my heart. Clouds covered the sky and they did not allow the sun rays to illuminate the city in the same way as bad thoughts were spread in my head and did not leave room for positivity. I was broken down in pieces and I did not know how to pick them up, nor how to put them back together. The months passed by and the weather did not improve even during summer making the atmosphere rather sad and depressive. But at some point an unexpected event made everything shift.

It was the beginning of July and I was in Berlin for the WAVES 2024 conference. I had been to Berlin once many years before during a school trip and, back then, I did not like the city: I found it ugly, grey and sad. I hoped that coming back years later could allow me to look at the city with different eyes and re-evaluate it, but upon my arrival I had the same feelings: I could not find beauty in the fragmented architectural style due to the co-presence of old and new buildings constructed after World War II, nor in the lack of a large spectrum of colors.

One morning I was walking from the hotel I was staying in to the metro station Konstanzer Straße to go to the Harnack House which was the conference venue. Very few clouds were floating in the sky and the Sun could illuminate with its warm rays the road and the buildings. The very same dark and tall buildings that I did not use to like were surrounded by an intense light that was capable of bringing out their beauty and enhance it.

I found beauty in a place I thought it was condemned to ugliness and sadness. I discovered that everything is beautiful and light simply helps this beauty to be caught by the eyes and to be appreciated. It was a very important lesson I was fascinated to learn and internalize: since beauty intrinsically belongs to everything and everyone, it is my duty to look for it, even when external light is not illuminating, even where beauty does not seem to have any hope to exist.

Consequently there is beauty in imperfections and in what I normally perceive as ugly: a plate that breaks down in pieces when falling from a table to the ground, a sky full of dense clouds that hide the Sun. Now I am able to kneel, collect the pieces, put them together and fix the fractures with golden sutures. Now I can raise my eyes to the sky and acknowledge that, despite the clouds, the Sun is always there and it always shines above them.

KINTSUGI

[...]

I always find beauty in what is ugly
the affection in a farewell letter
the love in the hate of the people
the smile among thousand lies
a sun ray between two buildings in Berlin

it's raining today too, but I'm going out anyway

[...]

trovo sempre il bello in ciò che è brutto
l'affetto in una lettera d'addio
l'amore nell'odio della gente
il sorriso tra le mille bugie
un raggio di sole tra due edifici a Berlino

anche oggi piove, ma esco lo stesso

Contents

Introduction étendue (Français)

Contexte général

Les méthodes d'éléments finis (FEM) ont été largement utilisées au cours des dernières décennies par l'étude des phénomènes de propagation des ondes en régime harmonique, couvrant un large éventail de contextes physiques, y compris les ondes acoustiques, électromagnétiques et élastiques. Cette polyvalence explique en grande partie leur adoption massive dans les milieux académiques et industriels, où la capacité à traiter des configurations complexes et réalistes constitue un atout majeur.

Les premières applications se sont concentrées sur des problèmes intérieurs avec des géométries complexes, comme les systèmes vibro-acoustiques, les guides d'ondes et les cavités résonantes, principalement pour l'analyse des vibrations forcées, la diffusion et la réponse en fréquence. Ces cas d'étude ont servi de base pour tester la robustesse des premières formulations numériques et ont permis d'identifier les difficultés spécifiques liées aux configurations tridimensionnelles ou à forte complexité géométrique.

Au fil du temps, des avancées significatives ont étendu l'applicabilité des FEM aux problèmes extérieurs dans des domaines non bornés ainsi que dans des milieux hétérogènes.

Principaux défis

Un défi important dans les problèmes extérieurs est la gestion de la nature non bornée du domaine. La FEM aborde généralement cela en introduisant une frontière artificielle qui tronque le domaine infini à un domaine de calcul borné. Cependant, la réduction de la taille de ce domaine doit être soigneusement équilibrée avec le besoin de minimiser les réflexions parasites à la frontière. Des traitements aux limites efficaces et flexibles sont nécessaires pour minimiser de telles erreurs sans augmenter le coût de calcul. Cet équilibre représente un enjeu central dans la conception de méthodes de troncature , car il conditionne directement la qualité des résultats et la faisabilité des simulations de très grande taille. Les approches classiques incluent les éléments infinis [14, 19], les couches absorbantes [35, 103] et les conditions limites absorbantes (éventuellement non locales) [45, 98].

Par ailleurs, il y a eu des efforts considérables pour améliorer la FEM pour les problèmes à haute fréquence, où la FEM traditionnelle a du mal à résoudre les champs d'ondes fortement oscillatoires de manière précise. En effet, à hautes fréquences, un phénomène typique des problèmes de propagation d'ondes, appelé effet de pollution, détériore la solution numérique en générant une erreur qui se propage dans tout le domaine [79, 81]. Ce phénomène nécessite un raffinement important du maillage, ce qui accroit notablement les coûts de calcul. Pour résoudre ce problème, de nombreuses FEM avancées ont été introduites au cours des dernières décennies. Celles-ci incluent les éléments finis d'ordres élevés [3, 13], les méthodes de Galerkin stabilisées [66], les formulations variationnelles multi-échelles [76], et des discrétisations basées sur les ondes qui reposent sur l'intuition que l'incorporation de la physique des ondes dans la formulation numérique peut améliorer sa précision et sa efficacité. Ces approches tirent parti des spécificités du problème physique pour améliorer le modèle numérique, ce qui constitue un axe de recherche particulièrement actif et prometteur.

Un autre problème lié à la FEM est que le système linéaire résultant est généralement complexe, non hermitien et indéfini. Les solveurs directs deviennent rapidement trop coûteux pour les problèmes à haute fréquence ou de grande taille, où le nombre de degrés de liberté peut être extrêmement grand, et leur mise en œuvre dans des environnements parallèles est délicate. L'utilisation de méthodes itératives dans ce contexte constitue également un défi : le mauvais conditionnement de la matrice peut gravement réduire la vitesse de convergence [49]. Ces limitations imposent de recourir à des outils numériques plus sophistiqués, capables de gérer la taille croissante des systèmes considérés. Pour surmonter ces difficultés, diverses stratégies ont été développées, notamment l'utilisation de préconditionneurs spécialement adaptés aux opérateurs de type Helmholtz [46, 47] et des méthodes de décomposition de domaine [57]. De plus, des techniques de condensation statique et d'hybridation se sont révélées efficaces pour améliorer la convergence des solveurs itératifs pour les problèmes à grande échelle et à haute fréquence. Ces stratégies ont montré leur potentiel pour améliorer l'efficacité des simulations, même si leur conception et leur analyse restent des sujets de recherche actifs.

Méthodes d'éléments finis discontinus de type Galerkin

Les méthodes par éléments finis discontinus de type Galerkin (DG) reposent sur l'utilisation de fonctions de base discontinues aux interfaces entre les éléments et la définition de flux numériques qui influencent fortement la précision des schémas DG, de multiples façons, ainsi que des fonctions de base non standard. Ces formulations offrent ainsi plus de flexibilité que les méthodes par éléments finis continus (CG) traditionnelles. Cette flexibilité ouvre la voie à un large éventail de formulations adaptées à différents types de problèmes, y compris ceux présentant des discontinuités ou des variations rapides des paramètres physiques. Elles offrent également un large cadre pour les formulations qui permettent l'utilisation d'algorithmes itératifs rapides. Cette caractéristique est d'autant plus importante que les simulations modernes exigent des méthodes capables de tirer pleinement parti des architectures parallèles actuelles.

Dans les méthodes hybridables par éléments finis discontinus standard (HDG) [8, 33], une variable supplémentaire correspondant à la trace numérique de la solution est introduite aux interfaces entre les éléments et sur la frontière du domaine de calcul. Cette variable hybride agit comme un multiplicateur de Lagrange qui découple les inconnues physiques entre les éléments. L'élimination des champs physiques est réalisée en résolvant des problèmes locaux à l'intérieur des éléments. En pratique, après avoir inversé des matrices locales par élément, un système réduit avec uniquement la variable hybride est formulé sur le squelette du maillage. Cette réduction de la taille du système global constitue un avantage notable, en particulier pour les problèmes de grande dimension. Une fois que la solution de ce système global hybride (et généralement plus petit) est obtenue, la solution du problème original peut être récupérée dans une étape suivante.

Lors de l'utilisation de solveurs directs, l'avantage de cette approche est immédiat, car le système HDG réduit présente souvent beaucoup moins de degrés de liberté que le système DG original tout en préservant sa structure creuse. Cela contribue à diminuer significativement les besoins en mémoire, un facteur critique dans les simulations à grande échelle. D'autre part, pour les solveurs itératifs, où la taille et la structure de la matrice ne sont plus les principaux facteurs de performance, la situation devient plus complexe. Cependant, plusieurs études ont montré une amélioration significative, notamment une convergence plus rapide [97]. Ces résultats suggèrent que les méthodes HDG peuvent constituer une alternative particulièrement attrayante dans certains cas, même si leur efficacité dépend du choix précis des paramètres numériques. De plus, puisque les problèmes locaux peuvent être résolus indépendamment, la méthode se parallélise facilement, ce qui peut conduire à d'autres améliorations. Cette propriété en fait une candidate idéale pour les supercalculateurs modernes, où la distribution efficace des tâches est essentielle.

La méthode de Trefftz est une méthode basée sur les ondes qui prend des solutions de l'équation de Helmholtz homogène comme fonctions de base et de test sur chaque élément [23, 72, 93], par exemple des ondes planes propagatives. Cela donne une convergence rapide (souvent exponentielle) [73] et nécessite moins de degrés de liberté que la FEM basée sur des polynômes pour une précision similaire, en particulier dans les domaines homogènes ou homogènes par morceaux [92]. Cette efficacité s'explique par le fait que les sfonctions de base intègrent déjà les caractéristiques essentielles du comportement ondulatoire. Le réglage adaptatif des directions de propagation aide à capturer le champ d'onde dominant [39]. Cependant, les ondes planes purement propagatives souffrent d'un mauvais conditionnement numérique [7] : à mesure que leur nombre augmente, elles deviennent presque linéairement dépendantes, deteriorant le conditionnement et risquant une instabilité sans régularisation [105].Ce problème constitue une limite importante à l'utilisation de bases trop riches, nécessitant l'élaboration de stratégies de stabilisation appropriées. Parce qu'elles ne peuvent pas représenter les modes évanescents, limitant son utilisation près des singularités ou dans les milieux à coefficient variable, des bases plus riches, comme les ondes évanescentes ou les fonctions quasi-Trefftz [83, 84, 119], sont actuellement étudiées. Ces travaux de recherche récents montrent que l'intégration de composantes supplémentaires pourrait permettre de contourner certaines limitations fondamentales et d'élargir encore le domaine d'application des méthodes de type Trefftz.

Objectif de l'étude

Cette thèse fait partie du projet *WavesDG*. Il s'agit d'un projet de quatre ans qui implique plusieurs chercheurs et doctorants et dont le but est d'étudier et de développer des méthodes de Galerkin discontinu spécifiques aux ondes pour les phénomènes de propagation des ondes.

Les principaux axes de recherche comprennent l'étude de traitements d'interface adaptés aux ondes, basés sur des conditions de transmission d'ordre bas et elevé, ainsi que sur des fonctions de base spécifiques aux ondes. L'intérêt de ces traitements d'interface réside dans leur capacité à réduire les erreurs numériques et à améliorer la stabilité des schémas, ce qui est crucial pour obtenir des solutions précises sur des maillages complexes. Ces directions de recherche ciblées sur les ondes sont motivées par l'objectif d'améliorer les propriétés du système linéaire associé aux schémas numériques, conduisant éventuellement à des algorithmes itératifs efficaces qui peuvent être exécutés sur des architectures parallèles permettant de faire face à des problèmes réalistes à grande échelle. La parallélisation et l'efficacité computationnelle sont des aspects essentiels, car les simulations à grande échelle impliquent un nombre très élevé de degrés de liberté et nécessitent des ressources importantes en calcul.

La première contribution au projet *WavesDG* est un article d'Axel Modave et Théophile Chaumont-Frelet [97] publié peu avant le début de ma thèse. Les auteurs ont proposé une nouvelle méthode hybridable par éléments finis discontinus avec des flux numériques upwind, appelée méthode CHDG, où, au lieu de traces numériques, les variables caractéristiques jouent le rôle de variables hybrides. Cette approche innovante permet de mieux exploiter les propriétés physiques des ondes, en conservant les informations essentielles sur la direction et la vitesse de propagation. La méthode a montré une amélioration globale de la vitesse de convergence des procédures itératives par rapport aux méthodes DG et même aux méthodes HDG standard. De plus, cette amélioration pourrait faciliter, à terme, l'application de la méthode dans certains contextes pratiques où la rapidité de calcul constitue un critère important.

Dans ce contexte, l'objectif de cette thèse consiste à développer certains des nombreux axes d'amélioration que nous venons d'énumérer.

- Inspiré par l'étude encourageante [97], nous visons à étendre la théorie de la méthode CHDG aux problèmes généraux en régime harmonique avec un accent pour les tests numériques sur l'aéroacoustique dans les milieux homogènes. Cette extension permettra de valider la robustesse de la méthode dans différents scénarios physiques.

- Inspiré par des études récentes concernant les méthodes de décomposition de domaine [98], nous ciblons une autre extension de la méthode CHDG qui inclut des variables de transmission d'ordre élevé avec un focus pour la théorie et les tests numériques sur l'acoustique dans les milieux homogènes et hétérogènes.

- Avec l'idée de combiner des traitements d'interface adaptés aux ondes avec une méthode de type Trefftz, nous étudions la combinaison de conditions aux limites absorbantes d'ordre élevé avec une formulation UWVF (ultra weak variational formulation) utilisant des ondes planes propagatives comme fonctions de base. Cette combinaison vise à réduire les réflexions parasites aux frontières et d'améliorer la précision des simulations sur des domaines finis. Elle peut être vue comme une première étape pour l'utilisation d'opérateurs de transmission d'ordre élevé aux interfaces entre les éléments.

Structure de la thèse

Cette thèse contient cinq chapitres dont le contenu et les objectifs sont brièvement présentés ci-dessous.

Dans le Chapitre 1, nous présentons le cadre mathématique et numérique de l'acoustique et de l'aéroacoustique. En particulier, dans la première partie, nous introduisons les équations physiques qui modélisent la propagation des ondes dans les régimes dépendant du temps et harmonique ainsi que les conditions aux limites appropriées qui nous permettent de compléter le modèle mathématique. Dans la deuxième partie, nous décrivons les principales méthodes numériques qui visent à traiter les phénomènes de propagation d'ondes, en mettant en évidence les défis et les stratégies qui ont été étudiés dans la littérature et qui sont encore explorés dans les recherches récentes.

Dans le Chapitre 2, nous proposons une extension de la méthode CHDG aux problèmes généraux d'ondes en régime harmonique avec des coefficients constants. La méthode repose sur l'introduction de variables de transmission dont la définition est déterminée par l'équation physique du problème spécifique. En particulier, nous utilisons une méthode de splitting qui est équivalente à la résolution d'un problème de Riemann défini localement. Les variables de transmission sont reliées aux variables caractéristiques associées à ce problème et connues dans le contexte des systèmes hyperboliques. Elles sont définies localement pour chaque face de chaque élément et peuvent être interprétées comme des informations se propageant vers l'extérieur et vers l'intérieur à travers la face elle-même.

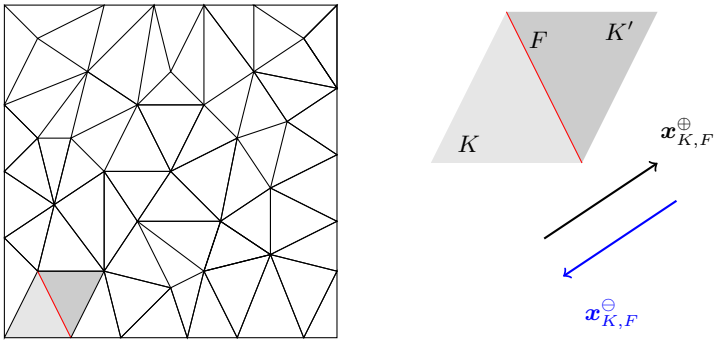


Figure 2: Exemple de maillage non stucturé et notation sur deux éléments voisins pour les variables de transmission.

La procédure d'hybridation consiste à éliminer les champs physiques grâce à la résolution de problèmes locaux et elle conduit à un système dont l'unique inconnue $\boldsymbol{x}$ correspond à la variable hybride définie sur le squelette du maillage. Le système hybridé peut être écrit sous la forme $(\mathrm{I}-\Pi\mathrm{S})\boldsymbol{x}=\boldsymbol{b}$ et, en prouvant que l'opérateur $\Pi\mathrm{S}$ est une contraction stricte, nous avons que le problème peut être résolu a priori par une itération à point fixe. En particulier, nous montrons que l'opérateur est une contraction stricte sous une hypothèse concernant les conditions aux limites qui peut être moralement interprétée par des questions énergétiques et qui peut être motivée par la nécessité d'avoir la frontière du domaine globalement passive (sans augmentation d'énergie possible).

Dans le Chapitre 3 nous appliquons la méthode CHDG au cadre aéroacoustique en régime harmonique en présence d'un écoulement moyen uniforme. En retraçant toutes les étapes théoriques du Chapitre précédent, nous mettons en évidence les difficultés inhérentes à l'aéroacoustique, nous donnons des détails sur les variables de transmission et la méthode numérique qui en résulte et nous énonçons l'hypothèse selon laquelle tous les résultats théoriques sont vrais.

En particulier, la présence et l'orientation de l'écoulement moyen jouent un rôle crucial dans la définition des variables de transmission, dans la représentation des modes de vorticité (ainsi que des modes acoustiques) et elles contrôlent la propriété de contraction de l'opérateur ΠS.

Les variables de transmission sont composées de deux composantes (normale et transverse) qui dépendent de l'écoulement moyen et correspondent à des modes acoustiques et hydrodynamiques. Il s'agit d'une différence (et généralisation) importante par rapport au contexte acoustique étudié dans [97] : en l'absence d'un écoulement moyen, les variables de transmission ne correspondent qu'à des modes acoustiques. En outre, pour conserver la propriété de contraction de l'opérateur ΠS, il est nécessaire que l'écoulement moyen soit sortant sur les portions de la frontière sur lesquelles des conditions aux limites de type Dirichlet ou Neumann sont imposées.

Des cas test bidimensionnels sont étudiés pour évaluer et comparer la performance numérique des algorithmes itératifs pour le système CHDG par rapport au système DG. En particulier, nous étudions l'effet de la présence d'un écoulement moyen sur la performance (et donc sur la vitesse de convergence) des schémas itératifs pour chaque cas test.

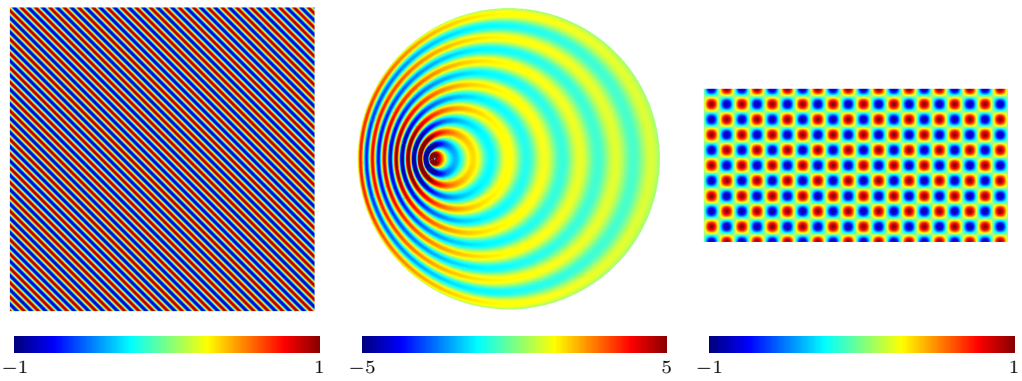


Figure 3: Solutions de référence pour les cas test du Chapitre 3 avec un écoulement moyen.

Dans le Chapitre 4 nous étendons la méthode CHDG pour traiter la propagation d'ondes dans des milieux hétérogènes avec des coefficients physiques constants par morceaux. Outre les flux numériques upwind, des flux symétriques généraux, incluant éventuellement un opérateur différentiel d'ordre élevé $\mathcal{A}$ dans leur définition, sont également considérés. Cet opérateur peut être interprété comme un opérateur d'impédance. Nous appelons ces flux " symétriques " car les images de l'opérateur $\mathcal{A}$, qui agit sur les deux côtés de chaque face du maillage, sont identiques.

Bien qu'il soit possible de prouver que le problème local est bien posé pour les deux flux numériques, nous ne pouvons prouver la propriété de contraction stricte de l'operator ΠS que dans le cas de flux numériques symétriques sous certaines hypothèses concernant l'opérateur $\mathcal{A}$, en particulier le fait qu'il soit positif et auto-adjoint.

Comme dans le Chapitre précédent, des cas test bidimensionnels simples sont utilisés pour étudier et comparer la performance numérique des algorithmes itératifs avec différents flux numériques pour les systèmes CHDG, DG et HDG standard. Dans l'étude numérique, nous sélectionnons en particulier deux opérateurs $\mathcal{A}$: un opérateur d'ordre zéro (une constante) et un opérateur d'ordre deux. Ces deux opérateurs constituent des approximations de l'opérateur Dirichlet-to-Neumann (DtN) associé à la troncature d'une frontière droite, qui permettent di modéliser l'absorption des ondes dans le cadre des conditions aux limites absorbantes. Ce choix est motivé par des travaux récents sur les méthodes de décomposition de domaine, lesquels ont montré qu'un opérateur de transmission d'ordre plus élevé le long des interfaces séparant les sous-domaines peut apporter une amélioration notable de la vitesse de convergence des méthodes itératives.

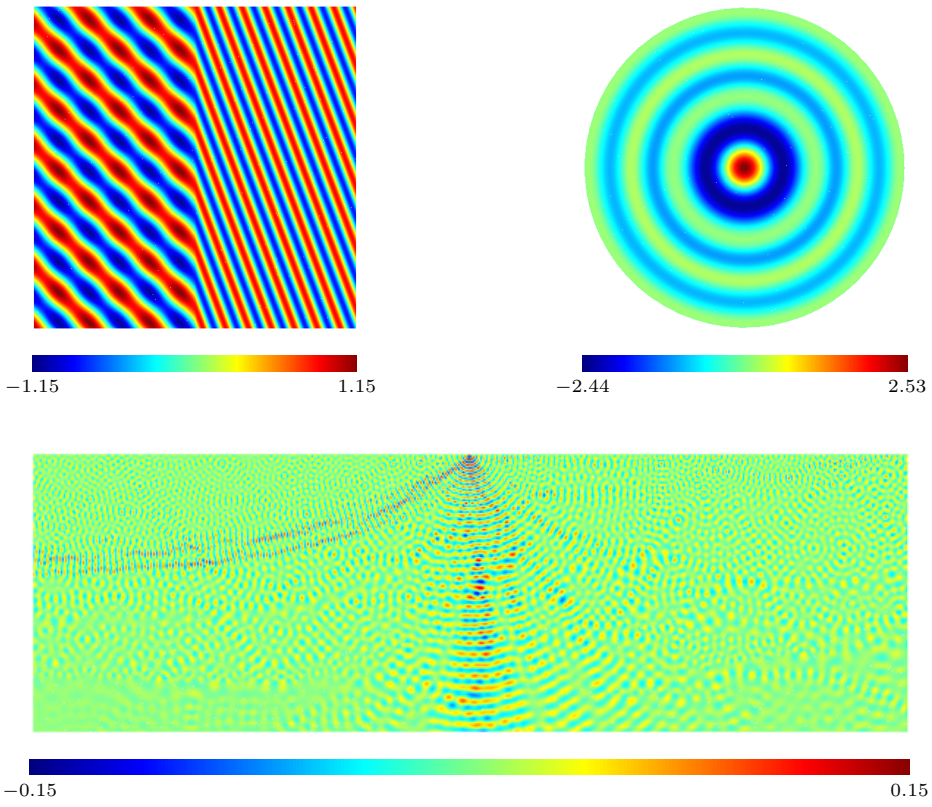


Figure 4: Solutions de référence pour les cas test du Chapitre 4 avec des milieux hétérogènes.

Dans le Chapitre 5, nous décrivons une méthode de Trefftz avec des fonctions de base d'ondes planes propagatives basée sur la formulation variationnelle ultra faible pour l'équation de Helmholtz complétée par des conditions aux limites absorbantes (ABC). Outre les ABC d'ordre faible, les ABC d'ordre élevé de type Padé sont également utilisées. Dans ce contexte, nous rappelons brièvement les motivations qui conduisent au choix de cette combinaison de techniques numériques, notamment leur capacité à gérer efficacement des régimes oscillatoires complexes et à réduire les réflexions parasites aux frontières artificielles du domaine. De plus, la présentation souligne comment ces outils s'intègrent naturellement dans les stratégies contemporaines de résolution d'équations aux dérivées partielles en haute fréquence.

Les ondes planes propagatives résolvent l'équation de Helmholtz homogène et sont caractérisées par une direction de propagation qui les distinguent. Cette propriété directionnelle joue un rôle essentiel dans la représentation locale des champs d'ondes, car elle permet une approximation ciblée des phénomènes de propagation dominants dans chaque sous-domaine. Cette observation justifie l'intérêt de sélectionner un ensemble de directions suffisamment riche pour capturer correctement la physique du problème, tout en maintenant un coût numérique raisonnable.

Des cas test bidimensionnels sont considerés pour étudier la performance numérique de la méthode, pour mettre en évidence les difficultés et problèmes possibles et pour explorer empiriquement l'impact à la fois des fonctions de base des ondes planes et des différents types d'ABC sur la précision. L'adaptivité directionnelle est également partiellement explorée pour comprendre son impact et son potentiel en termes d'amélioration de la précision de la méthode. Ces études numériques incluent une comparaison systématique des configurations testées, ce qui permet d'identifier plus clairement les situations dans lesquelles l'approche Trefftz offre un avantage notable. Les résultats obtenus servent ainsi de base à une réflexion plus large sur l'optimisation des choix directionnels et sur la pertinence des ABC utilisées dans des scénarios variés.

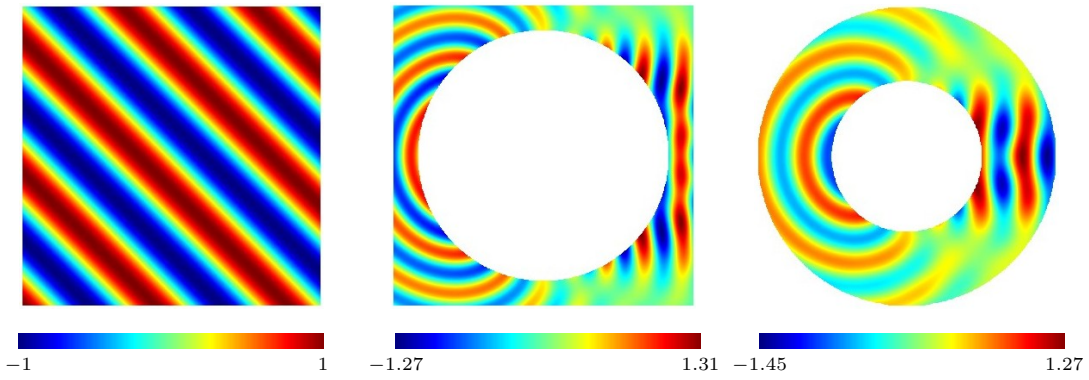


Figure 5: Solutions de référence pour les cas test du Chapitre 5.

Nous complétons la thèse avec des conclusions générales et en indiquant quelques directions naturelles pour les travaux de recherche futurs, en soulignant notamment les perspectives qui découlent des résultats obtenus et les pistes qui pourraient être explorées pour approfondir ou étendre les méthodes présentées à des scénarios de plus en plus réalistes et appliqués.

Code

Les expériences numériques présentées tout au long de cette thèse sont basées sur les méthodes d'éléments finis. Le solveur a été implémenté dans MATLAB, tandis que la génération de maillage a été effectuée à l'aide du logiciel gmsh.

Le code implémenté pour le Chapitres 3 et le Chapitre 4 correspond à des modules supplémentaires qui ont été ajoutés à un code existant appelé *WavesDGlab*. Il est maintenu par Axel Modave et d'autres chercheurs et doctorants ont contribué à son développement.

Pour le Chapitre 5, j'ai implémenté l'ensemble de solveur, à l'exception des fonctions de calcul de l'erreur numérique et pour tracer les solutions exactes et numériques, qui ont été fournies respectivement par Axel Modave et Gwénaël Gabard.

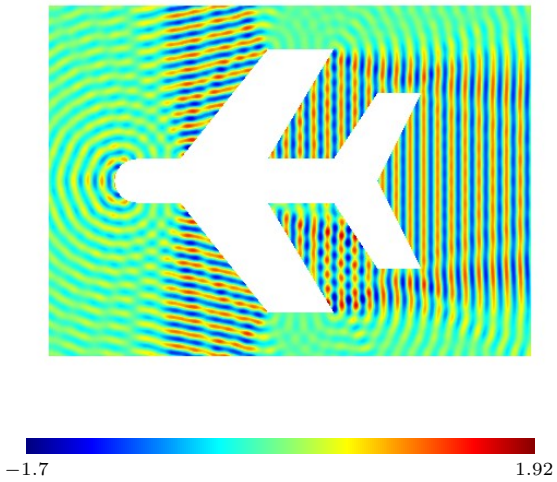


Figure 6: Solutions de référence pour un cas test complexe du Chapitre 5.

Publications et présentations à des conférences

Le contenu du Chapitre 4 a été publié dans un compte rendu de conférence et dans une revue à comité de lecture :

- S. Pescuma, G. Gabard, T. Chaumont-Frelet, A. Modave,
  A hybridizable discontinuous Galerkin method with transmission variables for time-harmonic wave problems in heterogeneous media,
  Proceeding of WAVES 2024 (2 pages) ;

- S. Pescuma, G. Gabard, T. Chaumont-Frelet, A. Modave,
  A hybridizable discontinuous Galerkin method with transmission variables for time-harmonic wave problems in heterogeneous media,
  Journal of Computational Physics, 534:114009, 2025 (19 pages).

De plus, la majorité de ces travaux a été présentée lors de plusieurs conférences internationales :

- *7th Chilean Workshop on Numerical Analysis for Partial Differential Equations*
  WONAPDE 2024 - Concepción, Chili ;

- *16th International Conference on Mathematical and Numerical Aspects of Wave Propagation*
  WAVES 2024 - Berlin, Allemagne ;

- *29th International Conference on Domain Decomposition Methods*
  DD29 2025 - Milan, Italie.

Introduction

General context

Finite element methods (FEM) have been extensively employed over the past four decades in the analysis of time-dependent and time-harmonic wave propagation, covering a wide range of physical contexts including acoustic, electromagnetic and elastic waves. Early applications focused on interior problems involving complex geometries, such as structural-acoustic systems, waveguides and resonant cavities, primarily for the analysis of forced vibrations, scattering, and frequency response. Over time, significant advancements have extended the applicability of FEM to exterior problems in unbounded domains as well as heterogeneous media.

Main challenges

A central challenge in exterior problems is managing the unbounded nature of the domain. FEM typically addresses this by introducing an artificial boundary that truncates the infinite domain to a finite one, creating a bounded computational domain. However, reducing the size of this domain must be carefully balanced against the need to prevent artificial reflections at the boundary. Efficient and flexible boundary treatments are required to minimize such errors without increasing computational cost. Classic approaches include infinite elements [14, 19], absorbing layers [35, 103] and (possibly nonlocal) absorbing boundary conditions [45, 98].

In addition, there have been considerable efforts to improve FEM for high-frequency problems, where traditional FEM struggles to resolve highly oscillatory wavefields in an accurate way. Indeed, at high frequencies, a phenomenon typical of wave-propagation problems, called pollution effect, deteriorates the numerical solution by generating an error that spreads throughout the domain [79, 81]. To address this issue, many advanced FEM have been introduced over the last decades. These include high-order finite elements [3, 13], stabilized Galerkin methods [66], multiscale variational formulations [76], and wave-based discretizations which rely on the intuition that the incorporation of wave physics into the numerical formulation can globally enhance accuracy and efficiency.

Another issue related to FEM is that the matrix associated to the resulting linear system is generally complex, non-Hermitian and indefinite. Direct solvers quickly become too costly for high-frequency or large three-dimensional problems, where the number of degrees of freedom can be extremely large. However, the use of iterative methods in this context is also a challenge: the poor conditioning of the matrix can severely affect the speed of convergence [49]. To overcome this difficulty, various strategies have been developed, including the use of preconditioners specifically tailored for Helmholtz-type operators [46, 47] and domain decomposition methods [57]. Additionally, hybridization techniques have proven to be effective in enhancing the convergence of iterative solvers for large-scale, high-frequency problems.

Discontinuous Galerkin finite element methods

Compared to continuous Galerkin finite element methods (CG), discontinuous Galerkin methods (DG) are very versatile because they allow the so-called numerical fluxes, that deeply influence the accuracy of the DG schemes, to be defined in multiple ways as well as non-standard basis functions within each element. Consequently, they offer a wide framework for formulations that facilitate the use of fast iterative algorithms.

In standard hybridizable discontinuous Galerkin methods (HDG) [8, 33], an additional variable corresponding to the numerical trace of the solution is introduced. This hybrid variable acts as a Lagrange multiplier that decouples the physical unknowns between the elements. The elimination of the physical fields is achieved by solving local problems within the elements. In practice, after inverting element-wise local matrices, a reduced system involving only the hybrid variable is formulated over the skeleton of the mesh. Once the solution of this hybridized (and usually smaller) global system is obtained, the solution of the original problem can be recovered in a post-processing step. When using direct solvers, the advantage of this approach is straightforward, as the reduced HDG system often features far less degrees of freedom than the original DG system while preserving its sparse pattern. On the other hand, for iterative solvers, where matrix size and structure are no longer the main performance factors, the situation becomes more complex. However several studies exhibited a significant improvement, namely a faster convergence [97]. Moreover, since the local problems can be solved independently, the method enables parallelism which can lead to further improvements.

The Trefftz method is a wave-based method that takes solutions of the homogeneous Helmholtz equation as trial and test functions on each element [23, 72, 93], propagative plane waves for instance. It achieves rapid (often exponential) convergence [73] and, in certain cases, needs fewer degrees of freedom than polynomial-based FEM for similar accuracy, especially in homogeneous or piecewise-homogeneous domains [92]. Adaptively tuning propagation directions helps capture the dominant wavefield [39]. However, purely propagative plane waves lead to numerical ill-conditioning [7]: as their number grows, they become nearly linearly dependent, inflating condition numbers and risking instability without regularization [105]. Because they cannot represent evanescent modes, limiting its use near singularities or in variable-coefficient media, richer bases, such as evanescent waves or quasi-Trefftz functions [83, 84, 119], are currently being investigated.

Goal of the study

This thesis is part of the *WavesDG* project. It is a four-year project that involves several researchers and PhD students and whose goal is to study and develop wave-specific discontinuous Galerkin methods for wave-propagation phenomena. The main directions of research include the study of wave-tailored interface treatments based on both low-order and high-order transmission conditions, as well as wave-specific basis functions. These directions of research targeted for waves are motivated by the aim of improving the properties of the linear system associated to the numerical schemes, possibly leading to efficient iterative algorithms that can be executed on parallel architectures that enable coping with realistic large-scale problems.

The first contribution to the *WavesDG* project is a paper by Axel Modave and Théophile Chaumont-Frelet [97] published shortly before the beginning of my PhD. The authors proposed a new hybridizable discontinuous Galerkin method with upwind numerical fluxes, called CHDG method, where, instead of numerical traces, characteristic variables play the role of hybrid variables. The method exhibited an overall improvement of the speed of convergence of iterative procedures once compared to DG and even standard HDG methods.

In this context, the goal of this thesis consists in developing some of the many axes of improvement we have just listed.

- Inspired by the encouraging study [97], we aim at extending the theory of the CHDG method using transmission variables to general linear time-harmonic problems with a focus on numerical tests on aeroacoustic problems with homogeneous media.
- Inspired by recent studies regarding domain decomposition methods [98], we target another extension of the CHDG method that includes high-order transmission variables with a focus on the theory and the numerical tests on acoustics for both homogeneous and heterogeneous media.
- With the idea of combining wave-tailored interface treatments with a wave-based method, we first concentrate on a Trefftz method with propagative plane-wave basis functions applied to Helmholtz problems complemented by high-order absorbing boundary conditions.

Structure of the thesis

This thesis contains five chapters whose contents and goals are briefly presented here.

In Chapter 1 we present the mathematical and numerical framework of acoustic and aeroacoustics. In particular, in the first part, we introduce the physical equations that model wave propagation in both time-dependent and time-harmonic regimes as well as appropriate boundary conditions that complete the mathematical model. In the second part, we describe the main numerical methods that aim at dealing with wave propagation phenomena, highlighting the challenges and the strategies that have been studied in the literature and that are still being investigated in recent research.

In Chapter 2 we propose an extension of the CHDG method to general time-harmonic wave problems with constant coefficients. The method relies on the introduction of transmission variables whose definition is driven by the physical equations of the specific problem. The hybridization procedure consists in the elimination of the physical fields thanks to the resolution of local problems and it leads to a system whose unique unknown $\boldsymbol{x}$ corresponds to the hybrid variable defined on the skeleton. The hybridized system can be written in the form $(\mathrm{I} - \Pi\mathrm{S})\boldsymbol{x} = \boldsymbol{b}$ and, by proving that the operator $\Pi\mathrm{S}$ is a strict contraction, we show that the problem can be a priori solved by a fixed-point iteration.

In Chapter 3 we apply the CHDG method to the propagation of time-harmonic aeroacoustic waves in the presence of a uniform mean flow. By retracing all the theoretical steps of the previous Chapter, we highlight the difficulties inherent to aeroacoustics, we give details on the transmission variables and the resulting numerical method and we state the hypothesis under which all the theoretical results hold. In particular, the presence and orientation of the base flow plays a crucial role in the definition of the transmission variables, in the representation of vorticity modes (as well as acoustic ones) and it controls the strict contraction property of the operator $\Pi\mathrm{S}$. Two-dimensional benchmarks are studied to investigate and compare the numerical performance of iterative algorithms for the CHDG system with respect to the DG one.

In Chapter 4 we extend the CHDG method to waves in heterogeneous media with piecewise constant physical coefficients. Besides upwind numerical fluxes, general symmetric ones, possibly including a high-order differential operator $\mathcal{A}$ in their definition, are considered too. Although it is possible to prove that the local problem is well-posed for both numerical fluxes, we can prove the strict contraction property only in the case of symmetric numerical fluxes under certain hypothesis regarding the operator $\mathcal{A}$. As in the previous Chapter, simple two-dimensional benchmarks are used to investigate and compare the numerical performance of iterative algorithms with different numerical fluxes for the CHDG, DG and standard HDG systems.

In Chapter 5 we describe a Trefftz method with propagative plane-wave basis functions based on the ultra weak variational formulation for the Helmholtz equation complemented by absorbing boundary conditions (ABCs). Besides low-order ABCs, Padé-type high-order ABCs are also taken into account. Two-dimensional benchmarks are studied to investigate the numerical performance of the method, to highlight possible difficulties and issues and to empirically explore the impact of both the plane-wave basis functions and the different kinds of ABCs on the accuracy.

We complete the thesis with general conclusions and indicating some natural directions for future research work.

Code

The numerical experiments presented throughout this thesis are based on finite element methods. The core solver was implemented in MATLAB, whereas mesh generation was performed using the software gmsh.

In particular, the code implemented for Chapter 3 and Chapter 4 consists in additional modules that have been added to an existing code repository called *WavesDGlab*. It is maintained by Axel Modave and some researchers and PhD students have been contributing to its development. For Chapter 5, I implemented all the code from scratch, except for the functions in charge of computing the numerical error and plotting the exact and numerical solutions, which were kindly provided respectively by Axel Modave and Gwénaël Gabard.

Publications and conference presentations

The content of Chapter 4 has been published in a conference proceeding and in a peer-reviewed journal:

- S. Pescuma, G. Gabard, T. Chaumont-Frelet, A. Modave,
  A hybridizable discontinuous Galerkin method with transmission variables for time-harmonic wave problems in heterogeneous media,
  Proceeding of WAVES 2024 (2 pages);

- S. Pescuma, G. Gabard, T. Chaumont-Frelet, A. Modave,
  A hybridizable discontinuous Galerkin method with transmission variables for time-harmonic wave problems in heterogeneous media,
  Journal of Computational Physics, 534:114009, 2025 (19 pages).

Moreover, the majority of this work has been presented at dedicated mini-symposia of some international conferences:

- *7th Chilean Workshop on Numerical Analysis for Partial Differential Equations*
  WONAPDE 2024 - Concepción, Chile;

- *16th International Conference on Mathematical and Numerical Aspects of Wave Propagation*
  WAVES 2024 - Berlin, Germany;

- *29th International Conference on Domain Decomposition Methods*
  DD29 2025 - Milan, Italy.

Chapter 1

Models and finite element solvers for time-harmonic acoustics and aeroacoustics

In this first Chapter, we present both a review of the current state of the art and a detailed exposition of the theoretical and numerical tools that constitute the foundation of this work. The aim is to provide a comprehensive overview of the key concepts, methods, and results from the literature, as well as to introduce the specific approaches and topics that are further developed throughout the thesis. Therefore, in Section 1.1 we introduce the necessary definitions and equations from fluid mechanics and, proceeding by steps, we retrieve the fundamental equations of acoustics and aeroacoustics. In Section 1.2, we review the main techniques and general framework for the numerical simulation of wave phenomena, including standard finite element methods, such as continuous and discontinuous Galerkin methods, and alternative ones, such as Trefftz methods. Lastly, in Section 1.3, we present some numerical algorithms for solving in an efficient way the algebraic linear systems associated to the finite element methods.

1.1 Mathematical models

In the first part of this Section, starting from fundamental laws of physics, we introduce the physical equations that model acoustic and aeroacoustic phenomena in the free field. In the last part, we complete them by describing appropriate and physical boundary conditions and, to make its numerical approximation possible, we will set it in a bounded domain.

1.1.1 Fundamental equations of fluid mechanics

Let t denote the time and $\boldsymbol{x}$ a point of the space.

Conservation laws

Equation (1.1) represents the *conservation of mass* (and it is called *continuity equation*) and involves the mass density ρ and the velocity $\boldsymbol{u}$ (their product $\rho\boldsymbol{u}$ represents the mass flux). It shows that any change in the density is balanced by an expansion or compression. It reads

$$\begin{aligned} &\frac{\partial \rho}{\partial t} + \boldsymbol{\nabla} \cdot (\rho \boldsymbol{u}) = 0, \qquad \text{(conservative form)} \\ &\frac{D\rho}{Dt} + \rho \boldsymbol{\nabla} \cdot \boldsymbol{u} = 0, \qquad \text{(non-conservative form)} \end{aligned} \tag{1.1}$$

where $D/Dt = \partial/\partial t + \boldsymbol{u} \cdot \boldsymbol{\nabla}$ denotes the material derivative.

If the fluid is assumed to be incompressible, namely ρ is constant in both time and space, then the continuity equation reduces to $\boldsymbol{\nabla} \cdot \boldsymbol{u} = 0$, which provides a kinematic condition on the fluid motion.

Equation (1.2) represents the *conservation of momentum* and involves the external force $\boldsymbol{f}$ acting on the fluid and the internal stress described by the Cauchy stress tensor $\boldsymbol{\sigma}$. It states that any change in the momentum $\rho\boldsymbol{u}$ of a fluid element is balanced by the action of external or internal forces. It reads

$$\begin{aligned} \frac{\partial \rho \boldsymbol{u}}{\partial t} + \boldsymbol{\nabla} \cdot (\rho \boldsymbol{u} \otimes \boldsymbol{u} - \boldsymbol{\sigma}) &= \boldsymbol{f}, \qquad \text{(conservative form)} \\ \rho \frac{D\boldsymbol{u}}{Dt} - \boldsymbol{\nabla} \cdot \boldsymbol{\sigma} &= \boldsymbol{f}. \qquad \text{(non-conservative form)} \end{aligned} \tag{1.2}$$

Equation (1.3) represents *conservation of energy* and involves the heat flux $\boldsymbol{q}$ (which we will assume to depend on the gradient of the temperature) and the work done by the external force $\boldsymbol{f}$ and the internal stress $\boldsymbol{\sigma}$ (which we will assume to depend on the pressure and on the gradient of the velocity $\boldsymbol{u}$). Denoting by e the specific energy density, it states that any change in the energy density ρe is balanced by an external supply of heat (we only consider thermal conduction) or mechanical power. It reads

$$\begin{aligned} \frac{\partial \rho e}{\partial t} + \boldsymbol{\nabla} \cdot (\rho e \boldsymbol{u} + \boldsymbol{q} - \boldsymbol{\sigma} \cdot \boldsymbol{u}) &= \boldsymbol{f} \cdot \boldsymbol{u}, \qquad \text{(conservative form)} \\ \rho \frac{De}{Dt} = \boldsymbol{f} \cdot \boldsymbol{u} + \boldsymbol{\nabla} \cdot (\boldsymbol{\sigma} \cdot \boldsymbol{u}) - \boldsymbol{\nabla} \cdot \boldsymbol{q}. &\qquad \text{(non-conservative form)} \end{aligned} \tag{1.3}$$

The Navier-Stokes equations

A general model for compressible flows consists in the *Navier–Stokes equations* which describe the flow of a compressible, viscous and heat conducting fluid. They are the set of the conservation of mass (1.1), of momentum (1.2) and of energy (1.3).

They are supplemented by a constitutive equation, namely an equation for $\boldsymbol{\sigma}$, that, for a Newtonian fluid, reads

$$\boldsymbol{\sigma} = -p\mathbf{I} + \boldsymbol{\tau}, \text{ with } \boldsymbol{\tau} = \mu \left[(\boldsymbol{\nabla} \boldsymbol{u}) + (\boldsymbol{\nabla} \boldsymbol{u})^T - \frac{1}{3} (\boldsymbol{\nabla} \cdot \boldsymbol{u}) \mathbf{I} \right] + \zeta (\boldsymbol{\nabla} \cdot \boldsymbol{u}) \mathbf{I},$$

where p is the pressure in the fluid, $\mathbf{I}$ is the identity tensor, $\boldsymbol{\tau}$ is the viscous strain tensor and μ and ζ are the shear and bulk viscosity, respectively.

The definition of the heat flux is based on Fourier's law

$$\boldsymbol{q} = -\lambda \boldsymbol{\nabla} T,$$

where λ is the heat conductivity that we assume to be independent of the temperature T.

We assume the fluid to be an ideal gas (we neglect the interactions between molecules) and, therefore, the following thermodynamic equation of state holds

$$p = \rho R T,$$

where R is the specific ideal gas constant.

1.1.2 The Euler and linearised Euler equations

The Euler equations

If molecular diffusion processes, namely viscosity and heat conduction, are neglected (i.e. $\boldsymbol{\tau} = \mathbf{0}$ and $\boldsymbol{q} = \mathbf{0}$), then the flow is said to be isentropic and the governing equations can be simplified to the *Euler equations*. In particular, the conservation of energy is rewritten in terms of pressure by exploiting the relation $Dp/Dt = c^2 D\rho/Dt$, where c denotes the sound speed, that holds for isentropic flows and the continuity equation. The Euler equations in non-conservative form read

$$\begin{cases} \dfrac{D\rho}{Dt} + \rho \boldsymbol{\nabla} \cdot \boldsymbol{u} = 0, \\ \rho \dfrac{D\boldsymbol{u}}{Dt} + \boldsymbol{\nabla} p = \boldsymbol{f}, \\ \dfrac{Dp}{Dt} + \rho c^2 \boldsymbol{\nabla} \cdot \boldsymbol{u} = 0. \end{cases}$$

The linearised Euler equations

In aeroacoustics, we generally consider the propagation of small-amplitude fluctuations superimposed on a steady base flow. In particular, the flow is decomposed into two parts as follows

$$\begin{aligned} \rho(\boldsymbol{x}, t) &= \rho_0(\boldsymbol{x}) + \rho'(\boldsymbol{x}, t), \\ \boldsymbol{u}(\boldsymbol{x}, t) &= \boldsymbol{u}_0(\boldsymbol{x}) + \boldsymbol{u}'(\boldsymbol{x}, t), \\ p(\boldsymbol{x}, t) &= p_0(\boldsymbol{x}) + p'(\boldsymbol{x}, t). \end{aligned} \tag{1.4}$$

Background flow quantities are denoted with a subscript 0. In the following, the background flow is assumed to be constant, uniform (i.e. ρ_0, p_0 and $\boldsymbol{u}_0$ are independent of $\boldsymbol{x}$ and t) and subsonic (i.e. $\|\boldsymbol{u}_0\| < c_0$, where c_0 is the sound speed in free field). The fluctuations are denoted with $'$ and are assumed to be small compared to the background flow quantities.

Classical acoustics, concerned with the propagation of small-amplitude sound waves in a uniform fluid at rest, is a special case of this flow decomposition and is obtained by removing the background flow, i.e. $\boldsymbol{u}_0 = \mathbf{0}$.

Using the assumption of small perturbations and of uniform background flow, it is possible to neglect any terms that contain products of fluctuating quantities which leads to the *linearised Euler equations* (LEE)

$$\begin{cases} \dfrac{D_0\rho'}{Dt} + \rho_0 \boldsymbol{\nabla} \cdot \boldsymbol{u}' = 0, \\ \rho_0 \dfrac{D_0\boldsymbol{u}'}{Dt} + \boldsymbol{\nabla} p' = \boldsymbol{f}, \\ \dfrac{D_0 p'}{Dt} + \rho_0 c_0^2 \boldsymbol{\nabla} \cdot \boldsymbol{u}' = 0. \end{cases} \tag{1.5}$$

where $D_0/Dt = \partial/\partial t + \boldsymbol{u}_0 \cdot \boldsymbol{\nabla}$ denotes the material derivative with respect to the mean velocity $\boldsymbol{u}_0$ and it represents the rate of change in time seen by an observer moving with the mean flow.

To summarize, these equations describe the propagation of any small-amplitude fluctuations on a steady uniform base flow of an inviscid fluid. From now on, we will focus on these fluctuations and we remove the $'$ to simplify the notation.

1.1.3 The acoustic wave, the convected wave and the vorticity equations

It is possible to eliminate the density ρ and the velocity $\boldsymbol{u}$ and obtain an equation with respect to the pressure p by taking the material derivative of the third equation of (1.5) and combining

it with the second equation of (1.5) (here we assume $\boldsymbol{f} = \boldsymbol{0}$ for simplicity), which yield to the so-called *convected wave equation*

$$\frac{1}{c_0^2}\frac{D_0^2 p}{Dt^2} - \Delta p = 0. \tag{1.6}$$

Equation (1.6) models the propagation of sound waves in a uniform mean flow. Moreover, when the mean flow is zero, i.e. $\boldsymbol{u}_0 = \boldsymbol{0}$, the material derivative D_0/Dt reduces to the standard time derivative $\partial/\partial t$, and we recover the *classical wave equation*

$$\frac{1}{c_0^2}\frac{\partial^2 p}{\partial t^2} - \Delta p = 0.$$

To obtain an equation that models the propagation of vorticity waves, we take the curl of the second equation of (1.5) (again, we assume $\boldsymbol{f} = \boldsymbol{0}$), we introduce the vorticity $\boldsymbol{\omega} := \boldsymbol{\nabla} \times \boldsymbol{u}$ and we get the *vorticity equation*

$$\left(\frac{\partial}{\partial t} + \boldsymbol{u}_0 \cdot \nabla\right) \boldsymbol{\omega} = \boldsymbol{0}.$$

1.1.4 Plane-wave analysis

To better understand the nature of the waves governed by the linearised Euler equations (1.5), we seek time-harmonic plane wave solutions of these equations, namely solutions in the form

$$\begin{bmatrix} \rho \\ \boldsymbol{u} \\ p \end{bmatrix} = \begin{bmatrix} \widehat{\rho} \\ \widehat{\boldsymbol{u}} \\ \widehat{p} \end{bmatrix} e^{-\imath\omega t + \imath \boldsymbol{\kappa}\cdot\boldsymbol{x}},$$

with angular frequency ω and wave vector $\boldsymbol{\kappa}$. This corresponds to applying Fourier transforms in both time and space. The variables $\widehat{\rho}$, $\widehat{\boldsymbol{u}}$ and $\widehat{p}$ represent the constant complex amplitudes of the fluctuating quantities.

Substituting this expression for the solution in the linearised Euler equations (1.5) leads to the following system of linear equations

$$\mathbf{A}(\boldsymbol{\kappa}, \omega)\boldsymbol{w} = \boldsymbol{0}, \quad \text{with } \mathbf{A} = \begin{bmatrix} \boldsymbol{\kappa}\cdot\boldsymbol{u}_0 - \omega & -\rho_o \boldsymbol{\kappa}^T & 0 \\ \boldsymbol{0} & \rho_0(\boldsymbol{\kappa}\cdot\boldsymbol{u}_0 - \omega)\mathbf{I} & -\boldsymbol{\kappa} \\ 0 & -\rho_0 c_o^2 \boldsymbol{\kappa}^T & \boldsymbol{\kappa}\cdot\boldsymbol{u}_0 - \omega \end{bmatrix}, \quad \boldsymbol{w} = \begin{bmatrix} \widehat{\rho} \\ \widehat{\boldsymbol{u}} \\ \widehat{p} \end{bmatrix}.$$

The previous dispersion relation of the linearised Euler equations corresponds to an eigenvalue problem involving the frequency ω, the wave vector $\boldsymbol{\kappa}$ and the eigenvector $\boldsymbol{w}$. We write the wave vector as $\boldsymbol{\kappa} = \kappa\boldsymbol{\theta}$ where $\boldsymbol{\theta}$ is a unit vector whose direction corresponds to the one of the wave vector. For a given angular frequency ω and direction $\boldsymbol{\theta}$, we aim at finding the wavenumbers that lead to non-trivial solutions $\boldsymbol{w}$. To do so, we solve $\det(\mathbf{A}) = 0$ for the wavenumber κ and we obtain the following four solutions

$$\kappa_a^{\pm} = \frac{\omega}{\boldsymbol{u}_0\cdot\boldsymbol{\theta} \pm c_0}, \quad \kappa_v = \kappa_e = \frac{\omega}{\boldsymbol{u}_0\cdot\boldsymbol{\theta}},$$

whose associated eigenvectors (in dimension 2) are

$$\boldsymbol{w}_a^{\pm} = \begin{bmatrix} \rho_0 \\ \pm c_0\theta_x \\ \pm c_0\theta_y \\ \rho_0 c_0^2 \end{bmatrix}, \quad \boldsymbol{w}_v = \begin{bmatrix} 0 \\ -\theta_y \\ \theta_x \\ 0 \end{bmatrix}, \quad \boldsymbol{w}_e = \begin{bmatrix} 1 \\ 0 \\ 0 \\ 0 \end{bmatrix}.$$

The solution $(\kappa_e, \boldsymbol{w}_e)$ corresponds to an entropy wave. The entropy is defined as $dS = \delta Q/T$. We do not give any further detail about this wave and, from now on, we assume homentropy: namely the entropy is constant with respect to both space and time. It yields

$$\rho = \frac{1}{c_0^2} p.$$

Vorticity waves are associated to the solution $(\kappa_v, \boldsymbol{w}_v)$ and generate perturbations only in velocity, leaving density and pressure unchanged and they are simply transported by the steady base flow. Moreover, from the LEE we obtain $\boldsymbol{\nabla} \cdot \boldsymbol{u} = 0$, namely the velocity field is solenoidal and it can only induce rotation on the fluid elements with no expansion or compression.

For sound waves associated to the solutions $(\kappa_a^{\pm}, \boldsymbol{w}_a^{\pm})$, the pressure field, whose wavefront is a plane surface with unit normal vector $\boldsymbol{\theta}$, can be written as

$$p(\boldsymbol{x}, t) = \widehat{p} e^{-\imath\omega t + \imath \boldsymbol{\kappa} \cdot \boldsymbol{x}}, \tag{1.7}$$

where $\boldsymbol{\kappa} = \omega \boldsymbol{\theta}/c_p$ denotes the wave vector, c_p is the phase velocity of the plane wave, namely the velocity of the wavefront in the direction $\boldsymbol{\theta}$ and $\widehat{p}$ is its complex amplitude. Other important parameters of the plane wave are its *wavenumber* κ and its *wavelength* λ defined as

$$\kappa = \|\boldsymbol{\kappa}\| = \frac{\omega}{c_p}, \quad \lambda = \frac{2\pi}{\kappa} = \frac{c_p}{f}.$$

The pressure field (1.7) solves the convected wave equation (1.6) if and only if the phase velocity c_p of the plane wave reads

$$c_p = c_0 \pm \boldsymbol{u}_0 \cdot \boldsymbol{\theta}. \tag{1.8}$$

Alternatively, plugging the pressure field (1.7) in the convected wave equation (1.6) gives

$$\frac{(\omega - \boldsymbol{u}_0 \cdot \boldsymbol{\kappa})^2}{c_0^2} = \|\boldsymbol{\kappa}\|^2.$$

This is again the dispersion relation of plane waves in a uniform flow. By solving it for κ, we get

$$\kappa = \frac{\omega}{\boldsymbol{u}_0 \cdot \boldsymbol{\theta} \pm c_0}$$

which is in accordance with (1.8) (and it coincides with $\kappa_a^{\pm}$).

The phase speed c_p is composed by two terms: c_0 that is the velocity of propagation of sound through the fluid and the convective term $\boldsymbol{u}_0 \cdot \boldsymbol{\theta}$ that is due to the fact that the fluid itself moves at a velocity $\boldsymbol{u}_0$.

We rewrite the above expressions as

$$c_p = (1 + M \cos\alpha) c_0, \quad \kappa = \frac{\kappa_0}{1 + M \cos\alpha}, \quad \lambda = \frac{c_0}{f}(1 + M \cos\alpha), \tag{1.9}$$

in which $M := \|\boldsymbol{u}_0\|/c_0$ is the *Mach number*, $\kappa_0 := \omega/c_0$ is the wavenumber of sound waves in free field and without flow and α is the angle between the wavefront direction $\boldsymbol{\theta}$ and the mean velocity $\boldsymbol{u}_0$. Since the flow is subsonic, we have that $|M| < 1$ and, as a consequence, the phase speed c_p is always positive for any angle α. Sound waves can propagate in every direction, even if sometimes very slowly.

When $0 < \alpha < \pi/2$, the wave propagates with the flow (*downstream propagation*). The phase speed is increased by the mean flow since $\boldsymbol{u}_0 \cdot \boldsymbol{\theta} > 0$, see Equation (1.8). Sound propagation and

the convective effect are interfering in a constructive way. This increase in phase speed results in a smaller wavenumber and a longer wavelength, as shown in (1.9). Vice versa, when $\pi/2 < \alpha < \pi$, the wave propagates against the flow (*upstream propagation*). The wave is decreased by the mean flow since $\boldsymbol{u}_0 \cdot \boldsymbol{\theta} < 0$. Sound propagation and the convective effect are interfering in a partly destructive way. This reduction of the phase speed results in a larger wavenumber and a shorter wavelength.

It is easy to check that the acoustic velocity and density induced by the plane wave defined in Equation (1.7) are given by

$$\boldsymbol{u}(\boldsymbol{x},t) = \frac{\boldsymbol{\theta}}{\rho_0 c_0} p(\boldsymbol{x},t), \quad \rho(\boldsymbol{x},t) = \frac{1}{c_0^2} p(\boldsymbol{x},t). \tag{1.10}$$

The expressions in (1.10) do not depend on the mean flow $\boldsymbol{u}_0$. In particular, the ratio between the acoustic pressure and the velocity, called *impedance* $\eta_0 = p/(\boldsymbol{\theta} \cdot \boldsymbol{u}) = \rho_0 c_0$, does not depend on $\boldsymbol{u}_0$. Moreover, the velocity field is parallel to the wavefront direction $\boldsymbol{\theta}$ and, consequently, a plane wave propagating in the direction $\boldsymbol{\theta}$ is always a longitudinal wave.

1.1.5 Acoustic energy

Energy is an important physical quantity associated to and transported by sound waves, therefore we define the density of energy e and the flux of energy $\boldsymbol{i}$ for an acoustic wave (see Section 1.11 of [109]) as follows

$$\begin{aligned} e &:= \frac{p^2}{2\rho_0 c_0^2} + \frac{1}{2}\rho_0 |\boldsymbol{u}|^2 + \frac{1}{c_0^2} p \boldsymbol{u}_0 \cdot \boldsymbol{u}, \\ \boldsymbol{i} &:= \left(\frac{p}{\rho_0} + \boldsymbol{u} \cdot \boldsymbol{u}_0 \right) \left(\rho_0 \boldsymbol{u} + \frac{1}{c_0^2} p \boldsymbol{u}_0 \right). \end{aligned}$$

These quantities are involved in the following conservation equation

$$\frac{\partial e}{\partial t} + \boldsymbol{\nabla} \cdot \boldsymbol{i} = 0.$$

The total energy E is then obtained by integrating the density e over the considered domain $\Omega \subset \mathbb{R}^d$ and it reads

$$E := \frac{\|p\|^2_{L^2(\Omega)}}{2\rho_0 c_0^2} + \frac{1}{2}\rho_0 \|\boldsymbol{u}\|^2_{L^2(\Omega)} + \left(\frac{1}{c_0^2} p \boldsymbol{u}_0, \boldsymbol{u} \right)_\Omega. \tag{1.11}$$

In classical acoustics, the density e, the flux $\boldsymbol{i}$ and the total energy E reduce to

$$e = \frac{1}{2\rho_0 c_0^2} p^2 + \frac{1}{2}\rho_0 |\boldsymbol{u}|^2, \qquad \boldsymbol{i} = p\boldsymbol{u}, \qquad E = \frac{\|p\|^2_{L^2(\Omega)}}{2\rho_0 c_0^2} + \frac{1}{2}\rho_0 \|\boldsymbol{u}\|^2_{L^2(\Omega)}.$$

The two terms in the density e correspond to the potential energy and kinetic energy, respectively. The flux $\boldsymbol{i}$ is the rate of work done by the acoustic pressure on the fluid elements.

1.1.6 Time-harmonic regime

In the frequency domain, we focus on a single angular frequency $\omega = 2\pi f$, where f is the frequency, and the acoustic physical fields are of the form

$$p(\boldsymbol{x},t) = \mathrm{Re}\{\tilde{p}(\boldsymbol{x}) e^{-\imath\omega t}\} \quad \text{and} \quad \boldsymbol{u}(\boldsymbol{x},t) = \mathrm{Re}\{\tilde{\boldsymbol{u}}(\boldsymbol{x}) e^{-\imath\omega t}\}, \tag{1.12}$$

where $\tilde{p}(\boldsymbol{x})$ and $\tilde{\boldsymbol{u}}(\boldsymbol{x})$ are the complex amplitudes. When we introduce the definition of the pressure field in (1.12) into the convected wave equation (1.6) we obtain the *convected Helmholtz equation*

$$\frac{1}{c_0^2}\frac{D_0^2\tilde{p}}{Dt^2} - \Delta\tilde{p} = 0, \quad \text{with } \frac{D_0}{Dt} = -\imath\omega + \boldsymbol{u}_0\cdot\boldsymbol{\nabla}. \tag{1.13}$$

With no flow, we recover the *classical Helmholtz equation*

$$-\Delta\tilde{p} - \kappa_0^2\tilde{p} = 0.$$

Rather than instantaneous quantities, we are often more interested in time averages. We denote by $\langle e\rangle$ the mean density of energy and by $\langle \boldsymbol{i}\rangle$ the mean flux of energy, which is also called *intensity* of the sound field. In the time-harmonic regime, these quantities read

$$\langle e\rangle = \frac{|\tilde{p}|^2}{4\rho_0 c_0^2} + \frac{1}{4}\rho_0|\tilde{\boldsymbol{u}}|^2 + \frac{1}{2}\,\mathrm{Re}\left\{\frac{1}{c_0^2}\tilde{p}\boldsymbol{u}_0\cdot\overline{\tilde{\boldsymbol{u}}}\right\},$$
$$\langle \boldsymbol{i}\rangle = \frac{1}{2}\,\mathrm{Re}\left\{\left(\frac{\tilde{p}}{\rho_0} + \tilde{\boldsymbol{u}}\cdot\boldsymbol{u}_0\right)\overline{\left(\rho_0\tilde{\boldsymbol{u}} + \frac{1}{c_0^2}\tilde{p}\boldsymbol{u}_0\right)}\right\}.$$

The total energy E (1.11) then becomes

$$E := \frac{\|\tilde{p}\|^2_{L^2(\Omega)}}{2\rho_0 c_0^2} + \frac{1}{2}\rho_0\|\tilde{\boldsymbol{u}}\|^2_{L^2(\Omega)} + \left(\frac{1}{2c_0^2}\tilde{p}\boldsymbol{u}_0, \tilde{\boldsymbol{u}}\right)_\Omega + \left(\frac{1}{2c_0^2}\tilde{\boldsymbol{u}}, \tilde{p}\boldsymbol{u}_0\right)_\Omega. \tag{1.14}$$

Assuming the background flow to be subsonic, i.e. $\|\boldsymbol{u}_0\| < c_0$, it is possible to prove that the energy E is a positive quantity that is equal to 0 if and only if the pressure p and the velocity $\boldsymbol{u}$ are null. Indeed, we have

$$\begin{aligned} E &= \int_\Omega \rho_0\left(\frac{1}{2\rho_0^2c_0^2}\tilde{p}\overline{\tilde{p}} + \frac{1}{2}\tilde{\boldsymbol{u}}\cdot\overline{\tilde{\boldsymbol{u}}} + \frac{1}{2\rho_0c_0^2}\tilde{p}(\boldsymbol{u}_0\cdot\overline{\tilde{\boldsymbol{u}}}) + \frac{1}{2\rho_0c_0^2}(\tilde{\boldsymbol{u}}\cdot\boldsymbol{u}_0)\overline{\tilde{p}}\right)d\boldsymbol{x} \\ &= \int_\Omega \rho_0\left(\frac{1}{2}\left|\tilde{\boldsymbol{u}} + \frac{1}{\rho_0c_0^2}\tilde{p}\boldsymbol{u}_0\right|^2 + \frac{c_0^2 - \|\boldsymbol{u}_0\|^2}{2\rho_0^2c_0^4}|\tilde{p}|^2\right)d\boldsymbol{x}. \end{aligned}$$

In classical acoustics, the mean density $\langle e\rangle$, the mean flux $\langle \boldsymbol{i}\rangle$ and the total energy E reduce to

$$\langle e\rangle = \frac{|\tilde{p}|^2}{4\rho_0c_0^2} + \frac{1}{4}\rho_0|\tilde{\boldsymbol{u}}|^2, \qquad \langle \boldsymbol{i}\rangle = \frac{1}{2}\,\mathrm{Re}\{\tilde{p}\overline{\tilde{\boldsymbol{u}}}\}, \qquad E = \frac{\|\tilde{p}\|^2_{L^2(\Omega)}}{2\rho_0c_0^2} + \frac{1}{2}\rho_0\|\tilde{\boldsymbol{u}}\|^2_{L^2(\Omega)}. \tag{1.15}$$

From now on, we focus on the time-harmonic regime and, consequently, we remove the $\sim$ to simplify the notation.

1.1.7 Boundary conditions and artificial boundary treatment

We need to complement the physical equation with appropriate boundary conditions. In the case of wave propagation in an unbounded domain, we distinguish between boundary conditions that are imposed on the physical boundaries of the domain and those that are imposed at infinity. Notice that, for the sake of clarity, in the following, we simply denote by κ the wavenumber instead of κ_0.

On the physical boundary, three types of conditions are typically considered:

- *Dirichlet* boundary condition (*sound soft* condition):

$$p = f.$$

  It corresponds to imposing the acoustic pressure on the boundary.

- *Neumann* boundary condition (*sound hard* condition):

$$\frac{\partial p}{\partial \boldsymbol{n}} = g.$$

It corresponds to imposing the normal derivative of the acoustic pressure on the boundary and therefore to impose the normal acoustic velocity.

- *Robin/impedance* boundary condition:

$$\alpha p + \beta \frac{\partial p}{\partial \boldsymbol{n}} = h,$$

with $\alpha, \beta \in \mathbb{C}$. A particular choice of the parameters, namely $\alpha = -\imath\kappa$ and $\beta = 1$, leads to

$$\frac{\partial p}{\partial \boldsymbol{n}} - \imath\kappa p = h,$$

and it corresponds to imposing the trace of an incoming wave on the boundary. Moreover, if $h \equiv 0$, we get

$$\frac{\partial p}{\partial \boldsymbol{n}} - \imath\kappa p = 0,$$

which is an approximation of an *absorbing* boundary condition (see later in this Section for further details concerning these two special cases).

At infinity, the following condition is typically imposed:

- *Sommerfeld* radiation condition [121]:

$$\lim_{\|\boldsymbol{x}\|\to\infty} \|\boldsymbol{x}\|^{\frac{d-1}{2}} \left(\nabla p \cdot \frac{\boldsymbol{x}}{\|\boldsymbol{x}\|} - \imath\kappa p \right) = 0.$$

It allows the existence of only outgoing waves which are proportional to $e^{\imath\kappa\|\boldsymbol{x}\|}$ at infinity and, at the same time, it forces the energy flux to be outgoing at infinity which ensures the uniqueness of the solution.

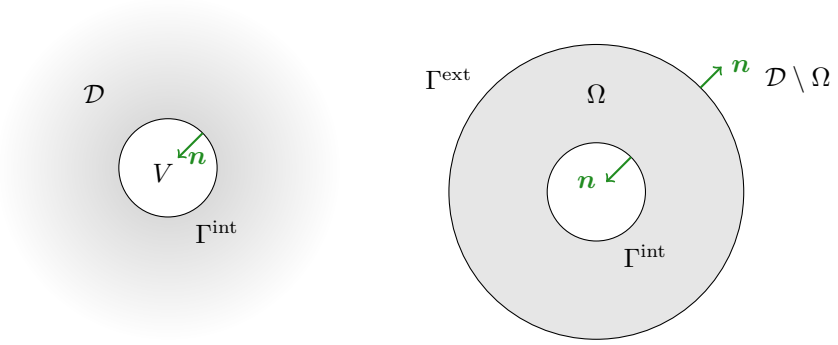
To better understand the complete physical framework, we consider a simple reference problem.

Let $V \subset \mathbb{R}^d$, $d = 2, 3$ be a compact set that represents an impenetrable object with boundary Γ^{int} and inward unit normal $\boldsymbol{n}$. We denote by $\mathcal{D}$ the complement of V, namely $\mathcal{D} = \mathbb{R}^d \setminus \overline{V}$, that represents the exterior domain where V is embedded (see Figure 1.1a). An incident plane-wave p_{inc} that illuminates the scatterer V generates a scattered acoustic field $p : \mathcal{D} \to \mathbb{C}$ that satisfies the following exterior boundary-value problem.

Problem 1.1.1. *Find $p \in L^2(\mathcal{D})$ such that*

$$\begin{cases} -\Delta p - \kappa^2 p = 0, & \text{in } \mathcal{D}, \\ p = f \text{ or } \dfrac{\partial p}{\partial \boldsymbol{n}} = g & \text{on } \Gamma^{int}, \\ \lim\limits_{\|\boldsymbol{x}\|\to\infty} \|\boldsymbol{x}\|^{\frac{d-1}{2}} \left(\nabla p \cdot \dfrac{\boldsymbol{x}}{\|\boldsymbol{x}\|} - \imath\kappa p \right) = 0, \end{cases} \tag{1.16}$$

with $f = p_{inc}$ and $g = -\dfrac{\partial p_{inc}}{\partial \boldsymbol{n}}$.



(a) Unbounded physical domain. (b) Bounded computational domain.

Figure 1.1: Example of a two-dimensional unbounded domain for the scattering problem and numerical setting.

Domain truncation

In the context of finite element methods, it is clearly impossible to consider unbounded domains and, on the contrary, because of computational effort reasons, numerical domains that are as small as possible are preferred.

A very common and natural choice consists in truncating the physical domain by the introduction of an artificial external boundary Γ^{ext} which divides the original unbounded domain in two regions: a bounded numerical domain Ω and its unbounded complement $\mathcal{D} \setminus \Omega$. The boundary condition at the artificial boundary needs to model the effect of the exterior complement $\mathcal{D} \setminus \Omega$. To pursue this aim, several techniques can be considered such as *absorbing boundary conditions* (ABC), *perfectly matched layers* (PML) ([12, 20, 35, 68, 103]), *infinite elements* ([14, 19]) and *finite element method-boundary element method* coupling (FEM-BEM) ([40, 58, 85, 129]). Infinite elements technique consists in a single layer of elements with infinite extent that are constructed with radial wave functions which automatically satisfy the Sommerfeld condition at infinity. The PML method consists in an exterior layer of finite thickness at an artificial interface such that outgoing waves are absorbed by the layer before reaching the artificial boundary. In particular, by splitting the pressure field into nonphysical components that satisfy equations that model decaying waves, and by properly selecting the coefficients associated to the PML, plane-wave reflection for an arbitrary angle of incidence is theoretically eliminated. The FEM-BEM coupling method combines the finite element method for modeling the interior (bounded) domain with the boundary element method to represent the unbounded exterior. BEM inherently satisfies the radiation condition at infinity by formulating equations only on the boundary of the domain, using fundamental solutions that inherently satisfy such a condition at infinity, allowing waves to exit the domain naturally without reflection. In this thesis we focus on the coupling with absorbing boundary conditions.

In the following, we proceed by steps to introduce the formulation of absorbing boundary conditions. For simplicity, from now on we limit ourselves to a two-dimensional reference case. However, all the considerations that follow extend to the three-dimensional case in a straightforward way. Let $\Omega \subset \mathbb{R}^2$ be a domain represented in Figure 1.1b. We consider the following model boundary-value problem set in terms of the Helmholtz equation for $p : \Omega \to \mathbb{C}$

$$\begin{cases} -\Delta p - \kappa^2 p = 0, & \text{in } \Omega, \\ p = f \text{ or } \dfrac{\partial p}{\partial \boldsymbol{n}} = g, & \text{on } \Gamma^{\text{int}}, \\ \dfrac{\partial p}{\partial \boldsymbol{n}} = \mathcal{B} p, & \text{on } \Gamma^{\text{ext}}, \end{cases} \tag{1.17}$$

where Γ^{ext} stands for the external boundary of Ω, Γ^{int} for its complementary, so that $\partial\Omega = \Gamma^{\text{ext}} \cup \Gamma^{\text{int}}$, and $\mathcal{B}$ is a suitable Dirichlet-to-Neumann (DtN) operator that permits to model absorption on Γ^{ext} and that we will characterize later in this Section.

The truncated problem (1.17) should have the same solutions as those of the full one (1.16) restricted to the smaller domain Ω. In the context of waves, this means that the ABCs are to be implemented so that all outgoing waves are allowed to pass the artificial boundary unaffected, in particular without being reflected, while no incoming wave is allowed into the domain from outside.

In order to understand the meaning and retrieve the boundary condition of problem (1.17), which is equivalent to characterize the operator $\mathcal{B}$, we first restrict ourselves to the one-dimensional case. A similar reasoning can be easily extended to the multidimensional one, as we do in the final part of this Section.

Exact absorbing boundary condition in 1D

To begin with, we consider the homogeneous Helmholtz equation in a one-dimensional and bounded domain $\Omega = (0, L)$

$$-p_{xx} - \kappa^2 p = 0, \quad \text{in } \Omega.$$

We can write the solution of the equation as a linear combination of the fundamental solutions of the equation, that is two waves moving in opposite directions

$$p(x) = \alpha e^{-\imath\kappa x} + \beta e^{+\imath\kappa x}, \quad x \in \{0, L\}, \tag{1.18}$$

where, because of the time-dependency convention with $e^{-\imath\omega t}$, $\alpha e^{-\imath\kappa x}$ propagates towards the negative x and $\beta e^{+\imath\kappa x}$ towards the positive x. Following the same idea as in [71], we want the ABC to be satisfied by every outgoing wave, namely it must be valid for any value of α, but disallow a non-zero β at $x = 0$ and the inverse at $x = L$. By imposing such conditions, we get

$$\frac{\partial p}{\partial \boldsymbol{n}} - \imath\kappa p = 0, \quad \text{on } \partial\Omega.$$

Exact absorbing boundary condition for a flat boundary in 2D

There exist many exact non-local boundary conditions for a flat boundary in 2D and the goal of this Section is to derive one of them. The method we analyze is based on the work presented in [45] and later exploited for example in [98].

We consider a Helmholtz problem in the whole 2D-space, that is

$$-\Delta p - \kappa^2 p = f, \quad \text{in } \mathbb{R}^2.$$

We decompose the unbounded domain $\mathbb{R}^2$ into the union of the two half-planes: the interior region $\Omega = \{(x, y) \in \mathbb{R}^2 : x < 0\}$ and the exterior region $\Omega^{\text{ext}} = \{(x, y) \in \mathbb{R}^2 : x > 0\}$ separated by the plane interface Γ whose equation is $x = 0$. The exterior medium is assumed to be homogeneous, that is κ is constant in Ω^{ext}, and free of source, that is f is compactly supported on Ω. The goal is to prescribe a boundary condition on Γ which is able to represent the outward propagation of waves leaving Ω.

We use the classical approach to obtain the exact non-reflecting boundary condition, which is to solve the exterior Helmholtz problem defined in Ω^{ext} for some Dirichlet data $p(y)$ on Γ. Applying the multidimensional transverse Fourier transform $\mathcal{F}_y$ in the y-direction to the Helmholtz equation yields

$$(\partial_x^2 + \lambda^+\lambda^-)\mathcal{F}_y[p](x, \xi_y) = 0, \quad x = 0, \quad \xi_y \in \mathbb{R},$$

where ξ_y is the wavenumber in the y-direction in the Fourier space and $\lambda^\pm$ is such that

$$\lambda^\pm = \begin{cases} \mp\sqrt{\kappa^2 - \xi_y^2}, & \text{if } |\kappa| > |\xi_y|, \\ \mp\imath\sqrt{\xi_y^2 - \kappa^2}, & \text{if } |\kappa| < |\xi_y|. \end{cases}$$

The solution that contains only outgoing traveling waves and bounded evanescent waves reads

$$\mathcal{F}_y[p](x,\xi_y) = \mathcal{F}_y[\overline{p}](\xi_y)e^{x\lambda^+(\xi_y)}.$$

Taking the derivative in x of this solution leads to

$$\partial_x\mathcal{F}_y[p](x,\xi_y) = \lambda^+(\xi_y)\mathcal{F}_y[p](x,\xi_y),$$

and then, using the inverse Fourier transform $\mathcal{F}_y^{-1}$, we have

$$\partial_x p(x,y) = \mathcal{F}_y^{-1}[\lambda^+(\xi_y)\mathcal{F}_y[p](x,\xi_y)].$$

Taking the restriction on Γ gives the exact non-reflective boundary condition for the interior problem,

$$\frac{\partial p}{\partial x} = \mathcal{B}p, \quad \text{on } \Gamma, \tag{1.19}$$

where $\mathcal{B}$ is the pseudo-differential operator defined as

$$\mathcal{B} = \mathcal{F}_y^{-1}[\lambda^+(\xi_y)\mathcal{F}_y] = \imath\kappa\sqrt{1+\partial_{yy}/\kappa^2}.$$

The operator $\mathcal{B}$ is the exact Dirichlet-to-Neumann (DtN) operator for this problem. The latter can be rewritten as follows in the more general case of flat boundary with outward unit normal $\boldsymbol{n}$ on Γ: the exact absorbing boundary condition reads as Equation (1.19) with the normal derivative $\partial/\partial\boldsymbol{n}$ in the left-hand side and the pseudo-differential operator $\mathcal{B}$ and the Laplace-Beltrami operator Δ_Γ defined as

$$\mathcal{B} = \imath\kappa\sqrt{1+\Delta_\Gamma/\kappa^2}, \quad \Delta_\Gamma = \Delta - \partial_{\boldsymbol{nn}}. \tag{1.20}$$

In the case of curved boundaries, the previous absorbing boundary condition is used in practice, but it necessarily introduces a source of error, since its formulation is exact only for flat boundaries.

Zeroth-order absorbing boundary condition in 2D

In higher dimensions there is no simple form like (1.18) for the solution close to the artificial boundary. In general, there will be waves propagating in every direction at each boundary point. An exact absorbing boundary condition of the type we derived in one dimension must necessarily be non-local: indeed, to perfectly absorb all incoming waves, the boundary condition must account for the interaction between waves coming from all parts of the boundary. This requires combining information from multiple points on the boundary, hence the ABC involves an integral over the entire boundary. Applying such non-local conditions numerically is expensive. Instead, it is preferred to use various local approximation of the ABC. Since they are not exact, they will in general cause spurious reflected waves when a wave hits the artificial boundary.

One common approach is to assume that the solution of the full problem, close to the artificial boundary, can be approximated by a plane wave that propagates in a direction normal to the boundary. Then the same argument as in one dimension can be used. Indeed, if

$$p(\boldsymbol{x}) \approx Ce^{+\imath\kappa\boldsymbol{n}\cdot\boldsymbol{x}},$$

where $\boldsymbol{n}$ is the normal of the boundary, then

$$\partial_{\boldsymbol{n}}p - \imath\kappa p \approx \imath\kappa Ce^{+\imath\kappa\boldsymbol{n}\cdot\boldsymbol{x}}\partial_{\boldsymbol{n}}(\boldsymbol{n}\cdot\boldsymbol{x}) - \imath\kappa Ce^{+\imath\kappa\boldsymbol{n}\cdot\boldsymbol{x}} = 0.$$

This motivates the simplest ABC, that takes the form

$$\frac{\partial p}{\partial\boldsymbol{n}} = \mathcal{B}_{\imath\kappa}p, \quad \text{on } \partial\Omega, \tag{1.21}$$

with $\mathcal{B}_{\imath\kappa} = \imath\kappa$.

As a final remark, we observe that when a plane wave hits the boundary at an oblique angle with direction $\boldsymbol{k} \neq \boldsymbol{n}$, ABC (1.21) gives an error, because of the artificially reflected wave, whose size is proportional to the difference $|\boldsymbol{k} - \boldsymbol{n}|$ for small deviations. In particular, denoting by θ the (small) deviation angle between $\boldsymbol{k}$ and $\boldsymbol{n}$, it is possible to show that the *reflection coefficient* $R(\theta)$, defined as the ratio between the amplitudes of the reflected and incident waves, reads

$$R(\theta) = \left| \frac{1 - \cos\theta}{1 + \cos\theta} \right| = \tan^2\left(\frac{\theta}{2}\right) \sim \frac{1}{4}\theta^2.$$

It is possible to derive higher order ABCs, which produce much smaller artificial reflections and this is the main goal of the following sections.

High-order absorbing boundary condition in 2D

Boundary condition (1.19) is non-local because of the square root in the symbol of $\mathcal{B}$ and this leads to issues whenever a finite element solver is exploited because of a strong coupling of the unknowns defined onto the fictitious boundary and thus a dense block in the matrix of the linear system for the associated entries. Nevertheless, local high-order boundary conditions can be derived by approximating the square root in (1.20).

In [45], using a Padé approximation of the square root, a family of local boundary conditions is derived. In particular, the $(2N+1)$th-order Padé approximation of the square root $f(X) = \sqrt{1+X}$ is written as the rational function

$$f^{\text{Padé}}_{2N+1}(X) = 1 + \frac{2}{M} \sum_{n=1}^{N} c_n \Big(1 - \frac{c_n + 1}{c_n + 1 + X}\Big),$$

with $c_n = \tan^2(n\pi/M)$ and $M = 2N + 1$. Nevertheless, Padé approximations lead to boundary conditions that are inappropriate for evanescent modes and inaccurate for grazing waves. In [95], a modified approximation with a change of variable to rotate the branch cut of the square root by some angle ϕ is proposed and leads to

$$f^{\phi}_{2N+1} = e^{\imath\phi/2} f^{\text{Padé}}_{2N+1}(e^{-\imath\phi}(1+X) - 1) = e^{\imath\phi/1}\Big[1 + \frac{2}{M} \sum_{n=1}^{N} c_n \Big(1 - \frac{e^{\imath\phi}(c_n + 1)}{e^{\imath\phi} c_n + 1 + X}\Big)\Big]. \tag{1.22}$$

The latter highly improves the accuracy of the high-order absorbing boundary condition (HABC) for evanescent modes, while it slightly increases the reflection of traveling modes. Letting the angle ϕ grow, its positive impact becomes more significant.

There exist more general approximations of the DtN operator. In particular, Higdon-type HABCs [70] turned out to be exact for waves that propagate with any angle of incidence of a given set $\{\theta_m\}_{m=1,\dots,M}$, but they remain inefficient for evanescent waves. Nevertheless, complete radiation boundary conditions (CRBCs) [64] are able to deal with both propagative and evanescent waves. Combining approximation (1.22) with the pseudo-differential operator (1.20), we have the approximate boundary condition

$$\partial_{\boldsymbol{n}} p = \mathcal{B}_{\text{Padé}} p, \quad \text{on } \Gamma, \tag{1.23}$$

where $\mathcal{B}_{\text{Padé}}$ is the pseudo-differential operator defined as

$$\mathcal{B}_{\text{Padé}} = \imath\kappa e^{\imath\phi/2} \left[1 + \frac{2}{M} \sum_{n=1}^{N} c_n \left(1 - \frac{e^{\imath\phi}(c_n + 1)}{e^{\imath\phi} c_n + 1 + \Delta_\Gamma/\kappa^2}\right)\right]. \tag{1.24}$$

Dealing with corners and curved boundaries in 2D

The ABC we have just derived is suitable only for flat boundaries since it consists in a non-local boundary condition expressed in terms of the DtN operator which is exact only for flat boundaries. In the case of a polygonal domain, that is a domain with straight sides and corners, a special treatment that deals with the local non-regularity of the boundary in the areas of the corners may be required ([98]).

In [86], a modified form of the ABC which is suitable for regular curved boundaries is derived. In two dimensions, the modified ABC for a regular curve $\Gamma \subset \mathbb{R}^2$ reads

$$\partial_{\boldsymbol{n}} p = (\mathcal{B} + \mathcal{S})p, \quad \text{on } \Gamma, \tag{1.25}$$

where $\mathcal{S}$ is the differential operator defined as

$$\mathcal{S} = -\frac{\gamma}{2} + \frac{\gamma^2}{8(\gamma - \imath\kappa)} - \partial_{\boldsymbol{\tau}} \left(\frac{\gamma}{2\kappa^2} \partial_{\boldsymbol{\tau}} \right), \tag{1.26}$$

with the local curvature γ which is defined as $\gamma = 1/\rho$, with ρ the radius of curvature that is the radius of the osculating circle and $\boldsymbol{\tau}$ denotes the unit vector which is counter-clockwise tangent to the boundary, see Figure 1.2.

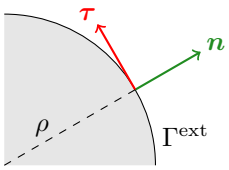


Figure 1.2: Notation for the normal $\boldsymbol{n}$ and tangent $\boldsymbol{\tau}$ unit vectors at the boundary Γ^{ext}.

Low-order and high-order absorbing boundary conditions (as well as their modified form for regular curved boundaries) are objects of investigation in Chapter 5. In particular, in the framework of a Trefftz discontinuous Galerkin method, we inspect the impact of these ABCs on the accuracy of the numerical solution.

1.2 Numerical methods

Reducing the domain to a bounded region represents the first important and necessary step that allows the application of numerical methods. In this section, we proceed with the steps leading to Galerkin finite element methods (FEMs). In particular, we present the framework, recall key references from the literature, and highlight difficulties and advantages of the most popular among them.

For the sake of clarity, we consider the Helmholtz equation complemented by a inhomogeneous Neumann boundary condition

$$\begin{cases} -\Delta p - \kappa^2 p = f, & \text{in } \Omega, \\ \dfrac{\partial p}{\partial \boldsymbol{n}} = g, & \text{on } \partial\Omega, \end{cases} \tag{1.27}$$

where $\Omega \subset \mathbb{R}^d$ is a bounded domain, $\kappa \in \mathbb{R}$ is the wavenumber, $f : \Omega \to \mathbb{C}$ is the volume source, $g : \partial\Omega \to \mathbb{C}$ is the boundary datum and with $d = 1, 2, 3$. The unknown is the acoustic pressure $p : \Omega \to \mathbb{C}$.

1.2.1 Continuous Galerkin methods

We multiply Equation (1.27) by the complex conjugate $\overline{q}$ of a sufficiently regular function $q \in \mathcal{V}$ (the space $\mathcal{V}$ is a proper Hilbert space, subspace of $H^1(\Omega)$), called *test function*, we integrate on Ω by parts and we get the following *variational formulation*

Problem 1.2.1. *Find $p \in V$ such that*

$$a(p,q) = F(q)$$

for all $q \in \mathcal{V}$, where $a : \mathcal{V} \times \mathcal{V} \to \mathbb{C}$ is a bilinear form and $F : \mathcal{V} \to \mathbb{C}$ is a linear form which are defined as

$$a(p,q) = \int_\Omega \nabla p \nabla \overline{q}\ d\boldsymbol{x} - \kappa^2 \int_\Omega p\overline{q}\ d\boldsymbol{x},$$
$$F(q) = \int_\Omega f\overline{q}\ d\boldsymbol{x} + \int_{\partial\Omega} g\overline{q}\ d\sigma(\boldsymbol{x}).$$

Aiming at an approximate solution of Problem 1.2.1, we introduce a family of spaces $\mathcal{V}_h$ that depend on a positive parameter h such that

$$\mathcal{V}_h \subset \mathcal{V}, \quad \dim \mathcal{V}_h = N_h < \infty, \quad \forall h > 0.$$

The most common choice for the approximate space $\mathcal{V}_h$ is obtained by considering a mesh $\mathcal{T}_h$ of the computational domain Ω, namely by approximating it with a segment ($d = 1$)/polygonal ($d = 2$)/polyhedron ($d = 3$) which is composed by a fixed number of non-overlapping intervals ($d = 1$)/triangles ($d = 2$)/tetrahedra ($d = 3$), called elements $K \in \mathcal{T}_h$.

The possibility of constructing non-uniform meshes that present a larger number of elements in certain areas than in others and the fact that the support of the basis functions is compact makes FEMs able to cope with complex geometries and fast oscillating solutions leading to sparse matrices which can be solved efficiently without the need of storing or computing null coefficients. If the basis functions are chosen to be polynomials of order smaller or equal to p ≥ 0 on each element $K \in \mathcal{T}_h$ and globally continuous on the domain, i.e.

$$\mathcal{V}_h := \{v \in \mathcal{C}^0(\Omega) | \forall K \in \mathcal{T}_h, v|_K \in \mathcal{P}_{\mathrm{p}}(K)\},$$

where $\mathcal{P}_{\mathrm{p}}(K)$ denotes the set of polynomials of order smaller or equal to p defined on the element $K \in \mathcal{T}_h$, the FEM is called *continuous Galerkin method* (CG) [44, 65]. Note that, from now on, $h := \max_{K \in \mathcal{T}_h} h_K$ is called mesh size and $h_K := \max_{\boldsymbol{x},\boldsymbol{y} \in K} \|\boldsymbol{x} - \boldsymbol{y}\|$ is called diameter of the element K.

The resulting discrete problem, called *Galerkin problem*, reads as follows.

Problem 1.2.2. *Find $p_h \in \mathcal{V}_h$ such that*

$$a(p_h, q_h) = F(q_h) \tag{1.28}$$

for all $q_h \in \mathcal{V}_h$.

Let $\{\boldsymbol{\varphi}_j\}_{j=1,\dots,N_h}$ a basis of $\mathcal{V}_h$, it is equivalent that (1.28) holds for each function of the basis, namely

$$a(p_h, \varphi_i) = F(\varphi_i), \quad i = 1, \dots, N_h. \tag{1.29}$$

Since $p_h \in \mathcal{V}_h$,

$$p_h(\boldsymbol{x}) = \sum_{j=1}^{N_h} p_j \varphi_j(\boldsymbol{x}),$$

where $\{p_j\}_{j=1,...,N_h}$ are the unknown coefficients. Equation (1.29) becomes

$$\sum_{j=1}^{N_h} a(\varphi_j, \varphi_i) p_j = F(\varphi_i), \quad i = 1, ..., N_h. \tag{1.30}$$

We denote by $\mathbf{A}$ the matrix with elements $\mathbf{A}_{ij} = a(\varphi_j, \varphi_i)$, by $\mathbf{b}$ the vector with components $b_i = F(\varphi_i)$ and by $\mathbf{x}$ the vector with the unknown coefficients p_j. Equation (1.30) can then be written as

$$\mathbf{A}\mathbf{x} = \mathbf{b}, \tag{1.31}$$

and finding a solution to the discrete problem reduces to computing the solution of Equation (1.31).

The algebraic properties of the matrix $\mathbf{A}$ impact the performance of both direct and iterative solvers exploited to solve Equation (1.31). These properties depend on both the bilinear form a (that is the intrinsic physical properties of the considered problem) and the chosen approximate space $\mathcal{V}_h$ (that is the numerical framework). In particular, in the case of large three-dimensional problems, the matrix $\mathbf{A}$ may become too large to be solved by direct solvers and consequently iterative ones are preferred (see Section 1.3).

1.2.2 Discontinuous Galerkin methods

In the *discontinuous Galerkin method* framework, the functions that belong to the space $\mathcal{V}_h$ are not continuous across the interface F shared by two elements K and K' and, in particular, the unknown p_h and the test function q_h belong to the space

$$\mathcal{V}_h := \prod_{K \in \mathcal{T}_h} \mathcal{P}_{\mathrm{p}}(K),$$

and, therefore, both the discrete unknown p_h and the test function q_h admit two traces on any interior face F. Consequently, interface quantities that appear in the weak formulation, such as the trace of p_h or its normal flux, are not uniquely defined and must be replaced by single-valued quantities, often referred to as *numerical flux*, constructed from the traces on the adjacent elements. In this case, the variational formulation reads as follows

Problem 1.2.3. *Find $p_h \in \mathcal{V}_h$ such that*

$$\tilde{a}(p_h, q_h) = F(q_h)$$

for all $q \in \mathcal{V}_h$, where $\tilde{a} : \mathcal{V}_h \times \mathcal{V}_h \to \mathbb{C}$ is a bilinear form which is defined as

$$\tilde{a}(p, q) = a(p, q) - \sum_{K \in \mathcal{T}_h} \sum_{F \in \mathcal{F}_K : F \not\subset \partial\Omega} \int_F \left(\frac{\partial p}{\partial \boldsymbol{n}} \right)_* \overline{q} \, d\sigma(\boldsymbol{x}),$$

where $\mathcal{F}_K$ denotes the set of faces of K.

The term $\left(\frac{\partial p_h}{\partial \boldsymbol{n}} \right)_*$ is the numerical flux that needs to be defined along an interior face. The most common choice for the numerical flux is to use a definition that takes into account only the values of the discrete fields defined on the two elements K and K' that share the interface F in terms of jumps and averages.

Problem 1.2.3 is based on the *second-order* formulation of the Helmholtz problem, that is (1.27). As we will see later, hybridizable discontinous Galerkin (HDG) methods are based on a *first order* formulation because it leads to a more straightforward definition of the numerical flux and to an easier implementation. The first order formulation of problem (1.27) is obtained by introducing

an additional variable $\boldsymbol{u}$ defined as $\boldsymbol{\nabla} p = \imath\kappa \boldsymbol{u}$. Computing the divergence of both sides of the previous relation we get $\Delta p = \imath\kappa \boldsymbol{\nabla}\cdot\boldsymbol{u}$ which, once plugged in (1.27), gives

$$\begin{cases} -\imath\kappa p + \boldsymbol{\nabla}\cdot\boldsymbol{u} = \tilde{f}, & \text{in } \Omega, \\ -\imath\kappa \boldsymbol{u} + \boldsymbol{\nabla} p = \boldsymbol{0}, & \text{in } \Omega, \\ \boldsymbol{n}\cdot\boldsymbol{u} = \tilde{g}, & \text{on } \partial\Omega, \end{cases} \tag{1.32}$$

with $f = -\imath\kappa\tilde{f}$ and $g = \imath\kappa\tilde{g}$.

After introducing the space

$$\boldsymbol{\mathcal{V}}_h := \prod_{K\in\mathcal{T}_h} \boldsymbol{\mathcal{P}}_{\mathrm{p}}(K),$$

where $\boldsymbol{\mathcal{P}}_{\mathrm{p}}(K)$ denotes the set of vectors whose components are polynomials of degree smaller or equal to p defined on the element $K \in \mathcal{T}_h$, we can derive the discontinuous Galerkin formulation of the problem: we multiply the two physical equations by two test functions $q_h \in \mathcal{V}_h$ and $\boldsymbol{v}_h \in \boldsymbol{\mathcal{V}}_h$, we integrate by parts on an element $K \in \mathcal{T}_h$, we sum all over the elements of $\mathcal{T}_h$ and we get the following problem.

Problem 1.2.4. *Find* $(p_h, \boldsymbol{u}_h) \in \mathcal{V}_h \times \boldsymbol{\mathcal{V}}_h$ *such that*

$$\begin{cases} -\imath\kappa(p_h, q_h)_{\mathcal{T}_h} - (\boldsymbol{u}_h, \boldsymbol{\nabla} q_h)_{\mathcal{T}_h} + \langle \boldsymbol{n}\cdot\widehat{\boldsymbol{u}}_h, q_h\rangle_{\partial\mathcal{T}_h} = (\tilde{f}, q_h)_{\mathcal{T}_h}, \\ -\imath\kappa(\boldsymbol{u}_h, \boldsymbol{v}_h)_{\mathcal{T}_h} - (p_h, \boldsymbol{\nabla}\cdot\boldsymbol{v}_h)_{\mathcal{T}_h} + \langle \widehat{p}_h, \boldsymbol{n}\cdot\boldsymbol{v}_h\rangle_{\partial\mathcal{T}_h} = 0, \end{cases}$$

for all $(q_h, \boldsymbol{v}_h) \in \mathcal{V}_h \times \boldsymbol{\mathcal{V}}_h$.

Here we use the following notation for volume and surface integrals, that is

$$(a,b)_{\mathcal{T}_h} = \sum_{K\in\mathcal{T}_h} \int_K a\overline{b}\, d\boldsymbol{x}, \quad \langle a,b\rangle_{\partial\mathcal{T}_h} = \sum_{K\in\mathcal{T}_h} \int_{\partial K} a\overline{b}\, d\sigma(\boldsymbol{x}).$$

Note that, in Problem 1.2.4, $\boldsymbol{n}\cdot\widehat{\boldsymbol{u}}_h$ and $\widehat{p}_h$ denote numerical fluxes that need to be defined. Well-known examples are *centered* and *upwind* numerical fluxes. As we mentioned before, the choice of the numerical flux is crucial for the performance of the numerical method [63, 75].

Many variations of the DG method have been introduced to cope with the difficulties that arise in the wave propagation framework. For example, the so-called *interior penalty discontinuous Galerkin* (IPDG) method consists in enriching the DG formulation with additional *penalty* terms on the jumps of the normal and tangential derivatives of the unknown and it results in an overall improvement of stability without any constraint on the mesh, see e.g. [25, 50, 128]. Other methods, called *hybrid high-order* (HHO) methods, require degrees of freedom defined in the elements (later eliminated by static condensation) and on the faces. They are based on a local gradient reconstruction and on a stabilization operator which allows the matching of the trace of the volume unknowns with the face ones in a weak sense. The advantages of this class of methods are multiple: they allow polyhedral meshes and consequently the use of hanging nodes during mesh refinements, they present optimal convergence rates and local conservation principles, see e.g. [17, 18, 48].

1.2.3 Dispersion and pollution errors

The following a priori error estimate in energy semi-norm for the numerical solution of continuous Galerkin FEM was proved in [81]

$$e_h \leq C_1 \left(\frac{\kappa h}{2\mathrm{p}}\right)^{\mathrm{p}} + C_2\kappa \left(\frac{\kappa h}{2\mathrm{p}}\right)^{2\mathrm{p}}, \tag{1.33}$$

where C_1 and C_2 are two constants independent of κ and h, and p is the polynomial degree of the approximation. Analogous estimates have been derived specifically for linear basis functions ($\mathrm{p} = 1$) in [80] and high-order basis functions ($\mathrm{p} \geq 2$) in [79, 122].

Therefore, numerical solutions can be damaged by a *best-approximation error* (the first term in (1.33)), namely an error due to interpolation, hence the element size h must be adapted according to the wavenumber so that the number of elements per wavelength stays below a resolution limit. The latter is computed by the non-dimensional wavenumber $\kappa h = 2\pi h/\lambda$, where λ is the wavelength, and a discrete dispersion analysis allows to quantify the upper bound of the mesh resolution.

In addition to interpolation, another source of error (the second term in (1.33)), known in literature as *pollution error*, damages the numerical solution. It originates from the deterioration of the stability constant at large wavenumbers [11] and it might have a significant impact even if the non-dimensional wavenumber κh is kept fixed below a resolution limit, for instance at high-frequencies. However, if it is reduced, both sources of error decrease.

Many studies have been carried out on this topic and to propose efficient techniques to reduce the impact of both interpolation and pollution errors. In particular, we have the following classes of numerical methods

- *high-order discontinuous Galerkin methods (DG)*: as we explained before, at the cost of more degrees of freedom, they allow piecewise discontinuous numerical solutions that are more appropriate to tackle non-smooth wave phenomena by limiting the errors. Several studies exhibit how high-order polynomial approximations significantly limit these errors [2–4, 28, 41, 122] (an intuition is already given by the a priori error estimate (1.33), indeed, if κh is sufficiently small, then the upper bound for e_h given by (1.33) monotonously decreases with p), especially in the case of high-frequency scenarios, see e.g. [13, 27, 37, 92].

- *stabilized Galerkin methods*: for low-order polynomial approximations, stabilized Galerkin methods, e.g. Galerkin least-squared (GLS) [66, 67] and generalized Galerkin methods in the variational multiscale (VMS) framework [76], which include the residual-free bubbles method [53], have been proposed because they exhibit an increasing accuracy with a very simple implementation.

- *flux reconstruction methods (FRM)*: they unify discontinuous Galerkin and spectral difference schemes into a single high-order framework. They rely on a strong formulation with no dependence on quadrature rules that ensure local conservation, allow energy-stable formulations and they minimize numerical dispersion and dissipation [106, 113].

- *wave-based/Trefftz methods*: they take as basis function solutions of the governing physical equations to embed the oscillatory behavior of the solution into the numerical method. This class of method is presented and discussed in the following Section 1.2.4.

1.2.4 Wave-based/Trefftz methods

The key idea of Trefftz methods is to include a priori knowledge about the local behavior of the solution into the space of solutions. This aim is achieved by considering an approximation space $\mathcal{V}_h$ composed by local canonical solutions of the equation [89, 90]. It might be assumed that Trefftz methods allow better accuracy than non-Trefftz methods, since in the latter case their approximation basis does not embed any physical feature and is not related to the equation itself. Nevertheless, the comparison between high-order polynomial finite element methods and wave-based methods is still an open discussion [92] and several studies are still being conducted.

In the context of the Helmholtz equation, we take the approximation space

$$\mathcal{V}_h = \prod_{K\in\mathcal{T}_h} \{v_K \in H^1(K) | -\Delta v_K - \kappa^2 v_K = 0\}. \tag{1.34}$$

In other words, $\mathcal{V}_h$ is the space of functions whose restriction on the element $K \in \mathcal{T}_h$ is a function v_K which solves the homogeneous Helmholtz equation with no boundary condition imposed. Many sets of basis functions can be considered: Green functions ([23]), Bessel functions ([93]) and plane waves are among the most popular choices.

A variant of Trefftz methods is the so-called *ultra weak variational formulation* (UWVF). It still requires a set of test functions given by solutions of the homogeneous Helmholtz equation, but this time volume integrals associated to the unknown are eliminated from the variational formulation leading to a problem whose unknown is defined only on the skeleton and its continuity across the faces that compose the skeleton itself is weakly prescribed. It was first introduced in 1998 [24] in the framework of the Helmholtz equation complemented by a zeroth-order absorbing boundary condition on the artificial boundary. In Chapter 4, we investigate an UWVF for the Helmholtz equation complemented by high-order absorbing boundary conditions with the aim of generalizing the results obtained in [24].

Refinement and adaptivity

In the case of propagative plane-wave basis functions, both h-refinement [61], that is mesh subdivision, p-refinement [72], that is local basis enrichment, and hp-refinement [37, 73], that is the combination of the previous two, are appropriate for improving the accuracy of the method. Once an element of the mesh is selected for a refinement, the type of the refinement needs to be decided. This choice depends on the nature of the solution: if it is smooth, a p-refinement will be more efficient at reducing the error, if that is not the case, an h-refinement should be preferred. Nevertheless, clearly, the previous considerations are relevant only when the exact solution is known a priori, while, if it is unknown, a posteriori error estimators do not indicate the type of refinement which should be undertaken, but some algorithms have been tested in the literature [96]. In [99] a very detailed and complete analysis concerning the plane-wave approximation of homogeneous Helmholtz solutions complemented by proofs is provided.

Besides the h and p refinements, which are common to all finite element methods, directional adaptivity finds application in this framework. The key aspect is that, in the simple case in which the exact solution is a plane wave, if the direction of one of the plane wave basis functions is aligned with the one of the exact solution, the method exactly recovers the analytical solution, up to rounding errors. Thus, the key idea is to rotate the basis associated to an element according to its dominant direction, which clearly needs to be numerically estimated a posteriori. An algorithm based on the study of the properties of the Hessian and its principal eigenvector of the numerical solution is presented in [39]. It is important to observe that directional adaptivity permits good results only in the case of a single dominant wave direction because multiple dominant wave directions are way harder to tackle by bare directional adaptivity.

Ill-conditioning

The linear system obtained by plane-wave discretization is prone to ill-conditioning whenever the trial space has high dimension because of a large number of elements in the mesh or of a large number of plane waves and this is due to the plane-wave basis [39, 56, 87, 105]. Indeed, if the elements have a relatively small size compared to the wavelength, that is $\|\boldsymbol{x}\| \ll \kappa^{-1}$, the plane-wave functions are almost constant, therefore they are nearly linearly dependent. Moreover, even in the case of elements whose size is smaller than the wavelength, plane waves become hard to distinguish whenever their density increases.

In particular, considering the mass matrix $\mathbf{M}_K$ of a single element K, it is possible to empirically observe that its spectral condition number increases like $\sim h_K^{-q}$ for element size $h_K \to 0$, where q is proportional to the number p_K of plane waves defined in the element K. Moreover, unfortunately, the condition number increases exponentially with q, that is $\mathrm{cond}(\mathbf{M}_K) \sim e^{\alpha q}$ for $q \to \infty$ and $\alpha > 0$. A similar growth of the condition number can be observed for the full matrix as the mesh is refined or more plane waves are used in each element.

Several heuristic algorithms have been tested in order to cope with the ill-conditioning. For example, limiting p_K by monitoring the condition numbers of element matrices seems also to lead to a control of the condition number of the full matrix [77]. Other techniques prescribe a priori the number of plane waves p_K, such as the estimate $\mathrm{p}_K = \mathrm{round}(\kappa h_K + C(\kappa h_K)^{1/3})$ with $3 \leq C \leq 14$ ([78]). A method based on the local orthogonalization of the plane-wave basis is effective in lowering the condition number of the global system matrix ([38]). Lastly, local preconditioners have been studied ([7, 105]) and they are equivalent to an orthonormalisation of the plane-waves basis with respect to an L^2 inner product computed on the boundary of the elements.

We highlight the fact that ill-conditioning is one of the main problems that the Trefftz method in the framework of Helmholtz presents, which may make any kind of refinement hard to pursue. Therefore, it is important to monitor it in order to avoid unreliable results.

In Chapter 5 many of the aspects we have just presented are taken into account and observed in the study. In particular, we keep track of possible ill-conditioning issues, we carefully perform h and p refinements as well as directional adaptivity.

1.3 Algorithms for the solution of linear systems

In this section we recall some methods to solve linear systems, namely direct and iterative methods, and we give details concerning those that are adopted in the rest of the thesis to solve the linear system obtained by the selected finite element methods. The content of the first part of this section mainly refers to Chapter 7 of [110]. In the second part of this section we focus on some accelerating techniques for iterative methods.

We consider a system of m linear equations in n unknowns

$$\sum_{j=1}^{n} a_{i,j} x_j = b_i, \quad i = 1, ..., m,$$

where x_j are the unknowns, $a_{i,j}$ the coefficients of the system and b_i given numbers. The previous system can be equivalently written in matrix form as

$$\mathbf{A}\mathbf{x} = \mathbf{b}, \tag{1.35}$$

where $\mathbf{A} = (a_{i,j}) \in \mathbb{C}^{m \times n}$ is the coefficient matrix, $\mathbf{x} = (x_i) \in \mathbb{C}^n$ is the unknown vector and $\mathbf{b} = (b_i) \in \mathbb{C}^m$ is the right-hand side vector. From now on, we assume that $m = n$ and that the matrix $\mathbf{A}$ is non-singular, i.e. $\det(\mathbf{A}) \neq 0$, so that the system admits a unique solution, namely there exists a unique vector $\mathbf{x}$ that verifies (1.35) and reads

$$\mathbf{x} = \mathbf{A}^{-1}\mathbf{b}.$$

However, the inversion of a matrix can be numerically very costly, especially in the case of a large matrix. Alternatively, Cramer's rule provides a closed formula for the solution of the linear system without the need of inverting the matrix, but it requires $(n+1)!$ operations for a dense matrix, which is clearly too computationally expensive even for small size matrices. Methods that aim at solving such linear systems without the need of inverting their matrix are called *direct* if they lead

to the solution after a finite number of steps and *iterative* if they require a theoretically infinite number of steps.

The difficulty of solving linear systems depends strongly on the structure of the matrix. Sparse matrices, which typically arise from finite differences or finite elements discretizations, have most entries equal to zero and can be stored efficiently; they allow the use of iterative solvers with preconditioning, making large-scale problems computationally feasible. In contrast, dense matrices often occur in boundary element methods; they require $\mathcal{O}(n^2)$ memory and $\mathcal{O}(n^3)$ operations for direct solvers, making them prohibitively expensive for large n. For Helmholtz problems, the situation is further complicated at high frequencies: both sparse and dense systems become harder to solve due to indefiniteness and oscillatory behavior, but the cost gap between them remains significant, with sparse systems generally being far more tractable for large-scale simulations.

1.3.1 Direct and iterative methods

An example of direct method is given by the *Gauss elimination method* (GEM): after reducing the initial system to an equivalent one of the form $\mathbf{A}^{(n)}\mathbf{x} = \mathbf{b}^{(n)}$, where $\mathbf{A}^{(n)} = \mathbf{U}$ is a non-singular upper triangular matrix and $\mathbf{b}^{(n)}$ is a new right-hand side vector, the latter system can be solved by a backward substitution algorithm thanks to the use of scalar quantities called multipliers. It can be shown that the algorithm requires about $2n^3/3$ operations for a dense matrix.

The GEM is equivalent to factorizing the matrix $\mathbf{A}$ as the product $\mathbf{LU}$ of two matrices: the matrix $\mathbf{U} = \mathbf{A}^{(n)}$ upper triangular and the matrix $\mathbf{L}$ lower triangular with diagonal elements equal to 1 and the remaining elements in the lower triangular section equal to the multipliers. After computing the matrices $\mathbf{L}$ and $\mathbf{U}$, the solution of the original linear is obtained by the successive solution of the two triangular systems $\mathbf{Ly} = \mathbf{b}, \mathbf{Ux} = \mathbf{y}$. Since the matrices $\mathbf{L}$ and $\mathbf{U}$ depend only on $\mathbf{A}$ and not on the right-hand side $\mathbf{b}$, they can be used to solve every linear system characterized by the matrix $\mathbf{A}$ regardless of the right-hand side $\mathbf{b}$.

Other direct methods based on particular factorizations of the matrix exist: they exploit and depend on the form of the matrix itself, e.g. block diagonal or tridiagonal matrices.

Iterative methods construct the solution $\mathbf{x}$ as the limit of a sequence $\{\mathbf{x}^{(k)}\}$ of approximate vectors. Starting from an initial guess $\mathbf{x}^{(0)}$, a key step is the computation of the *residue* vector $\mathbf{r}^{(k)}$ that is used to compute the approximate solution $\mathbf{x}^{(k)}$ at each iteration k

$$\mathbf{r}^{(k)} = \mathbf{b} - \mathbf{A}\mathbf{x}^{(k)}$$

which requires n^2 operations for a dense matrix. Therefore, iterative methods are competitive with direct methods (which require approximately $2n^3/3$ operations) if the convergence (with respect to a given criterion) is independent of n or is achieved in a sub-linear way (otherwise the overall number of operations exceeds that of direct methods).

Direct methods present some advantages over iterative ones: they guarantee convergence to the solution in a finite number of steps, the solution is exact up to numerical precision, they are relatively easy to implement (even though not in parallel), they do not depend on an initial guess and they are very efficient for matrices that present a particular form. Iterative methods do not typically lead to the exact solution, they require a careful choice of the stopping criterion, they may present a more complicated implementation and they can be sensitive with respect to the initial guess. Nevertheless, iterative methods also showcase some advantages over direct ones: they often require less memory, especially for large sparse matrices, in these cases they typically provide a solution with a faster convergence, they can be improved by preconditioning techniques and they are suitable for parallel implementations.

To sum up, in general, direct methods are preferred to iterative ones only in the case of small and dense matrices for which they lead to the exact solution in a finite number of steps, whereas in the case of big and sparse matrices they become prohibitively costly and iterative ones are to be chosen. Therefore, since wave propagation problems are very often associated to big, large and sparse matrices, we are rather interested in iterative methods.

They can be divided into *stationary iterative methods* and *Krylov subspace methods*. Stationary iterative methods generate at each iteration a new approximation from the previous one using a fixed formula, such as in Jacobi (J), Gauss-Seidel (GS), or successive over relaxation (SOR) methods. They are simple to implement and have low computational cost per iteration. However, they tend to converge slowly and require certain properties of the matrix, like diagonal dominance, making them less effective for large or ill-conditioned systems. On the other hand, Krylov subspace methods construct approximations within spaces generated by successive powers of the matrix applied to the initial residual. Examples include conjugate gradient (CG), generalized minimal residual (GMRES), and biconjugate gradient stabilized (BiCGSTAB) methods. They typically converge much faster, especially when combined with appropriate preconditioning, and are well suited for large, sparse systems. The drawbacks are higher memory and computational effort, and some methods may need restarting to control memory usage.

1.3.2 Stationary iterative methods

Richardson methods

We consider a splitting of the matrix $\mathbf{A}$, i.e. $\mathbf{A} = \mathbf{P} - \mathbf{N}$, where $\mathbf{P}$ is a non-singular matrix called *preconditioning matrix* or simply *preconditioner*. We can construct an iterative method in the following way: given $\mathbf{x}^{(0)}$, we compute $\mathbf{x}^{(k)}$ for each $k \geq 1$ by solving the system

$$\mathbf{P}\mathbf{x}^{(k+1)} = \mathbf{N}\mathbf{x}^{(k)} + \mathbf{b}, \quad k \geq 0,$$

or, equivalently,

$$\mathbf{x}^{(k+1)} = \mathbf{B}\mathbf{x}^{(k)} + \mathbf{P}^{-1}\mathbf{b}, \quad k \geq 0, \tag{1.36}$$

where $\mathbf{B} = \mathbf{P}^{-1}\mathbf{N}$ is called *iteration matrix*. For an iterative method to converge, we must have $\lim_{k\to\infty} \mathbf{e}^{(k)} = \mathbf{0}$, where $\mathbf{e}^{(k)} = \mathbf{x}^{(k)} - \mathbf{x}$ denotes the *error* vector. By recursion, we have that $\mathbf{e}^{k} = \mathbf{B}^{k}\mathbf{e}^{(0)}$, $k \geq 1$ and, consequently, the iterative method is convergent if and only if the spectral radius $\rho(\mathbf{B})$ of the iteration matrix (the maximum modulus of the eigenvalues $\{\lambda_i\}$ of $\mathbf{B}$) is strictly smaller than 1, i.e.

$$\rho(\mathbf{B}) = \max_{i=1,\dots,n} |\lambda_i| < 1.$$

Equation (1.36) can be rewritten as

$$\mathbf{x}^{(k+1)} = \mathbf{x}^{(k)} + \mathbf{P}^{-1}\mathbf{r}^{(k)}, \quad k \geq 0 \tag{1.37}$$

with the residual at the step k defined as

$$\mathbf{r}^{(k)} = \mathbf{b} - \mathbf{A}\mathbf{x}^{(k)},$$

or, equivalently,

$$\mathbf{x}^{(k+1)} = (\mathbf{I} - \mathbf{P}^{-1}\mathbf{A})\mathbf{x}^{(k)} + \mathbf{P}^{-1}\mathbf{b}, \quad k \geq 0$$

and highlights the fact that the matrix $\mathbf{P}$, besides being non-singular, must also be easily invertible. Note that $\mathbf{R_P} = \mathbf{I} - \mathbf{P}^{-1}\mathbf{A}$ denotes the iteration matrix of the iterative method.

It is possible to generalize Equation (1.37) by introducing a *relaxation/acceleration* parameter α_k which leads to the so-called *Richardson method*

$$\mathbf{x}^{(k+1)} = \mathbf{x}^{(k)} + \alpha_k\mathbf{P}^{-1}\mathbf{r}^{(k)}, \quad k \geq 0.$$

Gradient and conjugate gradient methods

If $\mathbf{A}$ is a symmetric positive definite matrix, solving system (1.35) is equivalent to finding the minimum $\mathbf{x} \in \mathbb{R}^n$ of a quadratic form called *energy* of the system. Therefore, the idea is to determine the minimum point $\mathbf{x}$ of energy starting from an initial point $\mathbf{x}^{(0)} \in \mathbb{R}^n$ and proceeding by choosing a suitable sequence of directions that leads to the solution $\mathbf{x}$ in the quickest possible way. The most intuitive idea consists in taking as direction the greatest increase of the energy and leads to the *gradient method*. The algorithm of its preconditioned version reads: given $\mathbf{x}^{(0)} \in \mathbb{R}^n$ and setting $\mathbf{r}^{(0)} = \mathbf{b} - \mathbf{A}\mathbf{x}^{(0)}$, $\mathbf{z}^{(0)} = \mathbf{P}^{-1}\mathbf{r}^{(0)}$ for $k = 0, 1, ...$ until convergence, we compute

$$\begin{aligned} \alpha_k &= \frac{\mathbf{z}^{(k)T}\mathbf{z}^{(k)}}{\mathbf{z}^{(k)T}\mathbf{A}\mathbf{z}^{(k)}}, \\ \mathbf{x}^{(k+1)} &= \mathbf{x}^{(k)} + \alpha_k \mathbf{z}^{(k)}, \\ \mathbf{r}^{(k+1)} &= \mathbf{r}^{(k)} - \alpha_k \mathbf{A}\mathbf{z}^{(k)}, \\ \mathbf{P}\mathbf{z}^{(k+1)} &= \mathbf{r}^{(k+1)}. \end{aligned}$$

A variant of this method, called *conjugate gradient method*, is even more effective and exploits different descent directions. In particular, setting $\mathbf{p}^{(0)} = \mathbf{r}^{(0)}$, we consider the so-called $\mathbf{A}$-orthogonal directions

$$\mathbf{p}^{(k+1)} = \mathbf{r}^{(k+1)} - \beta_k \mathbf{p}^{(k)}, \quad k = 0, 1, ...,$$

where the parameters $\beta_k \in \mathbb{R}$ are to be determined so that

$$(\mathbf{A}\mathbf{p}^{(j)})^T \mathbf{p}^{(k+1)} = 0, \quad j = 0, 1, ..., k.$$

The algorithm of its preconditioned version reads: given $\mathbf{x}^{(0)} \in \mathbb{R}^n$ and setting $\mathbf{r}^{(0)} = \mathbf{b} - \mathbf{A}\mathbf{x}^{(0)}$, $\mathbf{z}^{(0)} = \mathbf{P}^{-1}\mathbf{r}^{(0)}$, $\mathbf{p}^{(0)} = \mathbf{z}^{(0)}$ for $k = 0, 1, ...$ until convergence, we compute

$$\begin{aligned} \alpha_k &= \frac{\mathbf{p}^{(k)T}\mathbf{r}^{(k)}}{(\mathbf{A}\mathbf{p}^{(k)})^T \mathbf{p}^{(k)}}, \\ \mathbf{x}^{(k+1)} &= \mathbf{x}^{(k)} + \alpha_k \mathbf{p}^{(k)}, \\ \mathbf{r}^{(k+1)} &= \mathbf{r}^{(k)} - \alpha_k \mathbf{A}\mathbf{p}^{(k)}, \\ \mathbf{P}\mathbf{z}^{(k+1)} &= \mathbf{r}^{(k+1)}, \\ \beta_k &= \frac{(\mathbf{A}\mathbf{p}^{(k)})^T \mathbf{z}^{(k+1)}}{\mathbf{p}^{(k)T}\mathbf{A}\mathbf{p}^{(k)}}, \\ \mathbf{p}^{(k+1)} &= \mathbf{z}^{(k+1)} - \beta_k \mathbf{p}^{(k)}. \end{aligned}$$

If $\mathbf{A}$ is not symmetric, a very common approach, called *conjugate gradient normal residual method* (CGNR), consists in applying the conjugate gradient method to the normal system $\mathbf{A}^*\mathbf{A}\mathbf{x} = \mathbf{A}^*\mathbf{b}$, see e.g. [110, 116]. It requires one matrix-vector multiplication with $\mathbf{A}$ and one with $\mathbf{A}^*$ per iteration. Thus, each iteration costs roughly twice a single matrix-vector product plus a small overhead of vector operations $\mathcal{O}(n)$. For sparse matrices with a fixed number of nonzeros per row, the cost per iteration scales as $\mathcal{O}(n)$ while, for dense matrices, it scales as $\mathcal{O}(n^2)$.

1.3.3 Krylov subspace methods

Krylov methods are generalizations and extensions of the gradient method to the case where the matrix $\mathbf{A}$ is not symmetric and, in particular, they build a sequence of approximations to the solution by exploring a subspace generated from the initial residual and repeated applications of the matrix. In the following, we introduce only the *generalized minimal residual method* (GMRES).

GMRES

At each iteration, GMRES seeks the best approximation to the solution within this Krylov subspace by minimizing the Euclidean norm of the residual. To do this, it constructs an orthonormal basis for the subspace using the Arnoldi process, which ensures numerical stability through orthogonalization, typically via the Gram-Schmidt algorithm. Once the basis is built, GMRES solves a small least-squares problem to determine the combination of basis vectors that gives the smallest possible residual. This process is repeated at each step, refining the approximation and reducing the error.

Each iteration involves one matrix-vector multiplication with $\mathbf{A}$ and a growing number of inner products due to orthogonalization against all previously computed basis vectors. For sparse matrices, the orthogonalization cost can dominate if many iterations are needed; for dense matrices, the matrix-vector multiplication cost $\mathcal{O}(n^2)$ remains dominant. Moreover, for m iterations, the method demands memory $\mathcal{O}(mn)$. Therefore both the computational cost and memory usage increase with the number of iterations. To control this, a restarted version (GMRES(m)) is often used, where the algorithm resets after a fixed number m of steps and starts again from the current approximation.

Remark 1.3.1. In this Section we have talked about repeating a certain iteration "until convergence" multiple times, that is equivalent to defining a stopping criterion. To do so, we fix a small tolerance ε and choose a norm $\|\cdot\|$ for $\mathbb{R}^n$ (e.g. $\|\cdot\|_2$). If the exact solution $\mathbf{x}$ is known a priori, then at each iteration we can compute the absolute or relative error and stop the iterative scheme if such an error is below the tolerance, namely

$$\text{absolute error} = \|\mathbf{x} - \mathbf{x}^{(k)}\| < \varepsilon, \quad \text{relative error} = \frac{\|\mathbf{x} - \mathbf{x}^{(k)}\|}{\|\mathbf{x}\|} < \varepsilon.$$

However, in general, the exact solution is not known a priori. An alternative stopping criterion that does not rely on the exact solution, consists in a control on the norm of the absolute or relative residue, namely

$$\text{absolute residue} = \|\mathbf{r}^{(k)}\| = \|\mathbf{b} - \mathbf{A}\mathbf{x}^{(k)}\| < \varepsilon, \quad \text{relative residue} = \frac{\|\mathbf{r}^{(k)}\|}{\|\mathbf{b}\|} = \frac{\|\mathbf{b} - \mathbf{A}\mathbf{x}^{(k)}\|}{\|\mathbf{b}\|} < \varepsilon.$$

If the residue is hard to compute, another option consists in a control on the difference between two consecutive iterations (this stopping criterion might be very misleading in the case of non-uniform convergence, that is when slow and fast phases alternate during convergence, because it might stop the iterative scheme at a very early stage where its speed is temporarily low and the numerical solution is not accurate at all)

$$\text{relative change} = \frac{\|\mathbf{x}^{(k)} - \mathbf{x}^{(k-1)}\|}{\|\mathbf{x}^{(k)}\|} < \varepsilon.$$

Lastly, a stopping criterion based on a fixed number of iteration $k_{\max}$ can be considered to prevent an excessive computational effort. It consists in doing exactly $k_{\max}$ iterations regardless of the behavior of the numerical solution. It might be useful when there is no a priori knowledge of the behavior of the iterative scheme, but it clearly does not give any information on the quality of the numerical approximation.

We anticipate that iterative methods (in particular the Richardson/fixed-point iteration, CGNR and GMRES) are extensively exploited throughout the thesis. Indeed, they are the tools we use to solve the linear system associated to the finite element discretization in Chapter 3 and Chapter 4 and their performances with respect to different benchmarks and physical configurations are subject to comparison in our study.

1.3.4 Static condensation and hybridization

As we discussed in Section 1.2, discontinuous Galerkin methods combined with high-order polynomial basis functions can provide stable high-fidelity numerical solutions for complicated geometric and physical configurations. However, from a computational point of view, they require the solution of large sparse unstructured linear systems. On the one hand, direct solvers require huge amounts of computation and are complicated to run efficiently on parallel computers. On the other hand, iterative solvers require much less memory and allow efficient parallel implementations, but they may exhibit slow convergence due to intrinsic properties of the time-harmonic problems [49]. Indeed, for nonzero wavenumber κ, the Helmholtz operator $\mathcal{L} = -\Delta - \kappa^2$ and its discrete form are indefinite, with both positive and negative eigenvalues. This lack of positivity undermines the convergence properties of classical iterative methods: stationary schemes like Jacobi or Gauss-Seidel may stagnate or diverge. Krylov methods can be used, but their performance is often poor without carefully designed preconditioners, particularly at high frequencies or in the case of resonance phenomena where the system becomes highly oscillatory and ill-conditioned. These difficulties are a primary motivation for specialized algorithms tailored to the indefinite, dispersive nature of the Helmholtz problem. To speed up convergence, preconditioning techniques and domain decomposition methods (DDM) have been and are being intensively studied, see e.g. [10, 46, 47, 57].

Given the size of the sparse linear systems arising in wave-propagation problems, efficient parallelization is a critical requirement. Distributing both the computational work and the associated data across multiple processors is essential not only to reduce runtime, but also to enable simulations at the scales and resolutions demanded by high-frequency applications. Although parallel implementation strategies are not part of the contributions presented here, they remain an important contextual factor that strongly influences solver design and practical applicability.

Static condensation and hybridization are other two techniques to obtain a reduced system that is easier to solve with standard algebraic methods. They are morally different even if they are very similar in computational implementation.

Static condensation is an algebraic manipulation of the global matrix associated to FEMs and it aims at reducing the size of the global system by eliminating internal degrees of freedom at the element level. It involves solving local equations to express internal variables in terms of boundary variables using the so-called *Schur complement*. This results in a smaller global system that involves only interface unknowns and that benefits of a better conditioning. After solving the reduced system, internal variables can be recovered locally. See e.g. [21, 29, 51, 120].

To understand what static condensation consists of in practice in terms of algebraic steps, we consider the physical system we aim at solving

$$\mathbf{A}\mathbf{x} = \mathbf{b},$$

where $\mathbf{x}$ denotes the vector of the physical unknowns, $\mathbf{A}$ is the associated physical matrix and $\mathbf{b}$ is the associated physical right-hand side. Splitting the vector $\mathbf{x}$ in its volume $\mathbf{x}_v$ and surface $\mathbf{x}_s$ components, we can rewrite the original system as

$$\begin{bmatrix} \mathbf{A}_{11} & \mathbf{A}_{12} \\ \mathbf{A}_{21} & \mathbf{A}_{22} \end{bmatrix} \begin{bmatrix} \mathbf{x}_v \\ \mathbf{x}_s \end{bmatrix} = \begin{bmatrix} \mathbf{b}_1 \\ \mathbf{b}_2 \end{bmatrix}, \tag{1.38}$$

which consists in the physical system itself written with respect to the volume and the surface variables and expresses the coupling between the latters.

Assuming that the block $\mathbf{A}_{11}$ is invertible, by mimicking the Gaussian elimination method, we first solve the first equation of system (1.38) with respect to $\mathbf{x}_v$, namely

$$\mathbf{x}_v = \mathbf{A}_{11}^{-1}(\mathbf{b}_1 - \mathbf{A}_{12}\mathbf{x}_s). \tag{1.39}$$

Then, we plug this result in the second equation of system (1.38), which gives

$$(\mathbf{A}_{22} - \mathbf{A}_{21}\mathbf{A}_{11}^{-1}\mathbf{A}_{12})\mathbf{x}_s = \mathbf{b}_2 - \mathbf{A}_{21}\mathbf{A}_{11}^{-1}\mathbf{b}_1. \tag{1.40}$$

If the matrix $\mathbf{A}_{22} - \mathbf{A}_{21}\mathbf{A}_{11}^{-1}\mathbf{A}_{12}$, called the Schur complement, is invertible, we can retrieve the solution of system (1.38) by inverting it to first obtain $\mathbf{x}_s$ from Equation (1.40) and then, by substituting in Equation (1.39), $\mathbf{x}_v$.

Another approach that aims at speeding up the convergence of iterative procedures consists in the class of *hybridizable discontinuous Galerkin methods* (HDG). The principle of HDG methods is to introduce a hybrid variable on the mesh skeleton that acts as a Lagrange multiplier in order to decouple the physical variables at the interface between the elements. The elimination of the physical variables within every element then leads to a hybridized system where the only unknown is the hybrid variable itself. This approach modifies the size, structure, and conditioning of the global system that is actually solved, potentially allowing for more efficient solution procedures. Once the hybrid variables are calculated, the physical variables inside each element can be recovered with a simple and inexpensive post-processing step. In particular, since the hybrid variable is defined on the skeleton, HDG methods present hybrid systems that conserve the sparsity pattern of the original standard DG methods, but exhibit significantly less degrees of freedom. This large reduction leads to a significant improvement in terms of both computational time and memory storage when direct solvers are exploited. Regarding iterative solvers the improvements are not a priori so clear since the size of the system and its structure are no longer the main performance criteria.

The first hybridization of a finite element method was proposed in 1965 [52] in the context of linear elasticity. Probably because at that time it was considered as an implementation technique, it was often interpreted as static condensation, indeed from a computational point of view, denoting by $\mathbf{g}$ the vector of the hybrid variable, the elimination of the physical variable is achieved by computing the Schur complement of the matrix $\mathbf{A}$ with respect to the hybrid variable $\mathbf{g}$ instead of the physical surface variable $\mathbf{x}_s$. However, in 1985 [5], hybridization was proved to be a more powerful tool since it was demonstrated that the hybrid variable contains additional information regarding the exact solution. This intuition led to improvements in the accuracy of the approximation via a local postprocessing [15, 16, 36].

Another important step was taken in 2005 [30] with the characterization of the hybrid variable as a solution of a variational problem where the physical one does not appear at all in the framework of the Raviart-Thomas (RT) and Brezzi-Douglas-Marini (BDM) mixed methods. In particular, it was shown that the hybrid variable simplified the assembly of the stiffness matrix and led to the identification of structural connections between HDG, RT, and BDM formulations, revealing that methods previously considered unrelated share a common variational framework.

Later, the hybridization approach was exploited in the setting of stationary Stokes equations with divergence-free velocities to overcome the difficult construction of this space of velocities. Specifically, it was applied to a DG method [22] and to a mixed one for Stokes flow [31, 32].

In the wave-propagation framework, studies concerning the application of the hybridization method started around fifteen years ago and the so-called numerical trace was classically taken as the hybrid variable, see e.g. [8, 9, 26, 33, 60, 104, 108, 115]. A priori, other choices of hybrid variable can be made and, a new alternative version of the method, called CHDG and introduced in 2023 [97] in the framework of the Helmholtz equation for homogeneous media, takes as hybrid variable a characteristic variable, which can be interpreted as a transmission variable or a Robin trace. Similar approaches applied on different frameworks were studied in [55]. Extensions to general time-harmonic scalar wave problems with constant coefficients will be characterized in Chapter 2, to the convected Helmholtz equation in Chapter 3, to the Helmholtz equation for heterogeneous media in Chapter 4 (see also [107]) and its application to the Maxwell equation for homogeneous media has recently been carried out in [111].

1.4 Summary

In this Chapter, we have presented the physical models for acoustic and aeroacoustic phenomena with a focus on the time-harmonic regime and the framework of the associated numerical methods. We have proposed a review of the current state of the art and, in particular, we have introduced the main finite element methods emphasizing those developed specifically to limit the pollution error, typical of wave propagation problems, and to allow fast iterative solutions. The common thread of this thesis is the inclusion of the wave behavior of the problem in the numerical methods with the hope of improving them (in terms of accuracy and/or speed of convergence of iterative solvers).

In this context, a hybridizable DG method with transmission variables and a Trefftz DG method with propagative plane waves are proposed and tested with two-dimensional benchmark problems in Chapters 2-3-4 and Chapter 5, respectively.

Chapter 2

CHDG method for time-harmonic wave problems in homogeneous media

In this Chapter, we present a hybridizable discontinuous Galerkin method for general time-harmonic wave problems with constant coefficients. In particular, we extend the method originally developed in [97] in the framework of acoustics in homogeneous media and later exploited in [111] in the context of electromagnetism for homogeneous media. The method relies on the introduction of transmission variables, which are related to the well-known characteristic variables, and they play a key role in the definition of the upwind fluxes exploited in the method.

The theory presented in this Chapter holds for an extensive class of problems and it provides a recipe that allows to write the CHDG formulation of any problem belonging to such a class starting from its governing equation. However, to investigate the quality of the performance of the method and to compare it with standard approaches, we limit ourselves to a particular setting and this is the goal of Chapter 3 in which we consider the framework of aeroacoustics.

After presenting the general model problem, we define and solve an associated Riemann problem to give some details regarding the physics and introduce the key definition of characteristic variables in Section 2.1. Then, to solve the problem numerically, in Section 2.2, we introduce the DG formulation for the problem and we define appropriate upwind numerical fluxes thanks to a flux vector splitting method. In Section 2.3, we introduce the transmission variables and we present and study the CHDG method.

2.1 Hyperbolic system and characteristic variables

We consider an unknown vector field $\boldsymbol{U}(\boldsymbol{x},t) \in \mathbb{R}^l$, with $l \in \mathbb{N}$, where $\boldsymbol{x} \in \Omega$ denotes the space variable. The bounded domain is $\Omega \subset \mathbb{R}^d$ with $d = 2$ or 3, it has boundary $\partial\Omega = \Gamma$ and $t \in \mathbb{R}^+$ is the time. The unknown is governed by the following *symmetric hyperbolic system*

$$\frac{\partial \boldsymbol{U}}{\partial t} + \boldsymbol{A}_j \frac{\partial \boldsymbol{U}}{\partial x_j} = \boldsymbol{G}_v, \quad \text{in } \Omega. \tag{2.1}$$

$\boldsymbol{A}_j \in \mathbb{R}^{l\times l}$ are real, constant, symmetric matrices, $\boldsymbol{G}_v : \Omega \times \mathbb{R}^+ \to \mathbb{R}^l$ is a given volume source vector and we adopt Einstein summation convention for repeated indices.

System (2.1) is said to be *hyperbolic* if and only if $\nu_j \boldsymbol{A}_j$ is diagonalizable in $\mathbb{R}$ for all $\boldsymbol{\nu} \in \mathbb{R}^l$. This condition is automatically satisfied under the previous hypothesis regarding the matrices $\boldsymbol{A}_j$.

To better understand the physics of the problem and to introduce some key aspects and definitions, we consider the time-dependent system in one dimension. It corresponds to Equation (2.1) on the unbounded domain $\mathbb{R}$ without volume source, i.e. $\boldsymbol{G}_v = \boldsymbol{0}$, complemented by an initial condition. The system reads

$$\begin{cases} \dfrac{\partial \boldsymbol{U}}{\partial t}(x_1,t) + \boldsymbol{A}_1 \dfrac{\partial \boldsymbol{U}}{\partial x_1}(x_1,t) = \boldsymbol{0}, & (x_1,t) \in \mathbb{R} \times \mathbb{R}^+, \\ \boldsymbol{U}(x_1,0) = \boldsymbol{U}^0(x_1), & x_1 \in \mathbb{R}. \end{cases} \tag{2.2}$$

Since $\boldsymbol{A}_1$ is real and symmetric, it is diagonalizable in $\mathbb{R}$ and it admits the following eigendecomposition

$$\boldsymbol{A}_1 = \boldsymbol{W}\boldsymbol{\Lambda}\boldsymbol{W}^{-1} = \boldsymbol{W}\boldsymbol{\Lambda}\boldsymbol{W}^{\dagger},$$

where $\boldsymbol{W} = [\boldsymbol{w}_1, ..., \boldsymbol{w}_l] \in \mathbb{R}^{l\times l}$ is the orthogonal matrix (i.e. $\boldsymbol{W}^{-1} = \boldsymbol{W}^{\dagger}$) whose columns are the eigenvectors $\{\boldsymbol{w}_i\}_{i=1,...,l}$ and $\boldsymbol{\Lambda} = \mathrm{diag}(\lambda_i) \in \mathbb{R}^{l\times l}$ is the diagonal matrix whose elements are the eigenvalues $\{\lambda_i\}_{i=1,...,l}$.

By following the standard approach (see e.g. [117]), we left-multiply the first equation of (2.2) by $\boldsymbol{W}^{\dagger}$, we introduce the variable

$$\boldsymbol{V} := \boldsymbol{W}^{\dagger}\boldsymbol{U} \tag{2.3}$$

and we get the following system of uncoupled differential equations which is called *canonical form* or *characteristic form* of the first equation of (2.2)

$$\frac{\partial \boldsymbol{V}}{\partial t} + \boldsymbol{\Lambda}\frac{\partial \boldsymbol{V}}{\partial x_1} = \boldsymbol{0}, \tag{2.4}$$

complemented by the initial datum $\boldsymbol{V}^0 := \boldsymbol{W}^{\dagger}\boldsymbol{U}^0$. The components of the variable $\boldsymbol{V}$ are often called *characteristic variables* in the literature. They play a key role in the theory of Riemann problems [88, 94] and Riemann solvers [91, 100, 123]. Note that the definition of the characteristic variables is not unique, as they can differ by a multiplicative coefficient.

Equation (2.4) is equivalent to the following set of uncoupled differential equations

$$\frac{\partial V_i}{\partial t} + \lambda_i \frac{\partial V_i}{\partial x_1} = 0, \quad i = 1, ..., l. \tag{2.5}$$

The solution of system (2.5) is given by the well-known method of characteristics (see e.g. [117, 123]) and reads

$$V_i(x_1,t) = V_i^0(x_1 - \lambda_i t), \quad i = 1, ..., l. \tag{2.6}$$

Lastly, we can retrieve the solution $\boldsymbol{U}$ of the original problem by computing

$$\boldsymbol{U}(x_1,t) = \sum_{i=1}^{l} V_i^0(x_1 - \lambda_i t)\boldsymbol{w}_i. \tag{2.7}$$

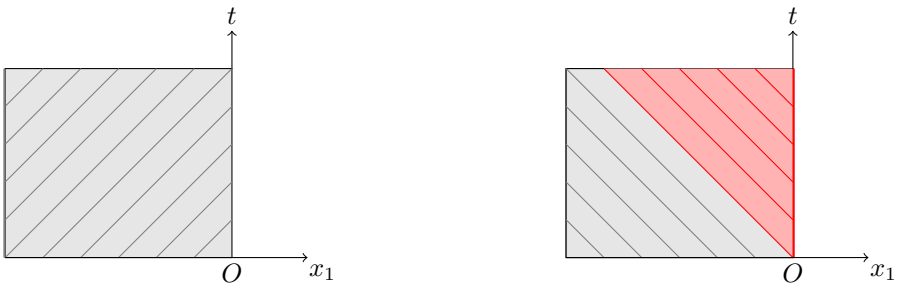
In other words, the function $V_i^0(x_1 - \lambda_i t)$ is the coefficient of $\boldsymbol{w}_i$ in an eigenvector expansion of the vector $\boldsymbol{U}$. The solution $\boldsymbol{U}$ can consequently be understood as the superposition of l waves where the i-th wave has shape $V_i^0(x_1 - \lambda_i t)\boldsymbol{w}_i$ and propagates with speed λ_i.

Remark 2.1.1. It is possible to prove (see e.g. [117]) that the line $x_1(t)$, called *characteristic line* and defined by

$$\frac{dx_1(t)}{dt} = \lambda_i,$$

represents the line in the plane (x_1, t) along which the solution is constant. If this line at $t = 0$ admits a point $x_1(0)$ that belongs to the support of the initial datum V_i^0, then the solution reads as (2.6), namely the solution V_i in (x_1, t) is given by the initial datum V_i^0 in $x_1(0) = x_1 - \lambda_i t$. However, if $x_1(0)$ does not belong to the support of the initial datum V_i^0, a boundary condition is needed.

For instance, let us consider the domain $\mathbb{R}^-$. If $\lambda_i \geq 0$, then for all $(x_1, t) \in \mathbb{R}^- \times \mathbb{R}^+$ we have that $x_1(0) \in \mathbb{R}^-$ (see Figure 2.1a). However if $\lambda_i < 0$, then the boundary condition $V_i(0, t) = g_i(t)$ for a given sufficiently regular $g_i : \mathbb{R}^+ \to \mathbb{R}$ is required for all $(x_1, t) \in \mathbb{R}^- \times \mathbb{R}^+$ such that $x_1 > \lambda_i t$ (see Figure 2.1b) and the solution reads $V_i(x_1, t) = g_i(t - x_1/\lambda_i)$.



(a) $\lambda_i > 0$: no boundary condition is needed.

(b) $\lambda_i < 0$: a boundary condition is needed at $x_1 = 0$.

Figure 2.1: Characteristic lines in the plane (x_1, t) depending on the sign of λ_i. The gray ones require an initial condition whereas the red ones require a boundary condition at $x_1 = 0$.

In other words, negative (positive, respectively) eigenvalues are associated to the propagation of characteristic variables towards the interior (exterior, respectively). Since the solution $\boldsymbol{U}$ consists in a linear combination of V_i, $i = 1, ..., l$ (see Equation (2.7)), if there are m negative eigenvalues, then exactly m boundary conditions, which can be interpreted as a necessary condition for the well-posedness of the physical problem, are needed (see [74, 125]).

2.1.1 Time-harmonic problem

By assuming $\boldsymbol{U} = \mathrm{Re}\{\tilde{\boldsymbol{U}} e^{-\imath\omega t}\}$ and $\boldsymbol{G}_v = \mathrm{Re}\{\tilde{\boldsymbol{G}}_v e^{-\imath\omega t}\}$, where ω is the angular frequency and $\tilde{\boldsymbol{U}}$ and $\tilde{\boldsymbol{G}}_v$ are the complex amplitudes, we can write the time-harmonic formulation associated to Equation (2.1) for the unknown $\tilde{\boldsymbol{U}}(\boldsymbol{x}) \in \mathbb{C}^l$

$$-\imath\omega\tilde{\boldsymbol{U}} + \boldsymbol{A}_j \frac{\partial \tilde{\boldsymbol{U}}}{\partial x_j} = \tilde{\boldsymbol{G}}_v, \quad \text{in } \Omega. \tag{2.8}$$

Equation (2.8) needs to be complemented by m boundary conditions that can be written in the following abstract form

$$\boldsymbol{B}\tilde{\boldsymbol{U}} = \tilde{\boldsymbol{G}}_s, \quad \text{on } \Gamma,$$

where $\boldsymbol{B} \in \mathbb{R}^{m \times l}$ is a matrix and $\tilde{\boldsymbol{G}}_s : \Gamma \to \mathbb{C}^m$ is a given surface source vector. Note that the number m of boundary conditions that need to be prescribed strongly depends on the physics of the problem and, in particular, on the matrices $\boldsymbol{A}_j$ as we explain in Remark 2.1.1 of Section 2.1.

Finally, we want to focus on the time-harmonic regime and, consequently, we remove the $\sim$ to simplify the notation and we consider the problem

Problem 2.1.2. *Find* $\boldsymbol{U} : \Omega \to \mathbb{C}^l$ *such that*

$$\begin{cases} -\imath\omega\boldsymbol{U} + \boldsymbol{A}_j \dfrac{\partial \boldsymbol{U}}{\partial x_j} = \boldsymbol{G}_v, & \text{in } \Omega, \\ \boldsymbol{B}\boldsymbol{U} = \boldsymbol{G}_s, & \text{on } \Gamma, \end{cases}$$

for some given data $\boldsymbol{G}_v : \Omega \to \mathbb{C}^l$ *and* $\boldsymbol{G}_s : \Gamma \to \mathbb{C}^m$.

2.2 Discontinuous Galerkin method

The goal of this section is to introduce the discontinuous Galerkin discretization of Problem 2.1.2.

2.2.1 Mesh, approximation spaces and inner products

We consider a conforming mesh $\mathcal{T}_h$ of the domain Ω consisting of simplicial elements K whose faces are denoted by the generic letter F. The collection of element boundaries is denoted by $\partial\mathcal{T}_h := \{\partial K \mid K \in \mathcal{T}_h\}$, and the collection of faces of the mesh is denoted by $\mathcal{F}_h$. The collection of faces of an element K is denoted by $\mathcal{F}_K$.

The approximate fields are piecewise polynomials. For simplicity, we fix a polynomial order $\mathrm{p} \geq 0$ and introduce

$$\boldsymbol{\mathcal{V}}_h^l := \prod_{K\in\mathcal{T}_h} \boldsymbol{\mathcal{P}}_{\mathrm{p}}^l(K),$$

where $\boldsymbol{\mathcal{P}}_{\mathrm{p}}^l(\cdot)$ denotes the space of vector complex-valued (in dimension l) polynomials of order smaller or equal to p. The restrictions of $\boldsymbol{U}_h \in \boldsymbol{\mathcal{V}}_h^l$ on K is denoted $\boldsymbol{U}_K$. For a face F of K, $\boldsymbol{n}_{K,F}$ is the unit outward normal to K on F.

We introduce the sesquilinear forms

$$(\boldsymbol{U},\boldsymbol{V})_K := \int_K \boldsymbol{U}\cdot\overline{\boldsymbol{V}}\,d\boldsymbol{x}, \qquad \langle\boldsymbol{U},\boldsymbol{V}\rangle_{\partial K} := \sum_{F\in\mathcal{F}_K}\int_F \boldsymbol{U}\cdot\overline{\boldsymbol{V}}\,d\sigma(\boldsymbol{x}),$$
$$(\boldsymbol{U},\boldsymbol{V})_{\mathcal{T}_h} := \sum_{K\in\mathcal{T}_h}(\boldsymbol{U},\boldsymbol{V})_K, \qquad \langle\boldsymbol{U},\boldsymbol{V}\rangle_{\partial\mathcal{T}_h} := \sum_{K\in\mathcal{T}_h}\langle\boldsymbol{U},\boldsymbol{V}\rangle_{\partial K},$$

where the quantities used in the surface integral $\langle\cdot,\cdot\rangle_{\partial K}$ correspond to the restriction of fields defined on K (e.g. $\boldsymbol{V}_K$) or quantities associated with the faces of K (e.g. $\boldsymbol{n}_{K,F}$ with $F \in \mathcal{F}_K$), unless explicitly specified.

2.2.2 Standard DG formulation

To obtain the variational formulation of Problem 2.1.2, we multiply the first equation of Problem 2.1.2 by a test function $\boldsymbol{V}_h \in \boldsymbol{\mathcal{V}}_h^l$, we integrate on each element $K \in \mathcal{T}_h$, we integrate by parts and we sum all over the elements of the mesh. In the following, for the sake of simplicity, we consider a null volume source, i.e. $\boldsymbol{G}_v = \boldsymbol{0}$, however a non-null volume source could be included without any difficulty.

Problem 2.2.1. *Find $\boldsymbol{U}_h \in \boldsymbol{\mathcal{V}}_h^l$ such that, for all $\boldsymbol{V}_h \in \boldsymbol{\mathcal{V}}_h^l$,*

$$-\imath\omega(\boldsymbol{U}_h,\boldsymbol{V}_h)_{\mathcal{T}_h} - \Big(\boldsymbol{A}_j\boldsymbol{U}_h, \frac{\partial\boldsymbol{V}_h}{\partial x_j}\Big)_{\mathcal{T}_h} + \langle\boldsymbol{F}\boldsymbol{U}_h,\boldsymbol{V}_h\rangle_{\partial\mathcal{T}_h} = 0.$$

The real symmetric matrix $\boldsymbol{F}$ is called *flux matrix* and it is defined as

$$\boldsymbol{F} := \boldsymbol{A}_j n_j.$$

Problem 2.2.1 is not written in conservative form. Indeed, the quantity $\boldsymbol{F}\boldsymbol{U}_h$ that appears in the boundary term represents the normal flux of a priori non-conserved quantities through each face $F \in \partial\mathcal{T}_h$ because $\boldsymbol{F}\boldsymbol{U}_K \neq \boldsymbol{F}\boldsymbol{U}_{K'}$. In other words, the quantity $\boldsymbol{F}\boldsymbol{U}_h$ is not uniquely defined across each face F because the numerical solution $\boldsymbol{U}_h$ is a priori discontinuous.

To overcome this issue and write the problem in conservative form, we rather consider the following standard discontinuous Galerkin (DG) formulation of Problem 2.1.2.

Problem 2.2.2. *Find* $\boldsymbol{U}_h \in \boldsymbol{\mathcal{V}}_h^l$ *such that, for all* $\boldsymbol{V}_h \in \boldsymbol{\mathcal{V}}_h^l$,

$$-\imath\omega(\boldsymbol{U}_h, \boldsymbol{V}_h)_{\mathcal{T}_h} - \Big(\boldsymbol{A}_j \boldsymbol{U}_h, \frac{\partial \boldsymbol{V}_h}{\partial x_j}\Big)_{\mathcal{T}_h} + \langle \boldsymbol{F}\widehat{\boldsymbol{U}}_h, \boldsymbol{V}_h\rangle_{\partial\mathcal{T}_h} = 0.$$

The *numerical flux* $\boldsymbol{F}\widehat{\boldsymbol{U}}_h$ is defined face by face. Its main purpose is to provide a consistent and stable way to approximate the flux across each face $F \in \partial\mathcal{T}_h$, where the numerical solution $\boldsymbol{U}_h$ is a priori discontinuous. It acts as a surrogate for the physical flux at interfaces and determines how information is exchanged between neighboring elements.

It is important to highlight the fact that, for an internal face F, the normal flux is a local interface quantity, and therefore it is influenced only by the physical fields defined on the two neighboring elements K and K' that share the face F. Consequently, the definition of the numerical flux only involves the physical fields $\boldsymbol{U}_K$ and $\boldsymbol{U}_{K'}$ (see Figure 2.2).

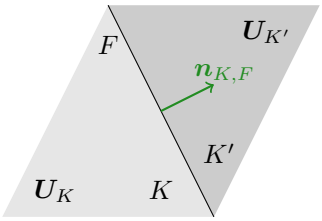


Figure 2.2: Notation for $\boldsymbol{U}_K$ and $\boldsymbol{U}_{K'}$ defined on two neighboring elements K and K' that share the face F.

Several types of numerical flux have been proposed in the literature, each with different properties and advantages (see e.g. [91, 110, 123]). See Figure 2.2 for the notation we use hereafter.

- *Central flux*: for each interior face $F \not\subset \partial\Omega$, it is defined as

  $$\boldsymbol{F}\widehat{\boldsymbol{U}}_F := \frac{1}{2}(\boldsymbol{F}\boldsymbol{U}_K + \boldsymbol{F}\boldsymbol{U}_{K'}),$$

  namely as the average of the fluxes of both sides of the face F. It is symmetric and dispersive, but it is not dissipative.

- *Lax-Friedrichs flux*: for each interior face $F \not\subset \partial\Omega$, it is defined as

  $$\boldsymbol{F}\widehat{\boldsymbol{U}}_F := \frac{1}{2}(\boldsymbol{F}\boldsymbol{U}_K + \boldsymbol{F}\boldsymbol{U}_{K'}) - \delta(\boldsymbol{U}_{K'} - \boldsymbol{U}_K),$$

  namely it consists in the central flux with an additional penalty term which is proportional to the jump across the face by the coefficient $\delta > 0$ which controls dissipation: the bigger δ is the more numerical dissipation is included, possibly at the expense of accuracy.

- *Upwind flux*: we consider the eigendecomposition $\boldsymbol{F} = \boldsymbol{W}\boldsymbol{\Lambda}\boldsymbol{W}^\dagger$ and a splitting of the matrix $\boldsymbol{\Lambda}$ in its blocks $\boldsymbol{\Lambda}^-$ and $\boldsymbol{\Lambda}^+$ and of the matrix $\boldsymbol{W}$ in its submatrices $\boldsymbol{W}^-$ and $\boldsymbol{W}^+$. We denote by $\boldsymbol{\Lambda}^- \in \mathbb{R}^{m\times m}$ and $\boldsymbol{\Lambda}^+ \in \mathbb{R}^{n\times n}$ the two diagonal blocks that compose the matrix $\boldsymbol{\Lambda}$ and that contain only the m strictly negative eigenvalues and the n strictly positive ones, respectively. In a similar way, we denote by $\boldsymbol{W}^- \in \mathbb{R}^{l\times m}$ and $\boldsymbol{W}^+ \in \mathbb{R}^{l\times n}$ the matrices whose columns are the eigenvectors associated only to the strictly negative eigenvalues and to the strictly positive ones, respectively.

  For interior face $F \not\subset \partial\Omega$, the upwind flux is defined as

  $$\boldsymbol{F}\widehat{\boldsymbol{U}}_F := \boldsymbol{F}^+\boldsymbol{U}_K + \boldsymbol{F}^-\boldsymbol{U}_{K'}, \tag{2.9}$$

where

$$\boldsymbol{F}^- = \boldsymbol{W}^-\boldsymbol{\Lambda}^-\boldsymbol{W}^{-\dagger} \quad \text{and} \quad \boldsymbol{F}^+ = \boldsymbol{W}^+\boldsymbol{\Lambda}^+\boldsymbol{W}^{+\dagger}.$$

It accounts for the direction of wave propagation by using information from the upstream side of each interface F. It mimics the physical flow of information, enhancing stability by adding dissipation only where needed and ensuring robustness in the numerical scheme. The splitting method we have just used to define the upwind numerical flux had been exploited before in [43, 119] in the framework of acoustics and electromagnetism, respectively.

Note that boundary faces $F \subset \partial\Omega$ are clearly not shared by two elements K and K', and therefore the definitions of the numerical fluxes we have just introduced can not be used. In practice, the numerical flux is chosen in such a way that $\boldsymbol{B}\widehat{\boldsymbol{U}}_F = \boldsymbol{G}_s$ for the sake of consistency.

2.3 Hybridization with transmission variables

We want to perform hybridization by introducing an additional variable. To do so, we define on each interior face F shared by two neighboring elements K and K' two trace quantities $\boldsymbol{x}^{\ominus}_{K,F} \in \mathbb{C}^m, \boldsymbol{x}^{\oplus}_{K,F} \in \mathbb{C}^n$ with $n \in \mathbb{N}$ written in terms of the physical fields defined only on K and K', respectively.

2.3.1 Transmission variables

Inspired by the standard approach detailed in Section 2.1, we left-multiply the variable (2.3) by $\sqrt{|\boldsymbol{\Lambda}|}$, that is the diagonal matrix whose elements are the square roots of the absolute value of the eigenvalues of the matrix $\boldsymbol{F}$. This particular scaling choice is necessary to write and prove the energy estimate presented in Lemma 2.3.11. A priori there could be other possible scalings (or, equivalently, other norms) with respect to which Lemma 2.3.11 holds. We obtain the variable

$$\boldsymbol{x} := \sqrt{|\boldsymbol{\Lambda}|}\boldsymbol{W}^\dagger\boldsymbol{U} = \begin{bmatrix} \boldsymbol{x}^{\ominus} \\ \boldsymbol{0} \\ \boldsymbol{x}^{\oplus} \end{bmatrix},$$

with components $\boldsymbol{x}^{\ominus} = \sqrt{-\boldsymbol{\Lambda}^-}\boldsymbol{W}^{-\dagger}\boldsymbol{U} \in \mathbb{C}^m$, $\boldsymbol{0} \in \mathbb{C}^{l-m-n}$ and $\boldsymbol{x}^{\oplus} = \sqrt{\boldsymbol{\Lambda}^+}\boldsymbol{W}^{+\dagger}\boldsymbol{U} \in \mathbb{C}^n$.

Note that in the following we refer to $\boldsymbol{x}^{\ominus} \in \mathbb{C}^m$ and $\boldsymbol{x}^{\oplus} \in \mathbb{C}^n$ as *transmission variables*, they can be interpreted as information propagating towards the interior and the exterior of an element, respectively, and they correspond to a simple scaling of the characteristic variables we introduced before. The scaling simply consists in multiplicative coefficients that are equal to the square root of the absolute value of the eigenvalues of the flux matrix $\boldsymbol{F}$. These eigenvalues correspond to the velocities of propagation of the information transported by the transmission variables themselves.

Motivated by the definition of the upwind numerical flux, we define the *outgoing transmission variable* $\boldsymbol{x}^{\oplus}_{K,F}$ and the *incoming transmission variable* $\boldsymbol{x}^{\ominus}_{K,F}$ as follows

$$\boldsymbol{x}^{\oplus}_{K,F} := \sqrt{\boldsymbol{\Lambda}^+}\boldsymbol{W}^{+\dagger}\boldsymbol{U}_{K,F}, \qquad \forall F, \tag{2.10}$$

$$\boldsymbol{x}^{\ominus}_{K,F} := \sqrt{-\boldsymbol{\Lambda}^-}\boldsymbol{W}^{-\dagger}\boldsymbol{U}_{K',F}, \qquad \text{if } F \not\subset \partial\Omega. \tag{2.11}$$

Note that $\boldsymbol{x}^{\ominus}_{K,F} = \boldsymbol{x}^{\oplus}_{K',F}$ if $F \not\subset \partial\Omega$.

We can use these definitions to rewrite the upwind flux (2.9) for interior faces $F \not\subset \partial\Omega$ as follows

$$\boldsymbol{F}\widehat{\boldsymbol{U}}_F = \boldsymbol{W}^+\sqrt{\boldsymbol{\Lambda}^+}\boldsymbol{x}^{\oplus}_{K,F} - \boldsymbol{W}^-\sqrt{-\boldsymbol{\Lambda}^-}\boldsymbol{x}^{\ominus}_{K,F}. \tag{2.12}$$

The following important property regarding the upwind numerical flux holds.

Property 2.3.1. *The upwind numerical flux $\boldsymbol{F}\widehat{\boldsymbol{U}}_F$ on interior faces $F \not\subset \partial\Omega$ is the same for both neighboring elements K and K'.*

Proof. This property is a direct consequence of the definition of the upwind numerical flux, see either Equation (2.9) or Equation (2.12). □

Property 2.3.1 states a necessary condition for the *consistency* of the numerical method to hold. Indeed, since the numerical flux is an approximation of the physical one which is continuous across the faces, it needs to be uniquely defined.

It is also possible to uniquely define the trace $\widehat{\boldsymbol{U}}_F$ and, in particular, to characterize its projections on the spaces generated by the columns of the matrices $\boldsymbol{W}^+$ and $\boldsymbol{W}^-$ which appear in the definition of the transmission variables as stated by the following Property.

Property 2.3.2. *The outgoing and incoming transmission variables verify*

$$\boldsymbol{x}^{\oplus}_{K,F} = \sqrt{\boldsymbol{\Lambda}^+}\boldsymbol{W}^{+\dagger}\widehat{\boldsymbol{U}}_F \quad \text{and} \quad \boldsymbol{x}^{\ominus}_{K,F} = \sqrt{-\boldsymbol{\Lambda}^-}\boldsymbol{W}^{-\dagger}\widehat{\boldsymbol{U}}_F$$

on interior faces $F \not\subset \partial\Omega$.

Proof. We left-multiply Equation (2.12) by $\sqrt{(\boldsymbol{\Lambda}^+)^{-1}}\boldsymbol{W}^{+\dagger}$ and we get

$$\boldsymbol{x}^{\oplus}_{K,F} - \sqrt{(\boldsymbol{\Lambda}^+)^{-1}}\boldsymbol{W}^{+\dagger}\boldsymbol{W}^-\sqrt{-\boldsymbol{\Lambda}^-}\boldsymbol{x}^{\ominus}_{K,F} = \sqrt{(\boldsymbol{\Lambda}^+)^{-1}}\boldsymbol{W}^{+\dagger}\boldsymbol{F}\widehat{\boldsymbol{U}}_F.$$

Since the matrix $\boldsymbol{W}$ is orthogonal, we have that $\boldsymbol{W}^{+\dagger}\boldsymbol{W}^- = \boldsymbol{0}$ and, consequently, the second term in the left-hand side cancels out and we are left with

$$\begin{aligned}\boldsymbol{x}^{\oplus}_{K,F} &= \sqrt{(\boldsymbol{\Lambda}^+)^{-1}}\boldsymbol{W}^{+\dagger}\boldsymbol{W}\boldsymbol{\Lambda}\boldsymbol{W}^{\dagger}\widehat{\boldsymbol{U}}_F,\\ &= \sqrt{\boldsymbol{\Lambda}^+}\boldsymbol{W}^{+\dagger}\widehat{\boldsymbol{U}}_F.\end{aligned}$$

The result regarding $\boldsymbol{x}^{\ominus}_{K,F}$ can be similarly obtained by left-multiplying Equation (2.12) by $\sqrt{(\boldsymbol{\Lambda}^-)^{-1}}\boldsymbol{W}^{-\dagger}$. □

Since the outgoing transmission variable depends only on the values of the physical fields defined on the current element K, its definition can be extended to boundary faces. On the contrary, the incoming transmission variable depends only on the values of the physical fields defined on the neighboring element K', therefore its definition can not be extended to boundary faces (they are not shared with any neighboring element), but it can be properly changed by taking into account the boundary conditions. In particular, combining Property 2.3.2 with $\boldsymbol{B}\widehat{\boldsymbol{U}}_F = \boldsymbol{G}_s$ (the prescription of boundary conditions corresponds to assigning some of the components of $\widehat{\boldsymbol{U}}_F$), it is possible to define the incoming transmission variable on boundary faces as follows

$$\boldsymbol{x}^{\ominus}_{K,F} := \boldsymbol{\alpha}\boldsymbol{x}^{\oplus}_{K,F} + \boldsymbol{\beta}\boldsymbol{G}_s, \qquad \text{if } F \subset \partial\Omega, \tag{2.13}$$

for two matrices $\boldsymbol{\alpha} \in \mathbb{R}^{m\times n}$ and $\boldsymbol{\beta} \in \mathbb{R}^{m\times m}$ that depend on the specific boundary condition in such a way that Equation (2.12) holds also for boundary faces $F \subset \partial\Omega$. In practice, the matrix $\boldsymbol{\alpha}$ can be interpreted as a reflection coefficient.

2.3.2 CHDG formulation

Two additional variables, denoted $\boldsymbol{x}^{\ominus}_h$ and $\boldsymbol{x}^{\oplus}$, corresponding to the incoming and outgoing transmission variables respectively, are introduced on the skeleton to rigorously define the CHDG formulation. They belong respectively to the spaces

$$\boldsymbol{\mathcal{X}}^-_h := \prod_{K\in\mathcal{T}_h}\prod_{F\in\mathcal{F}_K}\boldsymbol{\mathcal{P}}^{m_{K,F}}_{\mathrm{p}}(F) \quad \text{and} \quad \boldsymbol{\mathcal{X}}^+_h := \prod_{K\in\mathcal{T}_h}\prod_{F\in\mathcal{F}_K}\boldsymbol{\mathcal{P}}^{n_{K,F}}_{\mathrm{p}}(F).$$

Note that $m_{K,F}$ is the number of negative eigenvalues and $n_{K,F}$ is the number of positive ones of the flux matrix $\boldsymbol{F}$ associated to the face F of the element K and, therefore, they depend on the face F of the element K itself.

These spaces are equipped with the following norm

$$\|\boldsymbol{x}_h\| := \sqrt{\sum_{K\in\mathcal{T}_h}\sum_{F\in\mathcal{F}_K}\|\boldsymbol{x}_{K,F}\|_F^2},$$

where $\|\cdot\|_F := \sqrt{\langle\cdot,\cdot\rangle_F}$ is the usual $L^2(F)$ norm defined for vectors with the proper size.

In order to derive a weak formulation, the outgoing transmission variable $\boldsymbol{x}^{\oplus}_{K,F}$ that appears in the numerical flux (2.12) is rewritten in terms of the physical variables using its definition (2.10). Moreover, the definition of $\boldsymbol{x}^{\ominus}_{K,F}$ and the boundary conditions are weakly imposed by multiplying Equations (2.11)-(2.13) by a test function $\xi_{K,F} \in \boldsymbol{\mathcal{P}}_{\mathrm{p}}^{m_{K,F}}(F)$. In this way, we are led to the following CHDG formulation.

Problem 2.3.3. *Find* $(\boldsymbol{U}_h, \boldsymbol{x}_h^{\ominus}) \in \boldsymbol{\mathcal{V}}_h^l \times \boldsymbol{\mathcal{X}}_h^-$ *such that, for all* $(\boldsymbol{V}_h, \boldsymbol{\xi}_h) \in \boldsymbol{\mathcal{V}}_h^l \times \boldsymbol{\mathcal{X}}_h^-$,

$$-\imath\omega(\boldsymbol{U}_h, \boldsymbol{V}_h)_{\mathcal{T}_h} - \Big(\boldsymbol{A}_j\boldsymbol{U}_h, \frac{\partial \boldsymbol{V}_h}{\partial x_j}\Big)_{\mathcal{T}_h} + \langle \boldsymbol{F}\widehat{\boldsymbol{U}}_h(\boldsymbol{U}_h, \boldsymbol{x}_h^{\ominus}), \boldsymbol{V}_h\rangle_{\partial\mathcal{T}_h} = 0,$$

and

$$\langle \boldsymbol{x}_h^{\ominus} - \Pi(\boldsymbol{x}^{\oplus}(\boldsymbol{U}_h)), \boldsymbol{\xi}_h\rangle_{\partial\mathcal{T}_h} = \langle \boldsymbol{b}, \boldsymbol{\xi}_h\rangle_{\partial\mathcal{T}_h},$$

with $\boldsymbol{x}^{\oplus}(\boldsymbol{U}_h) := \sqrt{\boldsymbol{\Lambda}^+}\boldsymbol{W}^{+\dagger}\boldsymbol{U}_h$.

To simplify the presentation, we have introduced the *global exchange operator* $\Pi : \boldsymbol{\mathcal{X}}_h^+ \to \boldsymbol{\mathcal{X}}_h^-$ and the *global right-hand side* $\boldsymbol{b}$. For each face F of each element K, and for any $\boldsymbol{x}^{\oplus} \in \boldsymbol{\mathcal{X}}_h^+$, they are defined as

$$\Pi(\boldsymbol{x}^{\oplus})|_{K,F} = \begin{cases} \boldsymbol{x}^{\oplus}_{K',F} & \text{if } F \not\subset \partial\Omega, \\ \boldsymbol{\alpha}\boldsymbol{x}^{\oplus}_{K,F} & \text{if } F \subset \partial\Omega, \end{cases} \qquad \boldsymbol{b}|_{K,F} = \begin{cases} \boldsymbol{0} & \text{if } F \not\subset \partial\Omega, \\ \boldsymbol{\beta}\boldsymbol{G}_s & \text{if } F \subset \partial\Omega. \end{cases}$$

The operator Π allows to enforce the weak coupling of the element-wise problems at the interior faces and the boundary conditions at the boundary faces. At interior faces, it simply swaps the outgoing and incoming transmission variables of the two neighboring elements. Its definition is properly modified at boundary faces to account for boundary conditions.

From the definition of the global exchange operator Π, we can retrieve some considerations regarding the admissibility of the boundary conditions. Indeed, for interior faces $F \not\subset \partial\Omega$, we have

$$\|\boldsymbol{x}^{\ominus}_{K,F}\| = \|\boldsymbol{x}^{\oplus}_{K',F}\|,$$

namely, the norm of the incoming transmission variable is equal to the norm of the outgoing one. This fact can be interpreted as conservation of energy across the interface F. For boundary faces $F \subset \partial\Omega$ we have

$$\|\boldsymbol{x}^{\ominus}_{K,F}\| \le \|\boldsymbol{\alpha}\|\|\boldsymbol{x}^{\oplus}_{K,F}\|,$$

with

$$\|\boldsymbol{\alpha}\| := \max_{i=1,\dots,n}\sqrt{|\lambda_i(\boldsymbol{\alpha}^\dagger\boldsymbol{\alpha})|},$$

namely, the norm of the incoming transmission variable is smaller than the norm of the outgoing one if $\|\boldsymbol{\alpha}\| \leq 1$. In other words, we can interpret the face F as passive if $\|\boldsymbol{\alpha}\| \leq 1$.

For numerical reasons we avoid energy amplification (see later Theorem 2.3.13), therefore we will always impose boundary conditions that assure conservation or dissipation of energy across the boundary of the domain. To sum up, from now on, we will make the following assumption.

Assumption 2.3.4. *The boundary condition is such that*

$$\|\boldsymbol{\alpha}\| \leq 1.$$

2.3.3 Local element-wise discrete problems

The physical field $\boldsymbol{U}_h$ can be eliminated from Problem 2.3.3 by solving local element-wise problems where the incoming transmission variable $\boldsymbol{x}_h^{\ominus}$ is considered as a given datum.

For each element K, the local problem reads as follows.

Problem 2.3.5. *Find $\boldsymbol{U}_K \in \boldsymbol{\mathcal{P}}^l_{\mathrm{p}}(K)$ such that, for all $\boldsymbol{V}_K \in \boldsymbol{\mathcal{P}}^l_{\mathrm{p}}(K)$,*

$$-\imath\omega(\boldsymbol{U}_K, \boldsymbol{V}_K)_K - \left(\boldsymbol{A}_j \boldsymbol{U}_K, \frac{\partial \boldsymbol{V}_K}{\partial x_j}\right)_K + \langle \boldsymbol{F}^+ \boldsymbol{U}_K, \boldsymbol{V}_K \rangle_{\partial K} = \langle \boldsymbol{S}, \boldsymbol{V}_K \rangle_{\partial K}, \tag{2.14}$$

for a given surface datum $\boldsymbol{S} \in \boldsymbol{\mathcal{P}}^l_{\mathrm{p}}(K)$ for all $F \in \mathcal{F}_K$.

Remark 2.3.6. For each face F of an element K, the right-hand side $\boldsymbol{S}$ depends on the incoming transmission variable and reads

$$\boldsymbol{S}|_{K,F} = \boldsymbol{W}^- \sqrt{-\boldsymbol{\Lambda}^-}\, \boldsymbol{x}^{\ominus}_{K,F}.$$

Theorem 2.3.7 (Well-posedness of the local discrete problem)**.** *Problem 2.3.5 is well-posed.*

Proof. We simply have to prove that, if $\boldsymbol{S} = \boldsymbol{0}$ on ∂K, the unique solution of Problem 2.3.5 is $\boldsymbol{U}_K = 0$. For the sake of brevity, the subscripts K and F are omitted for the local fields, the test functions and the outgoing unit normal. Taking $\boldsymbol{V} = \boldsymbol{U}$ in Equation (2.14) gives

$$-\imath\omega(\boldsymbol{U}, \boldsymbol{U})_K - \left(\boldsymbol{A}_j \boldsymbol{U}, \frac{\partial \boldsymbol{U}}{\partial x_j}\right)_K + \langle \boldsymbol{F}^+ \boldsymbol{U}, \boldsymbol{U} \rangle_{\partial K} = 0. \tag{2.15}$$

We integrate by parts the second term

$$-\imath\omega(\boldsymbol{U}, \boldsymbol{U})_K + \left(\boldsymbol{A}_j \frac{\partial \boldsymbol{U}}{\partial x_j}, \boldsymbol{U}\right)_K - \langle \boldsymbol{F}^- \boldsymbol{U}, \boldsymbol{U} \rangle_{\partial K} = 0.$$

We take the complex conjugate of the previous equation

$$+\imath\omega(\boldsymbol{U}, \boldsymbol{U})_K + \left(\boldsymbol{U}, \boldsymbol{A}_j \frac{\partial \boldsymbol{U}}{\partial x_j}\right)_K - \langle \boldsymbol{U}, \boldsymbol{F}^- \boldsymbol{U} \rangle_{\partial K} = 0,$$

that, thanks to the symmetry of the matrices $\boldsymbol{A}_j$ and $\boldsymbol{F}^-$, can be rewritten as

$$+\imath\omega(\boldsymbol{U}, \boldsymbol{U})_K + \left(\boldsymbol{A}_j \boldsymbol{U}, \frac{\partial \boldsymbol{U}}{\partial x_j}\right)_K - \langle \boldsymbol{F}^- \boldsymbol{U}, \boldsymbol{U} \rangle_{\partial K} = 0. \tag{2.16}$$

We sum Equations (2.15) and (2.16) to get

$$\langle (\boldsymbol{F}^+ - \boldsymbol{F}^-)\boldsymbol{U}, \boldsymbol{U} \rangle_{\partial K} = \sum_{F \in \mathcal{F}_K} (\|\sqrt{\boldsymbol{\Lambda}^+} \boldsymbol{W}^{+\dagger} \boldsymbol{U}\|_F^2 + \|\sqrt{-\boldsymbol{\Lambda}^-} \boldsymbol{W}^{-\dagger} \boldsymbol{U}\|_F^2) = 0.$$

Hence, on each face $F \in \mathcal{F}_K$, we have

$$\sqrt{\boldsymbol{\Lambda}^+}\boldsymbol{W}^{+\dagger}\boldsymbol{U} = \sqrt{-\boldsymbol{\Lambda}^-}\boldsymbol{W}^{-\dagger}\boldsymbol{U} = \boldsymbol{0},$$

which implies that

$$\boldsymbol{F}^+\boldsymbol{U} = \boldsymbol{F}^-\boldsymbol{U} = \boldsymbol{0}.$$

Using this result in Problem 2.3.5, we see that $\boldsymbol{U}$ must satisfy

$$-\imath\omega(\boldsymbol{U},\boldsymbol{V})_K - \left(\boldsymbol{A}_j\boldsymbol{U}, \frac{\partial \boldsymbol{V}}{\partial x_j}\right)_K = 0$$

for all $\boldsymbol{V} \in \boldsymbol{\mathcal{P}}^l_{\mathrm{p}}(K)$, and integration by parts shows that $\boldsymbol{U}$ solves the system in strong form. Because there is no non-zero polynomial solution of the strong problem, it means that $\boldsymbol{U} = \boldsymbol{0}$. This yields the result. □

2.3.4 Abstract form of the hybridized system

We denote by $\boldsymbol{U}_h(\boldsymbol{x}_h^{\ominus})$ the collection of the local solutions $\boldsymbol{U}_K$ for all $K \in \mathcal{T}_h$ of the local problems with prescribed $\boldsymbol{x}_{K,F}^{\ominus}$ for all $F \in \mathcal{F}_K$ and for all $K \in \mathcal{T}_h$. We can write the following hybridized CHDG problem.

Problem 2.3.8. *Find $\boldsymbol{x}_h^{\ominus} \in \boldsymbol{\mathcal{X}}_h^-$ such that, for all $\boldsymbol{\xi}_h \in \boldsymbol{\mathcal{X}}_h^-$,*

$$\langle \boldsymbol{x}_h^{\ominus} - \Pi(\boldsymbol{x}^{\oplus}(\boldsymbol{U}_h(\boldsymbol{x}_h^{\ominus}))), \boldsymbol{\xi}_h\rangle_{\partial\mathcal{T}_h} = \langle \boldsymbol{b}, \boldsymbol{\xi}_h\rangle_{\partial\mathcal{T}_h},$$

with $\boldsymbol{x}^{\oplus}(\boldsymbol{U}_h(\boldsymbol{x}_h^{\ominus})) := \sqrt{\boldsymbol{\Lambda}^+}\boldsymbol{W}^{+\dagger}\boldsymbol{U}_h(\boldsymbol{x}_h^{\ominus})$.

The hybridized CHDG problem can be written in a convenient abstract form by introducing the *global scattering operator* $\mathrm{S} : \boldsymbol{\mathcal{X}}_h^- \to \boldsymbol{\mathcal{X}}_h^+$ defined such that, for each face F of each element K,

$$\mathrm{S}(\boldsymbol{x}_h^{\ominus})|_{K,F} := \boldsymbol{x}^{\oplus}|_{K,F},$$

with

$$\boldsymbol{x}^{\oplus}|_{K,F} := \sqrt{\boldsymbol{\Lambda}^+}\boldsymbol{W}^{+\dagger}\boldsymbol{U}_K(\boldsymbol{x}_h^{\ominus}), \tag{2.17}$$

where $\boldsymbol{U}_K(\boldsymbol{x}_h^{\ominus})$ is the solution of Problem 2.3.5 with the incoming transmission variable $(\boldsymbol{x}_{K,F}^{\ominus})_{F\in\mathcal{F}_K}$ contained in $\boldsymbol{x}_h^{\ominus}$ as a given surface datum. It can be interpreted as an "incoming transmission variable to outgoing transmission variable" operator that encodes the reflection-transmission phenomenon.

By using the operator S, Problem 2.3.3 is rewritten as follows.

Problem 2.3.9. *Find $\boldsymbol{x}_h^{\ominus} \in \boldsymbol{\mathcal{X}}_h^-$ such that, for all $\boldsymbol{\xi}_h \in \boldsymbol{\mathcal{X}}_h^-$,*

$$\left\langle \boldsymbol{x}_h^{\ominus}, \boldsymbol{\xi}_h\right\rangle_{\partial\mathcal{T}_h} - \left\langle \Pi(\mathrm{S}(\boldsymbol{x}_h^{\ominus})), \boldsymbol{\xi}_h\right\rangle_{\partial\mathcal{T}_h} = \left\langle \boldsymbol{b}, \boldsymbol{\xi}_h\right\rangle_{\partial\mathcal{T}_h}.$$

We introduce the *global projected right-hand sides* $\boldsymbol{b}_h := P_h\boldsymbol{b} \in \boldsymbol{\mathcal{X}}_h^-$, where $P_h : L^2(\partial\mathcal{T}_h) \to \boldsymbol{\mathcal{X}}_h^-$ is the projection operator defined such that $\left\langle P_h\boldsymbol{b}, \boldsymbol{\xi}_h\right\rangle_{\partial\mathcal{T}_h} = \left\langle \boldsymbol{b}, \boldsymbol{\xi}_h\right\rangle_{\partial\mathcal{T}_h}$ for all $\boldsymbol{\xi}_h \in \boldsymbol{\mathcal{X}}_h^-$. Problem 2.3.9 can then be rewritten as follows.

Problem 2.3.10. *Find $\boldsymbol{x}_h^{\ominus} \in \boldsymbol{\mathcal{X}}_h^-$ such that*

$$(\mathrm{I} - \Pi\mathrm{S})\boldsymbol{x}_h^{\ominus} = \boldsymbol{b}_h,$$

where I *is the identity operator on $\boldsymbol{\mathcal{X}}_h^-$.*

Problem 2.3.9 and Problem 2.3.10 are equivalent to Problem 2.3.3 because the element-wise local problems are well-posed.

2.3.5 Strict contraction of the operator ΠS

An interesting property of the CHDG formulation is that the operator ΠS is a strict contraction. As a consequence, Problem 2.3.10 is always well-posed, and it can be solved with the fixed-point iteration without relaxation.

Lemma 2.3.11. *(i) The solution of Problem 2.3.5 verifies*

$$\sum_{F\in\mathcal{F}_K} \|\sqrt{\boldsymbol{\Lambda}^+}\boldsymbol{W}^{+\dagger}\boldsymbol{U}\|_F^2 + \sum_{F\in\mathcal{F}_K} \|\sqrt{-\boldsymbol{\Lambda}^-}\boldsymbol{W}^{-\dagger}\boldsymbol{U} - \boldsymbol{x}_{K,F}^{\ominus}\|_F^2 = \sum_{F\in\mathcal{F}_K} \|\boldsymbol{x}_{K,F}^{\ominus}\|_F^2. \tag{2.18}$$

(ii) The second term on the left-hand side of (2.18) vanishes if and only if $\boldsymbol{x}_{K,F}^{\ominus} = \boldsymbol{0}$.

Proof. For brevity, the subscripts K and F are omitted for the local fields, the test functions, the unit outgoing normal and the surface data.

(i) We proceed as in the proof of Theorem 2.3.7: we take the equation of Problem 2.3.5 with $\boldsymbol{V} = \boldsymbol{U}$, we sum it with the one obtained from the previous equation by integrating by parts and taking the complex conjugate. Consequently, we have

$$\sum_{F\in\mathcal{F}_K} \langle(\boldsymbol{F}^+ - \boldsymbol{F}^-)\boldsymbol{U},\boldsymbol{U}\rangle_F = \sum_{F\in\mathcal{F}_K} \langle\boldsymbol{S},\boldsymbol{U}\rangle_F + \sum_{F\in\mathcal{F}_K} \langle\boldsymbol{U},\boldsymbol{S}\rangle_F.$$

Recalling that $\boldsymbol{S} = \boldsymbol{W}^-\sqrt{-\boldsymbol{\Lambda}^-}\boldsymbol{x}^{\ominus}$, we have

$$\sum_{F\in\mathcal{F}_K} \Big(\langle\boldsymbol{F}^+\boldsymbol{U},\boldsymbol{U}\rangle_F - \langle\boldsymbol{F}^-\boldsymbol{U},\boldsymbol{U}\rangle_F - \langle\boldsymbol{W}^-\sqrt{-\boldsymbol{\Lambda}^-}\boldsymbol{x}^{\ominus},\boldsymbol{U}\rangle_F - \langle\boldsymbol{U},\boldsymbol{W}^-\sqrt{-\boldsymbol{\Lambda}^-}\boldsymbol{x}^{\ominus}\rangle_F\Big) = 0.$$

Using the definition of $\boldsymbol{F}^+$ and $\boldsymbol{F}^-$ leads to

$$\begin{aligned}
&\sum_{F\in\mathcal{F}_K} \Big(\langle\boldsymbol{W}^+\boldsymbol{\Lambda}^+\boldsymbol{W}^{+\dagger}\boldsymbol{U},\boldsymbol{U}\rangle_F - \langle\boldsymbol{W}^-\boldsymbol{\Lambda}^-\boldsymbol{W}^{-\dagger}\boldsymbol{U},\boldsymbol{U}\rangle_F\Big)\\
&\quad + \sum_{F\in\mathcal{F}_K} \Big(-\langle\boldsymbol{x}^{\ominus},\sqrt{-\boldsymbol{\Lambda}^-}\boldsymbol{W}^{-\dagger}\boldsymbol{U}\rangle_F - \langle\sqrt{-\boldsymbol{\Lambda}^-}\boldsymbol{W}^{-\dagger}\boldsymbol{U},\boldsymbol{x}^{\ominus}\rangle_F\Big) = 0,\\
&\sum_{F\in\mathcal{F}_K} \Big(\langle\boldsymbol{\Lambda}^+\boldsymbol{W}^{+\dagger}\boldsymbol{U},\boldsymbol{W}^{\dagger}\boldsymbol{U}\rangle_F + \langle\sqrt{-\boldsymbol{\Lambda}^-}\boldsymbol{W}^{-\dagger}\boldsymbol{U} - \boldsymbol{x}^{\ominus},\sqrt{-\boldsymbol{\Lambda}^-}\boldsymbol{W}^{-\dagger}\boldsymbol{U} - \boldsymbol{x}^{\ominus}\rangle_F - \langle\boldsymbol{x}^{\ominus},\boldsymbol{x}^{\ominus}\rangle_F\Big) = 0,\\
&\sum_{F\in\mathcal{F}_K} \Big(\|\sqrt{\boldsymbol{\Lambda}^+}\boldsymbol{W}^{+\dagger}\boldsymbol{U}\|_F^2 + \|\sqrt{-\boldsymbol{\Lambda}^-}\boldsymbol{W}^{-\dagger}\boldsymbol{U} - \boldsymbol{x}^{\ominus}\|_F^2 - \|\boldsymbol{x}^{\ominus}\|_F^2\Big) = 0.
\end{aligned}$$

(ii) If the second term on the left-hand side of (2.18) vanishes, then $\boldsymbol{x}^{\ominus} = \sqrt{-\boldsymbol{\Lambda}^-}\boldsymbol{W}^{-\dagger}\boldsymbol{U}$ on ∂K. Using this relation in Problem 2.3.5, we see that $\boldsymbol{U}$ must satisfy

$$-\imath\omega(\boldsymbol{U},\boldsymbol{V})_K - \Big(\boldsymbol{A}_j\boldsymbol{U},\frac{\partial\boldsymbol{V}}{\partial x_j}\Big)_K + \langle\boldsymbol{F}\boldsymbol{U},\boldsymbol{V}\rangle_{\partial K} = 0$$

for all $\boldsymbol{V}\in\boldsymbol{\mathcal{P}}_{\mathrm{p}}^l(K)$, and integration by parts shows that $\boldsymbol{U}$ solves the system in strong form. Because there is no non-zero polynomial solution to the previous equation, meaning that $\boldsymbol{U} = \boldsymbol{0}$, and then $\boldsymbol{x}^{\ominus} = \boldsymbol{0}$, this yields the results. The converse statement is direct, because the local problem is well-posed. □

Theorem 2.3.12. *The scattering operator* S *is a strict contraction, i.e.*

$$\|\mathrm{S}(\boldsymbol{x}_h^{\ominus})\| < \|\boldsymbol{x}_h^{\ominus}\|, \quad \forall\boldsymbol{x}_h^{\ominus}\in\boldsymbol{\mathcal{X}}_h^-\backslash\{\boldsymbol{0}\}.$$

Proof. Let $\boldsymbol{x}_h^{\ominus} \in \boldsymbol{\mathcal{X}}_h^- \backslash \{\mathbf{0}\}$. By Lemma 2.3.11, one has

$$\sum_{F \in \mathcal{F}_K} \|\sqrt{\boldsymbol{\Lambda}^+} \boldsymbol{W}^{+\dagger} \boldsymbol{U}\|_F^2 \leq \sum_{F \in \mathcal{F}_K} \|\boldsymbol{x}_{K,F}^{\ominus}\|_F^2,$$

or, more explicitly, thanks to the definition of $\boldsymbol{x}_{K,F}^{\oplus}$ in Equation (2.10),

$$\sum_{F \in \mathcal{F}_K} \|\boldsymbol{x}_{K,F}^{\oplus}\|_F^2 \leq \sum_{F \in \mathcal{F}_K} \|\boldsymbol{x}_{K,F}^{\ominus}\|_F^2,$$

with the equality that holds if and only if $\boldsymbol{x}_{K,F}^{\ominus} = \mathbf{0}$ for all $F \in \mathcal{F}_K$.

Then, by using the definition of S in Equation (2.17), one has

$$\sum_{F \in \mathcal{F}_K} \|\mathrm{S}(\boldsymbol{x}_h^{\ominus})|_{K,F}\|_F^2 \leq \sum_{F \in \mathcal{F}_K} \|\boldsymbol{x}_{K,F}^{\ominus}\|_F^2.$$

Summing this estimate over all $K \in \mathcal{T}_h$ and noticing that the equality can not globally happen because $\boldsymbol{x}_h^{\ominus} \neq \mathbf{0}$ gives the result. □

Theorem 2.3.13. *The exchange operator* Π *is a contraction, i.e.*

$$\|\Pi(\boldsymbol{x}_h^{\ominus})\| \leq \|\boldsymbol{x}_h^{\ominus}\|, \quad \forall \boldsymbol{x}_h^{\ominus} \in \boldsymbol{\mathcal{X}}_h^-.$$

Proof. This result is a straightforward consequence of the definition of Π and Assumption 2.3.4. □

Corollary 2.3.14. *The operator* $\Pi\mathrm{S}$ *is a strict contraction, i.e.*

$$\|\Pi\mathrm{S}(\boldsymbol{x}_h^{\ominus})\| < \|\boldsymbol{x}_h^{\ominus}\|, \quad \forall \boldsymbol{x}_h^{\ominus} \in \boldsymbol{\mathcal{X}}_h^- \backslash \{\mathbf{0}\}.$$

Proof. Corollary 2.3.14 is a direct consequence of Theorems 2.3.12 and 2.3.13. □

Remark 2.3.15. The strict contraction of S is a numerical effect and it does not depend on the physics of the problem: the equation has no homogeneous polynomial solution (see the proof of Lemma 2.3.11). For the continuous case, Equation (2.18) corresponds to a conservation of energy. By contrast, the properties of Π are related to the boundary conditions, and would be preserved at the continuous level.

2.4 Summary

In this Chapter, we have extended and studied the CHDG method, originally proposed for Helmholtz problems in [97], in order to deal with general time-harmonic wave problems with constant physical coefficients. More specifically, the theory we developed covers both the acoustic and electromagnetic frameworks of earlier works [97] and [111] respectively and, in practice, it can also straightforwardly be used to deal with many other frameworks, including elasticity for instance.

We considered a discontinuous Galerkin formulation with standard upwind numerical fluxes. For the hybridization procedure, transmission variables, whose definition is supported by the theory of symmetric hyperbolic systems itself, are introduced at the interfaces between the elements.

The CHDG hybridized system can be written in the form $(\mathrm{I} - \Pi\mathrm{S})\boldsymbol{x} = \boldsymbol{b}$. Under certain hypothesis on the boundary conditions, we proved that the operator $\Pi\mathrm{S}$ is a strict contraction, which makes it possible to solve the system with a fixed-point iteration. These hypothesis are motivated by energy conservation reasons and, morally, they control the amount of energy that is brought into the computational domain by the boundary data in such a way that they disallow an increase in energy.

Chapter 3

CHDG method for time-harmonic aeroacoustics with a uniform mean flow

In this Chapter, we apply and study the CHDG method to an aeroacoustic problem. In particular, we consider the propagation of acoustic waves in a homogeneous medium and in the presence of a constant background flow. Compared to classical acoustics, additional difficulties in the numerical method are involved: the presence of a vorticity mode that needs to be represented by the transmission variables and a different number of outgoing and incoming transmission variables that makes the definition of the upwind numerical fluxes apparently more complex.

We follow the steps of the method introduced in Chapter 2: in Section 3.1 we present a model problem set in terms of the time-harmonic linearised Euler equations, in Section 3.2 we write its DG formulation and in Section 3.3, we write its CHDG formulation and we define the transmission variables. We highlight the difficulties that characterize the physical equation and the tools that allow us to cope with aeroacoustics. To assess the numerical aspects and performance of the CHDG method, in Section 3.4 we analyze the numerical results obtained with several benchmarks.

3.1 Problem

We consider the propagation of waves in the time-harmonic regime in a bounded domain $\Omega \subset \mathbb{R}^d$, with $d = 2$ or 3. The medium is homogeneous with constant density ρ_0, sound speed c_0 and mean flow $\boldsymbol{u}_0$. The pressure p and velocity $\boldsymbol{u}$ are governed by the time-harmonic version of the LEE, namely Equation (1.5) that we recall here below

$$\begin{cases} \dfrac{D_0\rho}{Dt} + \rho_0 \boldsymbol{\nabla}\cdot\boldsymbol{u} = 0, \\ \rho_0 \dfrac{D_0\boldsymbol{u}}{Dt} + \boldsymbol{\nabla} p = \boldsymbol{f}, \\ \dfrac{D_0 p}{Dt} + \rho_0 c_0^2 \boldsymbol{\nabla}\cdot\boldsymbol{u} = 0. \end{cases}$$

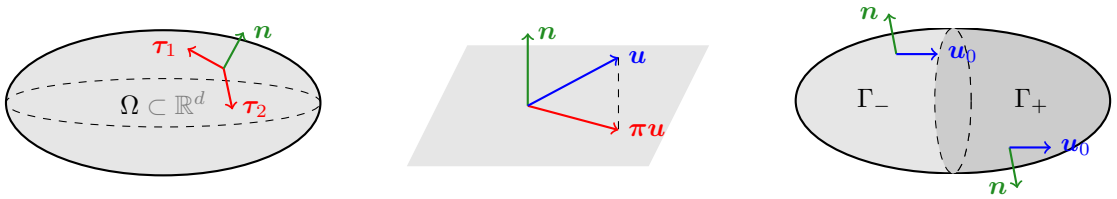
We denote by $\boldsymbol{n}$ the normal unit vector to the boundary $\partial\Omega = \Gamma$ and by $\boldsymbol{T} \in \mathbb{R}^{d\times(d-1)}$ the matrix whose columns form the set of linearly independent vectors $\{\boldsymbol{\tau}_j\}_{j=1,\dots,d-1}$ that generates the corresponding tangent subspace (see Figure 3.1a) and that, together with the vector $\boldsymbol{n}$, they form an orthonormal basis of $\mathbb{R}^d$. Notice that both $\boldsymbol{n}$ and $\boldsymbol{T}$ are locally defined on the boundary Γ. We introduce the projection operator $\boldsymbol{\pi} : \mathbb{R}^d \to \mathbb{R}^d$ defined as $\boldsymbol{\pi}\boldsymbol{u} = (\boldsymbol{I} - \boldsymbol{n}\boldsymbol{n}^\dagger)\boldsymbol{u} = \boldsymbol{T}\boldsymbol{T}^\dagger\boldsymbol{u}$ for all $\boldsymbol{u} \in \mathbb{R}^d$ (see Figure 3.1b) which let us decompose every vector $\boldsymbol{u} \in \mathbb{R}^d$ as $\boldsymbol{u} = (\boldsymbol{n}\cdot\boldsymbol{u})\boldsymbol{n} + \boldsymbol{\pi}\boldsymbol{u}$.

Remark 3.1.1. In dimension $d = 2$, the tangent space is reduced to a one dimensional subspace generated by the tangent vector $\boldsymbol{\tau} = (-n_y, n_x)$. In this case, the projection operator reads $\boldsymbol{\pi u} = (\boldsymbol{\tau} \cdot \boldsymbol{u})\boldsymbol{\tau}$ and the matrix $\boldsymbol{T}$ reduces to $\boldsymbol{T} = \boldsymbol{\tau}$.

In accordance with Remark 2.1.1, the number of boundary conditions depends on the number of negative eigenvalues associated to the flux matrix $\boldsymbol{F}$ and, in this case, it depends on the sign of $\boldsymbol{u}_0 \cdot \boldsymbol{n}$, as we see in Section 3.2.

Consequently, we consider two different partitions of Γ in non-overlapping polytopal Lipschitz subsets:

- $\Gamma_{\mathrm{D}} \cup \Gamma_{\mathrm{N}} \cup \Gamma_{\mathrm{R}}$ for imposing boundary conditions on p and/or $\boldsymbol{n} \cdot \boldsymbol{u}$ for acoustic waves;
- $\Gamma_+ \cup \Gamma_-$ for imposing boundary conditions on $\boldsymbol{T}^\dagger \boldsymbol{u}$ when $\boldsymbol{u}_0 \cdot \boldsymbol{n} < 0$ for hydrodynamic waves, where the outflow boundary Γ_+ (respectively the inflow boundary Γ_-) is the portion of Γ where $\boldsymbol{u}_0 \cdot \boldsymbol{n} \geq 0$ (respectively $\boldsymbol{u}_0 \cdot \boldsymbol{n} < 0$) (see Figure 3.1c).



(a) Normal and tangential unit vectors.

(b) Projection $\boldsymbol{\pi u}$ of a vector $\boldsymbol{u}$ on a plane surface.

(c) Outflow and inflow boundary.

Figure 3.1: Notation on the boundary (a)-(c) and on the projection operator (b).

The problem of interest can then be expressed as follows.

Problem 3.1.2. *Find* $(p, \boldsymbol{u}) : \Omega \to \mathbb{C} \times \mathbb{C}^d$ *such that*

$$\begin{cases} -\dfrac{\imath\omega}{\rho_0 c_0^2} p + \dfrac{1}{\rho_0 c_0^2} \boldsymbol{u}_0 \cdot \boldsymbol{\nabla} p + \boldsymbol{\nabla} \cdot \boldsymbol{u} = f, & \text{in } \Omega, \\ -\imath\omega\rho_0 \boldsymbol{u} + \boldsymbol{\nabla} p + \rho_0 (\boldsymbol{u}_0 \cdot \boldsymbol{\nabla}) \boldsymbol{u} = \boldsymbol{0}, & \text{in } \Omega, \\ p = s_{\mathrm{D}}, & \text{on } \Gamma_{\mathrm{D}}, \\ \boldsymbol{n} \cdot \boldsymbol{u} = s_{\mathrm{N}}, & \text{on } \Gamma_{\mathrm{N}}, \\ p - \rho_0 c_0 \boldsymbol{n} \cdot \boldsymbol{u} = s_{\mathrm{R}}, & \text{on } \Gamma_{\mathrm{R}}, \\ \boldsymbol{T}^\dagger \boldsymbol{u} = \boldsymbol{s}_-, & \text{on } \Gamma_-, \end{cases}$$

with the given volume datum $f : \Omega \to \mathbb{C}$ *and boundary data* $s_{\mathrm{D}} : \Gamma_{\mathrm{D}} \to \mathbb{C}$, $s_{\mathrm{N}} : \Gamma_{\mathrm{N}} \to \mathbb{C}$, $s_{\mathrm{R}} : \Gamma_{\mathrm{R}} \to \mathbb{C}$ *and* $\boldsymbol{s}_- : \Gamma_- \to \mathbb{C}^{d-1}$.

This problem can be expressed as Problem 2.1.2, namely

$$-\imath\omega \boldsymbol{U} + \boldsymbol{A}_j \frac{\partial \boldsymbol{U}}{\partial x_j} = \boldsymbol{G}_v,$$

with

$$\boldsymbol{U} := \begin{bmatrix} p \\ \rho_0 c_0 \boldsymbol{u} \end{bmatrix}, \quad \boldsymbol{A}_j := \begin{bmatrix} u_{0,x_j} & c_0 \boldsymbol{e}_j^\dagger \\ c_0 \boldsymbol{e}_j & u_{0,x_j} \boldsymbol{I} \end{bmatrix} \quad \text{and} \quad \boldsymbol{G}_v := \begin{bmatrix} f \\ \boldsymbol{0} \end{bmatrix},$$

where $\{\boldsymbol{e}_j\}_{j=1,\dots,d}$ is the canonical basis of $\mathbb{R}^d$ and with $u_{0,x_j} := \boldsymbol{u}_0 \cdot \boldsymbol{e}_j$.

As far as the boundary conditions are concerned, they can be written in the form of Problem 2.1.2, namely $\boldsymbol{BU} = \boldsymbol{G}_s$, where the matrix $\boldsymbol{B}$ and the right-hand side $\boldsymbol{G}_s$ clearly depend on the boundary conditions themselves. In particular, if $F \subset \Gamma_+$, they read

$$\boldsymbol{B} := \boldsymbol{B}_1 \quad \text{and} \quad \boldsymbol{G}_s := \mathrm{g}_1,$$

while, if $F \subset \Gamma_-$, they read

$$\boldsymbol{B} := \begin{bmatrix} \boldsymbol{B}_1 \\ \boldsymbol{B}_2 \end{bmatrix} \quad \text{and} \quad \boldsymbol{G}_s := \begin{bmatrix} \mathrm{g}_1 \\ \mathbf{g}_2 \end{bmatrix},$$

with

$$\boldsymbol{B}_1 := \begin{cases} \begin{bmatrix} 1 & \mathbf{0}^\dagger \end{bmatrix} & \text{if } F \subset \Gamma_{\mathrm{D}}, \\ \begin{bmatrix} 0 & \frac{1}{\rho_0 c_0} \boldsymbol{n}^\dagger \end{bmatrix} & \text{if } F \subset \Gamma_{\mathrm{N}}, \\ \begin{bmatrix} 1 & -\boldsymbol{n}^\dagger \end{bmatrix} & \text{if } F \subset \Gamma_{\mathrm{R}}, \end{cases} \qquad \mathrm{g}_1 := \begin{cases} s_{\mathrm{D}} & \text{if } F \subset \Gamma_{\mathrm{D}}, \\ s_{\mathrm{N}} & \text{if } F \subset \Gamma_{\mathrm{N}}, \\ s_{\mathrm{R}} & \text{if } F \subset \Gamma_{\mathrm{R}}, \end{cases}$$

$$\boldsymbol{B}_2 := \begin{bmatrix} \mathbf{0} & \dfrac{1}{\rho_0 c_0} \boldsymbol{T}^\dagger \end{bmatrix}, \qquad \mathbf{g}_2 := \boldsymbol{s}_-.$$

Notice that here $\boldsymbol{U} \in \mathbb{C}^l$, $\boldsymbol{A}_j \in \mathbb{R}^{l \times l}$, $\boldsymbol{G}_v \in \mathbb{C}^l$, $\boldsymbol{B} \in \mathbb{R}^{m \times l}$ and $\boldsymbol{G}_s \in \mathbb{C}^m$ with $l = d+1$ and $m = 1$ on the outflow boundary and $m = d$ on the inflow one.

3.2 Discontinuous Galerkin method

The goal of thi Section is to introduce the DG formulation of Problem 2.2.2 and give details about the numerical flux with respect to the physical parameters of the problem.

The DG formulation is given by Problem 2.2.2 with $\boldsymbol{\mathcal{V}}_h^l = \boldsymbol{\mathcal{V}}_h^1 \times \boldsymbol{\mathcal{V}}_h^d$. For the sake of brevity, in the following, we denote $\mathcal{V}_h = \boldsymbol{\mathcal{V}}_h^1$ and $\boldsymbol{\mathcal{V}}_h = \boldsymbol{\mathcal{V}}_h^d$.

It can be written in terms of the physical unknowns as follows

Problem 3.2.1. *Find $(p_h, \boldsymbol{u}_h) \in \mathcal{V}_h \times \boldsymbol{\mathcal{V}}_h$ such that, for all $(q_h, \boldsymbol{v}_h) \in \mathcal{V}_h \times \boldsymbol{\mathcal{V}}_h$,*

$$\begin{cases} -\imath\omega\Big(\dfrac{1}{\rho_0 c_0^2} p_h, q_h\Big)_{\mathcal{T}_h} - \Big(\dfrac{1}{\rho_0 c_0^2} p_h, \boldsymbol{u}_0 \cdot \boldsymbol{\nabla} q_h\Big)_{\mathcal{T}_h} - \big(\boldsymbol{u}_h, \boldsymbol{\nabla} q_h\big)_{\mathcal{T}_h} + \Big\langle \dfrac{\boldsymbol{u}_0 \cdot \boldsymbol{n}}{\rho_0 c_0^2} \widehat{p}_h + \boldsymbol{n} \cdot \widehat{\boldsymbol{u}}_h, q_h \Big\rangle_{\partial \mathcal{T}_h} = 0, \\ -\imath\omega\big(\rho_0 \boldsymbol{u}_h, \boldsymbol{v}_h\big)_{\mathcal{T}_h} - \big(\rho_0 \boldsymbol{u}_h, \boldsymbol{\nabla} \boldsymbol{v}_h \cdot \boldsymbol{u}_0\big)_{\mathcal{T}_h} - \big(p_h, \boldsymbol{\nabla} \cdot \boldsymbol{v}_h\big)_{\mathcal{T}_h} + \big\langle \widehat{p}_h \boldsymbol{n} + \rho_0 (\boldsymbol{u}_0 \cdot \boldsymbol{n}) \widehat{\boldsymbol{u}}_h, \boldsymbol{v}_h \big\rangle_{\partial \mathcal{T}_h} = 0. \end{cases}$$

Note that, for the sake of simplicity, we assume $f = 0$.

This formulation has been obtained by using the flux matrix $\boldsymbol{F}$ and the numerical flux $\boldsymbol{F}\widehat{\boldsymbol{U}}_F$ defined as

$$\boldsymbol{F} := \begin{bmatrix} u_{0,n} & c_0 \boldsymbol{n}^\dagger \\ c_0 \boldsymbol{n} & u_{0,n} \boldsymbol{I} \end{bmatrix} \quad \text{and} \quad \boldsymbol{F}\widehat{\boldsymbol{U}}_F := \begin{bmatrix} \dfrac{\boldsymbol{u}_0 \cdot \boldsymbol{n}_{K,F}}{\rho_0 c_0^2} \widehat{p}_F + \boldsymbol{n}_{K,F} \cdot \widehat{\boldsymbol{u}}_F \\ \widehat{p}_F \boldsymbol{n}_{K,F} + \rho_0 (\boldsymbol{u}_0 \cdot \boldsymbol{n}_{K,F}) \widehat{\boldsymbol{u}}_F \end{bmatrix}, \tag{3.1}$$

with $u_{0,n} := \boldsymbol{u}_0 \cdot \boldsymbol{n}$.

In the following, we consider a splitting of the flux matrix $\boldsymbol{F}$ to define the upwind numerical flux, as we explained in Section 2.2.2.

We start by computing the eigenvalues and eigenvectors of the matrix $\boldsymbol{F}$ by solving $\det(\boldsymbol{F}-\lambda\boldsymbol{I})=0$:

$$\begin{vmatrix} u_{0,n}-\lambda & c_0\boldsymbol{n}^\dagger \\ c_0\boldsymbol{n} & (u_{0,n}-\lambda)\boldsymbol{I} \end{vmatrix} = 0 \quad \Longleftrightarrow \quad (u_{0,n}-\lambda)^2((u_{0,n}-\lambda)^2-c_0^2)=0.$$

We obtain the following eigenvalues λ_j with $j=1,...,4$

$$\lambda_1 = u_{0,n}-c_0<0, \quad \lambda_2 = u_{0,n}+c_0>0, \quad \lambda_3=\lambda_4=u_{0,n},$$

and the associated eigenvectors $\boldsymbol{w}_j$ such that $\boldsymbol{F}\boldsymbol{w}_j=\lambda_j\boldsymbol{w}_j$ and $\|\boldsymbol{w}_j\|=1$ with $j=1,...,4$

$$\boldsymbol{w}_1 := \frac{1}{\sqrt{2}}\begin{bmatrix}1\\-\boldsymbol{n}\end{bmatrix}, \quad \boldsymbol{w}_2 := \frac{1}{\sqrt{2}}\begin{bmatrix}1\\\boldsymbol{n}\end{bmatrix}, \quad \boldsymbol{w}_3 := \begin{bmatrix}0\\\boldsymbol{\tau}_1\end{bmatrix}, \quad \boldsymbol{w}_4 := \begin{bmatrix}0\\\boldsymbol{\tau}_2\end{bmatrix}.$$

Depending on the sign of $u_{0,n}$, we have several scenarios with different number of positive, negative and null eigenvalues. Since the mean flow is subsonic (i.e. $\|\boldsymbol{u}_0\|<c_0$):

- if $u_{0,n}>0$ there are three positive eigenvalues and one negative,
- if $u_{0,n}<0$ there are three negative eigenvalues and one positive,
- if $u_{0,n}=0$ the spectrum consists of one positive, one negative, and two null eigenvalues.

With the splitting method, the numerical flux at each face $F\not\subset\partial\Omega$ is defined as Equation (2.9), namely

$$\boldsymbol{F}\widehat{\boldsymbol{U}}_F = \boldsymbol{F}^+\boldsymbol{U}_K + \boldsymbol{F}^-\boldsymbol{U}_{K'},$$

with $\boldsymbol{F}^+ = \boldsymbol{W}^+\boldsymbol{\Lambda}^+\boldsymbol{W}^{+\dagger}$ and $\boldsymbol{F}^- = \boldsymbol{W}^-\boldsymbol{\Lambda}^-\boldsymbol{W}^{-\dagger}$.

Depending on the subcase, the explicit terms of the decomposition are then given by

- if $u_{0,n}>0$:

$$\boldsymbol{\Lambda}^+ = \begin{bmatrix} u_{0,n}+c_0 & 0 & 0 \\ 0 & u_{0,n} & 0 \\ 0 & 0 & u_{0,n}\end{bmatrix}, \quad \boldsymbol{W}^+ = \begin{bmatrix} \frac{1}{\sqrt{2}} & 0 & 0 \\ \frac{1}{\sqrt{2}}\boldsymbol{n} & \boldsymbol{\tau}_1 & \boldsymbol{\tau}_2\end{bmatrix},$$

$$\boldsymbol{F}^+\boldsymbol{U}_K = \begin{bmatrix} \dfrac{c_0+u_{0,n}}{2}(p_K+\rho_0c_0\boldsymbol{n}_{K,F}\cdot\boldsymbol{u}_K) \\ \dfrac{c_0+u_{0,n}}{2}(p_K+\rho_0c_0\boldsymbol{n}_{K,F}\cdot\boldsymbol{u}_K)\boldsymbol{n}_{K,F}+\rho_0c_0u_{0,n}\boldsymbol{\pi}_{K,F}\boldsymbol{u}_K\end{bmatrix},$$

$$\boldsymbol{\Lambda}^- = \begin{bmatrix} u_{0,n}-c_0\end{bmatrix}, \quad \boldsymbol{W}^- = \begin{bmatrix} \frac{1}{\sqrt{2}} \\ -\frac{1}{\sqrt{2}}\boldsymbol{n}\end{bmatrix},$$

$$\boldsymbol{F}^-\boldsymbol{U}_{K'} = \begin{bmatrix} -\dfrac{c_0-u_{0,n}}{2}(p_{K'}-\rho_0c_0\boldsymbol{n}_{K,F}\cdot\boldsymbol{u}_{K'}) \\ \dfrac{c_0-u_{0,n}}{2}(p_{K'}-\rho_0c_0\boldsymbol{n}_{K,F}\cdot\boldsymbol{u}_{K'})\boldsymbol{n}_{K,F}\end{bmatrix};$$

- if $u_{0,n} < 0$:

$$\boldsymbol{\Lambda}^+ = \left[u_{0,n} + c_0\right], \quad \boldsymbol{W}^+ = \begin{bmatrix} \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}}\boldsymbol{n} \end{bmatrix},$$

$$\boldsymbol{F}^+\boldsymbol{U}_K = \begin{bmatrix} \frac{c_0 + u_{0,n}}{2}(p_K + \rho_0 c_0 \boldsymbol{n}_{K,F} \cdot \boldsymbol{u}_K) \\ \frac{c_0 + u_{0,n}}{2}(p_K + \rho_0 c_0 \boldsymbol{n}_{K,F} \cdot \boldsymbol{u}_K)\boldsymbol{n}_{K,F} \end{bmatrix},$$

$$\boldsymbol{\Lambda}^- = \begin{bmatrix} u_{0,n} - c_0 & 0 & 0 \\ 0 & u_{0,n} & 0 \\ 0 & 0 & u_{0,n} \end{bmatrix}, \quad \boldsymbol{W}^- = \begin{bmatrix} \frac{1}{\sqrt{2}} & 0 & 0 \\ -\frac{1}{\sqrt{2}}\boldsymbol{n} & \boldsymbol{\tau}_1 & \boldsymbol{\tau}_2 \end{bmatrix},$$

$$\boldsymbol{F}^-\boldsymbol{U}_{K'} = \begin{bmatrix} -\frac{c_0 - u_{0,n}}{2}(p_{K'} - \rho_0 c_0 \boldsymbol{n}_{K,F} \cdot \boldsymbol{u}_{K'}) \\ \frac{c_0 - u_{0,n}}{2}(p_{K'} - \rho_0 c_0 \boldsymbol{n}_{K,F} \cdot \boldsymbol{u}_{K'})\boldsymbol{n}_{K,F} + \rho_0 c_0 u_{0,n} \boldsymbol{\pi}_{K,F}\boldsymbol{u}_{K'} \end{bmatrix};$$

- if $u_{0,n} = 0$:

$$\boldsymbol{\Lambda}^+ = \left[c_0\right], \quad \boldsymbol{W}^+ = \begin{bmatrix} \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}}\boldsymbol{n} \end{bmatrix},$$

$$\boldsymbol{F}^+\boldsymbol{U}_K = \begin{bmatrix} \frac{c_0}{2}(p_K + \rho_0 c_0 \boldsymbol{n}_{K,F} \cdot \boldsymbol{u}_K) \\ \frac{c_0}{2}(p_K + \rho_0 c_0 \boldsymbol{n}_{K,F} \cdot \boldsymbol{u}_K)\boldsymbol{n}_{K,F} \end{bmatrix},$$

$$\boldsymbol{\Lambda}^- = \left[-c_0\right], \quad \boldsymbol{W}^- = \begin{bmatrix} \frac{1}{\sqrt{2}} \\ -\frac{1}{\sqrt{2}}\boldsymbol{n} \end{bmatrix},$$

$$\boldsymbol{F}^-\boldsymbol{U}_{K'} = \begin{bmatrix} -\frac{c_0}{2}(p_{K'} - \rho_0 c_0 \boldsymbol{n}_{K,F} \cdot \boldsymbol{u}_{K'}) \\ \frac{c_0}{2}(p_{K'} - \rho_0 c_0 \boldsymbol{n}_{K,F} \cdot \boldsymbol{u}_{K'})\boldsymbol{n}_{K,F} \end{bmatrix},$$

where $\boldsymbol{\pi}_{K,F}$ denotes the projection operator defined on the face F of an element K.

Note that the only term of the numerical flux $\boldsymbol{F}\widehat{\boldsymbol{U}}_F$ that depends on the sign of $u_{0,n}$ is the projection of the velocity field onto the tangent plane.

More explicitly, by writing the quantities $\widehat{p}_F$, $\boldsymbol{n}_{K,F} \cdot \widehat{\boldsymbol{u}}_F$ and $\boldsymbol{\pi}_{K,F}\widehat{u}_F$ in terms of the physical fields, this corresponds to the following choices of fluxes:

$$\widehat{p}_F = \frac{p_K + p_{K'}}{2} + \rho_0 c_0 \boldsymbol{n}_{K,F} \cdot \frac{\boldsymbol{u}_K - \boldsymbol{u}_{K'}}{2}, \qquad \text{if } F \not\subset \partial\Omega,$$

$$\boldsymbol{n}_{K,F} \cdot \widehat{\boldsymbol{u}}_F = \boldsymbol{n}_{K,F} \cdot \frac{\boldsymbol{u}_K + \boldsymbol{u}_{K'}}{2} + \frac{1}{\rho_0 c_0}\frac{p_K - p_{K'}}{2}, \qquad \text{if } F \not\subset \partial\Omega,$$

$$\boldsymbol{\pi}_{K,F}\widehat{\boldsymbol{u}}_F = \begin{cases} \boldsymbol{\pi}_{K,F}\boldsymbol{u}_K, & \text{if } u_{0,n} \geq 0, \\ \boldsymbol{\pi}_{K,F}\boldsymbol{u}_{K'}, & \text{if } u_{0,n} < 0, \end{cases} \qquad \text{if } F \not\subset \partial\Omega.$$

Moreover the following relationship, which is equivalent to Property 2.3.2, holds

$$\boldsymbol{n}_{K,F} \cdot \widehat{\boldsymbol{u}}_F = \boldsymbol{n}_{K,F} \cdot \boldsymbol{u}_K + \frac{1}{\rho_0 c_0}(p_K - \widehat{p}_F).$$

Notice that this relation is fundamental in standard HDG methods because it is used to rewrite $\boldsymbol{n}_{K,F} \cdot \widehat{\boldsymbol{u}}_F$ in terms of $\widehat{p}_F$, p_K and $\boldsymbol{u}_K$. In this way only the latter appear in the formulation of the problem and the numerical flux $\widehat{p}_F$ plays the role of unique hybrid variable.

It can be used to define the numerical fluxes on boundary faces: it is enough to plug one of the following relations (depending on the nature of the boundary face $F \subset \partial\Omega$) into the previous equation

$$\begin{aligned} \widehat{p}_F &= s_{\mathrm{D}}, && \text{if } F \subset \Gamma_{\mathrm{D}}, \\ \boldsymbol{n}_{K,F} \cdot \widehat{\boldsymbol{u}}_F &= s_{\mathrm{N}}, && \text{if } F \subset \Gamma_{\mathrm{N}}, \\ \widehat{p}_F - \rho_0 c_0 \boldsymbol{n}_{K,F} \cdot \widehat{\boldsymbol{u}}_F &= s_{\mathrm{R}}, && \text{if } F \subset \Gamma_{\mathrm{R}}, \\ \boldsymbol{\pi}_{K,F} \widehat{\boldsymbol{u}}_F &= \boldsymbol{T}\boldsymbol{s}_-, && \text{if } F \subset \Gamma_-, \end{aligned}$$

and manipulating it to get

$$\widehat{p}_F = \begin{cases} s_{\mathrm{D}}, & \text{if } F \subset \Gamma_{\mathrm{D}}, \\ p_K + \rho_0 c_0 \left(\boldsymbol{n}_{K,F} \cdot \boldsymbol{u}_K - s_{\mathrm{N}} \right), & \text{if } F \subset \Gamma_{\mathrm{N}}, \\ \dfrac{1}{2} \left(p_K + \rho_0 c_0 \boldsymbol{n}_{K,F} \cdot \boldsymbol{u}_K + s_{\mathrm{R}} \right), & \text{if } F \subset \Gamma_{\mathrm{R}}, \end{cases}$$

$$\boldsymbol{n}_{K,F} \cdot \widehat{\boldsymbol{u}}_F = \begin{cases} \boldsymbol{n}_{K,F} \cdot \boldsymbol{u}_K + \dfrac{1}{\rho_0 c_0} \left(p_K - s_{\mathrm{D}} \right), & \text{if } F \subset \Gamma_{\mathrm{D}}, \\ s_{\mathrm{N}}, & \text{if } F \subset \Gamma_{\mathrm{N}}, \\ \dfrac{1}{2\rho_0 c_0} \left(p_K + \rho_0 c_0 \boldsymbol{n}_{K,F} \cdot \boldsymbol{u}_K - s_{\mathrm{R}} \right), & \text{if } F \subset \Gamma_{\mathrm{R}}, \end{cases}$$

$$\boldsymbol{\pi}_{K,F} \widehat{\boldsymbol{u}}_F = \begin{cases} \boldsymbol{T}\boldsymbol{s}_-, & \text{if } F \subset \Gamma_-, \\ \boldsymbol{\pi}_{K,F} \boldsymbol{u}_K, & \text{if } F \subset \Gamma_+. \end{cases}$$

3.3 Hybridization with transmission variables

In this Section, with the goal of introducing the CHDG formulation of the problem, we follow the steps presented in Chapter 2. In particular, we introduce the transmission variables, we give details about the operators Π and S that appear in the CHDG formulation and we determine the hypothesis that allow Assumption 2.3.4 to hold.

3.3.1 Transmission variables

We recall that the outgoing and incoming transmission variables are defined (Equations (2.10)-(2.11)) as

$$\begin{aligned} \boldsymbol{x}^{\oplus}_{K,F} &:= \sqrt{\boldsymbol{\Lambda}^+} \boldsymbol{W}^{+\dagger} \boldsymbol{U}_{K,F}, && \forall F, \\ \boldsymbol{x}^{\ominus}_{K,F} &:= \sqrt{-\boldsymbol{\Lambda}^-} \boldsymbol{W}^{-\dagger} \boldsymbol{U}_{K',F}, && \text{if } F \not\subset \partial\Omega. \end{aligned}$$

We can rewrite them in terms of their normal $\mathrm{g}^{\oplus,n}_{K,F}$, $\mathrm{g}^{\ominus,n}_{K,F}$ and tangential $\mathbf{g}^{\oplus,t}_{K,F}$, $\mathbf{g}^{\ominus,t}_{K,F}$ components as follows

$$\boldsymbol{x}^{\oplus}_{K,F} = \begin{cases} \begin{bmatrix} \mathrm{g}^{\oplus,n}_{K,F} \\ \mathbf{g}^{\oplus,t}_{K,F} \end{bmatrix}, & \text{if } \boldsymbol{u}_0 \cdot \boldsymbol{n}_{K,F} > 0, \\ \begin{bmatrix} \mathrm{g}^{\oplus,n}_{K,F} \end{bmatrix}, & \text{if } \boldsymbol{u}_0 \cdot \boldsymbol{n}_{K,F} < 0, \\ \begin{bmatrix} \mathrm{g}^{\oplus,n}_{K,F} \end{bmatrix}, & \text{if } \boldsymbol{u}_0 \cdot \boldsymbol{n}_{K,F} = 0, \end{cases} \qquad \boldsymbol{x}^{\ominus}_{K,F} = \begin{cases} \begin{bmatrix} \mathrm{g}^{\ominus,n}_{K,F} \end{bmatrix}, & \text{if } \boldsymbol{u}_0 \cdot \boldsymbol{n}_{K,F} > 0, \\ \begin{bmatrix} \mathrm{g}^{\ominus,n}_{K,F} \\ \mathbf{g}^{\ominus,t}_{K,F} \end{bmatrix}, & \text{if } \boldsymbol{u}_0 \cdot \boldsymbol{n}_{K,F} < 0, \\ \begin{bmatrix} \mathrm{g}^{\ominus,n}_{K,F} \end{bmatrix}, & \text{if } \boldsymbol{u}_0 \cdot \boldsymbol{n}_{K,F} = 0. \end{cases}$$

Normal direction We define the *outgoing normal transmission variable* $\mathrm{g}_{K,F}^{\oplus,n}$, and the *incoming normal transmission variable* $\mathrm{g}_{K,F}^{\ominus,n}$, which correspond to *acoustic waves*, as follows

$$\begin{aligned}
\mathrm{g}_{K,F}^{\oplus,n} &:= \sqrt{\frac{c_0 + \boldsymbol{u}_0 \cdot \boldsymbol{n}_{K,F}}{2}}(p_K + \rho_0 c_0 \boldsymbol{n}_{K,F} \cdot \boldsymbol{u}_K), && \forall F,\\
\mathrm{g}_{K,F}^{\ominus,n} &:= \sqrt{\frac{c_0 - \boldsymbol{u}_0 \cdot \boldsymbol{n}_{K,F}}{2}}(p_{K'} - \rho_0 c_0 \boldsymbol{n}_{K,F} \cdot \boldsymbol{u}_{K'}), && \text{if } F \notin \partial\Omega.
\end{aligned}$$

Note that $\mathrm{g}_{K,F}^{\ominus,n} = \mathrm{g}_{K',F}^{\oplus,n}$ if $F \notin \partial\Omega$. For boundary faces, by taking into account the boundary condition, the incoming normal transmission variable reads

$$\mathrm{g}_{K,F}^{\ominus,n} := \begin{cases}
-\mathrm{g}_{K,F}^{\oplus,n}\sqrt{\dfrac{c_0 - \boldsymbol{u}_0 \cdot \boldsymbol{n}_{K,F}}{c_0 + \boldsymbol{u}_0 \cdot \boldsymbol{n}_{K,F}}} + s_{\mathrm{D}}\sqrt{2(c_0 - \boldsymbol{u}_0 \cdot \boldsymbol{n}_{K,F})}, & \text{if } F \subset \Gamma_{\mathrm{D}},\\
\mathrm{g}_{K,F}^{\oplus,n}\sqrt{\dfrac{c_0 - \boldsymbol{u}_0 \cdot \boldsymbol{n}_{K,F}}{c_0 + \boldsymbol{u}_0 \cdot \boldsymbol{n}_{K,F}}} - s_{\mathrm{N}}\rho_0 c_0\sqrt{2(c_0 - \boldsymbol{u}_0 \cdot \boldsymbol{n}_{K,F})}, & \text{if } F \subset \Gamma_{\mathrm{N}},\\
s_{\mathrm{R}}\sqrt{\dfrac{c_0 - \boldsymbol{u}_0 \cdot \boldsymbol{n}_{K,F}}{2}}, & \text{if } F \subset \Gamma_{\mathrm{R}}.
\end{cases}$$

Tangential direction We define the *outgoing tangential transmission variable* $\mathbf{g}_{K,F}^{\oplus,t}$ and *incoming tangential transmission variable* $\mathbf{g}_{K,F}^{\ominus,t}$, which correspond to *hydrodynamic waves*, as follows

$$\begin{aligned}
\mathbf{g}_{K,F}^{\oplus,t} &:= \rho_0 c_0 \sqrt{|\boldsymbol{u}_0 \cdot \boldsymbol{n}_{K,F}|}\boldsymbol{T}_{K,F}^{\dagger}\boldsymbol{u}_K, && \text{if } \boldsymbol{u}_0 \cdot \boldsymbol{n}_{K,F} > 0, && \forall F,\\
\mathbf{g}_{K,F}^{\ominus,t} &:= -\rho_0 c_0 \sqrt{|\boldsymbol{u}_0 \cdot \boldsymbol{n}_{K,F}|}\boldsymbol{T}_{K,F}^{\dagger}\boldsymbol{u}_{K'}, && \text{if } \boldsymbol{u}_0 \cdot \boldsymbol{n}_{K,F} < 0, && \text{if } F \notin \partial\Omega.
\end{aligned}$$

Note that the incoming transmission variable is defined only if $\boldsymbol{u}_0 \cdot \boldsymbol{n}_{K,F} < 0$ and the outgoing transmission variable is defined only if $\boldsymbol{u}_0 \cdot \boldsymbol{n}_{K,F} > 0$ and, in such a case, then $\mathbf{g}_{K,F}^{\ominus,t} = \mathbf{g}_{K',F}^{\oplus,t}$ if $F \notin \partial\Omega$. For inflow boundary faces, by taking into account the boundary condition, the incoming tangential transmission variable reads

$$\mathbf{g}_{K,F}^{\ominus,t} := -\rho_0 c_0 \sqrt{|\boldsymbol{u}_0 \cdot \boldsymbol{n}_{K,F}|}\boldsymbol{s}_-, \quad \text{if } F \subset \Gamma_-.$$

Now we can use the definitions we have just introduced to rewrite the upwind fluxes as follows

$$\begin{aligned}
\frac{\boldsymbol{u}_0 \cdot \boldsymbol{n}_{K,F}}{\rho_0 c_0^2}\widehat{p}_F + \boldsymbol{n}_{K,F} \cdot \widehat{\boldsymbol{u}}_F &= \frac{1}{\sqrt{2}\rho_0 c_0^2}(\sqrt{c_0 + \boldsymbol{u}_0 \cdot \boldsymbol{n}_{K,F}}\mathrm{g}_{K,F}^{\oplus,n} - \sqrt{c_0 - \boldsymbol{u}_0 \cdot \boldsymbol{n}_{K,F}}\mathrm{g}_{K,F}^{\ominus,n}),\\
\widehat{p}_F + \rho_0(\boldsymbol{u}_0 \cdot \boldsymbol{n}_{K,F})\boldsymbol{n}_{K,F} \cdot \widehat{\boldsymbol{u}}_F &= \frac{1}{\sqrt{2}c_0}(\sqrt{c_0 + \boldsymbol{u}_0 \cdot \boldsymbol{n}_{K,F}}\mathrm{g}_{K,F}^{\oplus,n} + \sqrt{c_0 - \boldsymbol{u}_0 \cdot \boldsymbol{n}_{K,F}}\mathrm{g}_{K,F}^{\ominus,n}),
\end{aligned}$$

and

$$\begin{aligned}
\rho_0(\boldsymbol{u}_0 \cdot \boldsymbol{n}_{K,F})\boldsymbol{\pi}_{K,F}\widehat{\boldsymbol{u}}_F &= \frac{\sqrt{|\boldsymbol{u}_0 \cdot \boldsymbol{n}_{K,F}|}}{c_0}\boldsymbol{T}_{K,F}\mathbf{g}_{K,F}^{\oplus,t}, && \text{if } \boldsymbol{u}_0 \cdot \boldsymbol{n}_{K,F} > 0,\\
\rho_0(\boldsymbol{u}_0 \cdot \boldsymbol{n}_{K,F})\boldsymbol{\pi}_{K,F}\widehat{\boldsymbol{u}}_F &= \frac{\sqrt{|\boldsymbol{u}_0 \cdot \boldsymbol{n}_{K,F}|}}{c_0}\boldsymbol{T}_{K,F}\mathbf{g}_{K,F}^{\ominus,t}, && \text{if } \boldsymbol{u}_0 \cdot \boldsymbol{n}_{K,F} < 0,\\
\rho_0(\boldsymbol{u}_0 \cdot \boldsymbol{n}_{K,F})\boldsymbol{\pi}_{K,F}\widehat{\boldsymbol{u}}_F &= \mathbf{0}, && \text{if } \boldsymbol{u}_0 \cdot \boldsymbol{n}_{K,F} = 0.
\end{aligned}$$

These relations can be used in Equation (3.1) to rewrite the upwind numerical flux in terms of the transmission variables.

3.3.2 CHDG formulation

To define the CHDG formulation, four additional variables, denoted $\mathrm{g}_h^{\ominus,n}$, $\mathbf{g}_h^{\ominus,t}$ and $\mathrm{g}^{\oplus,n}$, $\mathbf{g}^{\oplus,t}$, corresponding to the incoming normal and tangential transmission variables and to the outgoing normal and tangential ones are introduced on the skeleton and belong respectively to the spaces

$$\mathcal{G}_h^{-,n} := \prod_{K\in\mathcal{T}_h}\prod_{F\in\mathcal{F}_K} \mathcal{P}_{\mathrm{p}}(F) \quad \text{and} \quad \boldsymbol{\mathcal{G}}_h^{-,t} := \prod_{K\in\mathcal{T}_h}\prod_{F\in\mathcal{F}_K^-} \boldsymbol{\mathcal{P}}_{\mathrm{p}}^{d-1}(F),$$
$$\mathcal{G}_h^{+,n} := \prod_{K\in\mathcal{T}_h}\prod_{F\in\mathcal{F}_K} \mathcal{P}_{\mathrm{p}}(F) \quad \text{and} \quad \boldsymbol{\mathcal{G}}_h^{+,t} := \prod_{K\in\mathcal{T}_h}\prod_{F\in\mathcal{F}_K^+} \boldsymbol{\mathcal{P}}_{\mathrm{p}}^{d-1}(F),$$

where $\mathcal{F}_K^- = \{F \in \mathcal{F}_K, \boldsymbol{u}_0\cdot\boldsymbol{n}_{K,F} < 0\}$ and $\mathcal{F}_K^+ = \{F \in \mathcal{F}_K, \boldsymbol{u}_0\cdot\boldsymbol{n}_{K,F} > 0\}$ denote the collection of inflow and outflow faces of an element K, respectively. Lastly, we define $\boldsymbol{\mathcal{X}}_h^- := \mathcal{G}_h^{-,n}\times\boldsymbol{\mathcal{G}}_h^{-,t}$ and $\boldsymbol{\mathcal{X}}_h^+ := \mathcal{G}_h^{+,n}\times\boldsymbol{\mathcal{G}}_h^{+,t}$.

The CHDG formulation of the problem is given by Problem 2.3.9 where

$$\boldsymbol{x}_h^{\ominus} = \begin{bmatrix} \mathrm{g}_h^{\ominus,n} \\ \mathbf{g}_h^{\ominus,t}\end{bmatrix} \in \boldsymbol{\mathcal{X}}_h^-$$

and with:

- the *global exchange operator*

$$\Pi = \begin{bmatrix}\Pi_n \\ \Pi_t\end{bmatrix} : \boldsymbol{\mathcal{X}}_h^+ \to \boldsymbol{\mathcal{X}}_h^-$$

whose components, for each face F of each element K, and for any $\mathrm{g}^{\oplus,n}\in\mathcal{G}_h^{+,n}$, $\mathbf{g}^{\oplus,t}\in\boldsymbol{\mathcal{G}}_h^{+,t}$, are defined as follows

$$\Pi_n(\mathrm{g}^{\oplus,n})|_{K,F} = \begin{cases} \mathrm{g}_{K',F}^{\oplus,n} & \text{if } F\not\subset\partial\Omega, \\ -\mathrm{g}_{K,F}^{\oplus,n}\sqrt{\dfrac{c_0-\boldsymbol{u}_0\cdot\boldsymbol{n}_{K,F}}{c_0+\boldsymbol{u}_0\cdot\boldsymbol{n}_{K,F}}} & \text{if } F\subset\Gamma_{\mathrm{D}}, \\ \mathrm{g}_{K,F}^{\oplus,n}\sqrt{\dfrac{c_0-\boldsymbol{u}_0\cdot\boldsymbol{n}_{K,F}}{c_0+\boldsymbol{u}_0\cdot\boldsymbol{n}_{K,F}}} & \text{if } F\subset\Gamma_{\mathrm{N}}, \\ 0 & \text{if } F\subset\Gamma_{\mathrm{R}}, \end{cases}$$
$$\Pi_t(\mathbf{g}^{\oplus,t})|_{K,F} = \begin{cases} \mathbf{g}_{K',F}^{\oplus,t} & \text{if } F\not\subset\partial\Omega,\ \boldsymbol{u}_0\cdot\boldsymbol{n}_{K,F}<0, \\ \mathbf{0} & \text{if } F\subset\Gamma_-; \end{cases}$$

- the *global scattering operator*

$$\mathrm{S} : \boldsymbol{\mathcal{X}}_h^- \to \boldsymbol{\mathcal{X}}_h^+$$

that, for each face F of each element K, and for any $\boldsymbol{x}_h^{\ominus}\in\boldsymbol{\mathcal{X}}_h^-$, is defined as follows

$$\mathrm{S}(\boldsymbol{x}_h^{\ominus})|_{K,F} := \boldsymbol{x}_h^{\oplus}|_{K,F} = \begin{cases} \begin{bmatrix} \mathrm{g}_h^{\oplus,n}|_{K,F} \\ \mathbf{g}_h^{\oplus,t}|_{K,F}\end{bmatrix}, & \text{if } \boldsymbol{u}_0\cdot\boldsymbol{n}_{K,F}>0, \\ \mathrm{g}_h^{\oplus,n}|_{K,F}, & \text{if } \boldsymbol{u}_0\cdot\boldsymbol{n}_{K,F}\le 0, \end{cases}$$

with

$$\mathrm{g}_h^{\oplus,n}|_{K,F} := \sqrt{\frac{c_0+\boldsymbol{u}_0\cdot\boldsymbol{n}_{K,F}}{2}}(p_K(\boldsymbol{x}_h^{\ominus}) + \rho_0 c_0 \boldsymbol{n}_{K,F}\cdot\boldsymbol{u}_K(\boldsymbol{x}_h^{\ominus})),$$
$$\mathbf{g}_h^{\oplus,t}|_{K,F} := \rho_0 c_0\sqrt{|\boldsymbol{u}_0\cdot\boldsymbol{n}_{K,F}|}\boldsymbol{T}_{K,F}^{\dagger}\boldsymbol{u}_K(\boldsymbol{x}_h^{\ominus}), \qquad \text{if } \boldsymbol{u}_0\cdot\boldsymbol{n}_{K,F}>0,$$

where $\boldsymbol{U}_K = (p_K(\boldsymbol{x}_h^{\ominus}), \rho_0 c_0 \boldsymbol{u}_K(\boldsymbol{x}_h^{\ominus}))$ is the solution of the local problem with the incoming transmission variable $(\boldsymbol{x}_{K,F}^{\ominus})_{F\in\mathcal{F}_K}$ contained in $\boldsymbol{x}_h^{\ominus}$ as a given surface datum;

- the *global right-hand side*

$$\boldsymbol{b} = \begin{bmatrix} \mathrm{b}_n \\ \mathbf{b}_t \end{bmatrix} \in \boldsymbol{\mathcal{X}}_h^-$$

with components

$$\mathrm{b}_n|_{K,F} = \begin{cases} 0 & \text{if } F \not\subset \partial\Omega, \\ s_{\mathrm{D}}\sqrt{2(c_0 - \boldsymbol{u}_0\cdot\boldsymbol{n}_{K,F})} & \text{if } F \subset \Gamma_{\mathrm{D}}, \\ -s_{\mathrm{N}}\rho_0 c_0\sqrt{2(c_0 - \boldsymbol{u}_0\cdot\boldsymbol{n}_{K,F})} & \text{if } F \subset \Gamma_{\mathrm{N}}, \\ s_{\mathrm{R}}\sqrt{\dfrac{c_0 - \boldsymbol{u}_0\cdot\boldsymbol{n}_{K,F}}{2}} & \text{if } F \subset \Gamma_{\mathrm{R}}, \end{cases}$$

$$\mathbf{b}_t|_{K,F} = \begin{cases} \mathbf{0} & \text{if } F \not\subset \partial\Omega,\ \boldsymbol{u}_0\cdot\boldsymbol{n}_{K,F} < 0, \\ -\rho_0 c_0\sqrt{|\boldsymbol{u}_0\cdot\boldsymbol{n}_{K,F}|}\boldsymbol{s}_- & \text{if } F \subset \Gamma_-. \end{cases}$$

Verification of Assumption 2.3.4

We want to check if Assumption 2.3.4 is verified by all boundary faces, so that the whole boundary is passive. To do so, we first need to define the matrices $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ for which we can write a relation between $\boldsymbol{x}_{K,F}^{\ominus}$ and $\boldsymbol{x}_{K,F}^{\oplus}$ in the form of Equation (2.13).

If $F \subset \Gamma_+$, we have that $\boldsymbol{\alpha} \in \mathbb{R}^{1\times d}$ and $\boldsymbol{\beta} \in \mathbb{R}$ and, in particular, they read

$$\boldsymbol{\alpha} = \begin{bmatrix} \alpha_1 & \mathbf{0}^\dagger \end{bmatrix}, \quad \boldsymbol{\beta} = \begin{bmatrix} \beta_1 \end{bmatrix},$$

with

$$\alpha_1 := \begin{cases} -\sqrt{\dfrac{c_0 - \boldsymbol{u}_0\cdot\boldsymbol{n}_{K,F}}{c_0 + \boldsymbol{u}_0\cdot\boldsymbol{n}_{K,F}}}, & \text{if } F \subset \Gamma_{\mathrm{D}}, \\ \sqrt{\dfrac{c_0 - \boldsymbol{u}_0\cdot\boldsymbol{n}_{K,F}}{c_0 + \boldsymbol{u}_0\cdot\boldsymbol{n}_{K,F}}}, & \text{if } F \subset \Gamma_{\mathrm{N}}, \\ 0 & \text{if } F \subset \Gamma_{\mathrm{R}}, \end{cases} \qquad \beta_1 := \begin{cases} \sqrt{2(c_0 - \boldsymbol{u}_0\cdot\boldsymbol{n}_{K,F})}, & \text{if } F \subset \Gamma_{\mathrm{D}}, \\ -\rho_0 c_0\sqrt{2(c_0 - \boldsymbol{u}_0\cdot\boldsymbol{n}_{K,F})}, & \text{if } F \subset \Gamma_{\mathrm{N}}, \\ \sqrt{\dfrac{c_0 - \boldsymbol{u}_0\cdot\boldsymbol{n}_{K,F}}{2}} & \text{if } F \subset \Gamma_{\mathrm{R}}. \end{cases}$$

Whereas, if $F \subset \Gamma_-$, we have that $\boldsymbol{\alpha} \in \mathbb{R}^d$ and $\boldsymbol{\beta} \in \mathbb{R}^{d\times d}$ and, in particular, they read

$$\boldsymbol{\alpha} = \begin{bmatrix} \alpha_1 \\ \mathbf{0} \end{bmatrix}, \quad \boldsymbol{\beta} = \begin{bmatrix} \beta_1 & \mathbf{0}^\dagger \\ \mathbf{0} & \beta_2 \boldsymbol{I} \end{bmatrix},$$

with α_1 and β_1 defined as above and

$$\beta_2 := -\rho_0 c_0\sqrt{|\boldsymbol{u}_0\cdot\boldsymbol{n}_{K,F}|}.$$

Assumption 2.3.4 holds if and only if

$$\|\boldsymbol{\alpha}\| := \sqrt{\max_i |\lambda_i(\boldsymbol{\alpha}^\dagger\boldsymbol{\alpha})|} = |\alpha_1| \le 1.$$

For $F \subset \Gamma_{\mathrm{R}}$ we have

$$|\alpha_1| = 0,$$

and, therefore, the faces are non-reflective. Whereas, for $F \subset \Gamma_D \cup \Gamma_N$, we have

$$|\alpha_1| = \sqrt{\frac{c_0 - \boldsymbol{u}_0 \cdot \boldsymbol{n}_{K,F}}{c_0 + \boldsymbol{u}_0 \cdot \boldsymbol{n}_{K,F}}}.$$

In the case of outflow boundary, i.e. $\boldsymbol{u}_0 \cdot \boldsymbol{n}_{K,F} \geq 0$, the face is passive ($|\alpha_1| \leq 1$), while in the case of inflow boundary, i.e. $\boldsymbol{u}_0 \cdot \boldsymbol{n}_{K,F} < 0$, the face is not passive ($|\alpha_1| > 1$) and energy is brought inside the domain.

Consequently, the need of dissipation of energy obliges us to impose Robin boundary conditions for faces that belong to a portion of the inflow boundary (in the case of outflow boundary every boundary condition is dissipative). Therefore, in the context of Problem 3.1.2, Assumption 2.3.4 can be equivalently stated as follows.

Assumption 3.3.1. *The background flow is outgoing at Dirichlet and Neumann portions of the boundary, that is*

$$\boldsymbol{u}_0 \cdot \boldsymbol{n} \geq 0 \quad \text{on } \Gamma_D \cup \Gamma_N.$$

Theorem 2.3.7 holds and it ensures the well-posedness of the local problem. Moreover, under Assumption 3.3.1, all the results of Section 2.3.5 hold true and in particular Corollary 2.3.14, namely we have that the operator ΠS is a strict contraction.

3.4 Numerical results

In this Section, we aim at studying and comparing iterative solutions for the DG and CHDG methods by using simple two-dimensional benchmark problems.

In the first benchmark (Section 3.4.1) we consider a square domain in which the background flow is not necessarily parallel to the direction of propagation of the plane wave. In the second benchmark (Section 3.4.2) we deal with the propagation of a wave induced by a point source in a circular domain and in the presence of flow. In the third benchmark (Section 3.4.3) we address the propagation of a vorticity wave in a bounded rigid duct with flow.

The numerical simulations are performed with a dedicated MATLAB code. The mesh generation and the visualization are performed with gmsh and h is the element size specified in gmsh.

Iterative solvers. Three standard iterative solvers (presented in details in Section 1.3) are considered:

- fixed-point iteration (only for the CHDG system) (see Section 1.3.2): this choice naturally arises from the strict contraction property of the operator ΠS that assures the convergence of the method. Moreover, it requires a small memory storage;
- CGNR (*conjugate gradient normal residual*) iteration (see Section 1.3.2): it demands a small memory storage, but it requires at each iteration two matrix-vector products;
- GMRES (*generalized minimal residual*) iteration without restart (see Section 1.3.3): it requires at each iteration only one matrix-vector product, but it requires to store the history of all previous Krylov subspace vectors, which results in a memory storage that increases with the number of iterations.

Basis functions. The basis functions associated with the physical fields p_h and $\boldsymbol{u}_h$ are standard hierarchical basis functions obtained by tensor products of Lobatto functions (see e.g. [21, 120]). Concerning the field $\boldsymbol{x}_h^{\ominus}$, we simply use the restriction on the skeleton of the basis functions used for the physical fields. In all the cases, third-order polynomials (p = 3) are used so that small mesh sizes h are not needed to achieve an arbitrary accuracy.

Preconditioning. Similarly to the approach used in [97], we use a symmetric preconditioning with the mass matrix M associated to the faces of the elements. Denoting the Cholesky factorization of the mass matrix with $\mathsf{M} = \mathsf{L}\mathsf{L}^\top$, this leads to the system $\tilde{\mathsf{A}}\tilde{\mathsf{x}} = \tilde{\mathsf{b}}$ with $\tilde{\mathsf{A}} := \mathsf{L}^{-1}\mathsf{A}\mathsf{L}^{-\top}$, $\tilde{\mathsf{x}} := \mathsf{L}^\top \mathsf{x}$ and $\tilde{\mathsf{b}} := \mathsf{L}^{-1}\mathsf{b}$. The preconditioned system corresponds to the one that exploits orthonormal basis functions on the skeleton. Then, the 2-norm of an algebraic vector $\tilde{\mathsf{x}}$ is equal to the L^2-norm of the corresponding field, i.e. $\|\tilde{\mathsf{x}}\|_2 = \sqrt{\tilde{\mathsf{x}}^*\tilde{\mathsf{x}}} = \sqrt{\mathsf{x}^*\mathsf{M}\mathsf{x}} = \|\boldsymbol{x}_h\|$ with

$$\|\boldsymbol{x}_h\| = \sqrt{\sum_{K\in\mathcal{T}_h}\sum_{F\in\mathcal{F}_K}\langle\sum_{i=1}^{\mathrm{p}} x_{K,F,i}\phi_i, \sum_{j=1}^{\mathrm{p}} x_{K,F,j}\phi_j\rangle_F} = \sqrt{\sum_{K\in\mathcal{T}_h}\sum_{F\in\mathcal{F}_K}\mathsf{x}_{K,F}^*\mathsf{M}_{K,F}\mathsf{x}_{K,F}}.$$

Numerical error. We consider a relative numerical error based on the acoustic energy associated to the physical fields presented in Section 1.1.5. It is defined as

$$\text{relative error} := \frac{\|p_h - p_{\mathrm{ref}}, \boldsymbol{u}_h - \boldsymbol{u}_{\mathrm{ref}}\|_E}{\|p_{\mathrm{ref}}, \boldsymbol{u}_{\mathrm{ref}}\|_E},$$

with

$$\|p, \boldsymbol{u}\|_E^2 := \sum_{K\in\mathcal{T}_h}\left(\frac{\|p\|_{L^2(K)}^2}{2\rho_0 c_0^2} + \frac{1}{2}\rho_0\|\boldsymbol{u}\|_{L^2(K)}^2 + \left(\frac{1}{2c_0^2}p\boldsymbol{u}_0, \boldsymbol{u}\right)_K + \left(\frac{1}{2c_0^2}\boldsymbol{u}, p\boldsymbol{u}_0\right)_K\right),$$

and where $(p_h, \boldsymbol{u}_h)$ is the numerical solution, $(p_{\mathrm{ref}}, \boldsymbol{u}_{\mathrm{ref}})$ is the reference solution, and $\|.\|_E$ denotes the energy norm, see (1.14).

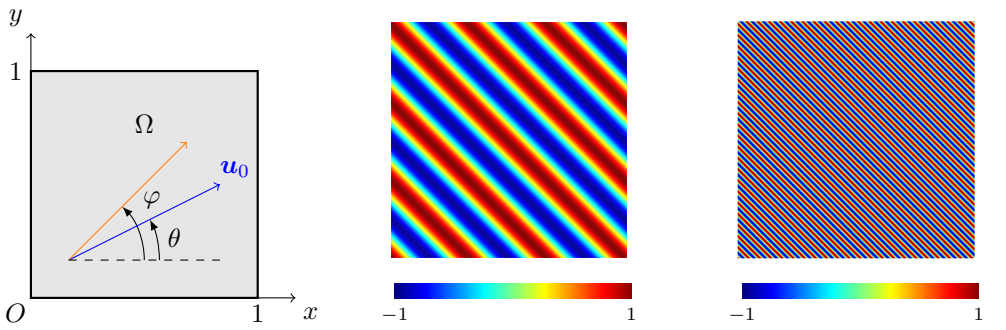
3.4.1 Propagative plane wave

The first benchmark consists in the propagation of a plane wave in a uniform background flow. We consider the square domain $\Omega = (0,1)^2$. The background flow $\boldsymbol{u}_0$ is oriented in the direction $\boldsymbol{\theta} = (\cos\theta, \sin\theta)$, with $\theta \in [0, 2\pi)$, i.e. $\boldsymbol{u}_0 = \|\boldsymbol{u}_0\|\cos\theta\boldsymbol{e}_x + \|\boldsymbol{u}_0\|\sin\theta\boldsymbol{e}_y$ (see Figure 3.2a) and we assume no volume source, i.e. $f \equiv 0$ in Ω.

The reference solution (see Figure 3.2b) reads (see Appendix 3.A.1 for further details)

$$p_{\mathrm{ref}}(\boldsymbol{x}) = \exp\left(\frac{\imath\kappa\boldsymbol{x}\cdot\boldsymbol{\varphi}}{1 + M(\boldsymbol{\theta}\cdot\boldsymbol{\varphi})}\right),$$
$$\boldsymbol{u}_{\mathrm{ref}}(\boldsymbol{x}) = \frac{p_{\mathrm{ref}}(\boldsymbol{x})}{\rho_0 c_0}\boldsymbol{\varphi},$$

where $M := \|\boldsymbol{u}_0\|/c_0$ is the *Mach number* with $0 \le M < 1$, $\kappa = \omega/c_0$ is the wavenumber and $\boldsymbol{\varphi} = (\cos\varphi, \sin\varphi)$, with $\varphi \in [0, 2\pi)$, is the direction of propagation of the plane wave.



(a) Square domain with a background flow.

(b) Real part of p_{ref} for $\varphi = \pi/4$, $\theta = \pi/4$ and $M = 0.75$.

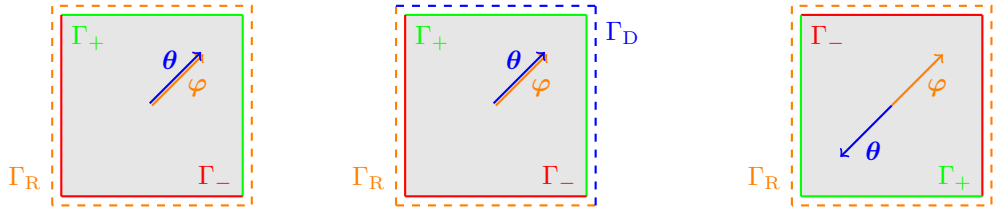
(c) Real part of p_{ref} for $\varphi = \pi/4$, $\theta = 5\pi/4$ and $M = 0.75$.

Figure 3.2: Domain and reference solution for the propagative plane wave benchmark.

We prescribe a non-homogeneous Robin boundary condition for the acoustic waves, besides a boundary condition for the hydrodynamic waves, on the inflow boundary (this choice satisfies Assumption 3.3.1) and we consider two configurations depending on the non-homogeneous boundary condition imposed on the outflow boundary: we pick $\Gamma_+ \subset \Gamma_R$ and $\Gamma_+ = \Gamma_D$, by exploiting the reference solution, i.e.

$$\begin{cases} p = p_{\text{ref}}, & \text{on } \Gamma_D, \\ p - \rho_0 c_0 \boldsymbol{n} \cdot \boldsymbol{u} = p_{\text{ref}} - \rho_0 c_0 \boldsymbol{n} \cdot \boldsymbol{u}_{\text{ref}}, & \text{on } \Gamma_R, \\ \boldsymbol{T}^\dagger \boldsymbol{u} = \boldsymbol{T}^\dagger \boldsymbol{u}_{\text{ref}}, & \text{on } \Gamma_-. \end{cases}$$

Moreover, we select two values of $\boldsymbol{\theta} \cdot \boldsymbol{\varphi}$, in particular, we pick $\boldsymbol{\theta} \cdot \boldsymbol{\varphi} = 1$ and $\boldsymbol{\theta} \cdot \boldsymbol{\varphi} = -1$, that are associated respectively to $\theta = \pi/4$ and $\theta = 5\pi/4$ that, combined with different choices of boundary conditions for the outflow boundary, lead to three cases (see Figure 3.3).



(i) Case i (ii) Case ii (iii) Case iii

Figure 3.3: Types of boundary (Γ_D = ——, Γ_R = – –, Γ_+ = —, Γ_- = —) and directions of propagation of the plane wave $\boldsymbol{\varphi}$ and of the background flow $\boldsymbol{\theta}$ in three cases.

For the numerical simulations, we choose the mesh size h in such a way that the relative error is of the order of 10^{-2}. We fix the density $\rho_0 = 1$, the sound speed $c_0 = 1$, the angular frequency $\omega = 15\pi$, the angle of the plane wave $\varphi = \pi/4$ and we pick two Mach numbers: $M = 0.25, 0.75$.

In Table 3.1 we gather the relative errors depending on the case. As expected, the smaller is $M\boldsymbol{\theta} \cdot \boldsymbol{\varphi}$, the smaller is the mesh size h. This is due to the fact that decreasing $M\boldsymbol{\theta} \cdot \boldsymbol{\varphi}$ leads to larger wavenumber and therefore to the high-frequency scenario which is known to require a bigger computational effort compared to the low-frequency scenario that can be obtained by increasing $M\boldsymbol{\theta} \cdot \boldsymbol{\varphi}$. In particular, this justifies the necessity of an extremely small mesh size h in the most computationally difficult configuration, that is the one with the smallest value of $M\boldsymbol{\theta} \cdot \boldsymbol{\varphi}$ (case iii with $\boldsymbol{\theta} \cdot \boldsymbol{\varphi} = -1$ and $M = 0.75$). Lastly, the presence of a Dirichlet portion of the boundary (case ii) leads to a slightly larger error (compared with case i).

Case	Γ_+	Γ_-	$\boldsymbol{\theta} \cdot \boldsymbol{\varphi}$	h	M	Relative error
i	$\subset \Gamma_R$	$\subset \Gamma_R$	1	1/13	0.25	$1.06 \cdot 10^{-2}$
				1/9	0.75	$1.28 \cdot 10^{-2}$
ii	$= \Gamma_D$	$= \Gamma_R$	1	1/13	0.25	$1.25 \cdot 10^{-2}$
				1/9	0.75	$1.29 \cdot 10^{-2}$
iii	$\subset \Gamma_R$	$\subset \Gamma_R$	-1	1/22	0.25	$1.10 \cdot 10^{-2}$
				1/75	0.75	$1.40 \cdot 10^{-2}$

Table 3.1: Parameters and relative error for the propagative plane wave benchmark, with p = 3, $\omega = 15\pi$, $\varphi = \pi/4$, $c_0 = 1$ and $\rho_0 = 1$.

In Table 3.2 we collect the number of degrees of freedom and the number of non-zero entries of the DG and CHDG systems with respect to the mesh size h.

		$h = 1/9$	$h = 1/13$	$h = 1/22$	$h = 1/75$
#dof	DG	5 880	12 240	34 560	394 080
	CHDG	4 704	9 792	27 648	315 264
#nnz	DG	233 580	489 628	1 393 032	16 001 760
	CHDG	78 944	165 968	472 896	5 442 992

Table 3.2: Number of degrees of freedom (#dof) and number of non-zero entries (#nnz) of the DG and CHDG systems for the propagative plane wave benchmark with respect to the mesh size h.

The CHDG system always presents less degrees of freedom than the DG system, even if the difference is sometimes less pronounced. For a more detailed insight regarding the memory storage for the two systems, we refer to Section 4.3 of [97].

Comparison with iterative solvers: DG and CHDG

The numerical results (see Figure 3.4) are consistent with the theory regarding the strict contraction property of the operator ΠS: the fixed-point iteration applied to the CHDG system always converges. Moreover, we can observe that its speed is generally very high thanks to the high physical dissipation introduced by the Robin boundary condition (case i and iii).

As far as CGNR and GMRES are concerned, they are always faster once they are applied to the CHDG system than to the DG one. GMRES is always slightly faster than CGNR in terms of number of iterations, but it requires a higher overall computational effort (especially since the algorithm excludes restart). However, the implementation of these solvers is not optimized (further major improvements could be implemented) and therefore these remarks can not lead to a precise choice of the iterative solver to use in practice and, therefore, a more detailed performance analysis is needed.

The use of a Dirichlet condition on part of the boundary (case ii) does not seem to have an important impact on the speed of convergence, which slightly reduces only for the fixed-point scheme applied to the CHDG system in the case of low Mach number ($M = 0.25$) and for the GMRES scheme applied to the DG system in the case of low Mach number.

Mixed boundary conditions (case ii) generally do not slow the convergence, neither for the CHDG system nor for the DG one, except for the GMRES scheme applied to the DG system in the case of low Mach number.

As a final remark, we can observe that in the most computationally difficult setting ($\boldsymbol{\theta} \cdot \boldsymbol{\varphi} = -1$ and $M = 0.75$) all the iterative solutions associated to the DG systems are extremely slow, whereas those associated to the CHDG system are simply slowed down compared to the other configurations.

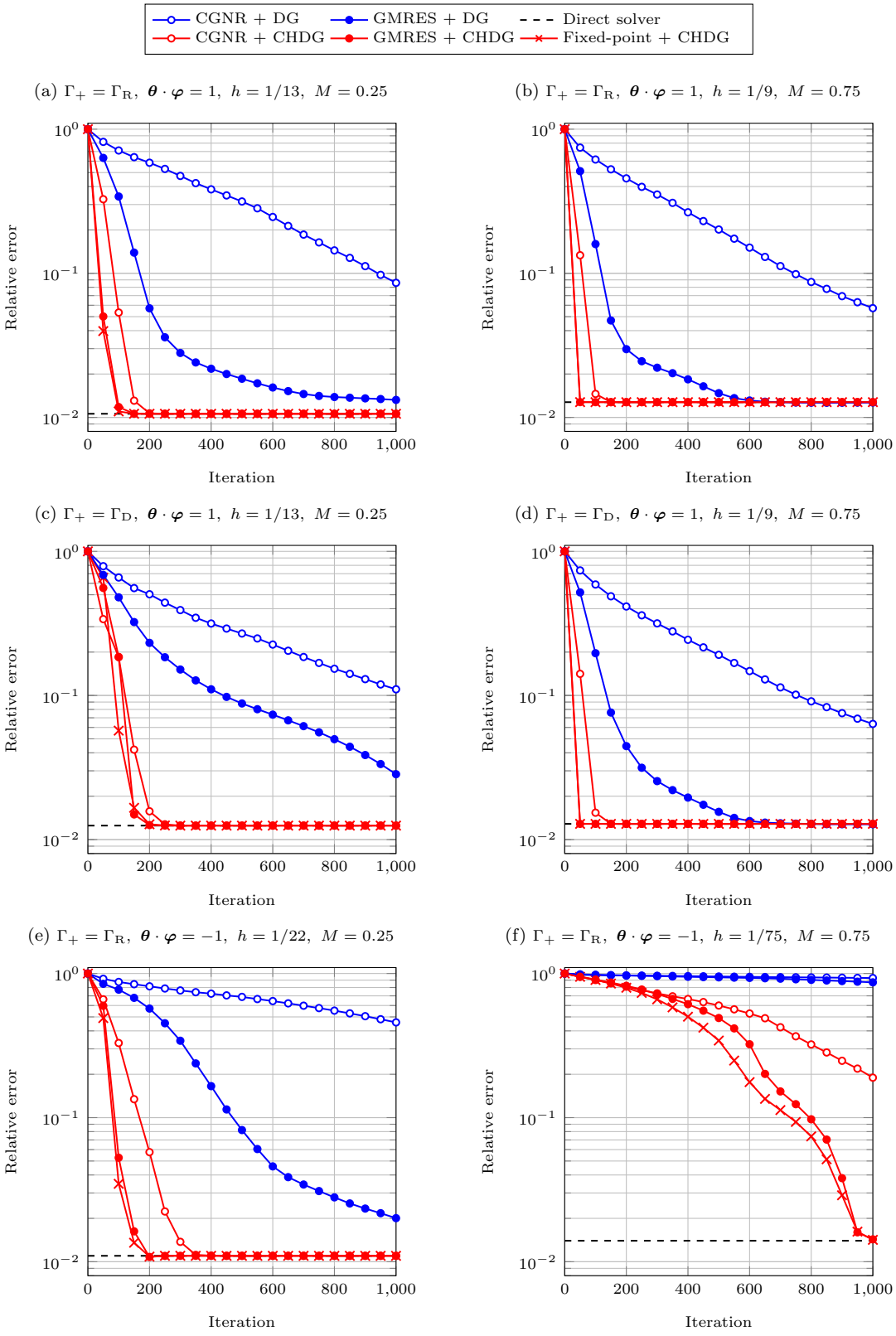


Figure 3.4: Error history with different iterative schemes and different DG methods for the propagative plane wave benchmark. The dashed horizontal lines correspond to the relative numerical errors obtained with a direct solver.

3.4.2 Acoustic point source

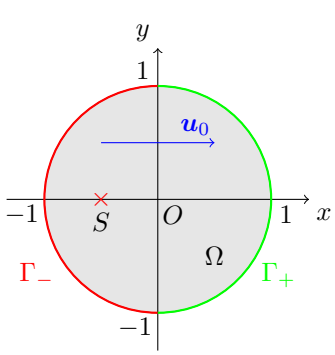
The second benchmark consists in the acoustic field generated in 2D by a point source in a uniform background flow. We consider the unit disc $\Omega = [0,1) \times [0, 2\pi)$ (given in polar coordinates (r, θ)). We assume a uniform horizontal background flow $\boldsymbol{u}_0$, i.e. $\boldsymbol{u}_0 = |\boldsymbol{u}_0|\boldsymbol{e}_x = Mc_0\boldsymbol{e}_x$, and an acoustic point source, i.e. $f(x,y) = \delta(x - x_s)\delta(y - y_s)$ where $S = (x_s, y_s)$ is the location of the unit source (see Figure 3.5a).

The reference solution (see Figure 3.5b) can be written in terms of the following Green function $\mathcal{G}$

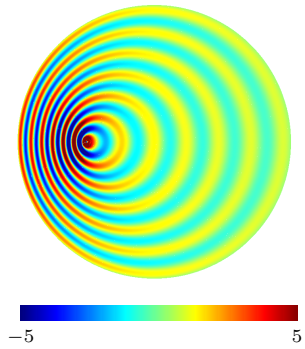
$$\mathcal{G}(x,y) = \frac{\imath}{4}\frac{1}{c_0^2\sqrt{1-M^2}} H_0^{(1)}\left[\kappa\frac{\sqrt{x^2+(1-M^2)y^2}}{1-M^2}\right] \exp\Big(-\imath\frac{M}{1-M^2}\kappa x\Big),$$

and reads (see Appendix 3.A.2 for further details)

$$\begin{aligned} p_{\text{ref}}(x,y) &= \rho_0 c_0^2(-\imath\omega + \boldsymbol{u}_0\cdot\boldsymbol{\nabla})\mathcal{G}(x-x_s, y-y_s),\\ \boldsymbol{u}_{\text{ref}}(x,y) &= -c_0^2\boldsymbol{\nabla}\mathcal{G}(x-x_s, y-y_s). \end{aligned}$$



(a) Circular domain with background flow and unit source.



(b) Real part of p_{ref} for $M = 0.5$.

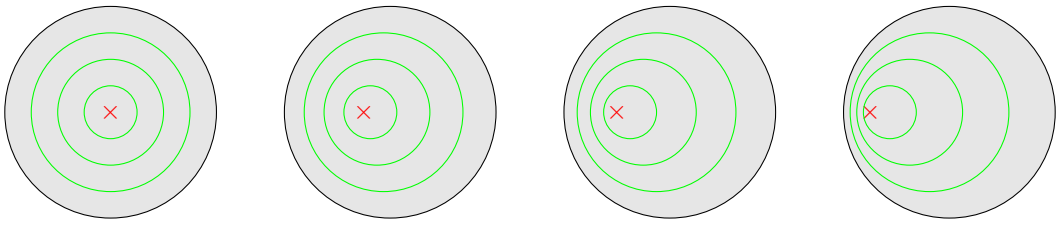
Figure 3.5: Domain and reference solution for the acoustic point source benchmark.

We impose a first-order absorbing boundary condition on $\partial\Omega$ and a boundary condition on the inflow boundary by exploiting the reference solution, i.e.

$$\begin{cases} p - \rho_0 c_0 \boldsymbol{n}\cdot\boldsymbol{u} = 0, & \text{on } \partial\Omega,\\ \boldsymbol{T}^\dagger\boldsymbol{u} = \boldsymbol{T}^\dagger\boldsymbol{u}_{\text{ref}}, & \text{on } \Gamma_-. \end{cases}$$

To reduce spurious reflections on $\partial\Omega$, we take $S = (-M, 0)$ so that the radiating waves hit the boundary with normal incidence.

For the numerical simulations, as in the previous benchmark, we choose the mesh size h in such a way that the relative error is of the order 10^{-2}. The location of the unit source is a singular point for the reference solution. We exclude the closest elements to the area close to the unit source from the computation of the numerical error and of the norm of the solution (we eliminate the elements whose centroid has a distance from the point source below the threshold $d = 0.05$, so that we exclude only the closest elements) because the reference solution in this area has a much bigger amplitude than that it has in areas far from it. Hence, if included, they would make the error bigger regardless of the possibly fairly good behavior of the discrete solution far from the unit source. We fix the density $\rho_0 = 1$, the sound speed $c_0 = 1$, the angular frequency $\omega = 40$. We consider four configurations depending on the value of the Mach number: $M = 0, 0.25, 0.5, 0.75$ (see Figure 3.6).



(i) $M = 0$ (ii) $M = 0.25$ (iii) $M = 0.5$ (iv) $M = 0.75$

Figure 3.6: Location of the point source ($\times$) and qualitative behavior of some wavefronts ($-$) in four cases.

In Table 3.3 we gather the relative errors depending on the case. The bigger M is, the smaller the mesh size h is. Similar to the propagative plane wave benchmark, the Mach number plays an important role in the computational difficulty of the problem. Indeed, increasing M, the location of the point source moves to the left and gets closer and closer to the boundary.

Case	h	M	Relative error
i	1/12	0	$1.23 \cdot 10^{-2}$
ii	1/25	0.25	$1.89 \cdot 10^{-2}$
iii	1/25	0.50	$2.48 \cdot 10^{-2}$
iv	1/50	0.75	$1.88 \cdot 10^{-2}$

Table 3.3: Parameters and relative error for the acoustic point source benchmark, with p = 3, $\omega = 40$, $c_0 = 1$ and $\rho_0 = 1$.

In Table 3.4 we collect the number of degrees of freedom and the number of non-zero entries of the DG and CHDG systems with respect to the mesh size h.

		$h = 1/12$	$h = 1/25$	$h = 1/50$
#dof	DG	51 660	161 820	548 430
	CHDG	41 328	129 456	438 744
#nnz	DG	2 085 238	6 568 682	22 293 772
	CHDG	409 632	2 232 480	7 582 176

Table 3.4: Number of degrees of freedom (#dof) and number of non-zero entries (#nnz) of the DG and CHDG systems for the acoustic point source benchmark with respect to the mesh size h.

Comparison with iterative solvers: DG and CHDG

Similar to the previous benchmark, the numerical results (see Figure 3.7) show that all iterative schemes perform better if applied to the CHDG system than to the DG one. However, this time, the fixed-point iteration is not as competitive as the GMRES method, despite presenting a fast convergence due to the high physical dissipation associated to the Robin boundary condition.

Also for this benchmark, in the most computationally difficult setting ($M = 0.75$), all iterative schemes applied to the DG system are extremely slow, whereas, when they are applied to the CHDG system, they are faster and, in particular, the GMRES and fixed-point iteration exhibit a good rate of convergence.

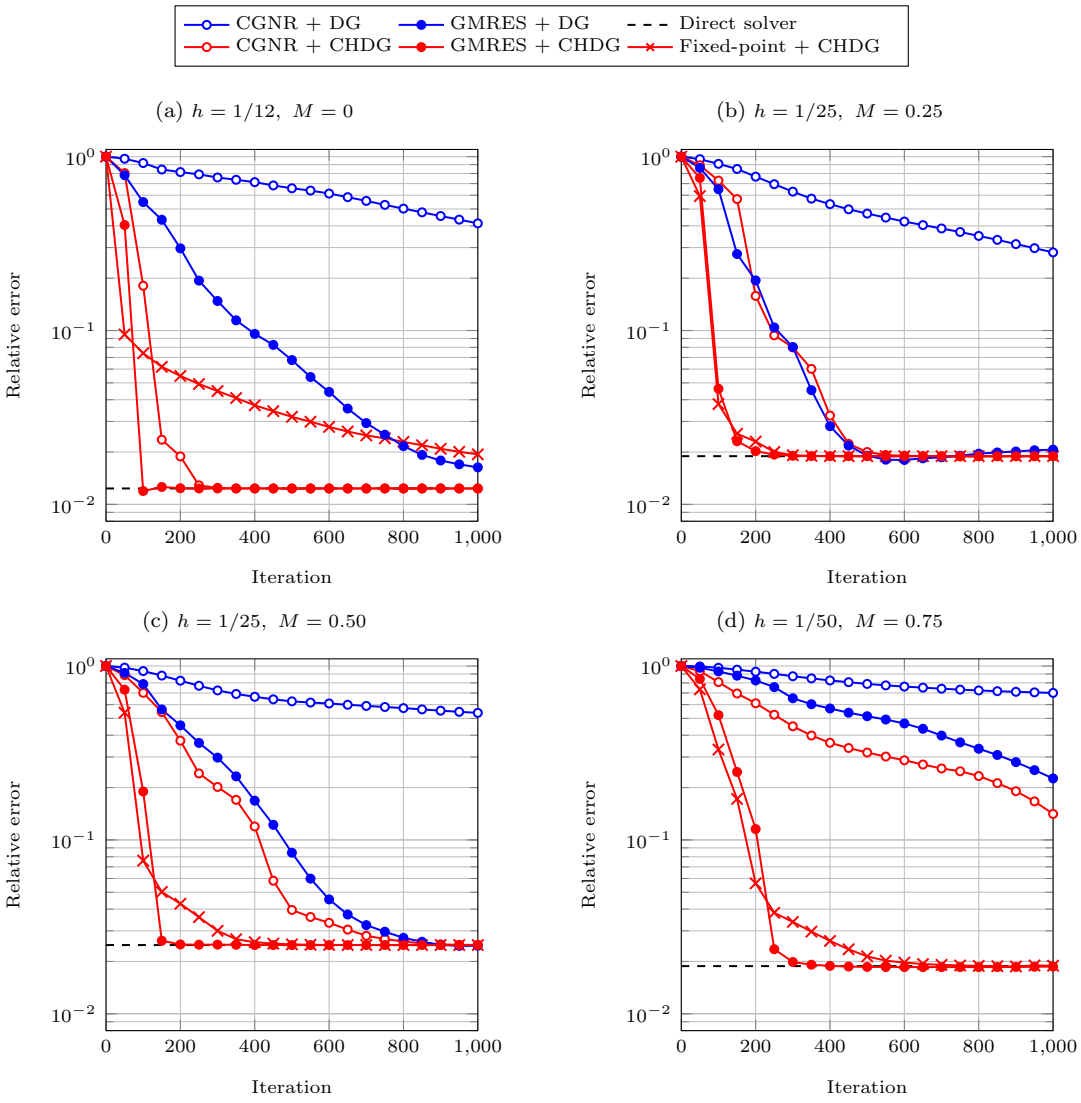


Figure 3.7: Error history with different iterative schemes and different DG methods for the acoustic point source benchmark. The dashed horizontal lines correspond to the relative numerical errors obtained with a direct solver.

3.4.3 Vorticity wave

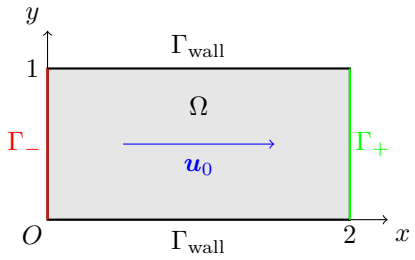
The third benchmark consists in the propagation of a vorticity wave inside a rigid duct in the presence of a uniform background flow. We consider a 2-dimensional infinite rigid duct whose walls are parallel to the x direction and whose width (i.e. the distance between the two walls) is 1, namely $\Gamma_{\text{wall}} = \mathbb{R} \times \{0, 1\}$. It corresponds to the unbounded domain $\Omega = \mathbb{R} \times (0, 1)$. We assume a uniform and horizontal background flow $\boldsymbol{u}_0$, i.e. $\boldsymbol{u}_0 = \|\boldsymbol{u}_0\| \boldsymbol{e}_x = M c_0 \boldsymbol{e}_x$ and no volume source, i.e. $f \equiv 0$ in Ω.

The reference solution reads (see Appendix 3.A.3 for further details)

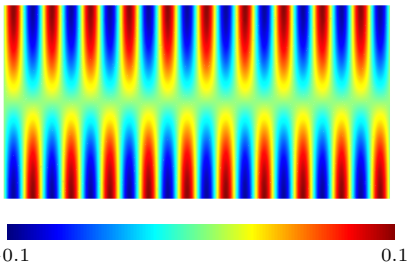
$$
\begin{aligned}
&p_{\text{ref}}(x, y) = 0, \\
&\boldsymbol{u}_{\text{ref}}(x, y) = \left(\frac{\imath n \pi M c_0}{\omega} \cos(n\pi y) \exp\left(\frac{\imath \omega}{M c_0} x \right), \sin(n\pi y) \exp\left(\frac{\imath \omega}{M c_0} x \right) \right),
\end{aligned}
$$

where $n \in \mathbb{N}$ denotes the n-th mode. While we have $\boldsymbol{\nabla} \times \boldsymbol{u} = \boldsymbol{0}$ for the acoustic wave considered before, we have $\boldsymbol{\nabla} \cdot \boldsymbol{u} = 0$ for the vorticity wave.

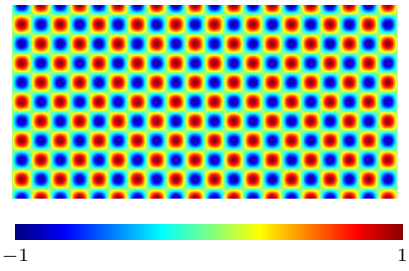
We consider the bounded domain $\Omega = (0, 2) \times (0, 1)$. In particular, we denote by Γ_- the inflow interface $\{0\} \times (0, 1)$ and by Γ_+ the outflow interface $\{2\} \times (0, 1)$ (see Figure 3.8a).



(a) Bounded rigid duct with a horizontal right-propagating background flow.



(b) Real part of $u_{\text{ref},x}$ for $n = 1$ and $M = 0.5$.



(c) Real part of $u_{\text{ref},x}$ for $n = 10$ and $M = 0.5$.

Figure 3.8: Domain and reference solution for the vorticity wave benchmark.

We prescribe a non-homogeneous Robin and inflow boundary condition on the inflow boundary and we consider two cases depending on the non-homogeneous boundary condition imposed on the outflow boundary: we pick $\Gamma_+ \subset \Gamma_{\text{R}}$ and $\Gamma_+ = \Gamma_{\text{D}}$, by exploiting the reference solution, and we impose a homogeneous Neumann boundary condition on the rigid walls, i.e.

$$
\begin{cases}
\boldsymbol{n} \cdot \boldsymbol{u} = 0, & \text{on } \Gamma_{\text{wall}}, \\
p = p_{\text{ref}}, & \text{on } \Gamma_{\text{D}}, \\
p - \rho_0 c_0 \boldsymbol{n} \cdot \boldsymbol{u} = p_{\text{ref}} - \rho_0 c_0 \boldsymbol{n} \cdot \boldsymbol{u}_{\text{ref}}, & \text{on } \Gamma_{\text{R}}, \\
\boldsymbol{T}^\dagger \boldsymbol{u} = \boldsymbol{T}^\dagger \boldsymbol{u}_{\text{ref}}, & \text{on } \Gamma_-.
\end{cases}
$$

For the numerical simulation, as in the previous benchmarks, we choose the mesh size h in such a way that the relative error is of the order 10^{-2}. We fix the density $\rho_0 = 1$, the sound velocity $c_0 = 1$, the angular frequency $\omega = 5\pi$ and the Mach number $M = 0.5$. We consider two configurations for each case depending on the number of the mode, in particular, we pick $n = 1$ and $n = 10$ (see Figures 3.8b-3.8c).

In Table 3.5 we gather the relative errors depending on the case. As expected, the higher-order mode ($n = 10$) compared to the low-order one ($n = 1$) is computationally more challenging and therefore a smaller mesh size is required to obtain a similar relative error. The boundary condition imposed on the outflow boundary does not seem to have a significant impact on the relative error.

Case	Γ_+	Γ_-	n	h	Relative error
i	$\subset \Gamma_R$	$\subset \Gamma_R$	1	1/11	$1.24 \cdot 10^{-2}$
			10	1/20	$1.12 \cdot 10^{-2}$
ii	$= \Gamma_D$	$= \Gamma_R$	1	1/11	$1.23 \cdot 10^{-2}$
			10	1/20	$1.12 \cdot 10^{-2}$

Table 3.5: Parameters and relative error for the vorticity wave benchmark, with p = 3, $\omega = 5\pi$, $c_0 = 1$ and $\rho_0 = 1$.

In Table 3.6 we collect the number of degrees of freedom and the number of non-zero entries of the DG and CHDG systems with respect to the mesh size h.

		$h = 1/11$	$h = 1/20$
#dof	DG	17 460	56 160
	CHDG	13 968	44 928
#nnz	DG	700 443	2 267 611
	CHDG	238 352	771 776

Table 3.6: Number of degrees of freedom (#dof) and number of non-zero entries (#nnz) of the DG and CHDG systems for the vorticity wave benchmark with respect to the mesh size h.

Comparison with iterative solvers: DG and CHDG

The numerical results (see Figure 3.9) confirm the main remarks and conclusions we were able to draw thanks to the previous two benchmarks. In particular, as far as the iterative schemes are concerned, the CHDG system systematically outperforms the DG one.

The boundary condition imposed on the outflow boundary Γ_+ does not have a significant impact on the iterative solutions for the mode $n = 10$. Indeed, all the corresponding curves that appear in Figures 3.9b-3.9d are almost identical, respectively. Whereas, considering the mode $n = 1$, we can observe that the iterative solutions that appear in Figure 3.9a and that are associated to a Robin outflow boundary result in being slower than those that appear in Figure 3.9c and that are associated to a Dirichlet outflow boundary. This comment is in contrast to the remark we made concerning the exchange operator according to which the presence of a Robin boundary should lead to faster fixed-point iteration, however other factors may have an impact on the speed of convergence, e.g. the behavior of the exact solution itself along this portion of the boundary.

CGNR is always the scheme that requires the most iterations to converge, whereas GMRES and the fixed-point iteration are often comparable ($n = 1$) or GMRES performs slightly better ($n = 10$).

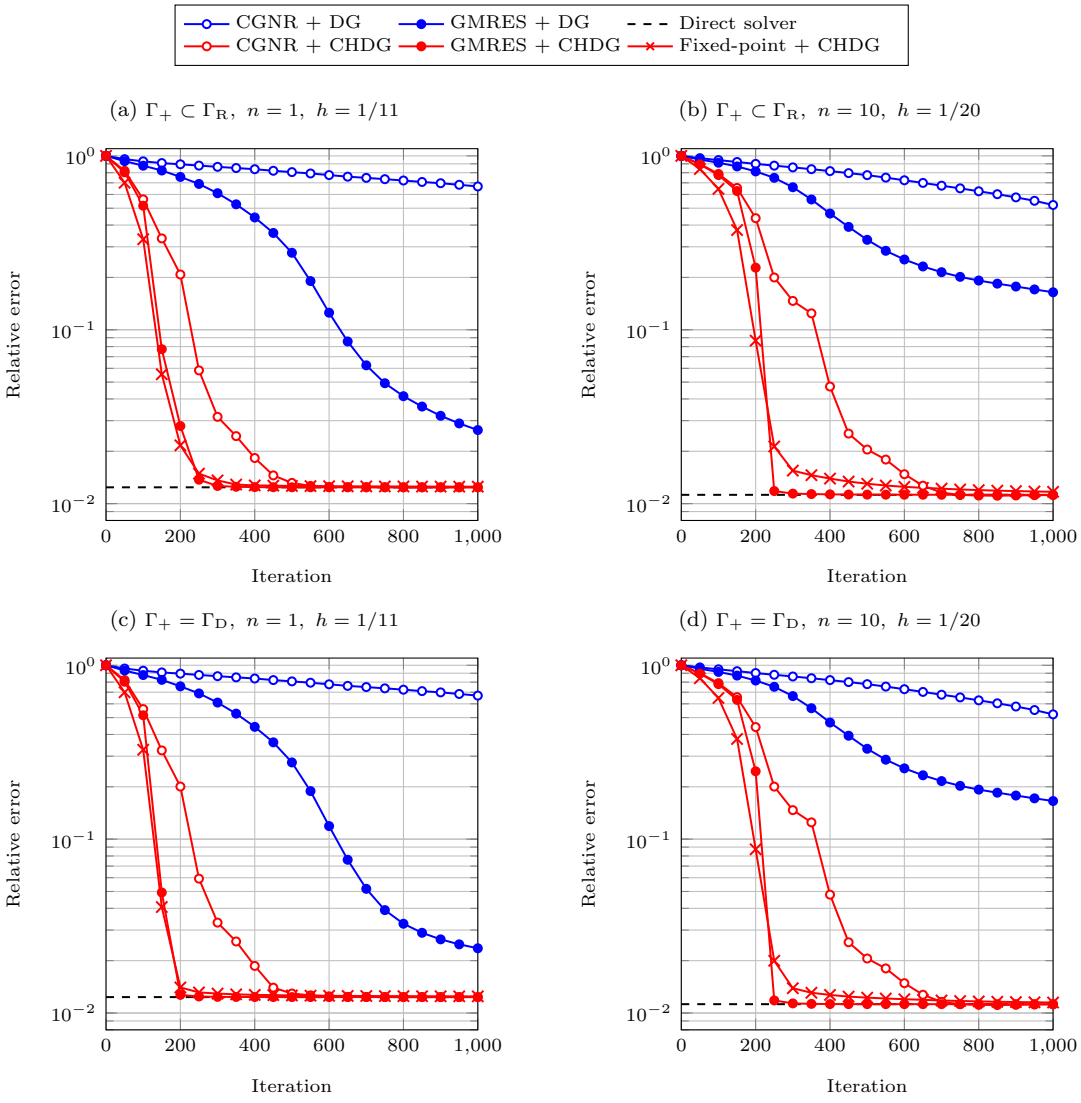


Figure 3.9: Error history with different iterative schemes and different DG methods for the vorticity wave benchmark. The dashed horizontal lines correspond to the relative numerical errors obtained with a direct solver.

3.5 Summary

This Chapter was devoted to the application of the CHDG method to aeroacoustics and, in particular, to the linearised Euler equations which model the propagation of acoustic and hydrodynamic waves in a homogeneous medium that moves at a constant subsonic velocity. We gave explicit definitions of the upwind numerical fluxes and of the transmission variables. The problems are defined with adequate boundary conditions: the number of boundary conditions is determined by the notion of outflow/inflow boundary, and choosing Robin boundary conditions on the outflow boundary guarantees that the operator ΠS is a contraction.

We have proposed three benchmark test-cases to illustrate the performance of the CHDG method comparing it to the performance of the standard DG method. To do so, we considered several iterative schemes in the context of 2D problems, probing different physical aspects and configurations to understand their impact on the numerical method.

- The fixed-point iteration applied to the CHDG system always converges in accordance with the theoretical results. Its speed of convergence in terms of number of iterations significantly depends on the considered benchmark and, in particular, on the physical parameters. The iteration is generally fast and it typically slows down in the case of high-frequency cases, which are well-known to be more difficult to deal with than low-frequency ones.
- The CGNR and GMRES iterations are always faster, in the sense that they require less iterations to achieve a given accuracy, with CHDG than with the standard DG method. The improvement in the performances is always remarkable and, even in the cases where the iterative schemes applied to the DG system turn out to be unreliable because of an extremely low speed of convergence, these very same schemes exhibit a way faster convergence once applied to the CHDG system.
- With some variability related to the benchmark, GMRES is always more or less faster than the fixed-point iteration, but demands a higher overall computational cost in terms of time and memory. Since the code is not fully optimized, this observation does not allow us to retrieve any conclusion regarding the most effective choice of the iterative scheme that should be used in practice, but a more focused study is worth being pursued.
- In light of the numerical results of the third benchmark, the CHDG method turns out to be effective also in the presence of vorticity waves. Iterative solutions are always faster for the CHDG system than for the DG one, even in the case of highly oscillatory behavior of the solution associated to big modes.

3.A Appendix: Analytical solutions of benchmarks

In this section, we retrieve the analytical expression of the solutions of the benchmarks presented in Section 3.4.

3.A.1 Propagative plane wave

We consider Equation (1.13) and we assume that the exact pressure field p is a plane wave with unit amplitude and wavenumber $\tilde{\kappa}$ that propagates in the direction $\boldsymbol{\varphi} = (\cos\varphi, \sin\varphi)$, with $\varphi \in [0, 2\pi)$

$$p(x, y) = \exp(\imath\tilde{\kappa}\boldsymbol{x} \cdot \boldsymbol{\varphi}) = \exp(\imath\tilde{\kappa}(x\cos\varphi + y\sin\varphi)). \tag{3.2}$$

To compute the wavenumber $\tilde{\kappa}$ with respect to M, $\boldsymbol{\theta}$ and $\boldsymbol{\phi}$, we consider Equation (1.13), which can be rewritten as

$$-\Delta p - \kappa^2 p - 2\imath\kappa M\Big(\frac{\partial p}{\partial x}\cos\theta + \frac{\partial p}{\partial y}\sin\theta\Big) + M^2\Big(\frac{\partial^2 p}{\partial x^2}\cos^2\theta + 2\frac{\partial^2 p}{\partial x \partial y}\sin\theta\cos\theta + \frac{\partial^2 p}{\partial y^2}\Big) = 0,$$

and we plug the expression of the exact pressure field given by Equation (3.2), so that we have

$$(1 - M^2(\boldsymbol{\theta}\cdot\boldsymbol{\varphi})^2)\tilde{\kappa}^2 + 2\kappa M(\boldsymbol{\theta}\cdot\boldsymbol{\varphi})\tilde{\kappa} - \kappa^2 = 0. \tag{3.3}$$

The quadratic Equation (3.3) has the following solutions

$$\tilde{\kappa}_{1,2} = \frac{-\kappa M(\boldsymbol{\theta}\cdot\boldsymbol{\varphi}) \pm \kappa}{1 - M^2(\boldsymbol{\theta}\cdot\boldsymbol{\varphi})^2} \implies \tilde{\kappa}_1 = -\frac{\kappa}{1 - M(\boldsymbol{\theta}\cdot\boldsymbol{\varphi})} < 0, \quad \tilde{\kappa}_2 = \frac{\kappa}{1 + M(\boldsymbol{\theta}\cdot\boldsymbol{\varphi})} > 0.$$

and, since the wavenumber must be positive, we have

$$\tilde{\kappa} = \frac{\kappa}{1 + M(\boldsymbol{\theta}\cdot\boldsymbol{\varphi})}.$$

It is easy to prove that the exact velocity field reads

$$\boldsymbol{u}(x, y) = \frac{\boldsymbol{\varphi}}{\rho_0 c_0} p(x, y).$$

To sum up, the exact pressure and velocity fields read

$$p(x, y) = \exp\Big(\frac{\imath\kappa}{1 + M(\boldsymbol{\theta}\cdot\boldsymbol{\varphi})}(x\cos\varphi + y\sin\varphi)\Big),$$
$$\boldsymbol{u}(x, y) = \left(\frac{\cos\varphi}{\rho_0 c_0}\exp\left(\frac{\imath\kappa(x\cos\varphi + y\sin\varphi)}{1 + M(\boldsymbol{\theta}\cdot\boldsymbol{\varphi})}\right), \frac{\sin\varphi}{\rho_0 c_0}\exp\left(\frac{\imath\kappa(x\cos\varphi + y\sin\varphi)}{1 + M(\boldsymbol{\theta}\cdot\boldsymbol{\varphi})}\right)\right).$$

3.A.2 Acoustic point source

We consider the first two equations of Problem 3.1.2 with the first equation multiplied by $\rho_0 c_0^2$ and $f(x,y) = \delta(x-x_s)\delta(y-y_s)$

$$-\imath\omega p + (\boldsymbol{u}_0 \cdot \boldsymbol{\nabla})p + \rho_0 c_0^2 \boldsymbol{\nabla}\cdot\boldsymbol{u} = \rho_0 c_0^2 \delta(x-x_s)\delta(y-y_s), \tag{3.4}$$

$$-i\omega\rho_0 \boldsymbol{u} + \boldsymbol{\nabla} p + \rho_0(\boldsymbol{u}_0\cdot\boldsymbol{\nabla})\boldsymbol{u} = \boldsymbol{0}. \tag{3.5}$$

We take the material derivative in the frequency domain $D_0/Dt = -\imath\omega + \boldsymbol{u}_0\cdot\boldsymbol{\nabla}$ of Equation (3.4) and the divergence of Equation (3.5)

$$\frac{D_0^2 p}{Dt^2} + c_0^2 \frac{D_0}{Dt}(\rho_0 \boldsymbol{\nabla}\cdot\boldsymbol{u}) = \rho_0 c_0^2 \frac{D_0}{Dt}\delta(x-x_s)\delta(y-y_s),$$
$$\boldsymbol{\nabla}\cdot\left(\rho_0 \frac{D_0\boldsymbol{u}}{Dt}\right) + \Delta p = 0.$$

By plugging the second equation into the first one, we eliminate the velocity field $\boldsymbol{u}$ and we get the following wave equation for the pressure field p

$$\frac{D_0^2 p}{Dt^2} - c_0^2\Delta p = \rho_0 c_0^2 \frac{D_0}{Dt}\delta(x-x_s)\delta(y-y_s). \tag{3.6}$$

The associated equation for the fundamental solution (Green's function) $\mathcal{G}$ is

$$\frac{D_0^2\mathcal{G}}{Dt^2} - c_0^2\Delta\mathcal{G} = \delta(x)\delta(y),$$

whose solution (see [6] for further details regarding the computations) reads

$$\mathcal{G}(x,y) = \frac{\imath}{4}\frac{1}{c_0^2\sqrt{1-M^2}} H_0^{(1)}\left[\kappa\frac{\sqrt{x^2+(1-M^2)y^2}}{1-M^2}\right]\exp\Big(-\imath\frac{M}{1-M^2}\kappa x\Big).$$

The solution $\tilde{\mathcal{G}}$ of the same equation with the shifted right-hand side $\delta(x-x_s)\delta(y-y_s)$, that is

$$\frac{D_0^2\tilde{\mathcal{G}}}{Dt^2} - c_0^2\Delta\tilde{\mathcal{G}} = \delta(x-x_s)\delta(y-y_s), \tag{3.7}$$

can be computed thanks to a spatial convolution product: $\tilde{\mathcal{G}}(x,y) = \delta(x-x_s)(y-y_s) * \mathcal{G}(x,y)$, which gives $\tilde{\mathcal{G}}(x,y) = \mathcal{G}(x-x_s, y-y_s)$.

Exploiting the latter result and comparing Equations (3.6) and (3.7), we have that

$$p(x,y) = \rho_0 c_0^2 \frac{D_0}{Dt}\mathcal{G}(x-x_s, y-y_s) \tag{3.8}$$

We introduce the velocity potential ϕ which is such that

$$\boldsymbol{u} = \boldsymbol{\nabla}\phi. \tag{3.9}$$

Plugging Equation (3.9) in Equation (3.5) and comparing the result with Equation (3.8), we have

$$p(x,y) = -\rho_0\frac{D_0\phi}{Dt}(x,y), \quad \phi(x,y) = -c_0^2\mathcal{G}(x-x_s, y-y_s).$$

To sum up, the exact pressure and velocity fields read

$$p(x,y) = \rho_0 c_0^2(-\imath\omega + \boldsymbol{u}_0\cdot\boldsymbol{\nabla})\mathcal{G}(x-x_s, y-y_s),$$
$$\boldsymbol{u}(x,y) = -c_0^2\boldsymbol{\nabla}\mathcal{G}(x-x_s, y-y_s).$$

3.A.3 Vorticity wave

We consider the first two equations of Problem 3.1.2 with $f \equiv 0$ and where $\Omega = \mathbb{R} \times (0, H)$ is a rigid duct whose width is $H > 0$. We denote by $\Gamma_{\text{wall}} = \mathbb{R} \times \{0, H\}$ its rigid walls and we impose a homogeneous Neumann boundary condition on them, namely

$$\boldsymbol{n} \cdot \boldsymbol{u} = 0, \quad \text{on } \Gamma_{\text{wall}},$$

and, being Γ_{wall} parallel to the x-axis, the latter is equivalent to

$$u_y = 0, \quad \text{on } \Gamma_{\text{wall}}. \tag{3.10}$$

We assume a null pressure field, namely $p \equiv 0$ in Ω, and therefore the previous equations reduce respectively to

$$\begin{aligned} \boldsymbol{\nabla} \cdot \boldsymbol{u} &= 0, \\ (-\imath\omega + \boldsymbol{u}_0 \cdot \boldsymbol{\nabla})\boldsymbol{u} &= \boldsymbol{0}. \end{aligned}$$

In the case of a horizontal background flow $\boldsymbol{u}_0$, i.e. $\boldsymbol{u}_0 = |\boldsymbol{u}_0|\boldsymbol{e}_x = Mc_0\boldsymbol{e}_x$, they read

$$\frac{\partial u_x}{\partial x} + \frac{\partial u_y}{\partial y} = 0, \tag{3.11}$$

$$\left(-\imath\omega + Mc_0 \frac{\partial}{\partial x}\right) u_x = 0, \tag{3.12}$$

$$\left(-\imath\omega + Mc_0 \frac{\partial}{\partial x}\right) u_y = 0. \tag{3.13}$$

Integrating Equation (3.13) from 0 to x, we can compute the y-component of the velocity field $\boldsymbol{u}$

$$u_y(x, y) = u_y(0, y) \exp\left(\frac{\imath\omega}{Mc_0} x\right).$$

By combining Equations (3.11)-(3.12), we can compute the x-component of the velocity field $\boldsymbol{u}$

$$\begin{aligned} u_x(x, y) &= -\frac{\imath Mc_0}{\omega} \frac{\partial u_x}{\partial x}(x, y) \\ &= \frac{\imath Mc_0}{\omega} \frac{\partial u_y}{\partial y}(x, y) \\ &= \frac{\imath Mc_0}{\omega} \frac{du_y}{dy}(0, y) \exp\left(\frac{\imath\omega}{Mc_0} x\right). \end{aligned}$$

Because of the homogeneous Neumann boundary condition on the rigid walls (3.10), we have that

$$u_y(0, 0) = u_y(0, H) = 0. \tag{3.14}$$

An example of function $u_y = u_y(y)$ that satisfies the constraint (3.14) is

$$u_y(y) = A \sin\left(n\pi \frac{y}{H}\right),$$

with $A \in \mathbb{R}^+$ and $n \in \mathbb{N}$. We will name *n-th mode* the solution associated to $n \in \mathbb{N}$.

To sum up, the exact pressure and velocity fields read

$$\begin{aligned} &p(x, y) = 0, \\ &\boldsymbol{u}(x, y) = \left(A \frac{\imath n\pi Mc_0}{\omega H} \cos\left(n\pi \frac{y}{H}\right) \exp\left(\frac{\imath\omega}{Mc_0} x\right), A \sin\left(n\pi \frac{y}{H}\right) \exp\left(\frac{\imath\omega}{Mc_0} x\right)\right). \end{aligned}$$

Chapter 4

CHDG method for time-harmonic acoustics in heterogeneous media

This Chapter is based on the paper [107]

S. Pescuma, G. Gabard, T. Chaumont-Frelet, and A. Modave
A hybridizable discontinuous Galerkin method with transmission variables for time-harmonic wave problems in heterogeneous media
Journal of Computational Physics, 534:114009, 2025

In this Chapter, we investigate extensions of the CHDG approach to scalar time-harmonic problems in heterogeneous media with material coefficients that are (possibly) discontinuous at the interfaces between the elements. Compared to the more classical problems of acoustics in homogeneous media, additional difficulties arise due to the heterogeneous setting: standard upwind numerical fluxes are no longer symmetric and, as a consequence, we are not able to prove the strict contraction property of the operator ΠS.

In Section 4.1 we introduce a model problem set in terms of the first-order formulation of the Helmholtz equation. In Section 4.2 we write its DG formulation, we define appropriate upwind numerical fluxes based on a Riemann solver and a general family of symmetric fluxes that rely on differential operators defined on the interfaces between elements. Besides the classic zeroth-order transmission variables, these operators can represent high-order transmission conditions. In Section 4.3 we present the standard HDG method and in Section 4.4 the CHDG method where, once again, a transmission variable, whose definition depends on the physical parameters of the problem and on the differential operator that appears in the numerical fluxes, are used as hybrid variable. In Section 4.5 we study and analyze the numerical results obtained with simple benchmarks.

4.1 Problem

Let $\Omega \subset \mathbb{R}^d$, with $d = 2$ or 3, be a Lipschitz polytopal domain. The boundary $\partial\Omega$ of the domain is partitioned into three non-overlapping polytopal Lipschitz subsets Γ_{D}, Γ_{N} and Γ_{R}, corresponding to the Dirichlet, Neumann and Robin boundaries, respectively. We consider the following time-harmonic wave propagation problem:

$$\begin{cases} -\imath\kappa\eta^{-1}p + \boldsymbol{\nabla}\cdot\boldsymbol{u} = 0, & \text{in } \Omega, \\ -\imath\kappa\eta\boldsymbol{u} + \boldsymbol{\nabla}p = \boldsymbol{0}, & \text{in } \Omega, \\ p = s_{\mathrm{D}}, & \text{on } \Gamma_{\mathrm{D}}, \\ \boldsymbol{n}\cdot\boldsymbol{u} = s_{\mathrm{N}}, & \text{on } \Gamma_{\mathrm{N}}, \\ p - \eta\boldsymbol{n}\cdot\boldsymbol{u} = s_{\mathrm{R}}, & \text{on } \Gamma_{\mathrm{R}}, \end{cases} \tag{4.1}$$

written for the unknown scalar field $p : \Omega \to \mathbb{C}$ and vector field $\boldsymbol{u} : \Omega \to \mathbb{C}^d$ corresponding to the acoustic pressure and velocity. On the boundary, we have introduced the source terms $s_{\mathrm{D}} : \Gamma_{\mathrm{D}} \to \mathbb{C}$, $s_{\mathrm{N}} : \Gamma_{\mathrm{N}} \to \mathbb{C}$ and $s_{\mathrm{R}} : \Gamma_{\mathrm{R}} \to \mathbb{C}$. The propagation medium is characterised by its density $\rho : \Omega \to \mathbb{R}$ and sound speed $c : \Omega \to \mathbb{R}$. These two quantities are strictly positive and assumed to be piecewise constant. Problem (4.1) is written in the frequency domain, assuming a time dependence $e^{-\imath\omega t}$ for the solution, where ω is the real-valued angular frequency. We also introduced the wavenumber $\kappa : \Omega \to \mathbb{R}$ and the acoustic impedance $\eta : \Omega \to \mathbb{R}$ such that $\kappa = \omega/c$ and $\eta = \rho c$. The vector field $\boldsymbol{n} : \partial\Omega \to \mathbb{R}^d$ is the unit outward normal to Ω. For brevity, we do not consider source terms on the right-hand sides of the first two Equations in (4.1), but these could be included without difficulty.

4.2 Discontinuous Galerkin methods

To solve Problem (4.1), as in Chapters 2 and 3, we use a conforming mesh $\mathcal{T}_h$ of the domain Ω consisting of simplicial elements K. We adopt the same notation for the mesh and the inner products (see Section 2.2.1). Lastly, we consider the approximation spaces $\mathcal{V}_h = \boldsymbol{\mathcal{V}}_h^1$ for scalar fields and $\boldsymbol{\mathcal{V}}_h = \boldsymbol{\mathcal{V}}_h^d$ for vector fields.

The coefficients κ and η are assumed to be constant on each element. Their values on K are denoted by κ_K and η_K, respectively. For a face F of K, $\boldsymbol{n}_{K,F}$ is the unit normal on F pointing outside K.

4.2.1 Variational formulation of the problem

The general discontinuous Galerkin (DG) formulation of system (4.1) reads:

Problem 4.2.1. *Find* $(p_h, \boldsymbol{u}_h) \in \mathcal{V}_h \times \boldsymbol{\mathcal{V}}_h$ *such that, for all* $(q_h, \boldsymbol{v}_h) \in \mathcal{V}_h \times \boldsymbol{\mathcal{V}}_h$,

$$\begin{cases} -\imath\big(\kappa\eta^{-1}p_h, q_h\big)_{\mathcal{T}_h} - \big(\boldsymbol{u}_h, \boldsymbol{\nabla}q_h\big)_{\mathcal{T}_h} + \big\langle \boldsymbol{n}\cdot\widehat{\boldsymbol{u}}, q_h\big\rangle_{\partial\mathcal{T}_h} = 0, \\ -\imath\big(\kappa\eta\boldsymbol{u}_h, \boldsymbol{v}_h\big)_{\mathcal{T}_h} - \big(p_h, \boldsymbol{\nabla}\cdot\boldsymbol{v}_h\big)_{\mathcal{T}_h} + \big\langle \widehat{p}, \boldsymbol{n}\cdot\boldsymbol{v}_h\big\rangle_{\partial\mathcal{T}_h} = 0, \end{cases} \tag{4.2}$$

with the numerical fluxes $\widehat{p}$ *and* $\boldsymbol{n}\cdot\widehat{\boldsymbol{u}}$ *to be defined.*

The numerical fluxes are defined face by face. At each interior face $F \not\subset \partial\Omega$, they depend on the values of the physical fields and coefficients associated to both neighboring elements. At each boundary face $F \subset \partial\Omega$, a specific definition is used to prescribe a boundary condition and, consequently, the numerical fluxes depend on the boundary data. Hereafter, two kinds of numerical fluxes are considered: standard upwind fluxes and general symmetric fluxes. Upwind fluxes are introduced because they are widely used in literature. However, in the framework of the CHDG method, we could not prove the main results regarding the property of contraction of the operator $\Pi\mathrm{S}$, hence the need of general symmetric fluxes for which the results of Section 4.4.4 hold.

Moreover, the use of high-order operators in transmission conditions has already been successfully experimented in the framework of DDM (see Section 4.2.3) which makes the comparison of these two fluxes interesting.

If F is a boundary face, we define $\eta_{K'} := \eta_K$ to simplify the presentation, even if there is no neighboring element K'.

4.2.2 Upwind numerical fluxes

Standard upwind numerical fluxes are obtained by solving a Riemann problem associated to each interface between two elements, see e.g. [91, 126]. For the wave propagation problem (4.1) with discontinuous coefficients between elements, the derivation is given in Section 9.9 of [91]. For a given face F, the *upwind numerical fluxes* are defined as

$$\begin{cases} \widehat{p}_F := \dfrac{1}{\eta_K + \eta_{K'}} \left(\eta_{K'} g^{\oplus}_{K,F} + \eta_K g^{\ominus}_{K,F} \right), \\ \boldsymbol{n}_{K,F} \cdot \widehat{\boldsymbol{u}}_F := \dfrac{1}{\eta_K + \eta_{K'}} \left(g^{\oplus}_{K,F} - g^{\ominus}_{K,F} \right), \end{cases} \tag{4.3}$$

where $g^{\oplus}_{K,F}$ and $g^{\ominus}_{K,F}$ are the outgoing and incoming transmission variables, respectively. In the standard terminology of Riemann solvers, the transmission variables are generally called characteristic variables, see e.g. [69, 123]. The *outgoing transmission variable* depends on the physical variables and coefficients of the considered element K. It is defined as

$$g^{\oplus}_{K,F} := p_K + \eta_K \boldsymbol{n}_{K,F} \cdot \boldsymbol{u}_K. \tag{4.4}$$

The *incoming transmission variable* depends on whether there is a neighboring element K' or not. In the latter case, it depends on the boundary condition. It is defined as

$$g^{\ominus}_{K,F} := \begin{cases} p_{K'} - \eta_{K'} \boldsymbol{n}_{K,F} \cdot \boldsymbol{u}_{K'} = g^{\oplus}_{K',F}, & \text{if } F \notin \partial\Omega, \\ 2 s_{\mathrm{D}} - g^{\oplus}_{K,F}, & \text{if } F \subset \Gamma_{\mathrm{D}}, \\ g^{\oplus}_{K,F} - 2\eta_K s_{\mathrm{N}}, & \text{if } F \subset \Gamma_{\mathrm{N}}, \\ s_{\mathrm{R}}, & \text{if } F \subset \Gamma_{\mathrm{R}}. \end{cases} \tag{4.5}$$

The upwind numerical fluxes can be written more explicitly as

$$\begin{cases} \widehat{p}_F = \dfrac{\eta_K \eta_{K'}}{\eta_K + \eta_{K'}} \left(\dfrac{1}{\eta_K} p_K + \dfrac{1}{\eta_{K'}} p_{K'} + \boldsymbol{n}_{K,F} \cdot (\boldsymbol{u}_K - \boldsymbol{u}_{K'}) \right), \\ \boldsymbol{n}_{K,F} \cdot \widehat{\boldsymbol{u}}_F = \dfrac{1}{\eta_K + \eta_{K'}} \left(\boldsymbol{n}_{K,F} \cdot (\eta_K \boldsymbol{u}_K + \eta_{K'} \boldsymbol{u}_{K'}) + p_K - p_{K'} \right), \end{cases} \quad \text{if } F \notin \partial\Omega,$$

$$\begin{cases} \widehat{p}_F = s_{\mathrm{D}}, \\ \boldsymbol{n}_{K,F} \cdot \widehat{\boldsymbol{u}}_F = \boldsymbol{n}_{K,F} \cdot \boldsymbol{u}_K + \dfrac{1}{\eta_K} \left(p_K - s_{\mathrm{D}} \right), \end{cases} \quad \text{if } F \subset \Gamma_{\mathrm{D}},$$

$$\begin{cases} \widehat{p}_F = p_K + \eta_K \left(\boldsymbol{n}_{K,F} \cdot \boldsymbol{u}_K - s_{\mathrm{N}} \right), \\ \boldsymbol{n}_{K,F} \cdot \widehat{\boldsymbol{u}}_F = s_{\mathrm{N}}, \end{cases} \quad \text{if } F \subset \Gamma_{\mathrm{N}},$$

$$\begin{cases} \widehat{p}_F = \dfrac{1}{2} \left(p_K + \eta_K \boldsymbol{n}_{K,F} \cdot \boldsymbol{u}_K + s_{\mathrm{R}} \right), \\ \boldsymbol{n}_{K,F} \cdot \widehat{\boldsymbol{u}}_F = \dfrac{1}{2\eta_K} \left(p_K + \eta_K \boldsymbol{n}_{K,F} \cdot \boldsymbol{u}_K - s_{\mathrm{R}} \right), \end{cases} \quad \text{if } F \subset \Gamma_{\mathrm{R}}.$$

These fluxes satisfy the relation

$$\widehat{p}_F + \eta_K \boldsymbol{n}_{K,F} \cdot \widehat{\boldsymbol{u}}_F = p_K + \eta_K \boldsymbol{n}_{K,F} \cdot \boldsymbol{u}_K, \tag{4.6}$$

which will be useful to derive the standard HDG formulation.

4.2.3 Symmetric numerical fluxes

We now consider a general family of numerical fluxes where the coefficients do not depend on the elements, but only on the faces. For each face F, we introduce a general operator $\mathcal{A}_F$ that verifies the following assumption where $\mathcal{P}_{\mathrm{p}}(F)$ denotes the space of polynomials of degree at most p defined on F.

Assumption 4.2.2. *For each face F, the linear operator $\mathcal{A}_F : \mathcal{P}_{\mathrm{p}}(F) \to \mathcal{P}_{\mathrm{p}}(F)$ is positive and self-adjoint with respect to the Hermitian inner product.*

The operator $\mathcal{A}_F$ acts between polynomial spaces and, therefore, it is a purely discrete operator. Moreover, Assumption 4.2.2 allows the definition of a scalar product ($\langle \mathcal{A}_F u, v\rangle_F$ for all $u, v \in L^2(F)$) and, consequently, of a norm induced by $\mathcal{A}_F$ ($\|u\|_F := \sqrt{\langle \mathcal{A}_F u, u\rangle_F}$ for all $u \in L^2(F)$), see Section 4.4.4. Lastly, Assumption 4.2.2is necessary to prove some of the results presented in Section 4.4.

For a given face F, the *symmetric numerical fluxes* are defined as

$$\begin{cases} \widehat{p}_F := \dfrac{1}{2}\mathcal{A}_F\left(g^{\oplus}_{K,F} + g^{\ominus}_{K,F}\right), \\ \boldsymbol{n}_{K,F}\cdot\widehat{\boldsymbol{u}}_F := \dfrac{1}{2}\left(g^{\oplus}_{K,F} - g^{\ominus}_{K,F}\right), \end{cases} \tag{4.7}$$

where the *outgoing transmission variable* is defined as

$$g^{\oplus}_{K,F} := \mathcal{A}_F^{-1} p_K + \boldsymbol{n}_{K,F}\cdot\boldsymbol{u}_K, \tag{4.8}$$

and the *incoming transmission variable* is defined as

$$g^{\ominus}_{K,F} := \begin{cases} \mathcal{A}_F^{-1} p_{K'} - \boldsymbol{n}_{K,F}\cdot\boldsymbol{u}_{K'} = g^{\oplus}_{K',F}, & \text{if } F \notin \partial\Omega, \\ 2\mathcal{A}_F^{-1} s_{\mathrm{D}} - g^{\oplus}_{K,F}, & \text{if } F \subset \Gamma_{\mathrm{D}}, \\ g^{\oplus}_{K,F} - 2s_{\mathrm{N}}, & \text{if } F \subset \Gamma_{\mathrm{N}}, \\ \mathcal{B}^{-1}_{K,F,+}\left(\mathcal{B}_{K,F,-} g^{\oplus}_{K,F} + 2\eta_K^{-1} s_{\mathrm{R}}\right), & \text{if } F \subset \Gamma_{\mathrm{R}}, \end{cases} \tag{4.9}$$

with $\mathcal{B}_{K,F,\pm} := 1 \pm \eta_K^{-1}\mathcal{A}_F$.

The symmetric numerical fluxes can be written more explicitly as

$$\begin{cases} \widehat{p}_F = \dfrac{p_K + p_{K'}}{2} + \mathcal{A}_F\left(\boldsymbol{n}_{K,F}\cdot\dfrac{\boldsymbol{u}_K - \boldsymbol{u}_{K'}}{2}\right), \\ \boldsymbol{n}_{K,F}\cdot\widehat{\boldsymbol{u}}_F = \boldsymbol{n}_{K,F}\cdot\dfrac{\boldsymbol{u}_K + \boldsymbol{u}_{K'}}{2} + \mathcal{A}_F^{-1}\left(\dfrac{p_K - p_{K'}}{2}\right), \end{cases} \quad \text{if } F \notin \partial\Omega,$$

$$\begin{cases} \widehat{p}_F = s_{\mathrm{D}}, \\ \boldsymbol{n}_{K,F}\cdot\widehat{\boldsymbol{u}}_F = \boldsymbol{n}_{K,F}\cdot\boldsymbol{u}_K + \mathcal{A}_F^{-1}\left(p_K - s_{\mathrm{D}}\right), \end{cases} \quad \text{if } F \subset \Gamma_{\mathrm{D}},$$

$$\begin{cases} \widehat{p}_F = p_K + \mathcal{A}_F\left(\boldsymbol{n}_{K,F}\cdot\boldsymbol{u}_K - s_{\mathrm{N}}\right), \\ \boldsymbol{n}_{K,F}\cdot\widehat{\boldsymbol{u}}_F = s_{\mathrm{N}}, \end{cases} \quad \text{if } F \subset \Gamma_{\mathrm{N}},$$

$$\begin{cases} \widehat{p}_F = \mathcal{B}^{-1}_{K,F,+}\left(p_K + \mathcal{A}_F\left(\boldsymbol{n}_{K,F}\cdot\boldsymbol{u}_K\right) + \eta_K^{-1}\mathcal{A}_F s_{\mathrm{R}}\right), \\ \boldsymbol{n}_{K,F}\cdot\widehat{\boldsymbol{u}}_F = \eta_K^{-1}\mathcal{B}^{-1}_{K,F,+}\left(p_K + \mathcal{A}_F\left(\boldsymbol{n}_{K,F}\cdot\boldsymbol{u}_K\right) - s_{\mathrm{R}}\right), \end{cases} \quad \text{if } F \subset \Gamma_{\mathrm{R}}.$$

For these symmetric fluxes as for the upwind fluxes, the following relation holds

$$\widehat{p}_F + \mathcal{A}_F\left(\boldsymbol{n}_{K,F}\cdot\widehat{\boldsymbol{u}}_F\right) = p_K + \mathcal{A}_F\left(\boldsymbol{n}_{K,F}\cdot\boldsymbol{u}_K\right). \tag{4.10}$$

It will be used in the derivation of the standard HDG formulation where the hybrid variable corresponds to the numerical trace $\widehat{p}_F$ of the pressure field.

Zeroth-order symmetric fluxes

If $\mathcal{A}_F = \mu_F$ is a strictly positive real parameter, the numerical fluxes simply read

$$\begin{cases} \widehat{p}_F := \dfrac{1}{2}\mu_F(g^{\oplus}_{K,F} + g^{\ominus}_{K,F}), \\ \boldsymbol{n}_{K,F} \cdot \widehat{\boldsymbol{u}}_F := \dfrac{1}{2}(g^{\oplus}_{K,F} - g^{\ominus}_{K,F}), \end{cases}$$

with

$$g^{\oplus}_{K,F} := \frac{1}{\mu_F} p_K + \boldsymbol{n}_{K,F} \cdot \boldsymbol{u}_K,$$

$$g^{\ominus}_{K,F} := \begin{cases} \dfrac{1}{\mu_F} p_{K'} - \boldsymbol{n}_{K,F} \cdot \boldsymbol{u}_{K'}, & \text{if } F \not\subset \partial\Omega, \\ -g^{\oplus}_{K,F} + \dfrac{2s_{\mathrm{D}}}{\mu_F}, & \text{if } F \subset \Gamma_{\mathrm{D}}, \\ g^{\oplus}_{K,F} - 2s_{\mathrm{N}}, & \text{if } F \subset \Gamma_{\mathrm{N}}, \\ \dfrac{\eta_K - \mu_F}{\eta_K + \mu_F} g^{\oplus}_{K,F} + \dfrac{2}{\eta_K + \mu_F} s_{\mathrm{R}}, & \text{if } F \subset \Gamma_{\mathrm{R}}. \end{cases}$$

Note that these fluxes are equivalent to the standard upwind fluxes if the impedance at the interfaces between elements is continuous, and μ_F is chosen as $\eta_K = \eta_{K'}$ for neighboring elements K and K', regardless of whether the wavenumber is continuous or not. We also point out that these transmission variables correspond to the lowest-order absorbing boundary conditions approximating Dirichlet-to-Neumann (DtN) maps.

Second-order symmetric fluxes

By analogy with non-overlapping domain decomposition methods (e.g. [98]), the operator $\mathcal{A}_F$ can be defined by using transmission operators based on more involved domain truncation techniques, such as high-order absorbing boundary conditions. We consider here an operator with second-order partial derivatives used in a second-order absorbing boundary condition in [42]. This operator is defined by

$$\langle \mathcal{A}_F \phi_F, v_F \rangle_F + \frac{1}{2\kappa_F^2} \langle \boldsymbol{\nabla}_F(\mathcal{A}_F \phi_F), \boldsymbol{\nabla}_F v_F \rangle_F = \mu_F \langle \phi_F, v_F \rangle_F, \quad \forall \phi_F, v_F \in \mathcal{P}_{\mathrm{p}}(F) \tag{4.11}$$

where the real parameter μ_F is strictly positive and $\boldsymbol{\nabla}_F$ is the gradient operator on F. The operator $\mathcal{A}_F$ is then a discrete realization of

$$\mathcal{A}_F := \mu_F \left(1 - \frac{1}{2\kappa_F^2} \Delta_F\right)^{-1},$$

where Δ_F is the Laplace–Beltrami operator on F, completed by homogeneous Neumann boundary conditions at the extremities of F. It can now be proved to satisfy Assumption 4.2.2.

Proposition 4.2.3 (Positivity and self-adjointness of the second-order operator)**.** *The operator $\mathcal{A}_F$ (4.11) with μ_F positive and piece-wise constant defined on an edge F is positive and self-adjoint.*

Proof. Because we are in a finite dimensional framework, it is sufficient to prove that $\mathcal{A}_F^{-1}$ is positive and self-adjoint. First, we prove that $\mathcal{A}_F^{-1}$ is positive. To do so, given $\phi_F \in \mathcal{P}_{\mathrm{p}}(F)$, we let $u_F := \mathcal{A}_F \phi_F$, and pick $v_F = u_F$ in (4.11). This gives

$$0 \leq \langle u_F, u_F \rangle_F + \frac{1}{2\kappa_F^2} \langle \boldsymbol{\nabla}_F u_F, \boldsymbol{\nabla}_F u_F \rangle_F = \mu_F \langle \mathcal{A}_F^{-1} u_F, u_F \rangle_F,$$

which is the desired inequality, as $\mu_F > 0$. Then, we prove that $\mathcal{A}_F^{-1}$ is self-adjoint, which follows from the fact that the left-hand side in (4.11) is an Hermitian form. Indeed,

$$\begin{aligned}\langle \mathcal{A}_F^{-1} u_F, v_F\rangle_F &= \frac{1}{\mu_F}\langle u_F, v_F\rangle_F + \frac{1}{2\kappa_F^2\mu_F}\langle \boldsymbol{\nabla}_F u_F, \boldsymbol{\nabla}_F v_F\rangle_F,\\ &= \overline{\frac{1}{\mu_F}\langle v_F, u_F\rangle_F + \frac{1}{2\kappa_F^2\mu_F}\langle \boldsymbol{\nabla}_F v_F, \boldsymbol{\nabla}_F u_F\rangle_F},\\ &= \overline{\langle \mathcal{A}_F^{-1} v_F, u_F\rangle_F} = \langle u_F, \mathcal{A}_F^{-1} v_F\rangle_F\end{aligned}$$

for all $u_F, v_F \in \mathcal{P}_{\mathrm{p}}(F)$. □

4.3 Hybridization with numerical trace

The goal of this Section is to introduce the standard hybridization technique with numerical trace. To do so, we first introduce a polynomial space for the hybrid variable, we write the HDG formulation of the problem, we define the local problem that needs to be solved to eliminate the physical fields and we prove that it is well-posed.

4.3.1 Standard HDG formulations

In the standard HDG methods, the hybrid variable, denoted $\widehat{p}_h$, corresponds to the numerical flux $\widehat{p}$, frequently called the *numerical trace* in the literature, see e.g. [62]. It belongs to the space of polynomials of order p on each face F of the mesh, denoted by

$$\widehat{\mathcal{V}}_h := \prod_{F\in\mathcal{F}_h} \mathcal{P}_{\mathrm{p}}(F).$$

For any field $\widehat{p}_h \in \widehat{\mathcal{V}}_h$, there is one set of scalar unknowns associated to each face of the mesh.

For the case with symmetric numerical fluxes, the variable $\widehat{p}_h$ is defined by Equation (4.7). Using Equation (4.10) in Problem 4.2.1 then leads to the following HDG formulation.

Problem 4.3.1 (HDG formulation with symmetric fluxes)**.** *Find* $(p_h, \boldsymbol{u}_h, \widehat{p}_h) \in \mathcal{V}_h \times \boldsymbol{\mathcal{V}}_h \times \widehat{\mathcal{V}}_h$ *such that, for all* $(q_h, \boldsymbol{v}_h, \widehat{q}_h) \in \mathcal{V}_h \times \boldsymbol{\mathcal{V}}_h \times \widehat{\mathcal{V}}_h$,

$$\begin{cases} -\imath\big(\kappa\eta^{-1}p_h, q_h\big)_{\mathcal{T}_h} - \big(\boldsymbol{u}_h, \boldsymbol{\nabla} q_h\big)_{\mathcal{T}_h} + \big\langle \boldsymbol{n}\cdot\boldsymbol{u}_h + \mathcal{A}^{-1}(p_h - \widehat{p}_h), q_h\big\rangle_{\partial\mathcal{T}_h} = 0,\\ -\imath\big(\kappa\eta\boldsymbol{u}_h, \boldsymbol{v}_h\big)_{\mathcal{T}_h} - \big(p_h, \boldsymbol{\nabla}\cdot\boldsymbol{v}_h\big)_{\mathcal{T}_h} + \big\langle \widehat{p}_h, \boldsymbol{n}\cdot\boldsymbol{v}_h\big\rangle_{\partial\mathcal{T}_h} = 0,\end{cases}$$

and

$$\begin{aligned}\big\langle \widehat{p}_h, \widehat{q}_h\big\rangle_{\mathcal{F}_h} &- \big\langle \tfrac{1}{2}(p_h + \mathcal{A}(\boldsymbol{n}\cdot\boldsymbol{u}_h)), \widehat{q}_h\big\rangle_{\partial\mathcal{T}_h\setminus\partial\Omega}\\ &- \big\langle p_h + \mathcal{A}(\boldsymbol{n}\cdot\boldsymbol{u}_h), \widehat{q}_h\big\rangle_{\Gamma_{\mathrm{N}}} - \big\langle (1+\eta^{-1}\mathcal{A})^{-1}(p_h + \mathcal{A}(\boldsymbol{n}\cdot\boldsymbol{u}_h)), \widehat{q}_h\big\rangle_{\Gamma_{\mathrm{R}}}\\ &= \big\langle s_{\mathrm{D}}, \widehat{q}_h\big\rangle_{\Gamma_{\mathrm{D}}} - \big\langle \mathcal{A}s_{\mathrm{N}}, \widehat{q}_h\big\rangle_{\Gamma_{\mathrm{N}}} + \big\langle (1+\eta^{-1}\mathcal{A})^{-1}(\eta^{-1}\mathcal{A}s_{\mathrm{R}}), \widehat{q}_h\big\rangle_{\Gamma_{\mathrm{R}}}.\end{aligned}$$

Remark 4.3.2 (Case with upwind fluxes). For the case with the upwind numerical fluxes, the variable $\widehat{p}_h$ is defined by Equation (4.3), and the HDG formulation is obtained similarly by using Equation (4.6) in Problem 4.2.1. We refer to [29, 34] for the complete description of the HDG formulation with upwind fluxes.

4.3.2 Local element-wise discrete problems

In the hybridization procedure, the physical fields p_h and $\boldsymbol{u}_h$ are eliminated by solving local element-wise problems, where the numerical trace $\widehat{p}_h$ is assumed to have been computed beforehand via the global hybridized system.

For each element K, the local problem corresponding to the symmetric fluxes reads as follows.

Problem 4.3.3. *Find* $(p_K, \boldsymbol{u}_K) \in \mathcal{P}_{\mathrm{p}}(K) \times \boldsymbol{\mathcal{P}}_{\mathrm{p}}(K)$ *such that, for all* $(q_K, \boldsymbol{v}_K) \in \mathcal{P}_{\mathrm{p}}(K) \times \boldsymbol{\mathcal{P}}_{\mathrm{p}}(K)$,

$$\begin{cases} -\imath\big(\kappa_K \eta_K^{-1} p_K, q_K\big)_K - \big(\boldsymbol{u}_K, \boldsymbol{\nabla} q_K\big)_K + \displaystyle\sum_{F \in \mathcal{F}_K} \big\langle (\boldsymbol{n}_{K,F} \cdot \boldsymbol{u}_K + \mathcal{A}_F^{-1} p_K), q_K \big\rangle_F \\ \qquad\qquad = \displaystyle\sum_{F \in \mathcal{F}_K} \big\langle \mathcal{A}_F^{-1} \widehat{p}_F, q_K \big\rangle_F, \\ -\imath\big(\kappa_K \eta_K \boldsymbol{u}_K, \boldsymbol{v}_K\big)_K - \big(p_K, \boldsymbol{\nabla} \cdot \boldsymbol{v}_K\big)_K = -\displaystyle\sum_{F \in \mathcal{F}_K} \big\langle \widehat{p}_F, \boldsymbol{n}_{K,F} \cdot \boldsymbol{v}_K \big\rangle_F, \end{cases}$$

for a given surface datum $\widehat{p}_F \in \mathcal{P}_{\mathrm{p}}(F)$ *for all* $F \in \mathcal{F}_K$.

This local problem can be interpreted as a discretized Helmholtz problem defined on K with a non-homogeneous Dirichlet boundary condition on ∂K. The discrete problem is well-posed without any condition, as stated and proved below. It is worth pointing out that at the continuous level, Helmholtz problems set in K with Dirichlet boundary conditions might not be well-posed, due to possible resonance frequencies. Well-posedness at the discrete level follows from the fact that the solutions for the resonances do not belong to $\mathcal{P}_{\mathrm{p}}(K)$ and $\boldsymbol{\mathcal{P}}_{\mathrm{p}}(K)$, as it can be seen at the end of the proof.

Theorem 4.3.4 (Well-posedness of the local discrete problem)**.** *If Assumption 4.2.2 holds, then Problem 4.3.3 is well-posed.*

Proof. We have to prove that, if $\widehat{p}_F = 0$ for all $F \in \mathcal{F}_K$, the unique solution of Problem 4.3.3 is $p_K = 0$ and $\mathbf{u}_K = \mathbf{0}$. For brevity, the subscripts K and F are omitted for the local fields, the test functions, the outgoing unit normal and the coefficients. Taking both equations of Problem 4.3.3 with $q = p$ and $\boldsymbol{v} = \boldsymbol{u}$ gives

$$-\imath(\kappa\eta^{-1} p, p)_K - (\boldsymbol{u}, \boldsymbol{\nabla} p)_K + \langle (\boldsymbol{n} \cdot \boldsymbol{u} + \mathcal{A}^{-1} p), p \rangle_{\partial K} = 0, \tag{4.12}$$

$$-\imath(\kappa\eta \boldsymbol{u}, \boldsymbol{u})_K - (p, \boldsymbol{\nabla} \cdot \boldsymbol{u})_K = 0. \tag{4.13}$$

Integrating by parts in both equations and taking the complex conjugate lead to

$$\imath(\kappa\eta^{-1} p, p)_K + (p, \boldsymbol{\nabla} \cdot \boldsymbol{u})_K + \langle p, \mathcal{A}^{-1} p \rangle_{\partial K} = 0, \tag{4.14}$$

$$\imath(\kappa\eta \boldsymbol{u}, \boldsymbol{u})_K + (\boldsymbol{u}, \boldsymbol{\nabla} p)_K - \langle \boldsymbol{n} \cdot \boldsymbol{u}, p \rangle_{\partial K} = 0. \tag{4.15}$$

Adding the four previous equations yields $\langle \mathcal{A}^{-1} p, p \rangle_{\partial K} = 0$. Combined with Assumption 4.2.2, this implies that $p = 0$ on ∂K. By using this result in Problem 4.3.3, one has

$$\begin{cases} -\imath(\kappa\eta^{-1} p, q)_K + (\boldsymbol{\nabla} \cdot \boldsymbol{u}, q)_K = 0, \\ -\imath(\kappa\eta \boldsymbol{u}, \boldsymbol{v})_K + (\boldsymbol{\nabla} p, \boldsymbol{v})_K = 0, \end{cases} \tag{4.16}$$

for all $(q, \boldsymbol{v}) \in \mathcal{P}_{\mathrm{p}}(K) \times \boldsymbol{\mathcal{P}}_{\mathrm{p}}(K)$. We conclude that

$$-\imath\kappa\eta^{-1} p + \boldsymbol{\nabla} \cdot \boldsymbol{u} = 0, \tag{4.17}$$

$$-\imath\kappa\eta \boldsymbol{u} + \boldsymbol{\nabla} p = 0, \tag{4.18}$$

in a strong sense. Because there is no non-zero polynomial solution to the previous equations, this yields the result. □

Remark 4.3.5 (Case with upwind fluxes). The well-posedness of the local problem in the case of upwind fluxes can be proved in a very similar way by following the same steps, see e.g. [29, 34].

4.4 Hybridization with transmission variables

Following the approach proposed in [97], the hybridization is performed by taking the incoming transmission variable as hybrid variable.

4.4.1 CHDG formulations

An additional variable, denoted $g_h^{\ominus}$, corresponding to the incoming transmission variable is introduced at each face of each element. This variable belongs to the space $\mathcal{G}_h = \mathcal{G}_h^{-,n}$ (see Section 3.3.2). Note that, since for each face F of each element K, exactly one incoming and one outgoing transmission variables are defined, the spaces $\mathcal{G}_h^{-,n}$ and $\mathcal{G}_h^{+,n}$ coincide.

Each field of this space is defined on each face of each element. Therefore, there are two values per interior face of the mesh, and one value per boundary face.

For the case with upwind fluxes, the hybrid variable is defined by Equation (4.5), and the numerical fluxes are defined by Equation (4.7). For the case with symmetric fluxes, Equations (4.9) and (4.3) are used instead. The following formulation is obtained in both cases.

Problem 4.4.1 (CHDG formulation with upwind/symmetric fluxes)**.** *Find* $(p_h, \boldsymbol{u}_h, g_h^{\ominus}) \in \mathcal{V}_h \times \boldsymbol{\mathcal{V}}_h \times \mathcal{G}_h$ *such that, for all* $(q_h, \boldsymbol{v}_h, \xi_h) \in \mathcal{V}_h \times \boldsymbol{\mathcal{V}}_h \times \mathcal{G}_h$*,*

$$\begin{cases} -\imath\big(\kappa\eta^{-1}p_h, q_h\big)_{\mathcal{T}_h} - \big(\boldsymbol{u}_h, \boldsymbol{\nabla} q_h\big)_{\mathcal{T}_h} + \big\langle \boldsymbol{n}\cdot\widehat{\boldsymbol{u}}(g^{\oplus}(p_h,\boldsymbol{u}_h), g_h^{\ominus}), q_h\big\rangle_{\partial\mathcal{T}_h} = 0, \\ -\imath\big(\kappa\eta\boldsymbol{u}_h, \boldsymbol{v}_h\big)_{\mathcal{T}_h} - \big(p_h, \boldsymbol{\nabla}\cdot\boldsymbol{v}_h\big)_{\mathcal{T}_h} + \big\langle \widehat{p}(g^{\oplus}(p_h,\boldsymbol{u}_h), g_h^{\ominus}), \boldsymbol{n}\cdot\boldsymbol{v}_h\big\rangle_{\partial\mathcal{T}_h} = 0, \end{cases}$$

and

$$\langle g_h^{\ominus} - \Pi(g^{\oplus}(p_h,\boldsymbol{u}_h)), \xi_h\rangle_{\partial\mathcal{T}_h} = \langle b, \xi_h\rangle_{\partial\mathcal{T}_h}, \tag{4.19}$$

with $g^{\oplus}(p_h,\boldsymbol{u}_h)|_{K,F} := p_K + \eta_K \boldsymbol{n}_{K,F}\cdot\boldsymbol{u}_K$ *in the upwind case and* $g^{\oplus}(p_h,\boldsymbol{u}_h)|_{K,F} := \mathcal{A}_F^{-1}p_K + \boldsymbol{n}_{K,F}\cdot\boldsymbol{u}_K$ *in the symmetric case.*

To emphasize important properties of the new method, we now introduce the *global exchange operator* $\Pi : \mathcal{G}_h \to \mathcal{G}_h$ and the *global right-hand side* b, whose definitions depend on the choice of numerical fluxes. For each face F of each element K, and for any $g^{\oplus} \in \mathcal{G}_h$, they are defined for the upwind fluxes as

$$\Pi(g^{\oplus})|_{K,F} = \begin{cases} g^{\oplus}_{K',F}, & \text{if } F \not\subset \partial\Omega, \\ -g^{\oplus}_{K,F}, & \text{if } F \subset \Gamma_{\mathrm{D}}, \\ g^{\oplus}_{K,F}, & \text{if } F \subset \Gamma_{\mathrm{N}}, \\ 0, & \text{if } F \subset \Gamma_{\mathrm{R}}, \end{cases} \quad \text{and} \quad b|_{K,F} = \begin{cases} 0, & \text{if } F \not\subset \partial\Omega, \\ 2s_{\mathrm{D}}, & \text{if } F \subset \Gamma_{\mathrm{D}}, \\ -2\eta_K s_{\mathrm{N}}, & \text{if } F \subset \Gamma_{\mathrm{N}}, \\ s_{\mathrm{R}}, & \text{if } F \subset \Gamma_{\mathrm{R}}, \end{cases}$$

and for the symmetric fluxes as

$$\Pi(g^{\oplus})|_{K,F} = \begin{cases} g^{\oplus}_{K',F}, & \text{if } F \not\subset \partial\Omega, \\ -g^{\oplus}_{K,F}, & \text{if } F \subset \Gamma_{\mathrm{D}}, \\ g^{\oplus}_{K,F}, & \text{if } F \subset \Gamma_{\mathrm{N}}, \\ \mathcal{B}^{-1}_{K,F,+}\mathcal{B}_{K,F,-}g^{\oplus}_{K,F}, & \text{if } F \subset \Gamma_{\mathrm{R}}, \end{cases} \quad \text{and} \quad b|_{K,F} = \begin{cases} 0, & \text{if } F \not\subset \partial\Omega, \\ 2\mathcal{A}_F^{-1}s_{\mathrm{D}}, & \text{if } F \subset \Gamma_{\mathrm{D}}, \\ -2s_{\mathrm{N}}, & \text{if } F \subset \Gamma_{\mathrm{N}}, \\ 2\eta_K^{-1}\mathcal{B}^{-1}_{K,F,+}s_{\mathrm{R}}, & \text{if } F \subset \Gamma_{\mathrm{R}}. \end{cases}$$

4.4.2 Local element-wise discrete problems

As before, the first important step to derive the CHDG formulation of the problem is to write the local problem and prove its well-posedness.

In the symmetric case, for each element K, the local problem reads as follows.

Problem 4.4.2. *Find* $(p_K, \boldsymbol{u}_K) \in \mathcal{P}_{\mathrm{p}}(K) \times \boldsymbol{\mathcal{P}}_{\mathrm{p}}(K)$ *such that, for all* $(q_K, \boldsymbol{v}_K) \in \mathcal{P}_{\mathrm{p}}(K) \times \boldsymbol{\mathcal{P}}_{\mathrm{p}}(K)$,

$$\begin{cases} -\imath\big(\kappa_K \eta_K^{-1} p_K, q_K\big)_K - \big(\boldsymbol{u}_K, \boldsymbol{\nabla} q_K\big)_K + \displaystyle\sum_{F \in \mathcal{F}_K} \frac{1}{2} \big\langle (\mathcal{A}_F^{-1} p_K + \boldsymbol{n}_{K,F} \cdot \boldsymbol{u}_K), q_K \big\rangle_F \\ \qquad = \displaystyle\sum_{F \in \mathcal{F}_K} \frac{1}{2} \big\langle g_{K,F}^{\ominus}, q_K \big\rangle_F, \\ -\imath\big(\kappa_K \eta_K \boldsymbol{u}_K, \boldsymbol{v}_K\big)_K - \big(p_K, \boldsymbol{\nabla} \cdot \boldsymbol{v}_K\big)_K + \displaystyle\sum_{F \in \mathcal{F}_K} \frac{1}{2} \big\langle (p_K + \mathcal{A}_F(\boldsymbol{n}_{K,F} \cdot \boldsymbol{u}_K)), \boldsymbol{n}_{K,F} \cdot \boldsymbol{v}_K \big\rangle_F \\ \qquad = - \displaystyle\sum_{F \in \mathcal{F}_K} \frac{1}{2} \big\langle \mathcal{A}_F g_{K,F}^{\ominus}, \boldsymbol{n}_{K,F} \cdot \boldsymbol{v}_K \big\rangle_F, \end{cases} \tag{4.20}$$

for a given surface datum $g_{K,F}^{\ominus} \in \mathcal{P}_{\mathrm{p}}(F)$ *for all* $F \in \mathcal{F}_K$.

This local discrete problem corresponds to a discretized Helmholtz problem defined on K with a non-homogeneous Robin boundary condition on ∂K. This problem is well-posed without any condition, as stated and proved hereafter. In contrast to the standard HDG framework, these local problems are also well-posed at the continuous level.

Theorem 4.4.3 (Well-posedness of the local discrete problem)**.** *If Assumption 4.2.2 holds, then Problem 4.4.2 is well-posed.*

Proof. We simply have to prove that, if $g_{K,F}^{\ominus} = 0$ for all $F \in \mathcal{F}_K$, the unique solution of Problem 4.4.2 is $p_K = 0$ and $\boldsymbol{u}_K = \boldsymbol{0}$. For brevity, the subscripts K and F are omitted for the local fields, the test functions, the outgoing unit normal and the coefficients. Taking both equations of Problem 4.4.2 with $q = p$ and $\boldsymbol{v} = \boldsymbol{u}$ gives

$$-\imath(\kappa\eta^{-1} p, p)_K - (\boldsymbol{u}, \boldsymbol{\nabla} p)_K + \sum_{F \in \mathcal{F}_K} \frac{1}{2} \langle (\mathcal{A}^{-1} p + \boldsymbol{n} \cdot \boldsymbol{u}), p \rangle_F = 0, \tag{4.21}$$

$$-\imath(\kappa\eta \boldsymbol{u}, \boldsymbol{u})_K - (p, \boldsymbol{\nabla} \cdot \boldsymbol{u})_K + \sum_{F \in \mathcal{F}_K} \frac{1}{2} \langle (p + \mathcal{A}(\boldsymbol{n} \cdot \boldsymbol{u})), \boldsymbol{n} \cdot \boldsymbol{u} \rangle_F = 0. \tag{4.22}$$

Integrating by parts in both equations and taking the complex conjugate lead to

$$\imath(\kappa\eta^{-1} p, p)_K + (p, \boldsymbol{\nabla} \cdot \boldsymbol{u})_K + \sum_{F \in \mathcal{F}_K} \frac{1}{2} \langle p, (\mathcal{A}^{-1} p - \boldsymbol{n} \cdot \boldsymbol{u}) \rangle_F = 0, \tag{4.23}$$

$$\imath(\kappa\eta \boldsymbol{u}, \boldsymbol{u})_K + (\boldsymbol{u}, \boldsymbol{\nabla} p)_K + \sum_{F \in \mathcal{F}_K} \frac{1}{2} \langle \boldsymbol{n} \cdot \boldsymbol{u}, (-p + \mathcal{A}(\boldsymbol{n} \cdot \boldsymbol{u})) \rangle_F = 0. \tag{4.24}$$

Adding the four previous equations yields

$$\sum_{F \in \mathcal{F}_K} \left(\frac{1}{2} \langle \mathcal{A}^{-1} p, p \rangle_F + \frac{1}{2} \langle \mathcal{A}(\boldsymbol{n} \cdot \boldsymbol{u}), \boldsymbol{n} \cdot \boldsymbol{u} \rangle_F + \frac{1}{2} \langle p, \mathcal{A}^{-1} p \rangle_F + \frac{1}{2} \langle \boldsymbol{n} \cdot \boldsymbol{u}, \mathcal{A}(\boldsymbol{n} \cdot \boldsymbol{u}) \rangle_F \right) = 0. \tag{4.25}$$

Since $\mathcal{A}$ and $\mathcal{A}^{-1}$ are self-adjoint, we have

$$\langle \mathcal{A}^{-1} p, p \rangle_{\partial K} + \langle \mathcal{A}(\boldsymbol{n} \cdot \boldsymbol{u}), \boldsymbol{n} \cdot \boldsymbol{u} \rangle_{\partial K} = 0, \tag{4.26}$$

which gives $p = 0$ and $\boldsymbol{n} \cdot \boldsymbol{u} = 0$ on ∂K, thanks to the positivity of $\mathcal{A}$ and $\mathcal{A}^{-1}$. By using these boundary conditions in Problem 4.4.2, we have that the fields should be a solution of the strong problem. Because there is no solution with both homogeneous Neumann and Dirichlet boundary conditions, this yields the result. □

Remark 4.4.4 (Case with upwind fluxes). It is possible to write a local element-wise discrete problem similar to Problem 4.4.2 in the case of upwind fluxes and prove its well-posedness by following the steps of the proof of Theorem 4.4.3. The task can be achieved with a similar reasoning. For brevity, the definition of the problem and the proof of its well-posedness are omitted.

4.4.3 Abstract form of the hybridized system

The hybridized CHDG problem can be written in a convenient abstract form by introducing the *global scattering operator* $\mathrm{S} : \mathcal{G}_h \to \mathcal{G}_h$ defined such that, for each face F of each element K,

$$\mathrm{S}(g_h^{\ominus})|_{K,F} := \mathcal{A}_F^{-1} p_K(g_h^{\ominus}) + \boldsymbol{n}_{K,F} \cdot \boldsymbol{u}_K(g_h^{\ominus}), \tag{4.27}$$

where $(p_K(g_h^{\ominus}), \boldsymbol{u}_K(g_h^{\ominus}))$ is the solution of Problem 4.4.2 with the incoming transmission variable $(g_{K,F}^{\ominus})_{F\in\mathcal{F}_K}$ contained in $g_h^{\ominus}$ as a given surface datum. This operator can be interpreted as an "*incoming transmission variable to outgoing transmission variable*" operator.

By using the operator S, Problem 4.4.2 is rewritten as:

Problem 4.4.5. *Find $g_h^{\ominus} \in \mathcal{G}_h$ such that, for all $\xi_h \in \mathcal{G}_h$,*

$$\left\langle g_h^{\ominus}, \xi_h \right\rangle_{\partial\mathcal{T}_h} - \left\langle \Pi(\mathrm{S}(g_h^{\ominus})), \xi_h \right\rangle_{\partial\mathcal{T}_h} = \left\langle b, \xi_h \right\rangle_{\partial\mathcal{T}_h}. \tag{4.28}$$

We introduce the *global projected right-hand side* $b_h := P_h b \in \mathcal{G}_h$, where $P_h : L^2(\partial\mathcal{T}_h) \to \mathcal{G}_h$ is the projection operator defined such that $\left\langle P_h b, \xi_h \right\rangle_{\partial\mathcal{T}_h} = \left\langle b, \xi_h \right\rangle_{\partial\mathcal{T}_h}$ for all $\xi_h \in \mathcal{G}_h$. Problem 4.4.5 can then be rewritten as:

Problem 4.4.6. *Find $g_h^{\ominus} \in \mathcal{G}_h$ such that*

$$(\mathrm{I} - \Pi\mathrm{S}) g_h^{\ominus} = b_h, \tag{4.29}$$

where I is the identity operator on $\mathcal{G}_h$. Problem 4.4.5 and Problem 4.4.6 are equivalent to Problem 4.4.1 because the element-wise local problems are well-posed.

4.4.4 Strict contraction of the operator ΠS *(symmetric fluxes case)*

An interesting property of the CHDG formulation when using symmetric fluxes satisfying Assumption 4.2.2 is that the operator ΠS is a strict contraction. As a consequence, Problem 4.4.6 is always well-posed, and it can be solved with the fixed-point iteration without relaxation.

Properties of S and Π are proved by using norms associated to $\mathcal{G}_h$ and $\bigoplus_{F\in\mathcal{F}_K} \mathcal{P}_{\mathrm{p}}(F)$ defined as

$$\|g_h^{\ominus}\| := \sqrt{\sum_{K\in\mathcal{T}_h} \sum_{F\in\mathcal{F}_K} \|g_{K,F}^{\ominus}\|_F^2} \qquad \text{and} \qquad \|u\|_{\partial K} := \sqrt{\sum_{F\in\mathcal{F}_K} \|u\|_F^2},$$

where $\|\cdot\|_F$ is the norm of $L^2(F)$ induced by $\mathcal{A}_F$, namely

$$\|u\|_F := \sqrt{\langle \mathcal{A}_F u, u \rangle_F}.$$

Lemma 4.4.7. *If Assumption 4.2.2 holds, then*
(i) The solution of Problem 4.4.2 verifies

$$\sum_{F\in\mathcal{F}_K} \|\mathcal{A}_F^{-1} p_K + \boldsymbol{n}_{K,F} \cdot \boldsymbol{u}_K\|_F^2 + \sum_{F\in\mathcal{F}_K} \|\mathcal{A}_F^{-1} p_K - \boldsymbol{n}_{K,F} \cdot \boldsymbol{u}_K - g_{K,F}^{\ominus}\|_F^2 = \sum_{F\in\mathcal{F}_K} \|g_{K,F}^{\ominus}\|_F^2. \tag{4.30}$$

(ii) The second term on the left-hand side of (4.30) vanishes if and only if $g_{K,F}^{\ominus} = 0$.

Proof. For brevity, the subscripts K and F are omitted for the local fields, the test functions, the unit outgoing normal, the surface data and the coefficients.

(i) Taking both equations of Problem 4.4.2 with $q = p$ and $\boldsymbol{v} = \boldsymbol{u}$ gives

$$-\imath(\kappa\eta^{-1}p, p)_K - (\boldsymbol{u}, \boldsymbol{\nabla} p)_K + \frac{1}{2}\big\langle \mathcal{A}(\mathcal{A}^{-1}p + \boldsymbol{n}\cdot\boldsymbol{u}), \mathcal{A}^{-1}p\big\rangle_{\partial K} = \frac{1}{2}\big\langle \mathcal{A}g^{\ominus}, \mathcal{A}^{-1}p\big\rangle_{\partial K},$$
$$-\imath(\kappa\eta\boldsymbol{u}, \boldsymbol{u})_K - (p, \boldsymbol{\nabla}\cdot\boldsymbol{u})_K + \frac{1}{2}\big\langle \mathcal{A}(\mathcal{A}^{-1}p + \boldsymbol{n}\cdot\boldsymbol{u}), \boldsymbol{n}\cdot\boldsymbol{u}\big\rangle_{\partial K} = -\frac{1}{2}\big\langle \mathcal{A}g^{\ominus}, \boldsymbol{n}\cdot\boldsymbol{u}\big\rangle_{\partial K}.$$

Integrating by parts in both equations and taking the complex conjugate lead to

$$\imath(\kappa\eta^{-1}p, p)_K + (p, \boldsymbol{\nabla}\cdot\boldsymbol{u})_K + \frac{1}{2}\big\langle \mathcal{A}^{-1}p, \mathcal{A}(\mathcal{A}^{-1}p - \boldsymbol{n}\cdot\boldsymbol{u})\big\rangle_{\partial K} = \frac{1}{2}\big\langle \mathcal{A}^{-1}p, \mathcal{A}g^{\ominus}\big\rangle_{\partial K},$$
$$\imath(\kappa\eta\boldsymbol{u}, \boldsymbol{u})_K + (\boldsymbol{u}, \boldsymbol{\nabla} p)_K - \frac{1}{2}\big\langle \boldsymbol{n}\cdot\boldsymbol{u}, \mathcal{A}(\mathcal{A}^{-1}p - \boldsymbol{n}\cdot\boldsymbol{u})\big\rangle_{\partial K} = -\frac{1}{2}\big\langle \boldsymbol{n}\cdot\boldsymbol{u}, \mathcal{A}g^{\ominus}\big\rangle_{\partial K}.$$

Adding the four previous equations and multiplying by two yield

$$\begin{aligned}\big\langle \mathcal{A}(\mathcal{A}^{-1}p + \boldsymbol{n}\cdot\boldsymbol{u}), \mathcal{A}^{-1}p + \boldsymbol{n}\cdot\boldsymbol{u}\big\rangle_{\partial K} + \big\langle \mathcal{A}^{-1}p - \boldsymbol{n}\cdot\boldsymbol{u}, \mathcal{A}(\mathcal{A}^{-1}p - \boldsymbol{n}\cdot\boldsymbol{u})\big\rangle_{\partial K}\\ = \big\langle \mathcal{A}g^{\ominus}, \mathcal{A}^{-1}p - \boldsymbol{n}\cdot\boldsymbol{u}\big\rangle_{\partial K} + \big\langle (\mathcal{A}^{-1}p - \boldsymbol{n}\cdot\boldsymbol{u}), \mathcal{A}g^{\ominus}\big\rangle_{\partial K}. \quad (4.31)\end{aligned}$$

Using the following identities for rewriting the right-hand side,

$$\begin{aligned}&\big\langle \mathcal{A}g^{\ominus}, \mathcal{A}^{-1}p - \boldsymbol{n}\cdot\boldsymbol{u}\big\rangle_{\partial K} + \big\langle (\mathcal{A}^{-1}p - \boldsymbol{n}\cdot\boldsymbol{u}), \mathcal{A}g^{\ominus}\big\rangle_{\partial K}\\ &\quad= \big\langle \mathcal{A}(\mathcal{A}^{-1}p - \boldsymbol{n}\cdot\boldsymbol{u}), \mathcal{A}^{-1}p - \boldsymbol{n}\cdot\boldsymbol{u}\big\rangle_{\partial K} + \big\langle \mathcal{A}g^{\ominus}, g^{\ominus}\big\rangle_{\partial K}\\ &\qquad\quad - \big\langle \mathcal{A}(\mathcal{A}^{-1}p - \boldsymbol{n}\cdot\boldsymbol{u} - g^{\ominus}), \mathcal{A}^{-1}p - \boldsymbol{n}\cdot\boldsymbol{u} - g^{\ominus}\big\rangle_{\partial K}\\ &\quad= \|\mathcal{A}^{-1}p - \boldsymbol{n}\cdot\boldsymbol{u}\|^2_{\partial K} - \|\mathcal{A}^{-1}p - \boldsymbol{n}\cdot\boldsymbol{u} - g^{\ominus}\|^2_{\partial K} + \|g^{\ominus}\|^2_{\partial K},\end{aligned}$$

Equation (4.31) becomes

$$\begin{aligned}\|\mathcal{A}^{-1}p + \boldsymbol{n}\cdot\boldsymbol{u}\|^2_{\partial K} + \|\mathcal{A}^{-1}p - \boldsymbol{n}\cdot\boldsymbol{u}\|^2_{\partial K}\\ = \|\mathcal{A}^{-1}p - \boldsymbol{n}\cdot\boldsymbol{u}\|^2_{\partial K} - \|\mathcal{A}^{-1}p - \boldsymbol{n}\cdot\boldsymbol{u} - g^{\ominus}\|^2_{\partial K} + \|g^{\ominus}\|^2_{\partial K},\end{aligned}$$

which gives the result (4.30).

(ii) If the second term on the left-hand side of (4.30) vanishes, then $g^{\ominus} = \mathcal{A}^{-1}p - \boldsymbol{n}\cdot\boldsymbol{u}$ on ∂K. Using this relation in Problem 4.4.2, we see that p and $\boldsymbol{u}$ must satisfy

$$-\imath\big(\kappa\eta^{-1}p, q\big)_K - \big(\boldsymbol{u}, \boldsymbol{\nabla} q\big)_K + \big\langle \boldsymbol{n}\cdot\boldsymbol{u}, q\big\rangle_{\partial K} = 0,$$
$$-\imath\big(\kappa\eta\boldsymbol{u}, \boldsymbol{v}\big)_K - \big(p, \boldsymbol{\nabla}\cdot\boldsymbol{v}\big)_K + \big\langle p, \boldsymbol{n}\cdot\boldsymbol{v}\big\rangle_{\partial K} = 0,$$

for all $q \in \mathcal{P}_{\mathrm{p}}(K)$ and $\boldsymbol{v} \in \boldsymbol{\mathcal{P}}_{\mathrm{p}}(K)$, and integration by parts shows that p and $\boldsymbol{u}$ solve the Helmholtz equation in strong form. Because there is no non-zero polynomial solution to the previous equations, meaning that $p = 0$ and $\boldsymbol{u} = \boldsymbol{0}$, and then $g^{\ominus} = 0$, this yields the result. The converse statement is direct, because the local problem is well-posed. □

Theorem 4.4.8. *The scattering operator* S *is a strict contraction, i.e.*

$$\|\mathrm{S}(g_h^{\ominus})\| < \|g_h^{\ominus}\|, \quad \forall g_h^{\ominus} \in \mathcal{G}_h\backslash\{0\}.$$

Proof. Let $g_h^{\ominus} \in \mathcal{G}_h\backslash\{0\}$. By Lemma 4.4.7, one has

$$\sum_{F\in\mathcal{F}_K} \|\mathcal{A}_F^{-1}p_K + \boldsymbol{n}_{K,F}\cdot\boldsymbol{u}_K\|^2_F < \sum_{F\in\mathcal{F}_K} \|g^{\ominus}_{K,F}\|^2_F.$$

Note that the equality cannot happen because $g_h^{\ominus} \neq 0$. Then, by using the definition of S in Equation (4.27), one has

$$\sum_{F\in\mathcal{F}_K} \|\mathrm{S}(g_h^{\ominus})|_{K,F}\|^2_F < \sum_{F\in\mathcal{F}_K} \|g^{\ominus}_{K,F}\|^2_F.$$

Summing this estimate over all $K \in \mathcal{T}_h$ gives the result. □

Theorem 4.4.9. *(i) If* $\Gamma_{\mathrm{R}} = \emptyset$, Π *is an involution, i.e.* $\Pi^2 = \mathrm{I}$, *and an isometry, i.e.*

$$\|\Pi(g_h^{\ominus})\| = \|g_h^{\ominus}\|, \quad \forall g_h^{\ominus} \in \mathcal{G}_h.$$

(ii) If $\Gamma_{\mathrm{R}} \neq \emptyset$, *the exchange operator* Π *is a contraction, i.e.*

$$\|\Pi(g_h^{\ominus})\| \leq \|g_h^{\ominus}\|, \quad \forall g_h^{\ominus} \in \mathcal{G}_h.$$

The inequality is strict for all $g_h^{\ominus} \in \mathcal{G}_h$ *such that there is at least one face* $F \subset \Gamma_{\mathrm{R}}$ *where* $g_{K,F}^{\ominus}$ *is non-zero.*

Proof. (i) The result is a straightforward consequence of the definition of Π for interior faces and for boundary faces belonging to Γ_{D} or Γ_{N}.

(ii) For every boundary face $F \subset \Gamma_{\mathrm{R}}$ belonging to an element K, we have to prove that

$$\|(1+\eta_K^{-1}\mathcal{A}_F)^{-1}(1-\eta_K^{-1}\mathcal{A}_F)g\|_F \leq \|g\|_F, \qquad \forall g \in \mathcal{P}_{\mathrm{p}}(F).$$

For any given $g \in \mathcal{P}_{\mathrm{p}}(F)$, the inequality holds if and only if

$$\|(1-\eta_K^{-1}\mathcal{A}_F)\xi\|_F \leq \|(1+\eta_K^{-1}\mathcal{A}_F)\xi\|_F,$$

with $\xi := (1+\eta_K^{-1}\mathcal{A}_F)^{-1}g$. Because

$$\|(1+\eta_K^{-1}\mathcal{A}_F)\xi\|_F^2 = \|\xi\|_F^2 + \|(\eta_K^{-1}\mathcal{A}_F)\xi\|_F^2 + 2\langle\eta_K^{-1}\mathcal{A}_F^2\xi,\xi\rangle_F,$$
$$\|(1-\eta_K^{-1}\mathcal{A}_F)\xi\|_F^2 = \|\xi\|_F^2 + \|(\eta_K^{-1}\mathcal{A}_F)\xi\|_F^2 - 2\langle\eta_K^{-1}\mathcal{A}_F^2\xi,\xi\rangle_F,$$

and $\langle\eta_K^{-1}\mathcal{A}_F^2\xi,\xi\rangle_F \geq 0$, the result holds true. In addition, if $g \neq 0$, then $\xi \neq 0$ and the inequality is strict. □

Corollary 4.4.10. *The operator* $\Pi\mathrm{S}$ *is a strict contraction, i.e.*

$$\|\Pi\mathrm{S}(g_h^{\ominus})\| < \|g_h^{\ominus}\|, \quad \forall g_h^{\ominus} \in \mathcal{G}_h\backslash\{0\}.$$

Proof. Corollary 4.4.10 is a direct consequence of Theorems 4.4.8 and 4.4.9. □

Remark 4.4.11 (Interpretation). The strict contraction of S is a numerical effect: we have used the fact that the Helmholtz equation has no homogeneous polynomial solution in the proof of Lemma 4.4.7. The version of Equation (4.30) for the continuous case corresponds to a conservation of energy. By contrast, the properties of Π are related to the boundary conditions, and would be preserved at the continuous level. This operator is contracting only with a boundary condition that does not preserve energy, i.e. the Robin boundary condition. More precisely, it is strictly contracting only for fields that are different from zero on Γ_{R}.

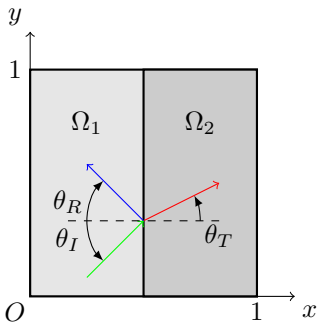
4.5 Numerical comparison of the methods

In this section, iterative solution procedures for the HDG and CHDG methods are studied and compared by using two simple benchmarks and a more realistic application. For the first two benchmarks, we consider configurations with homogeneous and heterogeneous media in Sections 4.5.2 and 4.5.3, respectively. For the last benchmark (Section 4.5.4), the complex heterogeneous medium illustrates a realistic subsurface medium for seismic wave propagation.

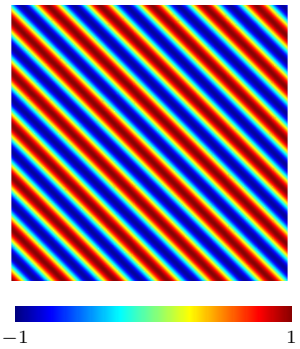
4.5.1 Benchmark and numerical setting

As in Chapter 3, the numerical simulations are performed with a MATLAB code, the mesh generation and the visualization are performed with gmsh and, in all the cases, third-order polynomial Lobatto basis functions ($\mathrm{p} = 3$) are used. For the symmetric fluxes, we take $\mu_F = \sqrt{\eta_K \eta_{K'}}$ and $\kappa_F = \sqrt{\kappa_K \kappa_{K'}}$ for an interior face F shared by two elements K and K'.

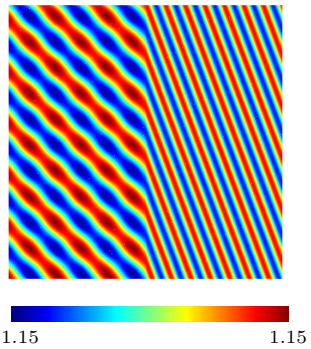
Benchmark 1 (Plane wave). The first benchmark consists in the phenomenon of reflection and transmission of an incident plane wave at the interface between two media. We consider the square domain $\Omega = (0,1)^2$ partitioned into two rectangular regions $\Omega_1 = (0,1/2) \times (0,1)$ and $\Omega_2 = (1/2,1) \times (0,1)$ corresponding to the two media. The wavenumber and impedance, which are constant in each region, are denoted by κ_1 and η_1 in Ω_1, and by κ_2 and η_2 in Ω_2 (see Figure 4.1a). We assume that the pressure p and the velocity $\boldsymbol{u}$ satisfy system (4.1) with a non-homogeneous Robin boundary condition on the boundary of Ω satisfied by the reference solution (see Appendix 4.A.1 for further details), i.e. $s_{\mathrm{R}} = p_{\mathrm{ref}} - \eta \boldsymbol{n} \cdot \boldsymbol{u}_{\mathrm{ref}}$ on $\partial\Omega = \Gamma_{\mathrm{R}}$.



(a) Square domain with a heterogeneous medium and scheme of incident (green), reflected (blue) and transmitted (red) plane waves.



(b) Real part of p_{ref} with a homogeneous medium.



(c) Real part of p_{ref} with a heterogeneous medium.

Figure 4.1: Domain and reference solution for the propagative plane wave benchmark.

The reference solution in the region Ω_1 is given by the superposition of the incident and reflected plane waves propagating with an angle θ_I and θ_R, respectively, with $\theta_I = \theta_R$, while in the region Ω_2 it corresponds to the transmitted plane wave propagating with an angle θ_T (see Figure 4.1a). The reference solution reads

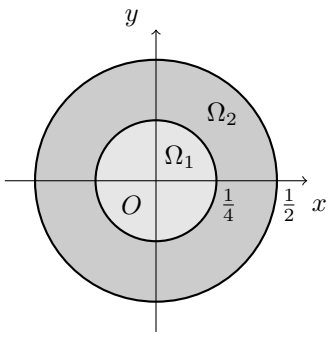
$$p_{\mathrm{ref}}(x,y) = \begin{cases} e^{\imath\kappa_1(x\cos\theta_I + y\sin\theta_I)} + Re^{\imath\kappa_1(-x\cos\theta_R + y\sin\theta_R)}, & \text{in } \Omega_1, \\ Te^{\imath\kappa_2(x\cos\theta_T + y\sin\theta_T)}, & \text{in } \Omega_2, \end{cases}$$

with the reflection coefficient R and transmission coefficient T given by

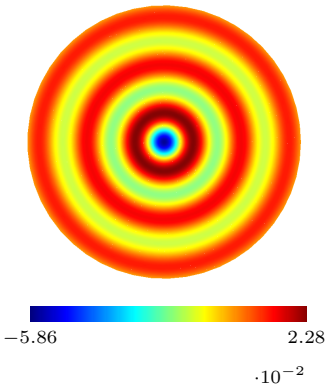
$$R = \frac{\eta_2\cos\theta_I - \eta_1\cos\theta_T}{\eta_1\cos\theta_T + \eta_2\cos\theta_I} e^{\imath\kappa_1\cos\theta_I} \qquad \text{and} \qquad T = \frac{2\eta_2\cos\theta_I}{\eta_1\cos\theta_T + \eta_2\cos\theta_I} e^{\imath(\kappa_1\cos\theta_I - \kappa_2\cos\theta_T)/2}.$$

The angle θ_I of the incident wave and θ_R of the reflected one are equal, i.e. $\theta_I = \theta_R$, and are set to $\theta_I = \theta_R = \pi/4$ in the following. The angle θ_T of the transmitted wave is $\theta_T = \arcsin(\frac{\kappa_1}{\kappa_2}\sin\theta_I)$. The reference solution is shown for cases with homogeneous and heterogeneous media in Figures 4.1b and 4.1c, respectively.

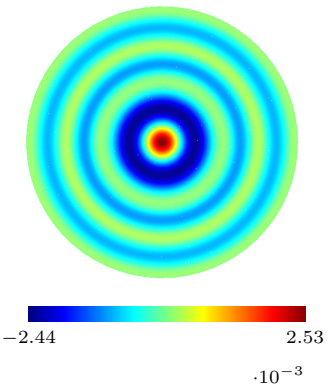
Benchmark 2 (Cavity). The second benchmark consists in the acoustic phenomenon generated by a uniform volume source. We consider a circular domain Ω with radius $1/2$ partitioned into the circular region Ω_1 with radius $1/4$ and the annulus Ω_2 with the radial coordinate $r \in (1/4, 1/2)$, all centered at the origin. The physical coefficients are constant in each region, with the same notation as in the first benchmark (see Figure 4.2a). We assume that the pressure p and the velocity $\boldsymbol{u}$ satisfy system (4.1) with a constant volume source term $f = -1/(\imath\kappa\eta)$ introduced on the right-hand side of the first equation and with a homogeneous Dirichlet boundary condition prescribed on the boundary of the domain, i.e. $\partial\Omega = \Gamma_{\mathrm{D}}$ and $s_{\mathrm{D}} = 0$.



(a) Circular domain with a heterogeneous medium.



(b) Real part of p_{ref} with a homogeneous medium.



(c) Real part of p_{ref} with a heterogeneous medium.

Figure 4.2: Domain and reference solution for the cavity benchmark.

If the physical coefficients do not correspond to a resonance mode, the solution of this benchmark is unique and real. In this case, the reference solution, which depends only on the radial coordinate r, reads

$$p_{\mathrm{ref},j}(r) = A_j \mathrm{J}_0(\kappa_j r) + B_j \mathrm{Y}_0(\kappa_j r) - \kappa_j^{-2}, \quad \text{in } \Omega_j, \quad j = 1, 2,$$

where A_j and B_j are constant coefficients, J_0 is the Bessel function of the first kind of order 0 and Y_0 is the Bessel function of the second kind of order 0 (see Appendix 4.A.2 for further details). The reference solution is shown for cases with homogeneous and heterogeneous media in Figures 4.2b and 4.2c, respectively.

Iterative solvers. As in Chapter 3, three standard iterations are considered: the fixed-point iteration (only for the CHDG system), the CGNR iteration, and the GMRES iteration without restart. Again, we have used a symmetric preconditioning with the mass matrix associated to the faces of the elements.

Numerical error. In this study, we consider a relative numerical error based on the acoustic energy associated to the physical fields. It is defined as

$$\text{relative error} := \frac{\|p_h - p_{\text{ref}}, \boldsymbol{u}_h - \boldsymbol{u}_{\text{ref}}\|_E}{\|p_{\text{ref}}, \boldsymbol{u}_{\text{ref}}\|_E}, \quad \text{with } \|p, \boldsymbol{u}\|_E^2 := \sum_{K\in\mathcal{T}_h} \frac{\|p_K\|^2_{L^2(K)}}{2\rho_K c_K^2} + \frac{1}{2}\rho_K \|\boldsymbol{u}_K\|^2_{L^2(K)},$$

where $(p_h, \boldsymbol{u}_h)$ is the numerical solution, $(p_{\text{ref}}, \boldsymbol{u}_{\text{ref}})$ is the reference solution, and $\|.\|_E$ denotes the energy norm (see (1.15)).

4.5.2 Comparison for homogeneous media

The benchmark problems are solved for constant physical parameters corresponding to two levels of difficulty: low- and high-frequency cases for the plane-wave benchmark, and wavenumbers close and very close to the resonance mode $\kappa_3 \approx 17.3075$ for the cavity benchmark (see Appendix 4.A.2). For the circular cavity with homogeneous medium, a resonance mode corresponds to $\kappa_j = 2x_j$, where $\{x_j\}_{j\in\mathbb{N}}$ are the zeros of $\mathrm{J}_0(x)$. In all the cases, the mesh sizes are chosen to achieve a relative error close to 10^{-2} when a direct solver is employed. The parameters are given in Table 4.1.

(a) Benchmark 1 (plane wave)

Case	κ	h	Numerical flux	$\rho(\boldsymbol{\Pi}\mathbf{S})$	Relative error
1	15π	1/16	Sym-0	$1 - 4.18 \cdot 10^{-3}$	$1.44 \cdot 10^{-2}$
			Sym-2	$1 - 7.97 \cdot 10^{-3}$	$1.70 \cdot 10^{-2}$
2	30π	1/34	Sym-0	$1 - 1.88 \cdot 10^{-3}$	$1.37 \cdot 10^{-2}$
			Sym-2	$1 - 3.48 \cdot 10^{-3}$	$1.59 \cdot 10^{-2}$

(b) Benchmark 2 (cavity)

Case	κ	h	Numerical flux	$\rho(\boldsymbol{\Pi}\mathbf{S})$	Relative error
1	16.5	0.04	Sym-0	$1 - 1.22 \cdot 10^{-8}$	$1.07 \cdot 10^{-2}$
			Sym-2	$1 - 1.71 \cdot 10^{-6}$	$1.07 \cdot 10^{-2}$
2	17	0.025	Sym-0	$1 - 1.15 \cdot 10^{-9}$	$1.16 \cdot 10^{-2}$
			Sym-2	$1 - 1.44 \cdot 10^{-7}$	$1.16 \cdot 10^{-2}$

Table 4.1: Parameters for the benchmark problems with homogeneous media, including the spectral radius $\rho(\boldsymbol{\Pi}\mathbf{S})$ of the matrix of the CHDG hybridized system, and the relative numerical error. In these cases, $\omega = \kappa$, $c = 1$, $\rho = 1$ and $\eta = 1$.

The history of relative error is plotted in Figure 4.3 during the fixed-point iterations (lines with marker $\times$, only for CHDG), the CGNR iterations (lines with marker $\circ$) and the GMRES iterations (lines with marker $\bullet$) applied to the HDG and CHDG hybridized systems. The zeroth-order symmetric flux (Sym-0) is considered for both methods, and the second-order symmetric flux (Sym-2) is used for CHDG. The relative errors obtained with a direct solver is indicated by the horizontal dashed lines for both numerical fluxes. Let us note that, since the impedance η is constant, the upwind flux is identical to the 0th-order symmetric flux with $\mathcal{A}_F = \eta$.

The following observations can be made:

- The fixed-point iteration applied to the CHDG system converges in all the cases. The convergence is fast for the plane-wave benchmark, and very slow for the cavity benchmark. In both cases, the convergence is slower for the parameters corresponding to the more difficult configuration (i.e. high frequency case or close to a resonance frequency). These results are in accordance with the spectral radii given in Table 4.1: the convergence is slower for radii closer to 1.
- Comparing the numerical fluxes used in the CHDG method, we observe that the convergence of the fixed point iteration is comparable with 'Sym-0' and 'Sym-2'. The results are similar for the GMRES iteration. In contrast, the CGNR iteration with 'Sym-2' is slower than with 'Sym-0', sometimes significantly so.

- The convergence of the CGNR and GMRES iterations is slower with HDG than with CHDG for the plane-wave benchmark. The difference is more important for the high-frequency case. For the cavity benchmark, the convergence of GMRES is similar with both methods, while the convergence of CGNR is slower with HDG than with CHDG.

In a nutshell, CHDG with 'Sym-0' is one of the best solution in all the cases. The fixed-point iteration applied to the CHDG system converges, but its performance is strongly influenced by the physical parameters, and it becomes unusable for cavities. The second-order numerical flux does not accelerate the convergence of the iterative procedures.

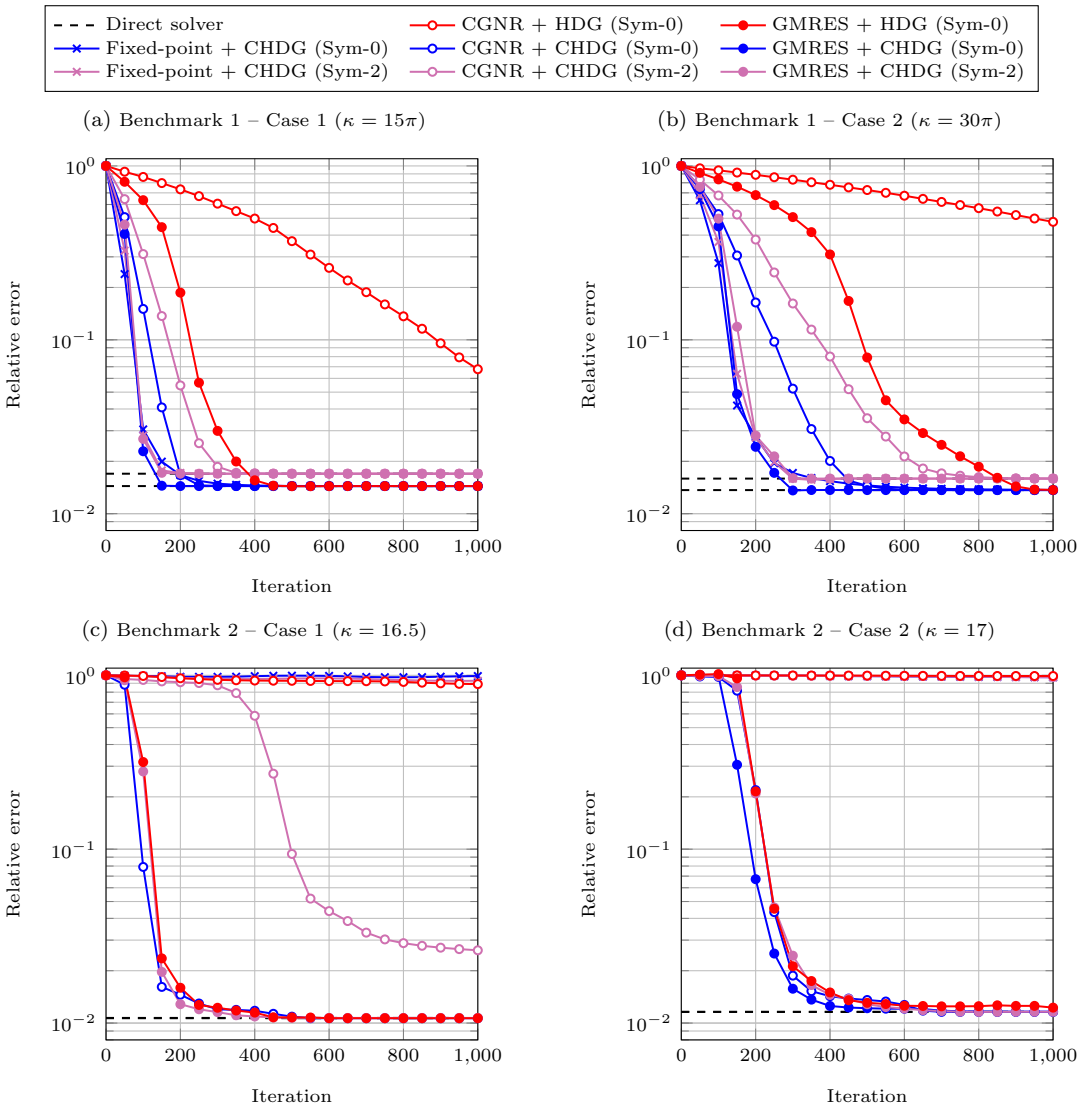


Figure 4.3: Results for the benchmark problems with homogeneous media. Error history with different iterative schemes and different DG methods. The dashed horizontal lines correspond to the relative numerical errors obtained with a direct solver for the different numerical fluxes.

4.5.3 Comparison for heterogeneous media

The benchmark problems are now tested with two sets of parameters. For both benchmarks, the impedance η is constant in the first set, it is not constant in the second set, and the wavenumber κ is not constant in all the sets. The parameters are given in Table 4.2. For all the cases, the mesh size is chosen to obtain a relative error close to 10^{-2} when a direct solver is employed.

(a) Benchmark 1 (plane wave)

Case	ω	c	ρ	κ	η	h	Numerical flux	$\rho(\mathbf{\Pi S})$	Relative error
1	15π	1 1/2	1 2	15π 30π	1 1	1/16 1/34	Upw	$1-1.63\cdot10^{-3}$	$1.01\cdot10^{-2}$
							Sym-0	$1-1.63\cdot10^{-3}$	$1.01\cdot10^{-2}$
							Sym-2	$1-4.32\cdot10^{-3}$	$1.20\cdot10^{-2}$
2	15π	1 1/2	1 1	15π 30π	1 1/2	1/16 1/34	Upw	$1+2.38\cdot10^{-2}$	$9.75\cdot10^{-3}$
							Sym-0	$1-1.55\cdot10^{-2}$	$9.74\cdot10^{-3}$
							Sym-2	$1-4.31\cdot10^{-3}$	$1.16\cdot10^{-2}$

(b) Benchmark 2 (cavity)

Case	ω	c	ρ	κ	η	h	Numerical flux	$\rho(\mathbf{\Pi S})$	Relative error
1	10π	1 2/3	1 3/2	10π 15π	1 1	1/12 1/16	Upw	$1-5.19\cdot10^{-5}$	$1.57\cdot10^{-2}$
							Sym-0	$1-5.19\cdot10^{-5}$	$1.57\cdot10^{-2}$
							Sym-2	$1-5.28\cdot10^{-5}$	$1.67\cdot10^{-2}$
2	10π	1 2/3	1 1	10π 15π	1 2/3	1/12 1/16	Upw	$1+3.45\cdot10^{-4}$	$8.59\cdot10^{-3}$
							Sym-0	$1-4.99\cdot10^{-5}$	$8.61\cdot10^{-3}$
							Sym-2	$1-5.29\cdot10^{-5}$	$1.01\cdot10^{-2}$

Table 4.2: Parameters of the benchmark problems with heterogeneous media, including spectral radius $\rho(\mathbf{\Pi S})$ of the matrix of the CHDG hybridized system, and the relative numerical error.

The history of relative error is plotted in Figure 4.4 for different combinations of hybridized methods, numerical fluxes and iterative schemes. For CHDG, we have considered the upwind fluxes (Upw) and both zeroth- and second-order symmetric fluxes (Sym-0 and Sym-2).

The following remarks can be made:

- The numerical fluxes 'Upw' and 'Sym-0' are identical when the impedance is constant. Nonetheless, the results are very similar when the impedance is not constant. For HDG, both fluxes lead to the same results, even when the impedance is discontinuous.
- We observe that the fixed-point iteration applied to the CHDG systems with the symmetric fluxes converges in all cases, while it fails to converge with the upwind fluxes when the impedance is discontinuous. The divergence can be seen in Figure 4.4b, and it also appears when increasing the number of iterations for the case corresponding to Figure 4.4d. These results are consistent with the theory and the spectral radii in Table 4.2: the contraction property of the operator ΠS was proved in Corollary 4.4.10 only for symmetric fluxes, and the convergence occurs numerically for radii strictly less than 1. The convergence is quite fast for the plane-wave benchmark, while it slows down dramatically for the cavity benchmark.
- Similarly to the cases with homogeneous media, when comparing the symmetric numerical fluxes used in the CHDG system, we observe that the convergence of CGNR and GMRES with 'Sym-2' is always either comparable or slower than with 'Sym-0' for both benchmarks. Once again, the convergence of the iterative schemes is much slower with HDG than with CHDG, regardless of the iterative procedure that is used.

In a nutshell, the CGNR and GMRES iterations combined with CHDG and either 'Upw' or 'Sym-0' are always the best options. The fixed-point iteration applied to CHDG with 'Sym-0' always converges, whereas it does not with HDG or CHDG with 'Upw'.

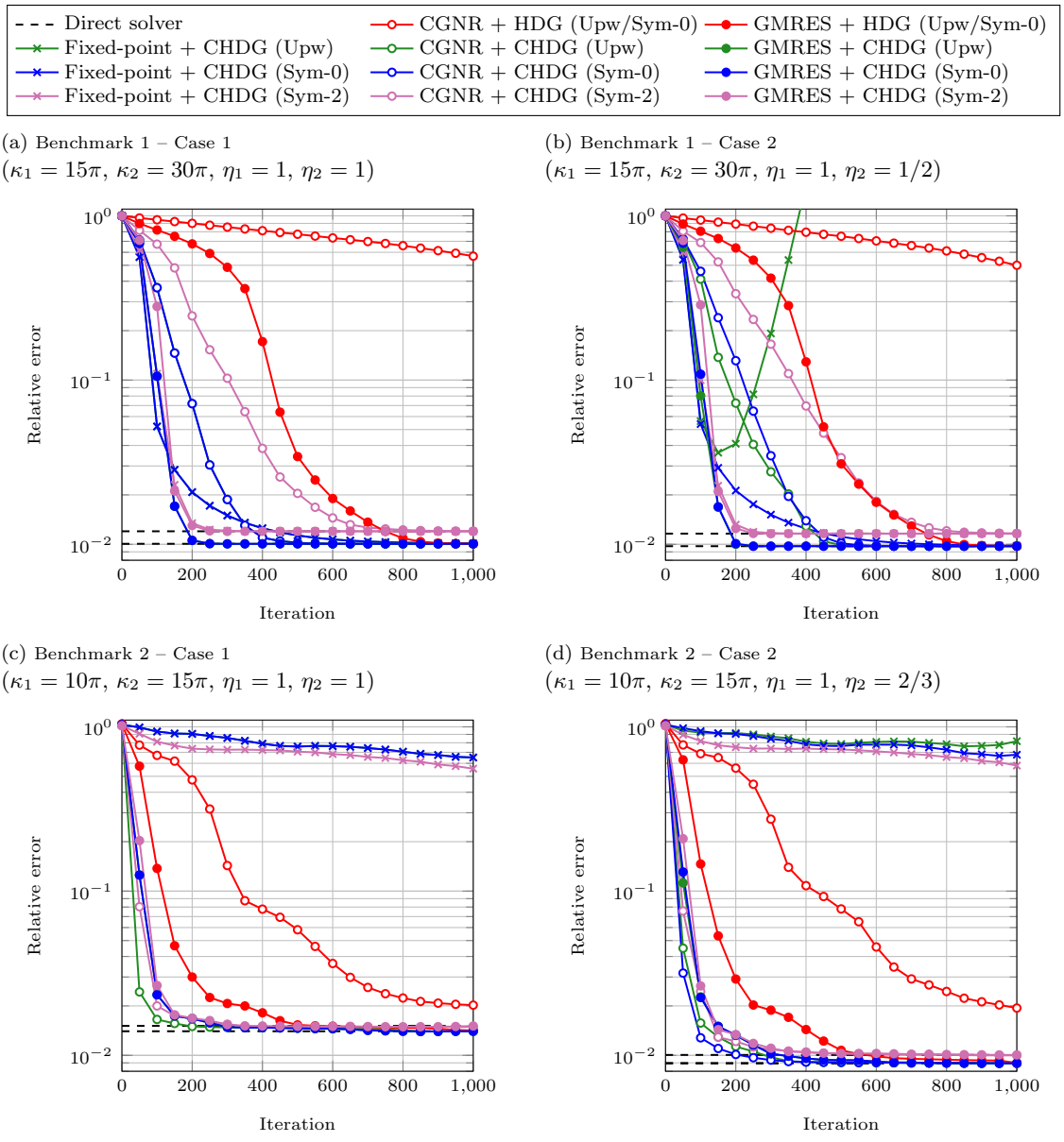


Figure 4.4: Results of the benchmark problems with heterogeneous media. Error history with different iterative schemes and different DG methods. The dashed horizontal lines correspond to the relative numerical errors obtained with a direct solver for the different numerical fluxes. In most graphs, the green lines (CHDG 'Upw') are hidden by the corresponding blue lines (CHDG 'Sym-0').

4.5.4 Realistic application

To compare the methods on a more illustrative case, we consider the well-known Marmousi benchmark problem, which consists of a realistic underground structure commonly used to assess numerical methods, see e.g. [124]. The variations of the velocity $c(\boldsymbol{x})$ and the density $\rho(\boldsymbol{x})$ are shown in Figure 4.5. The size and the highly heterogeneous character of the domain Ω (as detailed below) make this test case particularly challenging from a numerical point of view.

An acoustic field is generated by a source point at the coordinates $(4585\text{m}, -10\text{m})$, near the top of the rectangular domain $\Omega = (0\text{m}, 9192\text{m}) \times (-2094\text{m}, 0\text{m})$. The mesh is generated with gmsh [59], and includes 60 761 elements. The element size is adjusted following the empirical rule $h \approx \lambda(\boldsymbol{x})/n_\lambda$, with the wavelength $\lambda(\boldsymbol{x}) = c(\boldsymbol{x})/f$, the frequency $f = 30\,\text{Hz}$, the angular frequency $\omega = 2\pi f$, the mesh density $n_\lambda = 10/(\text{p}+1)$, and the polynomial degree $\text{p} = 3$. The physical coefficients are constant in each cell. The Dirichlet boundary condition $p = 0$ is applied at the top of the domain, and the absorbing boundary condition $p - \eta \boldsymbol{n} \cdot \boldsymbol{u} = 0$ is prescribed on the other sides of the domain. The mesh and the solution are shown in Figure 4.5.

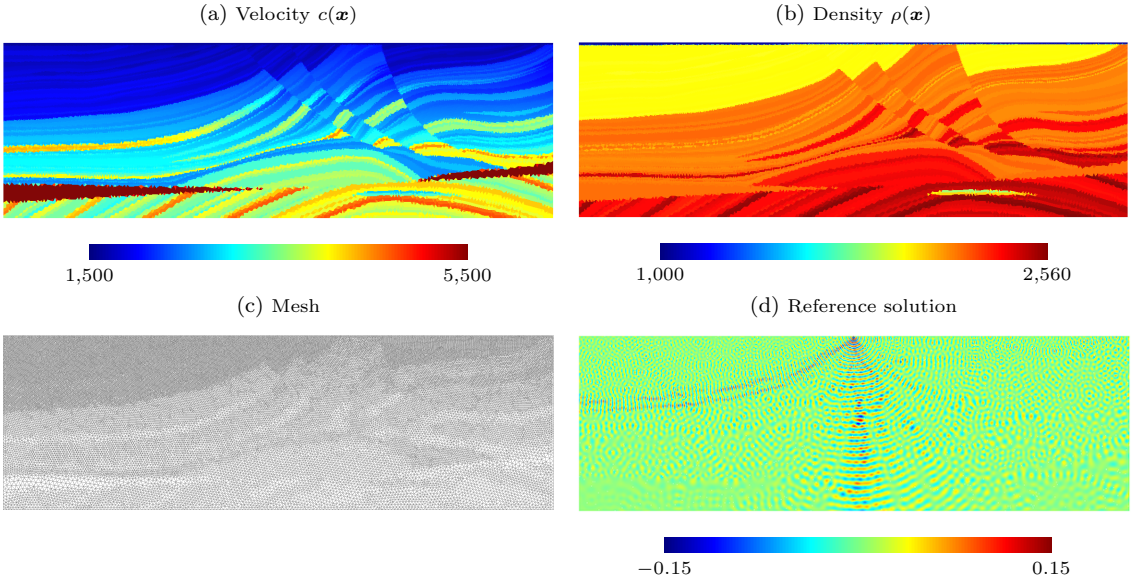


Figure 4.5: Marmousi benchmark problem. Profile of velocity $c(\boldsymbol{x})$ (a), of density $\rho(\boldsymbol{x})$(b), mesh (c) and real part of the reference pressure field obtained with a direct solver (d).

The problem is solved using HDG and CHDG with the zeroth-order symmetric flux (Sym-0) and the upwind flux (Upw). We consider the fixed-point iteration (only for CHDG with 'Sym-0'), the CGNR iteration, and the GMRES iterations (with a restart after every 10 iterations). The restart strategy, which is widely used in practice, helps to reduce computation and memory usage. The need of a restart is justified by the big number of degrees of freedom. In Table 4.3 we collect the number of degrees of freedom and the number of non-zero entries of the HDG and CHDG systems.

	HDG	CHDG
#dof	365 868	728 484
#nnz	7 276 130	11 449 779

Table 4.3: Number of degrees of freedom (#dof) and number of non-zero entries (#nnz) of the HDG and CHDG systems for the Marmousi benchmark.

Figure 4.6 shows the histories of residual and error for each combination of methods. We show the relative residual associated with the preconditioned hybridized system and the one corresponding to the physical system. In contrast to the previous sections, the relative error is computed by comparing the numerical solution obtained at each iteration with that obtained with a direct solver.

We observe that the iterative solution of HDG is considerably slower compared to that of CHDG, regardless of whether CGNR or GMRES is used, both in terms of residuals and errors. The choice of numerical flux has a negligible effect on the convergence of the iterations with HDG (only one curve is shown for clarity), while the iterations with 'Upw' are slightly slower than with 'Sym-0' when considering CHDG.

Comparing the different iterative strategies for solving the CHDG system with 'Sym-0', we observe that the decay of both error and residual is faster with GMRES than with CGNR. The slopes of the curves of error and residual are nonetheless similar. The fixed-point iteration performs well until approximately 4 000 iterations, where the decay of error slows down.

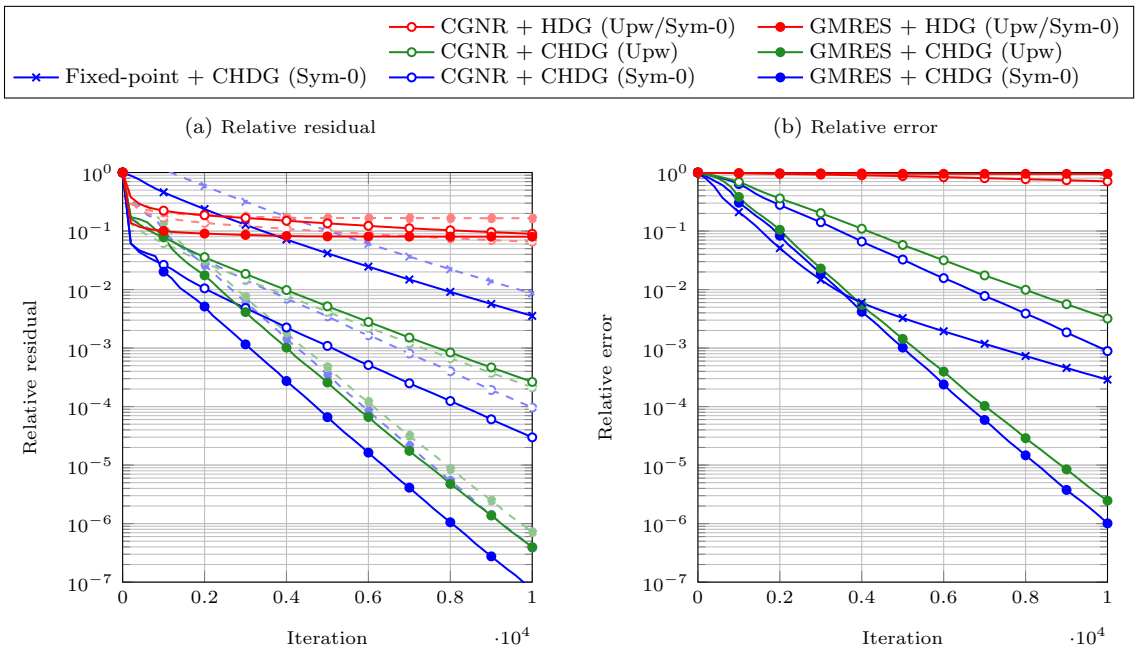


Figure 4.6: Marmousi benchmark problem. Relative residual (left) and relative error history (right) with different iterative schemes and DG methods. For the relative residual history, we consider for each case both the residual associated with the hybridized system (in ×, plain lines) and the one associated with the physical system (in physical norm, dashed lines). The relative error is computed by comparing the numerical physical solution obtained at each iteration with the reference numerical physical solution obtained with a direct solver.

One should note that these results do not give a final conclusion on which iterative scheme should be used in practice, since the runtime per iteration depends on the iterative scheme. The fixed-point iteration requires a single multiplication by the matrix of the system, while CGNR involves the multiplications by both the matrix and its adjoint at each iteration. The cost of GMRES increases at each iteration until the restart, and it is, overall, the most expensive method per iteration. Because these results are obtained with a MATLAB code that is not fully optimized, the observed runtimes are not reported here as they do not provide a representative comparison of efficiency.

4.6 Summary

In this Chapter, we have extended and studied the CHDG method, adapted in Chapter 2 to symmetric hyperbolic systems with constant physical coefficients, in order to address wave propagation in heterogeneous media with piecewise constant physical coefficients. In particular, we considered formulations with standard upwind numerical fluxes, which are widely used in the literature, and symmetric numerical fluxes involving a general impedance operator $\mathcal{A}$. For both numerical fluxes, transmission variables are introduced at the interfaces between the elements in the hybridization procedure. In contrast to [97], the definition of the transmission variables involves the physical coefficients and the operator $\mathcal{A}$.

The CHDG hybridized system can once again be written in the form $(\mathrm{I} - \Pi\mathrm{S})g = b$. In the case of symmetric fluxes, if $\mathcal{A}$ is a positive and self-adjoint operator, it is proven that the operator $\Pi\mathrm{S}$ is a strict contraction, which makes it possible to solve the system with a fixed-point iteration. This result also holds for the standard upwind fluxes if the impedance η is constant. Otherwise, $\Pi\mathrm{S}$ is not a contraction and the fixed-point iteration does not converge, as confirmed by the numerical results.

We have investigated and compared the different methods, numerical fluxes and iterative schemes by considering 2D benchmark problems.

- Most of the observations that were made in the homogeneous case in [97] remain valid. The fixed-point iteration applied to the CHDG system converges for the cases where it is proven that the operator $\Pi\mathrm{S}$ is a strict contraction. While the convergence of the fixed-point iteration is in general very fast for benchmarks with physical dissipation, it is very slow for cavity cases.
- Considering the CGNR and GMRES iterations, we observe that their convergence is nearly always faster with CHDG than with the standard HDG method. The difference is very significant for the benchmarks with impedance boundary conditions, and it is drastic for the Marmousi benchmark.
- We have also investigated the use of symmetric fluxes with a second-order differential operator $\mathcal{A}$, inspired by recent work on domain decomposition methods. While this operator satisfies the assumptions of the theoretical framework, it always results in a slower convergence of the iterative schemes compared to the symmetric fluxes with a scalar operator. The reason why a higher-order transmission condition slows the convergence remains to be interpreted.
- Comparing the upwind fluxes and the scalar symmetric fluxes, we have observed that, in the cases where they are not identical (i.e. when the impedance η is discontinuous at the interfaces), they provide similar results, except for the fixed-point iteration.

4.A Appendix: Analytical solutions of benchmarks

In this section, we retrieve the analytical expression of the exact solutions of the first two benchmarks presented in Section 4.5.

4.A.1 Propagative plane wave

We consider Equations (4.1) which are valid on each of the two regions where the physical coefficients are constant and we take $\partial\Omega = \Gamma_{\mathrm{R}}$

$$\begin{cases} -\imath\kappa_j\eta_j^{-1}p + \boldsymbol{\nabla}\cdot\boldsymbol{u} = -\dfrac{1}{\imath\kappa_j\eta_j}, & \text{in } \Omega_j, \quad j=1,2, \\ -\imath\kappa_j\eta_j\boldsymbol{u} + \boldsymbol{\nabla}p = \boldsymbol{0}, & \text{in } \Omega_j, \quad j=1,2, \\ p = 0, & \text{on } \Gamma_{\mathrm{D}}, \end{cases} \tag{4.32}$$

We re-write the second equation of (4.32) with respect to the velocity field $\boldsymbol{u}$

$$\boldsymbol{u} = \frac{1}{\imath\kappa_j\eta_j}\boldsymbol{\nabla}p, \quad \text{in } \Omega_j, \quad j=1,2, \tag{4.33}$$

We assume that an incident plane wave p_I propagates in the left subdomain Ω_1 towards the interface $x = 1/2$ with an incident angle θ_I with respect to the normal to the interface

$$p_I(x,y) = Ie^{\imath\kappa_1(x\cos\theta_I + y\sin\theta_I)},$$

where I denotes the amplitude of the plane wave.

The discontinuity in the physical coefficients along this interface generates a reflected plane wave p_R which propagates in Ω_1 with a reflected angle θ_R and a transmitted one p_T which propagates in Ω_2 with a transmitted angle θ_T (see Figure 4.7)

$$\begin{aligned} p_R(x,y) &= Re^{\imath\kappa_1(-x\cos\theta_R + y\sin\theta_R)}, \\ p_T(x,y) &= Te^{\imath\kappa_2(x\cos\theta_T + y\sin\theta_T)}, \end{aligned}$$

where R and T denote the reflection and transmission coefficient, respectively. The total pressure field is therefore given by the superposition of the incident and reflected field in Ω_1 and by the transmitted one in Ω_2, namely

$$p(x,y) = \begin{cases} p_I(x,y) + p_R(x,y), & \text{in } \Omega_1, \\ p_T(x,y), & \text{in } \Omega_2. \end{cases}$$

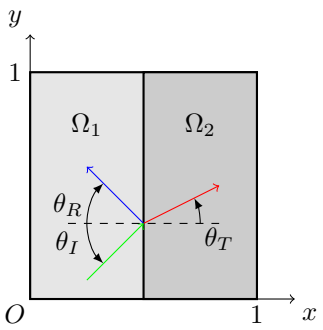


Figure 4.7: Square domain with heterogeneous medium and scheme of incident (green), reflected (blue) and transmitted (red) plane waves.

The relation between the three angles of propagation is established by the reflection law that reads

$$\theta_I = \theta_R,$$

and by Snell's law that claims

$$\frac{\sin\theta_I}{c_1} = \frac{\sin\theta_T}{c_2},$$

which gives

$$\theta_T = \arcsin\left(\frac{c_2}{c_1}\sin\theta_I\right) = \arcsin\left(\frac{\kappa_1}{\kappa_2}\sin\theta_I\right).$$

Note that if $c_1 > c_2$ there is always a transmitted angle θ_T, whereas if $c_1 \leq c_2$ the phenomenon of *total internal reflection* which consists in the lack of the transmitted plane wave occurs for incident angles that are bigger or equal than the so-called *critic angle* which is obtained by setting $(c_2/c_1)\sin\theta_C = 1$ and reads

$$\theta_C = \arcsin\left(\frac{c_1}{c_2}\right).$$

The relation between the amplitude of the incident plane wave and the reflection and transmission coefficients is established by imposing the continuity of the pressure p and of the normal component of the velocity $\boldsymbol{n}\cdot\boldsymbol{u}$ (recall that $\boldsymbol{u}$ is given be Equation (4.33)) along the interface $x = 1/2$. Imposing these two conditions is equivalent to solving the following system

$$\begin{cases} p_I\left(\frac{1}{2},y\right) + p_R\left(\frac{1}{2},y\right) = p_T\left(\frac{1}{2},y\right), \\ \frac{1}{\imath\kappa_1\eta_1}\left(\frac{\partial p_I}{\partial x} + \frac{\partial p_R}{\partial x}\right)\left(\frac{1}{2},y\right) = \frac{1}{\imath\kappa_2\eta_2}\frac{\partial p_T}{\partial x}\left(\frac{1}{2},y\right), \end{cases}$$

whose solution is

$$R = I\frac{\eta_2\cos\theta_I - \eta_1\cos\theta_T}{\eta_1\cos\theta_T + \eta_2\cos\theta_I}e^{\imath\kappa_1\cos\theta_I} \qquad \text{and} \qquad T = I\frac{2\eta_2\cos\theta_I}{\eta_1\cos\theta_T + \eta_2\cos\theta_I}e^{\imath(\kappa_1\cos\theta_I - \kappa_2\cos\theta_T)/2}.$$

4.A.2 Cavity

We consider Equations (4.1) which are valid on each of the two regions where the physical coefficients are constant and we take $\partial\Omega = \Gamma_{\mathrm{D}}$, $f = -1/(\imath\kappa\eta)$ (right-hand side in the first equation) and $s_{\mathrm{D}} = 0$

$$\begin{cases} -\imath\kappa_j\eta_j^{-1}p + \boldsymbol{\nabla}\cdot\boldsymbol{u} = -\dfrac{1}{\imath\kappa_j\eta_j}, & \text{in } \Omega_j, \quad j = 1,2, \\ -\imath\kappa_j\eta_j\boldsymbol{u} + \boldsymbol{\nabla}p = \boldsymbol{0}, & \text{in } \Omega_j, \quad j = 1,2, \\ p = 0, & \text{on } \Gamma_{\mathrm{D}}, \end{cases} \tag{4.34}$$

We re-write the second equation of (4.34) with respect to the velocity field $\boldsymbol{u}$

$$\boldsymbol{u} = \frac{1}{\imath\kappa_j\eta_j}\boldsymbol{\nabla}p, \quad \text{in } \Omega_j, \quad j = 1,2, \tag{4.35}$$

we take the divergence in both sides and we plug the result into the first equation of (4.34), we eliminate $\boldsymbol{u}$ and we get the following Helmholtz equation with the pressure field p as unique unknown

$$\begin{cases} -\Delta p - \kappa_j^2 p = 1, & \text{in } \Omega_j, \quad j = 1,2, \\ p = 0, & \text{on } \Gamma_{\mathrm{D}}. \end{cases} \tag{4.36}$$

Because of the circular geometry of the domain Ω, we can rewrite Equations (4.36) in terms of polar coordinates (r,θ) with $r \in [0, R_2)$ and $\theta \in [0, 2\pi)$ as

$$\begin{cases} -\dfrac{\partial^2 p}{\partial r^2} - \dfrac{1}{r}\dfrac{\partial p}{\partial r} - \dfrac{1}{r^2}\dfrac{\partial^2 p}{\partial\theta^2} - \kappa_j^2 p = 1, & \text{in } \Omega_j, \quad j = 1,2, \\ p(R_2) = 0. \end{cases} \tag{4.37}$$

Additionally, because of the circular symmetry of the boundary conditions, we can assume the solution to be independent of the second coordinate θ, i.e. $p = p(r)$, so that we can simplify the first equation of (4.37) and get

$$\begin{cases} -\dfrac{d^2 p}{dr^2} - \dfrac{1}{r}\dfrac{dp}{dr} - \kappa_j^2 p = 1, & \text{in } \Omega_j, \quad j = 1,2, \\ p(R_2) = 0, & \text{on } \Gamma_{\mathrm{D}}. \end{cases} \tag{4.38}$$

If the physical coefficients do not correspond to a resonance mode (see later), the solution of the first equation of (4.38) is unique, real and reads

$$p_{\mathrm{ref},j}(r) = A_j \mathrm{J}_0(\kappa_j r) + B_j \mathrm{Y}_0(\kappa_j r) - \kappa_j^{-2}, \quad \text{in } \Omega_j, \quad j = 1,2,$$

where A_j and B_j are constant coefficients, J_0 is the Bessel function of the first kind of order 0 and Y_0 is the Bessel function of the second kind of order 0. To preclude singular solutions at the origin, one has to set $B_1 = 0$. The other three coefficients can be determined by enforcing the continuity of p and $\boldsymbol{n}\cdot\boldsymbol{u}$ across the interface $r = R_1$ and the homogeneous Dirichlet boundary condition at $r = R_2$. This leads to the following linear system:

$$\begin{bmatrix} 0 & \mathrm{J}_0(\kappa_2 R_2) & \mathrm{Y}_0(\kappa_2 R_2) \\ \mathrm{J}_0(\kappa_1 R_1) & -\mathrm{J}_0(\kappa_2 R_1) & -\mathrm{Y}_0(\kappa_2 R_1) \\ \mathrm{J}_1(\kappa_1 R_1)/\eta_1 & -\mathrm{J}_1(\kappa_2 R_1)/\eta_2 & -\mathrm{Y}_1(\kappa_2 R_1)/\eta_2 \end{bmatrix} \begin{bmatrix} A_1 \\ A_2 \\ B_2 \end{bmatrix} = \begin{bmatrix} \kappa_2^{-2} \\ \kappa_1^{-2} - \kappa_2^{-2} \\ 0 \end{bmatrix}.$$

Resonance phenomena occur whenever the values of $\kappa_1 R_1$ and $\kappa_2 R_2$ are such that the determinant of the matrix associated to the linear system is zero, namely the matrix is not invertible.

Once the exact pressure field p_{ref} has been determined, the exact velocity field $\boldsymbol{u}_{\mathrm{ref}}$ is retrieved thanks to Equation (4.35).

If the medium is homogeneous, then the computation of the resonance modes is straightforward. We first compute the eigenvalues by solving

$$\begin{cases} -\dfrac{d^2p}{dr^2} - \dfrac{1}{r}\dfrac{dp}{dr} - \lambda p = 0, & \text{in } \Omega, \\ p = 0, & \text{on } \partial\Omega. \end{cases}$$

Reasoning as before, we are led to the following solution

$$p(r) = A\mathrm{J}_0(\sqrt{\lambda}r),$$

and, by imposing the homogeneous Dirichlet boundary condition, we are led to the additional condition

$$p(R_2) = A\mathrm{J}_0(\sqrt{\lambda}R_2) = 0.$$

Since we are looking for the corresponding (non identically null) eigenfunction, we must prescribe $A \neq 0$, which implies $\mathrm{J}_0(\sqrt{\lambda}R_2) = 0$. Denoting by $\{\tilde{x}_i\}_i$ the sequence of zeros of the Bessel function J_0, the sequence of eigenvalues $\{\lambda_i\}_i$ are such that $\lambda_i = (\tilde{x}_i/R_2)^2$. The associated sequence of resonance wavenumbers $\{\kappa_i\}_i$ is such that $\kappa_i = \tilde{x}_i/R_2$. Here below, we display some zeros $\{\tilde{x}_i\}_{2\leq i\leq 10}$ of the Bessel function J_0 and the associated resonance wavenumbers $\{\kappa_i\}_{2\leq i\leq 10}$ picking $R_2 = 1/2$.

	2	3	4	5	6	7	8	9	10
$\tilde{x}$	5.5201	8.6537	11.7915	14.9309	18.0711	21.2116	24.3525	27.4935	30.6346
κ	11.0402	17.3075	23.5831	29.8618	36.1421	42.4233	48.7049	54.9870	61.2692

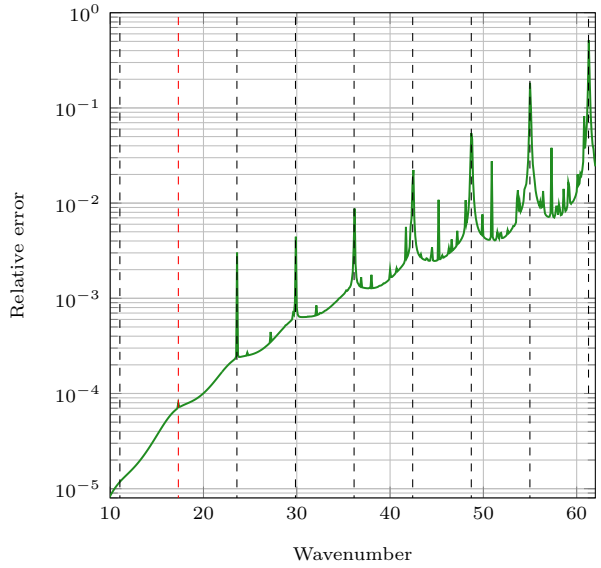


Figure 4.8: Relative error with respect to the wavenumber, the red dashed line corresponds to κ_3.

In Figure 4.8 it is showed the relative error with respect to the wavenumber which is let vary between 10 and 62, i.e. $\kappa \in [10, 62]$, computed by using a mesh of constant size $h = 1/20$. The closer to a resonance wavenumber is the wavenumber, the bigger is the relative error which is a manifestation of the numerical difficulty that arises in the context of resonance. Note that the smaller peaks are associated to the zeros of Bessel functions J_j of higher order $j > 0$.

Chapter 5

Investigation of a Trefftz discontinuous Galerkin method with absorbing boundary conditions for Helmholtz problems

In line with the overarching goal of this thesis of developing wave-specific numerical methods, this Chapter explores the combination of a wave-based Trefftz DG method with wave-tailored interface treatments, such as those we introduced in terms of transmission variables in the previous Chapters in the context of the CHDG method. The work reported in this Chapter was initiated during my MSc internship and it was continued during the PhD.

In particular, we describe a Trefftz method for the Helmholtz equation set in two-dimensional homogeneous media complemented by absorbing boundary conditions. It is based on a particular variational formulation of the problem called *ultra-weak variational formulation* (UWVF). The latter has the important property of presenting exclusively boundary integrals, except for possibly volume integrals associated to volume sources. Both trial and test discrete spaces are chosen to be spaces of propagative plane waves to embed the oscillating aspect that characterizes the problem in the numerical method and to improve the accuracy and efficiency. Lastly, both low-order ABCs and Padé-type high-order ABCs are exploited to model absorption and to test different approximations of the ABC. To the best of our knowledge, the combination of a Trefftz method with high-order ABCs represents a novelty that has never been studied before.

In Section 5.1 we introduce a model problem set in terms of the second-order formulation of the time-harmonic Helmholtz equation on a bounded computational domain complemented by absorbing boundary conditions. In Section 5.2 we define a general Trefftz approximation space, we introduce physically motivated interface conditions and we write the corresponding UWVF of the problem. In Section 5.3 we consider the approximation space of propagative plane waves, we give details about the coupling between these functions and the ABCs and we present the algebraic formulation of the discretized problem. In Section 5.4 we investigate the performance of the numerical scheme with several benchmarks.

The numerical results presented in Section 5.4 exhibit some unexpected issues. This Chapter is therefore intended as a first investigation and it is not a definitive analysis of these methods. Additional research effort is needed to clarify such issues and try to answer to open questions. In light of recent works, some directions of research on this topic will be given in the final Chapter of the thesis dedicated to conclusions and perspectives.

5.1 Problem

Let $\Omega \subset \mathbb{R}^2$ be a Lipschitz polytopal domain. We consider the following time-harmonic wave propagation problem set in terms of the Helmholtz equation:

$$\begin{cases} -\Delta p - \kappa^2 p = 0, & \text{in } \Omega, \\ \dfrac{\partial p}{\partial \boldsymbol{n}} = g, & \text{on } \Gamma^{\text{int}}, \\ \dfrac{\partial p}{\partial \boldsymbol{n}} = \mathcal{B} p, & \text{on } \Gamma^{\text{ext}}, \end{cases} \tag{5.1}$$

written for the unknown scalar field $p : \Omega \to \mathbb{C}$ corresponding to the acoustic pressure. Γ^{ext} stands for the external boundary of Ω, Γ^{int} for its complementary, so that $\Gamma := \partial\Omega = \Gamma^{\text{ext}} \cup \Gamma^{\text{int}}$, the vector field $\boldsymbol{n} : \partial\Omega \to \mathbb{R}^2$ is the unit outward (inward, respectively) normal to Ω along Γ^{ext} (Γ^{int}, respectively) (see Figure 5.1) and $\mathcal{B}$ is a suitable Dirichlet-to-Neumann (DtN) operator that permits to model absorption on Γ^{ext} (see Section 1.1.7). The wavenumber κ is assumed to be real and constant in Ω.

On the interior boundary, we have introduced the source term $g : \Gamma^{\text{int}} \to \mathbb{C}$ to define Neumann boundary condition. Note that it can be replaced by a different one, such as a Dirichlet or Robin boundary condition.

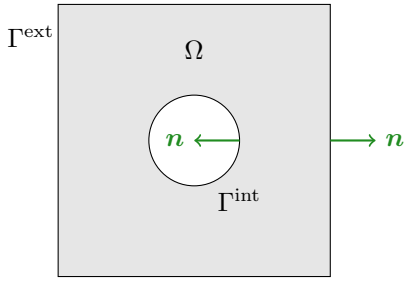


Figure 5.1: Example of a two-dimensional bounded computational domain for the scattering problem.

5.2 Ultra-weak variational formulation

The goal of this Section is to introduce the ultra-weak variational formulation of the problem. To do so, we first define a Trefftz space and interface conditions on the discrete solution.

5.2.1 Approximation space

To write the variational formulation of the problem, we define a Trefftz space $\mathcal{V}_h$ for the approximate solution and the test functions. As we already explained in Section 1.2.4, wave-based/Trefftz methods relies on the space of functions whose restriction on the element $K \in \mathcal{T}_h$ is a function v_K solving locally the homogeneous Helmholtz equation with no boundary condition imposed

$$\mathcal{V}_h = \prod_{K \in \mathcal{T}_h} \{v_K \in H^1(K) | -\Delta v_K - \kappa^2 v_K = 0\}.$$

5.2.2 Interface conditions

We impose the following interface conditions on the local solutions p_K defined in an element $K \in \mathcal{T}_h$ and $p_{K'}$ defined in its neighboring element $K' \in \mathcal{T}_h$

$$p_K = p_{K'}, \qquad \text{on } F, \tag{5.2}$$

$$\frac{\partial p_K}{\partial \boldsymbol{n}_{K,F}} + \frac{\partial p_{K'}}{\partial \boldsymbol{n}_{K',F}} = 0, \qquad \text{on } F. \tag{5.3}$$

Note that the previous conditions are to be understood in a point-wise sense and guarantee the continuity of the acoustic pressure and normal particle velocity, respectively, across the interfaces of the elements.

In Figure 5.2, we recall the notation used for two neighboring elements.

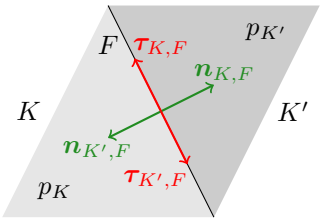


Figure 5.2: Notation for p_K and $p_{K'}$ defined on two neighboring elements K and K' that share the face F.

The interface conditions we have just introduced on the interfaces and the boundary conditions are variationally expressed in terms of outgoing and incoming traces that need to be defined. Let us proceed by introducing the concept of outgoing and incoming trace, hence we give the definition of outgoing operator $\mathcal{O}_{K,F}$ and incoming operator $\mathcal{I}_{K,F}$ that need to be applied to a function p_K in order to retrieve, respectively, its outgoing and incoming trace on the face F of the boundary ∂K of any element $K \in \mathcal{T}_h$

$$\mathcal{O}_{K,F} = \frac{1}{2}\left(\frac{1}{\imath\kappa}\frac{\partial}{\partial \boldsymbol{n}_{K,F}} + 1\right),$$
$$\mathcal{I}_{K,F} = \frac{1}{2}\left(-\frac{1}{\imath\kappa}\frac{\partial}{\partial \boldsymbol{n}_{K,F}} + 1\right).$$

These two operators require a minimum regularity of the local solution which they are applied to so that $\mathcal{O}_{K,F}p_K$, $\mathcal{I}_{K,F}p_K \in L^2(F)$ and all the integrals that will appear in the variational formulation are well defined. Thus, they are such that $\mathcal{O}_{K,F}p_K$, $\mathcal{I}_{K,F}p_K : L^2(F) \to L^2(F)$. Let us observe that conditions 5.2 and 5.3 can be equivalently written in terms of outgoing traces as

$$\mathcal{O}_{K,F}p_K = \mathcal{O}_{K,F}p_{K'},$$
$$\mathcal{O}_{K',F}p_K = \mathcal{O}_{K',F}p_{K'}.$$

In order to understand the meaning of the outgoing and incoming trace, we follow the example presented in [7], hence we consider a superposition of two plane waves $\alpha_K^+ \exp(+\imath\kappa\boldsymbol{x}\cdot\boldsymbol{n}_{K,F})$ and $\alpha_K^- \exp(-\imath\kappa\boldsymbol{x}\cdot\boldsymbol{n}_{K,F})$ propagating at normal incidence to an edge F of ∂K

$$p_K(\boldsymbol{x}) = \alpha_K^+ \exp(+\imath\kappa\boldsymbol{x}\cdot\boldsymbol{n}_{K,F}) + \alpha_K^- \exp(-\imath\kappa\boldsymbol{x}\cdot\boldsymbol{n}_{K,F}).$$

The two plane waves are respectively propagating towards the exterior and the interior of K. The outgoing and incoming traces

$$\mathcal{O}_{K,F}p_K = \frac{1}{2}\left(\frac{1}{\imath\kappa}\frac{\partial p_K}{\partial \boldsymbol{n}_{K,F}} + p_K\right),$$
$$\mathcal{I}_{K,F}p_K = \frac{1}{2}\left(-\frac{1}{\imath\kappa}\frac{\partial p_K}{\partial \boldsymbol{n}_{K,F}} + p_K\right),$$

exactly coincide with the trace of the outgoing wave and the one of the incoming wave through the edge F of K respectively. This fact justifies the terminology we use.

5.2.3 Variational formulation

In this Section, we state and prove the main theorem of this Chapter, namely the ultra-weak variational formulation of the problem. It will be the focus of discretization in Section 5.3.3.

Theorem 5.2.1 (Ultra-Weak Variational Formulation)**.** *Let Ω be a bounded domain and $\mathcal{T}_h$ one of its non-overlapping partitions composed by triangular elements K. Let the solution $p \in H^1(\Omega)$ of problem (5.1) be such that $\mathcal{B}p \in L^2(\Gamma^{ext})$ and $\partial p/\partial \boldsymbol{n} \in L^2(\partial K)$ for all $K \in \mathcal{T}_h$, let $g \in L^2(\partial\Gamma^{int})$ and let $\mathcal{V}_h$ as in (1.34). Then, $\forall q \in \mathcal{V}_h$, the solution p satisfies*

$$\begin{aligned}
&\sum_{K\in\mathcal{T}_h}\sum_{F\in\mathcal{F}_K}\langle\mathcal{O}_{K,F}p_K,\mathcal{O}_{K,F}q_K\rangle_F-\sum_{K\in\mathcal{T}_h}\sum_{F\in\mathcal{F}_K,F\subset\Gamma^{ext}}\langle-\frac{1}{2\imath\kappa}\mathcal{B}p_K,\mathcal{I}_{K,F}q_K\rangle_F\\
&\quad-\sum_{K\in\mathcal{T}_h}\sum_{F\in\mathcal{F}_K,F\not\subset\Gamma}\langle\mathcal{O}_{K',F}p_{K'},\mathcal{I}_{K,F}q_K\rangle_F-\sum_{K\in\mathcal{T}_h}\sum_{F\in\mathcal{F}_K,F\subset\Gamma}\langle\frac{1}{2}p_K,\mathcal{I}_{K,F}q_K\rangle_F \qquad (5.4)\\
&\quad=\sum_{K\in\mathcal{T}_h}\sum_{F\in\mathcal{F}_K,F\subset\Gamma^{int}}\langle-\frac{1}{2\imath\kappa}g,\mathcal{I}_{K,F}q_K\rangle_F.
\end{aligned}$$

Proof. Because p is a solution of problem (5.1) and $q \in \mathcal{V}_h$ by hypothesis, we have that

$$-\Delta p-\kappa^2 p=0, \qquad \text{in } \Omega, \qquad (5.5)$$

$$-\Delta\overline{q}_K-\kappa^2\overline{q}_K=0, \qquad \text{in } K, \qquad (5.6)$$

for all $K \in \mathcal{T}_h$ and for all $q \in \mathcal{V}_h$. We multiply Equation (5.5) by $\overline{q}$ and Equation (5.6) by p on K, we integrate the resulting equations by parts and we get

$$\begin{aligned}
-(\nabla p_K,\nabla q_K)_K+\kappa^2(p_K,q_K)_K&=-\sum_{F\in\mathcal{F}_K}\langle\frac{\partial p_K}{\partial\boldsymbol{n}_{K,F}},q_K\rangle_F,\\
-(\nabla p_K,\nabla q_K)_K+\kappa^2(p_K,q_K)_K&=-\sum_{F\in\mathcal{F}_K}\langle p_K,\frac{\partial q_K}{\partial\boldsymbol{n}_{K,F}}\rangle_F,
\end{aligned}$$

that gives

$$\sum_{F\in\mathcal{F}_K}\left(\langle p_K,\frac{\partial q_K}{\partial\boldsymbol{n}_{K,F}}\rangle_F-\langle\frac{\partial p_K}{\partial\boldsymbol{n}_{K,F}},q_K\rangle_F\right)=0. \qquad (5.7)$$

Moreover, we have

$$\begin{aligned}
&\sum_{F\in\mathcal{F}_K}\langle\frac{1}{2}\left(\frac{1}{\imath\kappa}\frac{\partial p_K}{\partial\boldsymbol{n}_{K,F}}+p_K\right),\frac{1}{2}\left(\frac{1}{\imath\kappa}\frac{\partial q_K}{\partial\boldsymbol{n}_{K,F}}+q_K\right)\rangle_F\\
-&\sum_{F\in\mathcal{F}_K}\langle\frac{1}{2}\left(-\frac{1}{\imath\kappa}\frac{\partial p_K}{\partial\boldsymbol{n}_{K,F}}+p_K\right),\frac{1}{2}\left(-\frac{1}{\imath\kappa}\frac{\partial q_K}{\partial\boldsymbol{n}_{K,F}}+q_K\right)\rangle_F\\
&=-\frac{1}{2\imath\kappa}\sum_{F\in\mathcal{F}_K}\left(\langle p_K,\frac{\partial q_K}{\partial\boldsymbol{n}_{K,F}}\rangle_F-\langle\frac{\partial p_K}{\partial\boldsymbol{n}_{K,F}},q_K\rangle_F\right)=0,
\end{aligned}$$

where the last step is an immediate consequence of Equation (5.7).

The continuity of p on F and the boundary conditions of system (5.1) write

$$\begin{aligned}
\frac{1}{2}\left(-\frac{1}{\imath\kappa}\frac{\partial}{\partial\boldsymbol{n}_{K,F}}+1\right)p_K&=\frac{1}{2}\left(\frac{1}{\imath\kappa}\frac{\partial}{\partial\boldsymbol{n}_{K',F}}+1\right)p_{K'}, && \text{on } F\not\subset\Gamma,\\
\frac{\partial p_K}{\partial\boldsymbol{n}_{K,F}}&=\mathcal{B}p_K, && \text{on } F\subset\Gamma^{\text{ext}},\\
\frac{\partial p_K}{\partial\boldsymbol{n}_{K,F}}&=g, && \text{on } F\subset\Gamma^{\text{int}}.
\end{aligned}$$

By summing all over the elements, recalling the definition of $\mathcal{O}_{K,F}$ and $\mathcal{I}_{K,F}$ and taking into account the previous relations, we get the result. □

Remark 5.2.2. The formulation we have just presented is a generalisation of the one shown in [24], indeed the DtN operator that we have exploited is a generalization of $\mathcal{B} = \imath\kappa$.

5.3 Plane-wave discretization

The goal of this Section is to provide the discrete form of the ultra-weak variational formulation presented in the previous Section. To do so, we give details about the Trefftz space $\mathcal{T}_h$ which will consist of propagative plane waves, we carefully present the discrete realizations of the different approximations of the ABCs that we want to compare and we eventually give details about the resulting discrete linear system.

5.3.1 Propagative plane-wave shape functions

With the notation from [7], we apply the Galerkin wave-based method considering a propagative plane-wave basis, that is

$$\{v_p\}_{1\leq p\leq \mathrm{p}}, \quad v_p(\boldsymbol{x}) = \exp(\imath\kappa \boldsymbol{d}_p \cdot \boldsymbol{x}), \tag{5.8}$$

where $\boldsymbol{d}_p$ with $p = 1, ..., \mathrm{p}$ belongs to a finite subset

$$\mathbb{S}_1^{\mathrm{p}} = \{\boldsymbol{d}_1, ..., \boldsymbol{d}_{\mathrm{p}}\},$$

consisting of p distinct directions of the unit circle $\mathbb{S}_1$ of $\mathbb{R}^2$ and $\boldsymbol{x}$ is a generic point of the plane. The parameter p refers to the number of plane waves used in each $K \in \mathcal{T}_h$. A priori it can be different for each element K, but we use a fixed number of plane waves p for each element K of the triangulation $\mathcal{T}_h$. Thus, the solution in K has the following discretized form

$$p_K^{\mathrm{p}}(\boldsymbol{x}) = \sum_{p=1}^{\mathrm{p}} \alpha_{K,p} v_p(\boldsymbol{x}) \tag{5.9}$$

where $\alpha_{K,p} \in \mathbb{C}$ are the coefficients of the linear combination of the plane-wave basis elements.

It is important to highlight once again the fact that on each element K of the triangulation $\mathcal{T}_h$ we can define a different set of shape functions, which means that the cardinality p of the set may be different, as well as the way the directions are distributed over $\mathbb{S}_2$. This is an important aspect whose potential will be partially exploited in some of the numerical simulations. Indeed, it is possible to exploit this property to select element wise the best set of shape functions according to criteria that need to be chosen with respect to the properties of the numerical solution itself and the problem we aim at numerically solving.

Finally, for the moment, we consider a system of directions for the plane-wave basis uniformly distributed over $\mathbb{S}_2$

$$\boldsymbol{d}_p = \begin{pmatrix} \cos\theta_p \\ \sin\theta_p \end{pmatrix}, \qquad \theta_p = \theta_1 + \frac{2\pi(p-1)}{\mathrm{p}}, \qquad p = 1, ..., \mathrm{p}, \tag{5.10}$$

where p denotes the number of directions. The first angle θ_1 is called *shift*. An example is shown in Figure 5.3.

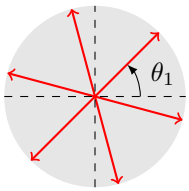


Figure 5.3: Example of $\mathrm{p} = 6$ uniformly spaced directions, with $\theta_1 = \pi/4$.

5.3.2 Absorbing boundary conditions treatment

Discretizing formulation (5.4) requires the implementation of the operator $\mathcal{B}$. We denote by $\mathcal{B}_*$ an approximation of the operator $\mathcal{B}$ and here we present some of these approximations (see Section 1.1.7) that we will compare from a numerical point of view later in Section 5.4.

We first consider the low-order approximation of the operator $\mathcal{B}$ that we denote "$\imath\kappa$". It simply consists in the zeroth-order ABC (1.21), namely

$$\mathcal{B}_* = \mathcal{B}_{\imath\kappa} = \imath\kappa. \tag{5.11}$$

A fundamental aspect that characterizes the use of plane-waves instead of polynomials in the Trefftz method is that it allows to avoid the issue concerning the approximation of the operator $\mathcal{B}$: indeed, it is possible to exactly compute its application to plane waves. Let us recall (1.19) and (1.20) which have to be verified by a plane wave $p(\boldsymbol{x}) = \exp(\imath\kappa\boldsymbol{\varphi}\cdot\boldsymbol{x})$.

$$\frac{\partial p}{\partial \boldsymbol{n}} = \nabla p \cdot \boldsymbol{n} = \imath\kappa(\boldsymbol{\varphi}\cdot\boldsymbol{n})p,$$
$$\mathcal{B}p = \imath\kappa\sqrt{1+\frac{\Delta_\Gamma}{\kappa^2}}p,$$

whose comparison let us retrieve that

$$\sqrt{1+\frac{\Delta_\Gamma}{\kappa^2}} = \boldsymbol{\varphi}\cdot\boldsymbol{n}.$$

Because the left hand side of the previous equation is positive, we deduce that the latter, in order to hold true, needs to present a positive right hand side as well, $\boldsymbol{\varphi}\cdot\boldsymbol{n} \geq 0$, which is satisfied whenever the plane wave is outgoing. The previous consideration turns out to be crucial because an ABC is prescribed onto the portions of the boundary $\partial\Omega$ for which there is an a priori knowledge concerning the direction of the exact solution, which specifically needs to be outgoing. Therefore, in terms of discretizion, an ABC is equivalent to prescribing a non homogeneous Neumann boundary condition that holds true only for those plane waves of the basis that are outgoing. Further manipulating the equation, we get the final expression of the Laplace-Beltrami operator as application to plane waves $p_p(\boldsymbol{x}) = \exp(\imath\kappa\boldsymbol{d}_p\cdot\boldsymbol{x})$, $p = 1, ..., \mathrm{p}$, that is

$$\Delta_\Gamma = -\kappa^2(1-(\boldsymbol{d}_p\cdot\boldsymbol{n})^2) = -\kappa^2(\boldsymbol{d}_p\cdot\boldsymbol{\tau})^2, \tag{5.12}$$

where $\boldsymbol{\tau}$ is the tangent unit vector (see Figure 5.2).

Exploiting the expression of the discretized Laplace-Beltrami operator (5.12), we have

$$\sqrt{1+\frac{\Delta_\Gamma}{\kappa^2}} = |\boldsymbol{d}_p\cdot\boldsymbol{n}|. \tag{5.13}$$

By using (5.13), we obtain the final form (5.14) of the second approximation of the operator $\mathcal{B}$ that we denote "sqrt"

$$\mathcal{B}_* = \mathcal{B}_{\mathrm{sqrt}} = \imath\kappa|\boldsymbol{d}_p\cdot\boldsymbol{n}|. \tag{5.14}$$

To obtain a high-order Padé-type approximation of the operator $\mathcal{B}$, we apply the Padé pseudo-differential operator (1.24) to the unknown local solution (5.9) and we have

$$\begin{aligned}\frac{\partial p_K^{\mathrm{p}}}{\partial \boldsymbol{n}_{K,F}} = \mathcal{B}_{\mathrm{Padé}}p_K^{\mathrm{p}} &= \imath\kappa e^{\imath\phi/2}\left[1+\frac{2}{M}\sum_{n=1}^{N}c_n\left(1-\frac{e^{\imath\phi}(c_n+1)}{e^{\imath\phi}c_n+1+\Delta_\Gamma/\kappa^2}\right)\right]p_K^{\mathrm{p}}\\ &= \imath\kappa e^{\imath\phi/2}\left(\sum_{p=1}^{\mathrm{p}}\alpha_{K,p}v_p+\frac{2}{M}\sum_{n=1}^{N}c_n\left(\sum_{p=1}^{\mathrm{p}}\alpha_{K,p}v_p-\tilde{p}\right)\right)\end{aligned}$$

where $\tilde{p}$ needs to be defined implicitly (because of the presence of the Laplace-Beltrami operator in the denominator) as follows

$$(e^{\imath\phi}c_n + 1 + \Delta_\Gamma/\kappa^2)\tilde{p} = e^{\imath\phi}(c_n+1)\sum_{p=1}^{\mathrm{p}} \alpha_{K,p} v_p. \tag{5.15}$$

Let us observe that

$$\Delta_\Gamma = \Delta - \partial_{\boldsymbol{nn}} = \partial_{\boldsymbol{\tau\tau}},$$

where $\partial_{\boldsymbol{nn}}$ and $\partial_{\boldsymbol{\tau\tau}}$ denote the second normal and tangent derivatives, respectively, and that

$$\begin{aligned}\Delta_\Gamma p_K^{\mathrm{p}}(\boldsymbol{x}) = \Delta_\Gamma \sum_{p=1}^{\mathrm{p}} \alpha_{K,p} v_p(\boldsymbol{x}) &= \sum_{p=1}^{\mathrm{p}} \alpha_{K,p} \partial_{\boldsymbol{\tau\tau}} v_p(\boldsymbol{x}) \\ &= -\kappa^2 \sum_{p=1}^{\mathrm{p}} \alpha_{K,p} (\boldsymbol{d}_p \cdot \boldsymbol{\tau})^2 v_p(\boldsymbol{x}) \\ &= -\kappa^2 \sum_{p=1}^{\mathrm{p}} \alpha_{K,p} (1 - (\boldsymbol{d}_p \cdot \boldsymbol{n})^2) v_p(\boldsymbol{x}).\end{aligned}$$

Given the differential equation (5.15), $\tilde{p}$ must have the form

$$\tilde{p}(\boldsymbol{x}) = \sum_{p=1}^{\mathrm{p}} \tilde{p}_p \alpha_{K,p} v_p(\boldsymbol{x}),$$

so we have

$$\sum_{p=1}^{\mathrm{p}} (e^{\imath\phi}c_n + (\boldsymbol{d}_p \cdot \boldsymbol{n})^2)\tilde{p}_p \alpha_{K,p} v_p(\boldsymbol{x}) = \sum_{p=1}^{\mathrm{p}} e^{\imath\phi}(c_n+1)\alpha_{K,p} v_p(\boldsymbol{x}),$$

$$\tilde{p}_p = \frac{e^{\imath\phi}(c_n+1)}{e^{\imath\phi}c_n + (\boldsymbol{d}_p \cdot \boldsymbol{n})^2},$$

and therefore

$$\mathcal{B}_* = \mathcal{B}_{\text{Padé}} = \imath\kappa e^{\imath\phi/2}\left(1 + \frac{2}{M}\sum_{n=1}^{N} c_n \frac{(\boldsymbol{d}_p \cdot \boldsymbol{n})^2 - e^{\imath\phi}}{e^{\imath\phi}c_n + (\boldsymbol{d}_p \cdot \boldsymbol{n})^2}\right). \tag{5.16}$$

5.3.3 Algebraic formulation

Formulation (5.4) is associated to the following algebraic linear system

$$(\mathbf{M} - \mathbf{E})\boldsymbol{\alpha} = \mathbf{c}. \tag{5.17}$$

In order to give an explicit form to the matrices $\mathbf{M}$, $\mathbf{E}$ and the vector $\mathbf{c}$, we consider their blocks related to an element K and summing on all the elements we get the final result. We take into account each integral that appears in (5.4) and rewrite it in terms of plane-wave discretization thanks to (5.9). In the following we denote by v_q with $q = 1, ..., \mathrm{p}$ the elements of the plane-wave basis.

Each outgoing $\mathcal{O}_{K,F}p_K^{\mathrm{p}}$ and incoming trace $\mathcal{I}_{K,F}p_K^{\mathrm{p}}$ is approximated by

$$\mathcal{O}_{K,F}p_K^{\mathrm{p}}(\boldsymbol{x}) = \sum_{p=1}^{\mathrm{p}} \alpha_{K,p} \frac{1}{2}\left(\frac{1}{\imath\kappa}\frac{\partial}{\partial \boldsymbol{n}_{K,F}} + 1\right) v_p(\boldsymbol{x}),$$
$$\mathcal{I}_{K,F}p_K^{\mathrm{p}}(\boldsymbol{x}) = \sum_{p=1}^{\mathrm{p}} \alpha_{K,p} \frac{1}{2}\left(-\frac{1}{\imath\kappa}\frac{\partial}{\partial \boldsymbol{n}_{K,F}} + 1\right) v_p(\boldsymbol{x}).$$

The previous expressions can be written in matrix form as

$$\mathcal{O}_{K,F}p_K^{\mathrm{p}}(\boldsymbol{x}) = \begin{bmatrix}\Lambda_{K,1}^{+,F} v_1(\boldsymbol{x}) & \dots & \Lambda_{K,\mathrm{p}}^{+,F} v_{\mathrm{p}}(\boldsymbol{x})\end{bmatrix} \begin{bmatrix}\alpha_{K,1}\\ \vdots \\ \alpha_{K,\mathrm{p}}\end{bmatrix},$$
$$\mathcal{I}_{K,F}p_K^{\mathrm{p}}(\boldsymbol{x}) = \begin{bmatrix}\Lambda_{K,1}^{-,F} v_1(\boldsymbol{x}) & \dots & \Lambda_{K,\mathrm{p}}^{-,F} v_{\mathrm{p}}(\boldsymbol{x})\end{bmatrix} \begin{bmatrix}\alpha_{K,1}\\ \vdots \\ \alpha_{K,\mathrm{p}}\end{bmatrix},$$

$$\Lambda_{K,p}^{\pm,F} := \frac{1}{2}\left(1 \pm \boldsymbol{d}_p \cdot \boldsymbol{n}_{K,F}\right).$$

Let us recall that we need to take into account the boundary condition prescribed on Γ_{int}. In (5.1), it consisted in a Neumann boundary condition, hence we develop the computation for this case. Its extension to Dirichlet or Robin boundary condition is straightforward.

To simplify the presentation, we adopt a lighter notation: we denote by K_i the $i-$th element of the triangulation $\mathcal{T}_h$ composed by K elements and we no longer write $\boldsymbol{n}_{K,F}$, but simply $\boldsymbol{n}_{i,F}$ to refer to the outward unit normal to the face F of an element K_i that, if written as a subscript, also becomes simply i.

Proposition 5.3.1 (Algebraic formulation)**.** *The matrices* M, $\mathsf{E} := \mathsf{E}_1 + \mathsf{E}_2 + \mathsf{E}_3$ *and the vector* c *of the linear system* (5.17) *associated to the plane wave discretization of* (5.4) *read*

$$\mathsf{M}_{ii,pq} = \sum_{F\in\mathcal{F}_{K_i}} \Lambda_{i,q}^{+,F}\Lambda_{i,p}^{+,F}\langle v_p, v_q\rangle_F,$$
$$\mathsf{E}_{1,ij,pq} = \sum_{F\in\mathcal{F}_{K_i}\cap\mathcal{F}_{K_j}} \Lambda_{i,q}^{-,F}\Lambda_{j,p}^{+,F}\langle v_p, v_q\rangle_F,$$
$$\mathsf{E}_{2,ii,pq} = -\frac{1}{2\imath\kappa}\sum_{F\in\mathcal{F}_{K_i}\cap\Gamma^{ext}} \Lambda_{i,q}^{-,F}\mathcal{B}_*\langle v_p, v_q\rangle_F,$$
$$\mathsf{E}_{3,ii,pq} = \frac{1}{2}\sum_{F\in\mathcal{F}_{K_i}\cap\Gamma} \Lambda_{i,q}^{-,F}\langle v_p, v_q\rangle_F,$$
$$\mathsf{c}_{i,q} = -\frac{1}{2\imath\kappa}\sum_{F\in\mathcal{F}_{K_i}\cap\Gamma^{int}} \Lambda_{i,q}^{-,F}\langle g, v_q\rangle_F,$$

with $\mathcal{B}_*$ *given by* (5.11), (5.14), (5.16) *depending on the chosen approximation of* $\mathcal{B}$ *and where the indices* $i, j = 1, ..., \mathrm{K}$ *identify the block related to the boundary of an element* K_i *or the face shared by two elements* K_i, K_j *and the indices* $p, q = 1, ..., \mathrm{p}$ *identify one of its elements.*

Proof. We are going to consider each integral that appears in (5.4).

Matrix $\mathbf{M}$: Let us start with the first integral whose manipulation consists in the matrix $\mathbf{M}$.

$$\sum_{F\in\mathcal{F}_{K_i}} \langle \frac{1}{2}\left(\frac{1}{\imath\kappa}\frac{\partial p_i^{\mathrm{p}}}{\partial \boldsymbol{n}_i} + p_i^{\mathrm{p}}\right), \frac{1}{2}\left(\frac{1}{\imath\kappa}\frac{\partial v_q}{\partial \boldsymbol{n}_i} + v_q\right)\rangle_F = \sum_{F\in\mathcal{F}_{K_i}} \Lambda_{i,q}^{+,F} \sum_{p=1}^{\mathrm{p}} \Lambda_{i,p}^{+,F} \alpha_{i,p} \langle v_p, v_q\rangle_F.$$

Since $\mathbf{M}$ is a block-diagonal matrix, we can use three indices i,p,q instead of four to represent each of its elements as

$$\mathbf{M}_{ii,pq} = \sum_{F\in\mathcal{F}_{K_i}} \Lambda_{i,q}^{+,F} \Lambda_{i,p}^{+,F} \langle v_p, v_q\rangle_F.$$

Matrix $\mathbf{E}_1$: Let us proceed with the third integral whose discrete evaluation appears as a part of the matrix $\mathbf{E}$, that we denote by $\mathbf{E}_1$.

$$\sum_{F\in\mathcal{F}_{K_i}\cap\mathcal{F}_{K_j}} \langle \frac{1}{2}\left(\frac{1}{\imath\kappa}\frac{\partial p_j^{\mathrm{p}}}{\partial \boldsymbol{n}_j} + p_j^{\mathrm{p}}\right), \frac{1}{2}\left(-\frac{1}{\imath\kappa}\frac{\partial v_q}{\partial \boldsymbol{n}_i} + v_q\right)\rangle_F = \sum_{F\in\mathcal{F}_{K_i}\cap\mathcal{F}_{K_j}} \Lambda_{i,q}^{-,F} \sum_{p=1}^{\mathrm{p}} \Lambda_{j,p}^{+,F} \alpha_{j,p} \langle v_p, v_q\rangle_F.$$

Since $\mathbf{E}_1$ is not a block-diagonal matrix, we need to use all four indices i,j,p,q to represent each of its elements as

$$\mathbf{E}_{1,ij,pq} = \sum_{F\in\mathcal{F}_{K_i}\cap\mathcal{F}_{K_j}} \Lambda_{i,q}^{-,F} \Lambda_{j,p}^{+,F} \langle v_p, v_q\rangle_F.$$

Matrix $\mathbf{E}_2$: Let us continue with the second integral whose discrete evaluation appears as a part of the matrix $\mathbf{E}$, that we denote by $\mathbf{E}_2$. For the numerical scheme, $\mathcal{B}_*$ is given by one of the following (5.11), (5.14), (5.16) approximations of $\mathcal{B}$.

$$\sum_{F\in\mathcal{F}_{K_i}\cap\Gamma^{\mathrm{ext}}} \langle -\frac{1}{2\imath\kappa}\mathcal{B}_* p_i^{\mathrm{p}}, \frac{1}{2}\left(-\frac{1}{\imath\kappa}\frac{\partial v_q}{\partial \boldsymbol{n}} + v_q\right)\rangle_F = -\frac{1}{2\imath\kappa} \sum_{F\in\mathcal{F}_{K_i}\cap\Gamma^{\mathrm{ext}}} \Lambda_{i,q}^{-} \sum_{p=1}^{\mathrm{p}} \alpha_{i,p} \mathcal{B}_* \langle v_p, v_q\rangle_F.$$

Thus, each element of the block-diagonal matrix $\mathbf{E}_2$ can be expressed as

$$\mathbf{E}_{2,ii,pq} = -\frac{1}{2\imath\kappa} \sum_{F\in\mathcal{F}_{K_i}\cap\Gamma^{\mathrm{ext}}} \Lambda_{i,q}^{-,F} \mathcal{B}_* \langle v_p, v_q\rangle_F.$$

Matrix $\mathbf{E}_3$: It corresponds to the last contribution that completes the matrix $\mathbf{E}$, consists in the following plane wave discretization of the fourth integral

$$\sum_{F\in\mathcal{F}_{K_i}\cap\Gamma} \langle \frac{1}{2} p_i^{\mathrm{p}}, \frac{1}{2}\left(-\frac{1}{\imath\kappa}\frac{\partial v_q}{\partial \boldsymbol{n}_{i,F}} + v_q\right)\rangle_F = \frac{1}{2} \sum_{F\in\mathcal{F}_{K_i}\cap\Gamma} \Lambda_{i,q}^{-,F} \sum_{p=1}^{\mathrm{p}} \alpha_{i,p} \langle v_p, v_q\rangle_F.$$

Hence, each element of the block-diagonal matrix $\mathbf{E}_3$ reads

$$\mathbf{E}_{3,ii,pq} = \frac{1}{2} \sum_{F\in\mathcal{F}_{K_i}\cap\Gamma} \Lambda_{i,q}^{-,F} \langle v_p, v_q\rangle_F.$$

Vector $\mathbf{c}$: Finally, vector $\mathbf{c}$ consists in the following plane-wave discretization of the right-hand-side of (5.4)

$$-\frac{1}{2\imath\kappa} \sum_{F\in\mathcal{F}_{K_i}\cap\Gamma^{\mathrm{int}}} \int_F g \overline{\frac{1}{2}\left(-\frac{1}{\imath\kappa}\frac{\partial v_q}{\partial \boldsymbol{n}_{i,F}} + v_q\right)} = -\frac{1}{2\imath\kappa} \sum_{F\in\mathcal{F}_{K_i}\cap\Gamma^{\mathrm{int}}} \Lambda_{i,q}^{-,F} \langle g, v_q\rangle_F = \mathbf{c}_{i,q}.$$

□

Let us observe again that $\mathbf{M}$ is a block-diagonal matrix, whose diagonal blocks are the matrices $\mathbf{M}_i$ related to the elements $K_i \in \mathcal{T}_h$. Conversely, although $\mathbf{E}_2$ and $\mathbf{E}_3$ are block-diagonal matrices, $\mathbf{E}$ can not be block-diagonal, since $\mathbf{E}_1$ is not block-diagonal, because of the presence of non-null blocks related to adjacent elements to each K_i.

Let us remark that the matrices $\mathbf{M}$ and $\mathbf{E}$ have dimension $(\mathrm{pK}, \mathrm{pK})$ and the vector $\mathbf{c}$ has dimension pK, where p is the number of basis functions and K is the number of elements of the mesh.

Let un conclude with a compact formula for the Hermitian form $\langle v_p, v_q \rangle_F$ computed along the face F with vertices $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$:

$$\langle v_p, v_q \rangle_F = \begin{cases} \dfrac{\exp(\imath\kappa(\boldsymbol{d}_p - \boldsymbol{d}_q)\cdot\boldsymbol{x}_2) - \exp(\imath\kappa(\boldsymbol{d}_p - \boldsymbol{d}_q)\cdot\boldsymbol{x}_1)}{\imath\kappa(\boldsymbol{d}_p - \boldsymbol{d}_q)\cdot(\boldsymbol{x}_2 - \boldsymbol{x}_1)}|F| & \text{if } (\boldsymbol{d}_p - \boldsymbol{d}_q)\cdot(\boldsymbol{x}_2 - \boldsymbol{x}_1) \neq 0, \\ |F| & \text{if } (\boldsymbol{d}_p - \boldsymbol{d}_q)\cdot(\boldsymbol{x}_2 - \boldsymbol{x}_1) = 0, \end{cases}$$

where $|F|$ denotes the length of the face F.

5.4 Numerical results

In this Section, we aim at investigating the influence of the ABC on the accuracy of the numerical method. In particular, we are interested in testing the different approximations of the ABC we presented in Section 5.3.2 and the propagative plane waves that compose the approximation space.

In the first benchmark (Section 5.4.1) we consider the propagation of a plane wave in a square domain that corresponds to the unique element of the mesh. In the second benchmark (Section 5.4.2) we deal with the scattering of a plane wave by a circle. The third benchmark (Section 5.4.3) is an illustration of the scattering of a plane wave by an obstacle with complicated boundary.

The numerical simulations and the visualization are performed with a dedicated MATLAB code that was implemented from scratch. The mesh generation is performed with gmsh and h is the element size specified in gmsh.

The global FEM system $(\mathbf{M} - \mathbf{E})\boldsymbol{\alpha} = \mathbf{c}$ is solved in MATLAB by using the backslash operator, i.e. $\boldsymbol{\alpha} = (\mathbf{M} - \mathbf{E}) \setminus \mathbf{c}$.

We consider a relative numerical error based on the L^2-norm of the pressure field. It is defined as

$$\text{relative error} := \frac{\|p_h - p_{\text{ref}}\|_{L^2(\Omega)}}{\|p_{\text{ref}}\|_{L^2(\Omega)}},$$

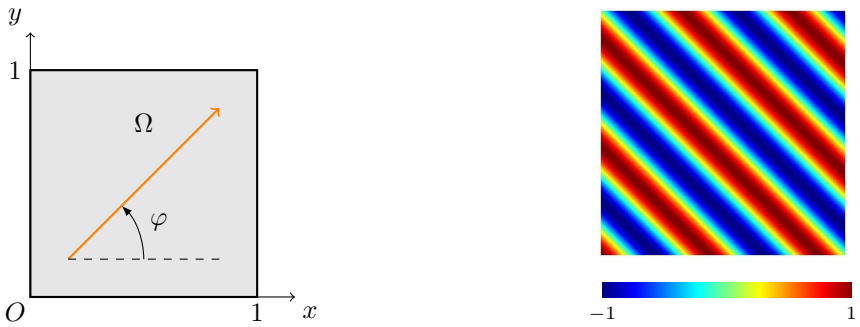
where p_h is the numerical solution and p_{ref} is the reference solution.

5.4.1 Preliminary numerical study

The first benchmark consists in the propagation of a plane wave. We consider the square domain $\Omega = (0,1)^2$ and we denote by $\boldsymbol{\varphi} = (\cos\varphi, \sin\varphi)$ with $\varphi \in [0, \pi/2)$ the direction of propagation (see Figure 5.4a).

The reference solution reads (see Figure 5.4b)

$$p_{\text{ref}}(\boldsymbol{x}) = \exp\left(\imath\kappa \boldsymbol{x}\cdot\boldsymbol{\varphi}\right).$$



(a) Square domain.

(b) Real part of p_{ref} for $\varphi = \pi/4$ and $\kappa = 20$.

Figure 5.4: Domain and reference solution for the propagative plane wave benchmark.

We prescribe a non-homogeneous Neumann boundary condition on the left and bottom edges by exploiting the reference solution and we consider two configurations depending on the boundary condition imposed on the right and top edges: a non-homogeneous Neumann boundary condition (again by exploiting the reference solution) (see Figure 5.5i) and an absorbing boundary condition (see Figure 5.5ii), i.e.

$$\begin{cases} \dfrac{\partial p}{\partial \boldsymbol{n}} = \dfrac{\partial p_{\text{ref}}}{\partial \boldsymbol{n}}, & \text{on } \Gamma_{\text{N}}, \\ \dfrac{\partial p}{\partial \boldsymbol{n}} = \mathcal{B}_* p, & \text{on } \Gamma_{\text{ABC}}. \end{cases}$$

For the ABC, we consider the expressions of the operator $\mathcal{B}_*$ presented in Section 5.3.2.

(i) Case i

(ii) Case ii

Figure 5.5: Types of boundary ($\Gamma_{\text{N}} = -$, $\Gamma_{\text{ABC}} = -$) in two cases.

For the numerical simulations, we consider a single element of the mesh. We fix the propagation angle $\varphi = \pi/4$ and the wavenumber $\kappa = 20$. In the expression of $\mathcal{B}_{\text{Padé}}$ (5.16) we take a null angle of rotation of the branch cut, that is $\phi = 0$. Even though in other studies the choice of this angle had a meaningful impact on the numerical performances (see e.g. [98]), we have performed some tests which did not exhibit a significant sensitivity on this parameter.

We consider two sets of shape functions (see Figure 5.6) composed by p plane waves (5.8) whose directions of propagation $\boldsymbol{d}_p = (\cos\theta_p, \sin\theta_p)^T$, $p = 1, ..., \mathrm{p}$ are uniformly spaced in the whole plane and in a quarter plane, respectively:

- "full" basis: $0 \leq \theta_p \leq 2\pi$

$$\theta_p = \pi\frac{\sqrt{2}}{2} + \frac{2\pi(p-1)}{\mathrm{p}}, \qquad p = 1, ..., \mathrm{p},$$

  that is (5.10) with $\theta_1 = \pi\frac{\sqrt{2}}{2}$ (so that no plane-wave direction corresponds to the one of the reference solution);

- "quarter" basis: $0 \leq \theta_p \leq \pi/2$

$$\theta_p = \frac{\pi(p-1)}{2(\mathrm{p}-1)}, \qquad p = 1, ..., \mathrm{p}.$$

(a) "Full" basis. (b) "Quarter" basis.

Figure 5.6: Shape functions for the propagative plane wave benchmark with $\mathrm{p} = 4$.

The use of the "quarter" basis is motivated by a principle of directional adaptivity. Indeed, the propagation angle of the reference plane wave being $\varphi \in (0, \pi/2) \subset (0, 2\pi)$, we can intuitively claim that only the plane waves whose direction belongs to the first quarter of the plane are useful. Consequently, a higher density of these directions in a neighborhood of the unknown direction $\boldsymbol{\varphi}$ should lead to a better approximation.

Besides comparing the results obtained in the different configurations, we also consider the L^2-projection $P_h p_{\mathrm{ref}}$ of the reference solution onto the approximation space by solving the following linear system (normal equation) for $\{\alpha_p\}_{p=1,...,\mathrm{p}}$

$$\sum_{p=1}^{\mathrm{p}} \alpha_p (v_p, v_q)_\Omega = (p_{\mathrm{ref}}, v_q)_\Omega, \quad q = 1, ..., \mathrm{p}.$$

Once the coefficients α_p, $p = 1, ..., \mathrm{p}$ of the projected function $P_h p_{\mathrm{ref}} = \sum_{p=1}^{\mathrm{p}} \alpha_p v_p$ are determined, we compute the relative L^2-error, which corresponds to the best approximation error,

$$\text{best approximation error} = \frac{||p_{\mathrm{ref}} - P_h p_{\mathrm{ref}}||_{L^2(\Omega)}}{||p_{\mathrm{ref}}||_{L^2(\Omega)}}.$$

For any configuration, we perform a p-refinement letting p vary from 1 to 80. The results with the relative error and the condition number of the matrix $\mathbf{M}-\mathbf{E}$ are showed in Figures 5.7-5.8-5.9-5.10.

Note that in the following the "$\imath\kappa$" ABC is equivalent to the Padé ABC with $N = 0$ and $\phi = 0$ (in the Figures they correspond to the red lines).

The following observations regarding the results obtained with a "full" basis and Neumann boundary condition, "$\imath\kappa$" and "sqrt" ABC can be made:

- In Figure 5.7a, we can see that after a certain threshold $\mathrm{p}_{\mathrm{MAX}} \approx 60$ results are not accurate anymore. This is due to the poor conditioning of the matrix $\mathbf{M}-\mathbf{E}$ associated to the linear system (5.17). Indeed, in Figure 5.7b, we see that, after an initial slow increase of the condition number, it then grows faster until the matrix becomes ill-conditioned around the same threshold $\mathrm{p}_{\mathrm{MAX}}$. The conditioning strongly depends on p, while the two types of ABC provide similar results.
- The numerical error associated to the Neumann case is close to the best approximation one. Indeed, in such a case, the boundary condition is exact (in the sense that it is verified by the exact solution) and the unique source of error is the numerical one associated to the plane-wave discretization.
- The "$\imath\kappa$" case is the one that presents the biggest error, since it also incorporates the source of error associated to the approximate model of the ABC.
- The "sqrt" case has an error that is intermediate between the previous two. However, we expected it to be closer to the one of the Neumann case since the ABC is exact with no approximation.

Concerning the results obtained with a "full" basis and Padé-type ABC, we can make the following remarks:

- In Figure 5.8a, we see that increasing the order N of the Padé approximation, we are able to retrieve a smaller and smaller error until $N = 3$. For $N > 3$ the error does not decrease anymore even though it tends towards the one associated to the "sqrt" case by increasing N. This fact is unexpected: Padé being an approximation of the exact formulation of the ABC, it should lead to a bigger error that gets smaller by increasing N and that eventually converges to the one of the "sqrt" case.
- The condition number increases in a very similar way for all orders N of the Padé approximation, see Figure 5.8b.

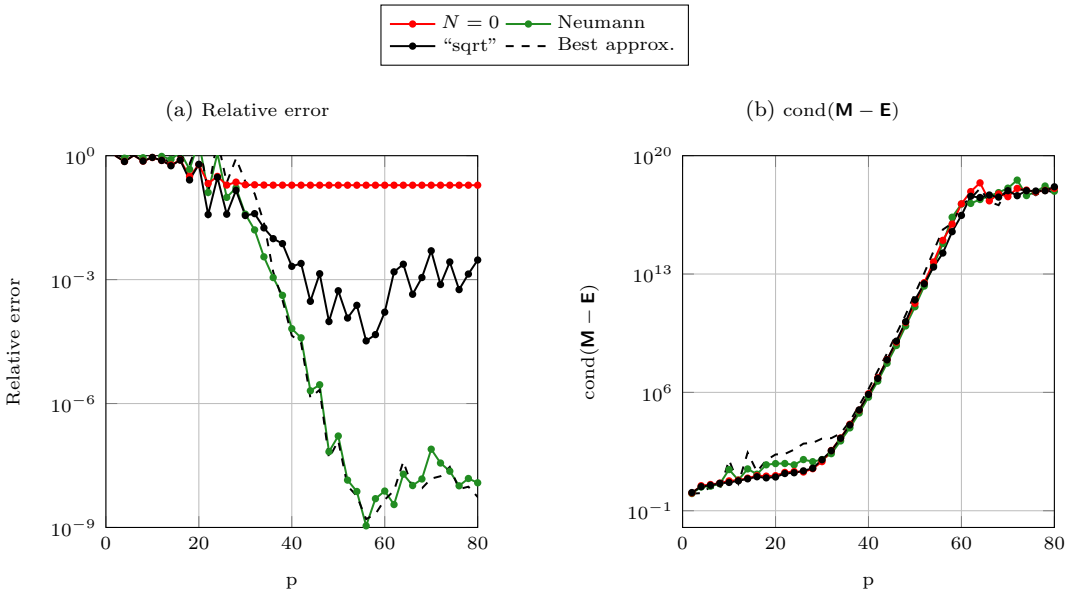


Figure 5.7: Error and condition number with respect to number of shape functions ("full" basis) with different configurations and ABCs for the preliminary test case.

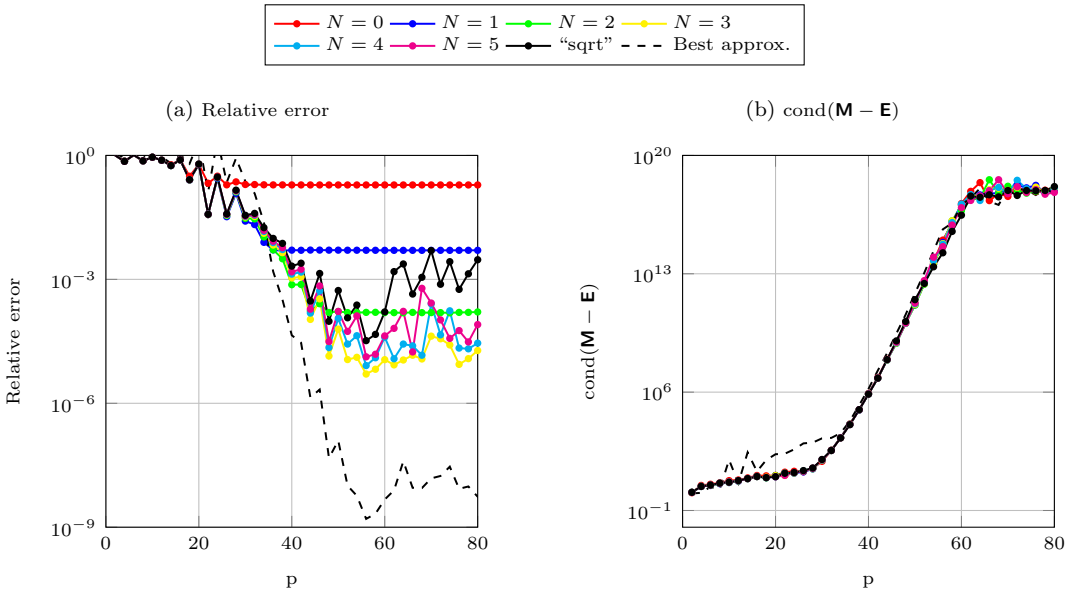


Figure 5.8: Error and condition number with respect to number of shape functions ("full" basis) with high-order ABCs for the preliminary test case.

Using the "quarter" basis, we can claim similar results regarding both the conditioning and the behavior of the numerical solutions. However, the oscillations in the error plots (Figures 5.9a-5.9b) and the bad conditioning (Figures 5.10a-5.10b) manifest at $p_{MAX} \approx 20$. This is in accordance with the intuition about the density of directions of the plane waves. Indeed, for a given accuracy, this adapted basis requires less plane-wave shape functions also in the case of Padé-type ABC and leads to better results: until the threshold p_{MAX} the error decreases monotonously and, after that, high-order approximations exhibits smaller oscillations. Moreover, the use of an adapted basis allows to achieve a smaller error in the best case: of order 10^{-7} with $N = 5$ and $p = 5$ for the "quarter" basis, against 10^{-5} with $N = 3$ and $p = 56$ for the "full" basis.

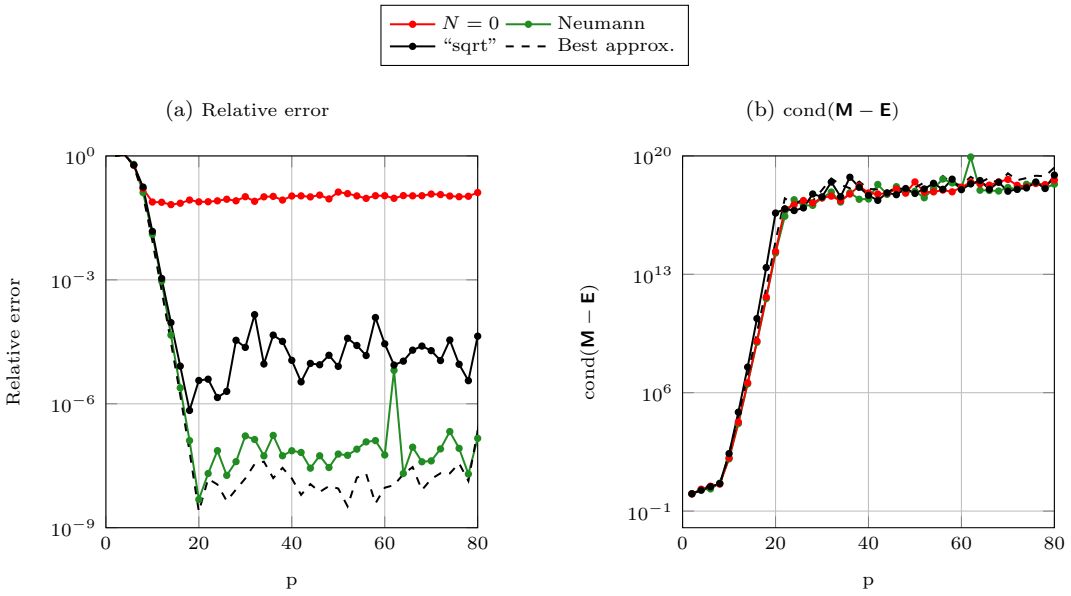


Figure 5.9: Error and condition number with respect to number of shape functions ("quarter" basis) with different configurations and ABCs for the preliminary test case.

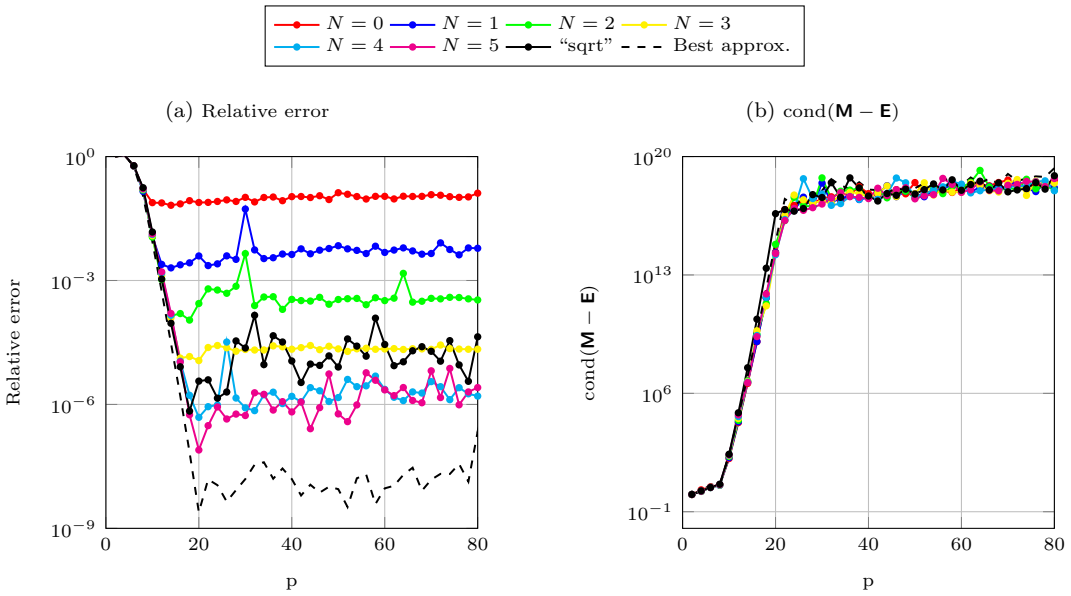


Figure 5.10: Error and condition number with respect to number of shape functions ("quarter" basis) with high-order ABCs for the preliminary test case.

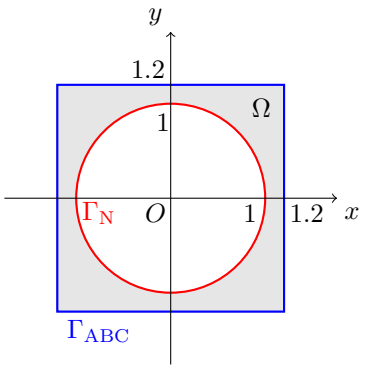
5.4.2 Study of a scattering benchmark

We consider a classic reference benchmark, that is the scattering of an incident plane wave by a circular cylinder. The computational domain is a square with a circular hole placed in its center, that is $\Omega = \Omega_{\text{square}} \setminus \Omega_{\text{circle}} = (-1.2, 1.2)^2 \setminus \{\boldsymbol{x} \in \mathbb{R}^2 : |\boldsymbol{x}| \leq 1\}$ (see Figure 5.11a).

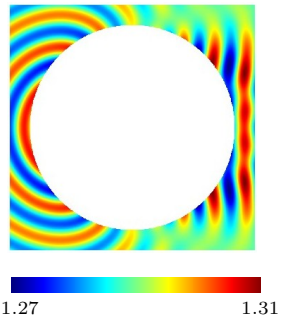
The scattering of the incident plane wave $p_{\text{inc}}(\boldsymbol{x}) = \exp(\imath\kappa\boldsymbol{\varphi} \cdot \boldsymbol{x})$ with direction of propagation $\boldsymbol{\varphi} = (1, 0)$, wavenumber κ, by the circular cylinder of radius R centered at the origin generates the scattered field (see Figure 5.11b)

$$p_{\text{ref}}(r, \theta) = -\sum_{m=0}^{\infty} \epsilon_m \imath^m \frac{\mathrm{J}'_m(\kappa R)}{\mathrm{H}^{(1)'}_m(\kappa R)} \mathrm{H}^{(1)}_m(\kappa r) \cos(m\theta), \quad r \geq R,$$

where (r, θ) are the polar coordinates, J_m is the mth-order Bessel's function, H_m is the mth-order first-kind Hankel function, and ϵ_m is the Neumann function which is equal to 1 for $m = 0$ and 2 otherwise.



(a) Square domain with cylinder.



(b) Real part of p_{ref} for $\varphi = 0$ and $\kappa = 20$.

Figure 5.11: Domain and reference solution for the scattering benchmark.

We prescribe a non-homogeneous Neumann boundary condition on the interior boundary, that is the boundary of the cylinder $\Gamma_{\text{N}} := \partial\Omega_{\text{circle}}$, by exploiting the incident plane wave and an absorbing boundary condition on the exterior boundary, that is the boundary of the square $\Gamma_{\text{ABC}} := \partial\Omega_{\text{square}}$, i.e.

$$\begin{cases} \dfrac{\partial p}{\partial \boldsymbol{n}} = -\dfrac{\partial p_{\text{inc}}}{\partial \boldsymbol{n}}, & \text{on } \Gamma_{\text{N}}, \\ \dfrac{\partial p}{\partial \boldsymbol{n}} = \mathcal{B}_* p, & \text{on } \Gamma_{\text{ABC}}. \end{cases}$$

For the numerical simulations, we fix the wavenumber $\kappa = 20$ and we consider a set of shape functions composed by $\mathrm{p} = 8$ plane waves whose directions of propagation are uniformly spaced in the whole plane for each element of the mesh with a null shift, that is (5.10) with $\theta_1 = 0$. Again, in the expression of $\mathcal{B}_{\text{Padé}}$ (5.16) we take a null angle of rotation of the branch cut, that is $\phi = 0$.

We perform a h-refinement. The results with the relative error with respect to the square root of the number of nodes of the mesh are showed in Figure 5.12.

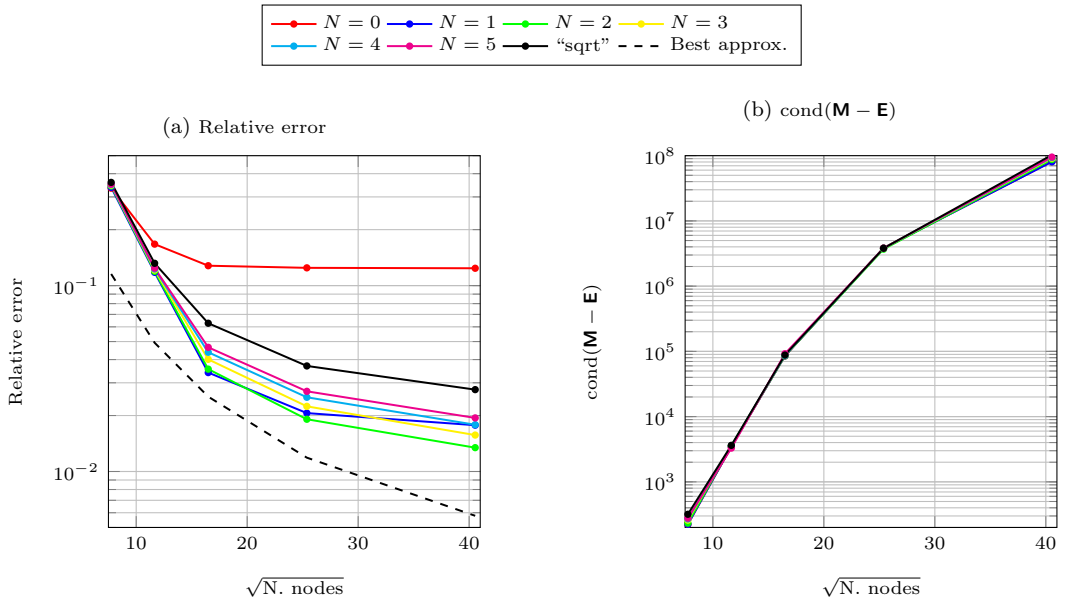


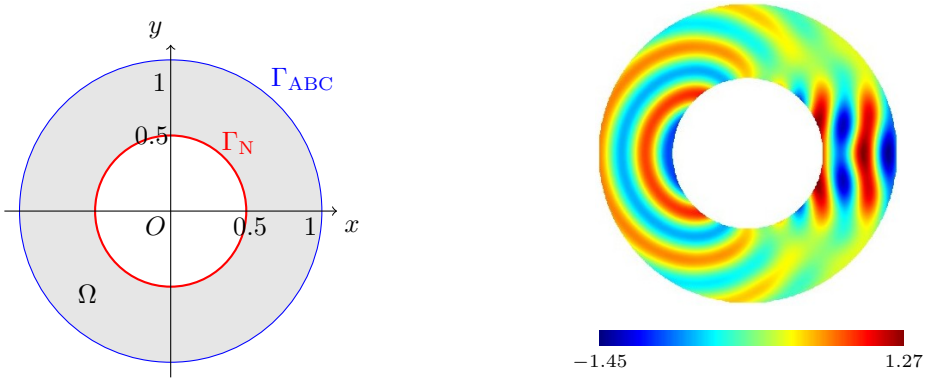
Figure 5.12: Error and condition number with respect to the square root of the number of nodes with different configurations and ABCs for the scattering benchmark with square boundary.

The results obtained by the use of the exact computation of the ABC are better than those related to the zeroth-order approximation. Nevertheless, they are way worse than the best approximation.

What we can observe from Figure 5.12a is that, refining the mesh, the numerical error decreases in all cases and is led to a plateau. Moreover, for $N \geq 1$ the numerical error is close to the one of the "sqrt" case. Again, a problem concerning the Padé approximation arises, indeed we expected the error to decrease in a significant way increasing N. Again, the condition number gets bigger by increasing the number of nodes in a similar way for all N, see Figure 5.12b.

We recall that the ABC $\partial p/\partial \boldsymbol{n} = \mathcal{B}p$ with $\mathcal{B}$ given by (1.20) is exact only for infinite straight boundaries. Hence, its use to bounded edges with corners consists in a source of error in the numerical scheme.

To investigate the influence of the presence of corners in the boundary of the domain, we consider the same scattering problem with the same parameters ($\kappa = 20$, p = 8), but the domain we take into account is an annulus $\Omega = \{\boldsymbol{x} \in \mathbb{R}^2 : 0.5 < |\boldsymbol{x}| < 1\}$ (see Figure 5.13a). Again, ABCs are prescribed on the external boundary of the domain and a non-homogeneous Neumann boundary condition is prescribed on the inner boundary of the domain.



(a) Square domain with cylinder.

(b) Real part of p_{ref} for $\varphi = 0$ and $\kappa = 20$.

Figure 5.13: Domain and reference solution for the scattering benchmark with square boundary.

The results with the relative error with respect to the square root of the number of nodes of the mesh are showed in Figure 5.14.

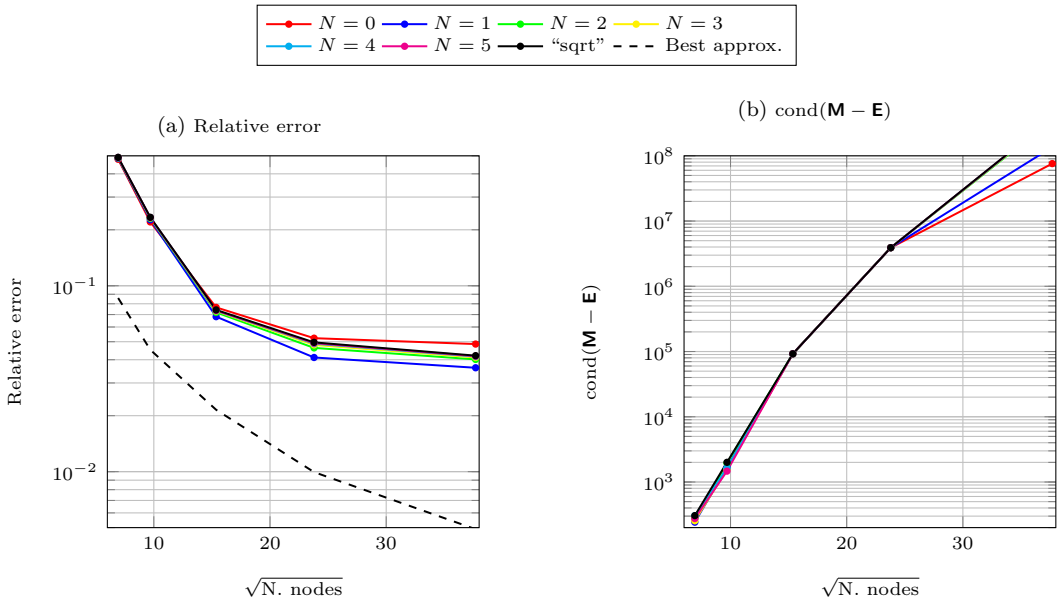


Figure 5.14: Error and condition number with respect to the square root of the number of nodes with different configurations and ABCs for the scattering benchmark with circular boundary.

In Figure 5.14a, we see that the numerical error related to the "sqrt" case is not close to the one associated to the best approximation and for $N \geq 1$ the numerical error is close to the one of the "sqrt" case. Again, since the ABC $\partial p/\partial \boldsymbol{n}$ is exact only for infinite straight boundaries, we know that an additional modeling error is present here because of the curved boundary. As we explained in Section 1.1.7, there exists a modified form of the ABC that holds in the case of regular curved boundaries given by (1.25)-(1.26). However, we performed some tests and they did not show any significant improvement in the accuracy of the numerical solution. Lastly, in Figure 5.14b, we can see the usual behavior of the condition number, which does not depend on N.

5.4.3 Illustration of an application

In this very last section, we take into account a numerical simulation for the scattering problem by an obstacle whose geometry is complex. In particular, we consider the scattering of a plane wave by an aircraft. We limit ourselves to an experiment with the "sqrt" approximation of the ABC on the external boundary and we comment the plot relative to the numerical solution, being the exact one not available.

The domain is $\Omega = (-4,7)\times(-4,4)\setminus\Omega_{\text{aircraft}}$. The scattering of the incident plane wave $p_{\text{inc}}(\boldsymbol{x}) = \exp(\imath\kappa\boldsymbol{\varphi}\cdot\boldsymbol{x})$ with direction of propagation $\boldsymbol{\varphi} = (1,0)$, wavenumber $\kappa = 20$, by the aircraft generates a scattered field that can not be represented in a closed form because of the complexity of the obstacle. As before, an ABC is prescribed on the external boundary of the domain and a non-homogeneous Neumann boundary condition is prescribed on the boundary of the aircraft. We take an unstructured mesh composed by $\mathrm{K} = 4316$ elements and with mesh size $h = 0.05$.

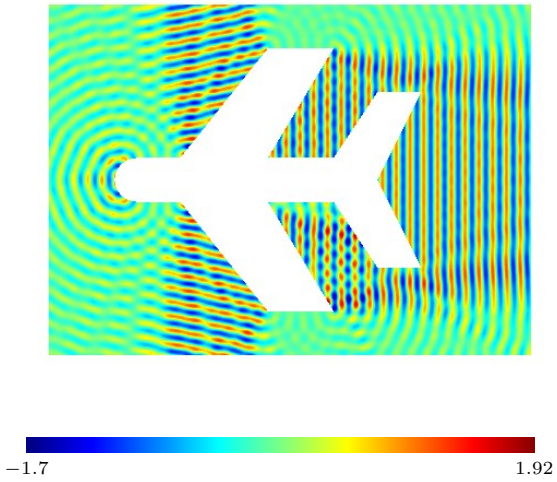


Figure 5.15: Real part of p_{ref} for $\varphi = 0$ and $\kappa = 20$.

Observing the numerical solution shown in Figure 5.15, we can claim a good behavior of the numerical scheme even in the case of non-trivial geometry. Indeed, we can observe a wave-shadow which is typically recurrent in similar examples with aircraft-shape scatterers. This shows that the extension of the study to every two dimensional scenario is straightforward with very little effort in the code: indeed it is enough to import the mesh associated to the benchmark we want to consider and no further modification is needed in the code.

5.5 Summary

In this Chapter, we investigated the combination of the Ultra Weak Variational Formulation with absorbing boundary conditions in the framework of a DG Trefftz method with propagative plane waves applied to the Helmholtz equation. In particular, we have provided an extension of the formulation introduced in [24] and we have implemented a finite element solver to perform numerical simulations.

- The simulations performed in the one-cell numerical domain show a better behavior of the absorbing boundary condition when it is expressed by the "sqrt" form rather than its zeroth-order approximation "$\imath\kappa$". Moreover, directional adaptivity turns out to be very effective in improving the results. Indeed, once it is applied to either the "sqrt" form and the high-order

Padé approximation in a proper way, we can recover results that are very similar to the ones obtained by projection. The Padé approximation behaves in the correct way: increasing the order N of the approximation, we get better results that tend to the ones obtained by the "sqrt" form.

- The simulations performed in the multi-cell numerical domain keep on showing a better behavior of the absorbing boundary condition when it is expressed by the "sqrt" form rather than its zeroth-order approximation "$\imath\kappa$". Nevertheless, the use of Padé approximation presents some issues and unexpected effects. In particular, in the case of approximation of order $N = 1$, the results are better than the ones associated to $N = 0$, while for a bigger value of N, we can not record any significant further improving. However, we can confirm the converging behavior of the Padé approximation to the exact formulation given by the "sqrt" form with respect to the order N of the approximation.

Conclusion

In this final Chapter, we provide a summary of the main contributions of this thesis, emphasizing the significance of the results obtained. We also highlight the methodological and conceptual advances achieved, discuss their limitations and possible applications, and point out perspectives and several directions for future research.

Main contributions and conclusions

Discontinuous Galerkin methods are very versatile and they offer a wide framework for alternative and non-standard formulations. In particular, they allow the development of DG methods specifically for wave-propagation phenomena. In this thesis we focused on the study of wave-tailored interface treatments based on transmission conditions, as well as wave-specific basis functions. This choice was motivated by the intuition that the embedding of the physics of the problem in the numerical method should improve the properties of the associated linear system. In particular, we aimed at efficient iterative algorithms that enable dealing with realistic large-scale problems and high-frequency phenomena.

We have proposed an extension of a hybridizable discontinuous Galerkin method, called CHDG method, to general time-harmonic wave problems with constant physical coefficients. In particular, the theory we developed covers the acoustics and electromagnetic frameworks studied by recent works, as well as linear water waves and linear elasticity. The method is based on a DG formulation with standard upwind numerical fluxes. Hybridization is performed by the introduction of transmission variables at the interfaces between the elements. Their definition naturally arises from the theory of symmetric hyperbolic systems. The CHDG hybridized system can be written in the form $(\mathrm{I} - \Pi\mathrm{S})\boldsymbol{x} = \boldsymbol{b}$. Under certain hypothesis on the boundary conditions based on energy conservation, we proved that the operator $\Pi\mathrm{S}$ is a strict contraction, which makes it possible to solve the system with a fixed-point iteration.

We then applied the CHDG method to aeroacoustics and, in particular, to the linearised Euler equations which model the propagation of acoustic and hydrodynamic waves in a homogeneous medium that moves at a constant subsonic velocity. Besides giving the definitions of the upwind numerical fluxes and of the transmission variables, we properly imposed the boundary conditions of the problem in such a way that the operator $\Pi\mathrm{S}$ is a contraction. In particular we had to impose Dirichlet or Neumann boundary conditions on the outflow portions of the boundary. The performance of the CHDG method was compared to that of the standard DG method by considering 2D benchmark problems and several standard iterative solvers (fixed-point, CGNR and GMRES). With some dependency on the benchmark and on the physical parameters, the iterative solvers require less iterations if applied to the CHDG system rather than to the DG one to achieve a given accuracy.

We proposed an additional extension of the CHDG method to address acoustics in heterogeneous media with piecewise constant physical coefficients. We considered a DG formulation based on both standard upwind numerical fluxes and symmetric fluxes involving a general impedance operator $\mathcal{A}$. The CHDG hybridized system can once again be written in the form $(\mathrm{I}-\Pi\mathrm{S})g = b$. In the case of symmetric fluxes, if $\mathcal{A}$ is a positive and self-adjoint operator, we proved that the operator $\Pi\mathrm{S}$ is a strict contraction. This property was not proved for standard upwind numerical fluxes. The performance of the CHDG method was once again studied by considering 2D benchmark problems and the same iterative solvers as before. The fixed-point iteration applied to the CHDG system with upwind numerical fluxes does not always converge, unlike the cases with symmetric fluxes. However, upwind and low-order symmetric numerical fluxes perform in a similar way when they both converge. High-order symmetric fluxes lead to slightly slower iterative solutions.

Lastly, with the idea of combining wave-tailored interface treatments with a wave-based method, we investigated a Trefftz DG method with propagative plane-wave shape functions for acoustics. In particular, it was applied to the Helmholtz equation complemented by absorbing boundary conditions. The method is based on the ultra weak variational formulation of the problem. Besides the classical low-order approximation of the ABC, also high-order Padé-type versions and an exact formulation are taken into account. To compare the accuracy of the method, we considered 2D benchmark problems and we performed h-refinements. The low-order approximation always perform worse than the exact one. Unexpectedly, high-order Padé-type ABCs perform better than the exact one, while still converging with respect to the order of the approximation to the exact formulation. In any case, enriching the set of shape functions eventually leads to ill-conditioning issues that deteriorate the numerical solution.

Perspectives

The theory we developed for the CHDG method covers problems that belong to a huge class of wave-propagation frameworks with homogeneous media. Consequently, a natural direction of research for future works is the application of the method to other scenarios: elasticity, for instance. It is enough to write the physical equations in the form of Problem 2.1.2, write its standard DG formulation with upwind numerical fluxes defined as (2.9) and the associated CHDG formulation thanks to the outgoing transmission variables defined as (2.10) and the incoming ones defined as (2.11) and (2.13) for interior and boundary faces, respectively. Also in this case, boundary conditions must be dealt with carefully so that Assumption 2.3.4 and, consequently, all the theoretical results (including Corollary 2.3.14), hold.

We also made a first step towards the applicability of the method to heterogeneous media by considering piecewise constant physical coefficients for acoustic problems. An additional step could be done by considering media that present non-constant possibly discontinuous physical coefficients, see e.g. [82]. The theory of hyperbolic PDEs and Riemann problems that allowed us to define both the upwind numerical fluxes and the transmission variables holds also in the framework of such heterogeneous media. Therefore, the extension of the CHDG method to space-dependent physical coefficients for acoustics looks straightforward.

Moreover, the structure of the method and the resulting linear system make the combination with domain decomposition methods and parallelization techniques very natural and, together with preconditioners, may lead to even faster iterative solutions. Recent works on domain decomposition methods, parallelization and preconditioners specifically proposed for HDG systems could inspire the development of a general approach, see e.g. [1, 101, 127], [114, 118] and [54, 102, 112], respectively.

Concerning the Trefftz DG method, further investigations should be led to address the unexpected behavior of the Padé-type ABC and poor conditioning. In particular, directional adaptivity, that turned out to be partially effective in simple benchmarks, should be tested to deal with more complex benchmark problems, see e.g. [39]. Moreover, poor conditioning has turned out to be an issue whenever the approximation space is enriched too much. Recent studies have showed that it is related to the exclusive use of propagative plane waves as shape functions and an alternative choice of shape functions, such as a combination of both propagative and evanescent plane waves (see e.g. [56, 105]), can limit the problem.

Bibliography

[1] E. Agullo, L. Giraud, A. Gobé, M. Kuhn, S. Lanteri, and L. Moya. High order HDG method and domain decomposition solvers for frequency-domain electromagnetics. *International Journal of Numerical Modelling: Electronic Networks, Devices and Fields*, 33(2):e2678, 2020.

[2] M. Ainsworth. Dispersive and dissipative behaviour of high order discontinuous Galerkin finite element methods. *Journal of Computational Physics*, 198(1):106–130, 2004.

[3] M. Ainsworth, P. Monk, and W. Muniz. Dispersive and dissipative properties of discontinuous Galerkin finite element methods for the second-order wave equation. *Journal of Scientific Computing*, 27(1):5–40, 2006.

[4] G. Alvarez, A. Loula, E. Carmo, and F. Rochinha. A discontinuous finite element formulation for Helmholtz equation. *Computer Methods in Applied Mechanics and Engineering*, 195: 4018–4035, 2006.

[5] D. N. Arnold and F. Brezzi. Mixed and nonconforming finite element methods : implementation, postprocessing and error estimates. *ESAIM: M2AN*, 19(1):7–32, 1985.

[6] C. Bailly and D. Juvé. Numerical solution of acoustic propagation problems using linearized Euler equations. *AIAA Journal*, 38:22–29, 01 2000. doi: 10.2514/2.949.

[7] H. Barucq, A. Bendali, J. Diaz, and S. Tordeux. Local strategies for improving the conditioning of the plane-wave ultra-weak variational formulation. *Journal of Computational Physics*, 441:110449, 2021.

[8] H. Barucq, J. Diaz, R.-C. Meyer, and H. Pham. Implementation of hybridizable discontinuous Galerkin method for time-harmonic anisotropic poroelasticity in two dimensions. *International Journal for Numerical Methods in Engineering*, 122(12):3015–3043, 2021.

[9] H. Barucq, N. Rouxelin, and S. Tordeux. Construction and analysis of a HDG solution for the total-flux formulation of the convected Helmholtz equation. *Mathematics of Computation*, 92(343):2097–2131, 2023.

[10] A. Bayliss, C. I. Goldstein, and E. Turkel. An iterative method for the Helmholtz equation. *Journal of Computational Physics*, 49:443–457, 1983.

[11] A. Bayliss, C. Goldstein, and E. Turkel. On accuracy conditions for the numerical computation of waves. *Journal of Computational Physics*, 59(3):396–404, 1985.

[12] J.-P. Berenger. A perfectly matched layer for the absorption of electromagnetic waves. *Journal of Computational Physics*, 114(2):185–200, 1994.

[13] H. Bériot, A. Prinn, and G. Gabard. Efficient implementation of high-order finite elements for Helmholtz problems. *International Journal for Numerical Methods in Engineering*, 106 (3):213–240, 2015.

[14] P. Bettess. Infinite elements. *International Journal for Numerical Methods in Engineering*, 11(1):53–64, 1977.

[15] J. H. Bramble and J. Xu. A local post-processing technique for improving the accuracy in mixed finite-element approximations. *SIAM Journal on Numerical Analysis*, 26(6):1267–1275, 1989.

[16] F. Brezzi, J. Douglas, and L. Marini. Two families of mixed finite elements for second order elliptic problems. *Numerische Mathematik*, 47:217–235, 1985.

[17] E. Burman, O. Duran, A. Ern, and M. Steins. Convergence analysis of hybrid high-order methods for the wave equation. *Journal of Scientific Computing*, 87(3):91, 2021.

[18] E. Burman, O. Duran, and A. Ern. Hybrid high-order methods for the acoustic wave equation in the time domain. *Communications on Applied Mathematics and Computation*, 4(2):597–633, 2022.

[19] D. S. Burnett and R. L. Holford. An ellipsoidal acoustic infinite element. *Computer Methods in Applied Mechanics and Engineering*, 164(1):49–76, 1998.

[20] H. Bériot and A. Modave. An automatic perfectly matched layer for acoustic finite element simulations in convex domains of general shape. *International Journal for Numerical Methods in Engineering*, 122(5):1239–1261, 2021.

[21] H. Bériot, A. Prinn, and G. Gabard. Efficient implementation of high-order finite elements for Helmholtz problems. *International Journal for Numerical Methods in Engineering*, 106 (3):213–240, 2016.

[22] J. Carrero, B. Cockburn, and D. Schötzau. Hybridized globally divergence-free LDG methods. Part I: The Stokes problem. *Mathematics of Computation*, 75(254):533–563, 2006.

[23] J. Caruthers, J. French, and G. Raviprakash. Green function discretization for numerical solution of the Helmholtz equation. *Journal of Sound and Vibration*, 187(4):553–568, 1995.

[24] O. Cessenat and B. Després. Application of an ultra weak variational formulation of elliptic PDEs to the two-dimensional Helmholtz problem. *SIAM Journal on Numerical Analysis*, 35 (1):255–299, 1998.

[25] T. Chaumont-Frelet. Duality analysis of interior penalty discontinuous Galerkin methods under minimal regularity and application to the a priori and a posteriori error analysis of Helmholtz problems. *ESAIM: M2AN*, 58(3):1087–1106, 2024.

[26] H. Chen, P. Lu, and X. Xu. A hybridizable discontinuous Galerkin method for the Helmholtz equation with high wave number. *SIAM Journal on Numerical Analysis*, 51(4):2166–2188, 2013.

[27] K. Christodoulou, O. Laghrouche, M. S. Mohamed, and J. Trevelyan. High-order finite elements for the solution of Helmholtz problems. *Computers & Structures*, 191:129–139, 2017.

[28] E. T. Chung and B. Engquist. Optimal discontinuous Galerkin methods for wave propagation. *SIAM Journal on Numerical Analysis*, 44(5):2131–2158, 2006.

[29] B. Cockburn. *Static Condensation, Hybridization, and the Devising of the HDG Methods*, pages 129–177. Springer International Publishing, 2016.

[30] B. Cockburn and J. Gopalakrishnan. A characterization of hybridized mixed methods for second order elliptic problems. *SIAM Journal on Numerical Analysis*, 42(1):283–301, 2005.

[31] B. Cockburn and J. Gopalakrishnan. Incompressible finite elements via hybridization. Part I: The Stokes system in two space dimensions. *SIAM Journal on Numerical Analysis*, 43(4): 1627–1650, 2006.

[32] B. Cockburn and J. Gopalakrishnan. Incompressible finite elements via hybridization. Part II: The Stokes system in three space dimensions. *SIAM Journal on Numerical Analysis*, 43 (4):1651–1672, 2006.

[33] B. Cockburn, B. Dong, and J. Guzmán. A superconvergent LDG-hybridizable Galerkin method for second-order elliptic problems. *Mathematics of Computation*, 77(264):1887–1916, 2008.

[34] B. Cockburn, J. Gopalakrishnan, and R. Lazarov. Unified hybridization of discontinuous Galerkin, mixed, and continuous Galerkin methods for second order elliptic problems. *SIAM Journal on Numerical Analysis*, 47(2):1319–1365, 2009.

[35] F. Collino and P. Monk. Optimizing the perfectly matched layer. *Computer Methods in Applied Mechanics and Engineering*, 164(1):157–171, 1998.

[36] M. I. Comodi. The Hellan-Herrmann-Johnson method: Some new error estimates and post-processing. *Mathematics of Computation*, 52(185):17–29, 1989.

[37] S. Congreve, J. Gedicke, and I. Perugia. Robust adaptive *hp*-discontinuous Galerkin finite element methods for the Helmholtz equation. *SIAM Journal on Scientific Computing*, 41 (2):A1121–A1147, 2019.

[38] S. Congreve, J. Gedicke, and I. Perugia. Numerical investigation of the conditioning for plane wave discontinuous Galerkin methods. In *Numerical mathematics and advanced applications ENUMATH 2017*, Lecture Notes in Computational Science and Engineering, pages 493–500. Springer, 2019.

[39] S. Congreve, P. Houston, and I. Perugia. Adaptive refinement for *hp*-version Trefftz discontinuous Galerkin methods for the homogeneous Helmholtz problem. *Advances in Computational Mathematics*, 45(1):361–393, 2019.

[40] A. de La Bourdonnaye. Some formulations coupling finite element and integral equation methods for Helmholtz equation and electromagnetism. *Numerische Mathematik*, 69(3): 527–268, 1995.

[41] A. Deraemaeker, I. Babuška, and P. Bouillard. Dispersion and pollution of the FEM solution for the Helmholtz equation in one, two and three dimensions. *International Journal for Numerical Methods in Engineering*, 46(4):471–499, 1999.

[42] B. Després, A. Nicolopoulos, and B. Thierry. Corners and stable optimized domain decomposition methods for the Helmholtz problem. *Numerische Mathematik*, 149:779–818, 2021.

[43] I. Djiba. *Méthode de décomposition de domaine basée sur une formulation de Trefftz pour l'acoustique anisotrope*. PhD thesis, 2025. URL https://theses.fr/s342267.

[44] E. D. do Carmo, R. Leal, and W. Mansur. A new scheme to obtain pollution-free solution for plane waves and to generate a continuous Petrov-Galerkin method for the Helmholtz problem. *Computer Methods in Applied Mechanics and Engineering*, 359:112674, 2020.

[45] B. Engquist and A. Majda. Absorbing boundary conditions for numerical simulation of waves. *Proceedings of the National Academy of Sciences*, 74(5):1765–1766, 1977.

[46] B. Engquist and L. Ying. Sweeping preconditioner for the Helmholtz equation: moving perfectly matched layers. *Multiscale Modeling & Simulation*, 9(2):686–710, 2011.

[47] Y. A. Erlangga, C. Vuik, and C. W. Oosterlee. On a class of preconditioners for solving the Helmholtz equation. *Applied Numerical Mathematics*, 50(3):409–425, 2004.

[48] A. Ern and M. Steins. Convergence analysis for the wave equation discretized with hybrid methods in space (HHO, HDG and WG) and the Leapfrog scheme in time. *Journal of Scientific Computing*, 101(1):7, 2024.

[49] O. G. Ernst and M. J. Gander. *Why it is difficult to solve Helmholtz problems with classical iterative methods*, pages 325–363. Springer Berlin Heidelberg, 2012.

[50] X. Feng and H. Wu. Discontinuous Galerkin methods for the Helmholtz equation with large wave number. *SIAM Journal on Numerical Analysis*, 47(4):2872–2896, 2009.

[51] P. Fernández, A. Christophe, S. Terrana, N. Nguyen, and J. Peraire. Hybridized discontinuous Galerkin methods for wave propagation. *Journal of Scientific Computing*, 77(3): 1566–1604, 2018.

[52] B. Fraeijs de Veubeke. Displacement and equilibrium models in the finite element method. In *Stress analysis*. John Wiley Sons, 1965.

[53] L. Franca, A. Nesliturk, and M. Stynes. On the stability of residual-free bubbles for convection-diffusion problems and their approximation by a two-level finite element method. *Computer Methods in Applied Mechanics and Engineering*, 166(1):35–49, 1998.

[54] G. Fu and W. Kuang. Optimal geometric multigrid preconditioners for HDG-P0 schemes for the reaction-diffusion equation and the generalized Stokes equations. *ESAIM: M2AN*, 57(3):1553–1587, 2023.

[55] H. S. Fure, S. Pernet, M. Sirdey, and S. Tordeux. A discontinuous Galerkin Trefftz type method for solving the two dimensional Maxwell equations. *SN Partial Differential Equations and Applications*, 1(4):23, 2020.

[56] N. Galante, A. Moiola, and E. Parolin. Stable approximation of Helmholtz solutions in the ball using evanescent plane waves, 2024. URL https://arxiv.org/abs/2401.04016.

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

[58] G. N. Gatica. Variational formulations of transmission problems via FEM, BEM and DtN mappings. *Computer Methods in Applied Mechanics and Engineering*, 182(3):341–354, 2000.

[59] C. Geuzaine and J.-F. Remacle. Gmsh: A 3-D finite element mesh generator with built-in pre- and post-processing facilities. *International Journal for Numerical Methods in Engineering*, 79(11):1309–1331, 2009.

[60] G. Giorgiani, S. Fernández-Méndez, and A. Huerta. Hybridizable discontinuous Galerkin p-adaptivity for wave propagation problems. *International Journal for Numerical Methods in Fluids*, 72(12):1244–1262, 2013.

[61] C. J. Gittelson, R. Hiptmair, and I. Perugia. Plane wave discontinuous Galerkin methods: analysis of the h-version. *ESAIM: Modélisation mathématique et analyse numérique*, 43(2): 297–331, 2009.

[62] M. Griesmaier and P. Roland. Error analysis for a hybridizable discontinuous Galerkin method for the Helmholtz equation. *Journal of Scientific Computing*, 49(3):291–310, 2011.

[63] M. J. Grote, A. Schneebeli, and D. Schötzau. Discontinuous Galerkin finite element method for the wave equation. *SIAM Journal on Numerical Analysis*, 44(6):2408–2431, 2006.

[64] T. Hagstrom and T. Warburton. Complete radiation boundary conditions: Minimizing the long time error growth of local methods. *SIAM J. Numerical Analysis*, 47:3678–3704, 2009.

[65] I. Harari and T. J. Hughes. Finite element methods for the Helmholtz equation in an exterior domain: Model problems. *Computer Methods in Applied Mechanics and Engineering*, 87(1): 59–96, 1991.

[66] I. Harari and T. J. Hughes. Galerkin/least-squares finite element methods for the reduced wave equation with non-reflecting boundary conditions in unbounded domains. *Computer Methods in Applied Mechanics and Engineering*, 98(3):411–454, 1992.

[67] I. Harari and F. Magoulès. Numerical investigations of stabilized finite element computations for acoustics. *Wave Motion*, 39(4):339–349, 2004.

[68] I. Harari, E. Turkel, and M. Slavutin. Analytical and numerical studies of a finite element PML for the Helmholtz equation. *Journal of Computational Acoustics*, 8:121–137, 2000.

[69] J. S. Hesthaven and T. Warburton. *Nodal discontinuous Galerkin methods: algorithms, analysis, and applications*. Springer Science & Business Media, 2007.

[70] R. L. Higdon. Absorbing boundary conditions for difference approximations to the multi-dimensional wave equation. *Mathematics of Computation*, 47(176):437–459, 1986.

[71] R. L. Higdon. Absorbing boundary conditions for acoustic and elastic waves in stratified media. *Journal of Computational Physics*, 101(2):386–418, 1992.

[72] R. Hiptmair, A. Moiola, and I. Perugia. Plane wave discontinuous Galerkin methods for the 2D Helmholtz equation: Analysis of the p-version. *SIAM Journal on Numerical Analysis*, 49(1):264–284, 2011.

[73] R. Hiptmair, A. Moiola, and I. Perugia. Plane wave discontinuous Galerkin methods: Exponential convergence of the hp-version. *Foundations of Computational Mathematics*, 16(3): 637–675, 2019.

[74] C. Hirsch. *Numerical computation of internal and external flows: the fundamentals of computational fluid dynamics, volume 1*. John Wiley Sons, Ltd, 2007.

[75] F. Q. Hu, M. Hussaini, and P. Rasetarinera. An analysis of the discontinuous Galerkin method for wave propagation problems. *Journal of Computational Physics*, 151(2):921–946, 1999.

[76] T. J. Hughes, G. R. Feijóo, L. Mazzei, and J.-B. Quincy. The variational multiscale method – a paradigm for computational mechanics. *Computer Methods in Applied Mechanics and Engineering*, 166(1):3–24, 1998.

[77] T. Huttunen, P. Monk, and J. P. Kaipio. Computational aspects of the ultra-weak variational formulation. *Journal of Computational Physics*, 182(1):27–46, 2002.

[78] T. Huttunen, P. Gamallo, and R. J. Astley. Comparison of two wave element methods for the Helmholtz problem. *Communications in Numerical Methods in Engineering*, 25(1):35–52, 2009.

[79] F. Ihlenburg. *Finite element analysis of acoustic scattering*. Applied mathematical sciences. Springer, 1998.

[80] F. Ihlenburg and I. Babuška. Finite element solution of the Helmholtz equation with high wave number part I: The h-version of the FEM. *Computers & Mathematics with Applications*, 30(9):9–37, 1995.

[81] F. Ihlenburg and I. Babuška. Finite element solution of the Helmholtz equation with high wave number part II: The hp-version of the FEM. *SIAM Journal on Numerical Analysis*, 34(1):315–358, 1997.

[82] L.-M. Imbert-Gérard and P. Monk. Numerical simulation of wave propagation in inhomogeneous media using generalized plane waves. *ESAIM: M2AN*, 51(4):1387–1406, 2017.

[83] L.-M. Imbert-Gérard and G. Sylvand. Three types of quasi-Trefftz functions for the 3D convected Helmholtz equation: construction and approximation properties. *IMA Journal of Numerical Analysis*, 45(4):2274–2328, 2024.

[84] L.-M. Imbert-Gérard, A. Moiola, C. Perinati, and P. Stocker. Polynomial quasi-Trefftz DG for PDEs with smooth coefficients: elliptic problems. *IMA Journal of Numerical Analysis*, 2025.

[85] C. Johnson and J. C. Nedelec. On the coupling of boundary integral and finite element methods. *Mathematics of Computation*, 35(152):1063–1079, 1980.

[86] R. Kechroud, X. Antoine, and A. Soulaimani. Numerical accuracy of a Padé-type non-reflecting boundary condition for the finite element solution of acoustic scattering problems at high-frequency. *International Journal for Numerical Methods in Engineering*, 64:1275 – 1302, 2005.

[87] O. Laghrouche, P. Bettess, and R. Astley. Modelling of short wave diffraction problems using approximating systems of plane waves. *International Journal for Numerical Methods in Engineering*, 54:1501–1533, 2002.

[88] P. D. Lax. *Hyperbolic systems of conservation laws and the mathematical theory of shock waves*. Society for Industrial and Applied Mathematics, 1973.

[89] P. L. Lederer, C. Lehrenfeld, P. Stocker, and I. Voulis. A unified framework for Trefftz-like discretization methods, 2024. URL https://arxiv.org/abs/2412.00806.

[90] C. Lehrenfeld and P. Stocker. Embedded Trefftz discontinuous Galerkin methods. *International Journal for Numerical Methods in Engineering*, 124(17):3637–3661, 2023.

[91] R. J. LeVeque. *Finite volume methods for hyperbolic problems*. Cambridge Texts in Applied Mathematics. Cambridge University Press, 2002.

[92] A. Lieu, G. Gabard, and H. Bériot. A comparison of high-order polynomial and wave-based methods for Helmholtz problems. *Journal of Computational Physics*, 321:105–125, 2016.

[93] T. Luostari, T. Huttunen, and P. Monk. The ultra weak variational formulation using Bessel basis functions. *Communications in Computational Physics*, 11, 2011.

[94] A. Majda. *Compressible fluid flow and systems of conservation laws in several space variables*, volume 53 of *Applied Mathematical Sciences*. Springer-Verlag, 1984.

[95] F. Milinazzo, C. A. Zala, and G. H. Brooke. Rational square-root approximations for parabolic equation algorithms. *Journal of the Acoustical Society of America*, 101:760–766, 1997.

[96] W. F. Mitchell and M. A. McClain. A comparison of hp-adaptive strategies for elliptic partial differential equations. *ACM Trans. Math. Softw.*, 41(1), 2014.

[97] A. Modave and T. Chaumont-Frelet. A hybridizable discontinuous Galerkin method with characteristic variables for Helmholtz problems. *Journal of Computational Physics*, 493: 112459, 2023.

[98] A. Modave, A. Royer, X. Antoine, and C. Geuzaine. A non-overlapping domain decomposition method with high-order transmission conditions and cross-point treatment for Helmholtz problems. *Computer Methods in Applied Mechanics and Engineering*, 368:113162, 2020.

[99] A. Moiola, R. Hiptmair, and I. Perugia. Plane wave approximation of homogeneous Helmholtz solutions. *Zeitschrift für angewandte Mathematik und Physik*, 62(5):809–837, 2011.

[100] G. Montecinos, C. Castro, M. Dumbser, and E. Toro. Comparison of solvers for the generalized Riemann problem for hyperbolic systems with source terms. *Journal of Computational Physics*, 231(19):6472–6494, 2012.

[101] S. Muralikrishnan, M.-B. Tran, and T. Bui-Thanh. iHDG: An iterative HDG framework for partial differential equations. *SIAM Journal on Scientific Computing*, 39(5):S782–S808, 2017.

[102] S. Muralikrishnan, S. Shannon, T. Bui-Thanh, and J. N. Shadid. A multilevel block preconditioner for the hdg trace system applied to incompressible resistive MHD. *Computer Methods in Applied Mechanics and Engineering*, 404:115775, 2023.

[103] F. Nataf. Absorbing boundary conditions and perfectly matched layers in wave propagation problems. *Radon Series on Computational and Applied Mathematics*, 11, 2013.

[104] N. C. Nguyen, J. Peraire, and B. Cockburn. High-order implicit hybridizable discontinuous Galerkin methods for acoustics and elastodynamics. *Journal of Computational Physics*, 230 (10):3695–3718, 2011.

[105] E. Parolin, D. Huybrechs, and A. Moiola. Stable approximation of Helmholtz solutions in the disk by evanescent plane waves. *ESAIM: M2AN*, 57(6):3499–3536, 2023.

[106] C. A. Pereira and B. C. Vermeire. Hybridized implicit-explicit flux reconstruction methods for geometry-induced stiffness. *Journal of Computational Physics*, 528:113819, 2025.

[107] S. Pescuma, G. Gabard, T. Chaumont-Frelet, and A. Modave. A hybridizable discontinuous Galerkin method with transmission variables for time-harmonic wave problems in heterogeneous media. *Journal of Computational Physics*, 534:114009, 2025.

[108] H. Pham, F. Faucher, and H. Barucq. Numerical investigation of stabilization in the hybridizable discontinuous Galerkin method for linear anisotropic elastic equation. *Computer Methods in Applied Mechanics and Engineering*, 428:117080, 2024.

[109] A. Pierce. *Acoustics: an introduction to its physical principles and applications*, volume 34. Springer, 1989.

[110] A. Quarteroni. *Numerical models for differential problems*. Springer Publishing Company, Incorporated, 2nd edition, 2013.

[111] A. E. Rappaport, T. Chaumont-Frelet, and A. Modave. A hybridizable discontinuous Galerkin method with transmission variables for time-harmonic electromagnetic problems. *HAL preprint hal-05016261*, 2025.

[112] S. Rhebergen and G. N. Wells. Preconditioning for a pressure-robust HDG discretization of the stokes equations. *SIAM Journal on Scientific Computing*, 44(1):A583–A604, 2022.

[113] M. Rivet, S. Pernet, and S. Tordeux. Optimised correction polynomial functions for the flux reconstruction method in time-harmonic electromagnetism. *Applied Mathematics Letters*, 157:109187, 2024. ISSN 0893-9659. doi: https://doi.org/10.1016/j.aml.2024.109187.

[114] X. Roca, C. Nguyen, and J. Peraire. Scalable parallelization of the hybridized discontinuous Galerkin method for compressible flow. In *21st AIAA Computational Fluid Dynamics Conference*, page 2939, 2013.

[115] N. Rouxelin. *Methodes numeriques mixtes condensees pour l'etude de la propagation des ondes acoustiques en ecoulement. Applications en heliosismologie.* PhD thesis, 2021. URL http://www.theses.fr/2021PAUU3045.

[116] Y. Saad. *Iterative methods for sparse linear systems.* SIAM, 2003.

[117] S. Salsa. *Partial differential equations in action: from modelling to theory.* Universitext. Springer, 3rd edition, 2016.

[118] A. Samii, C. Michoski, and C. Dawson. A parallel and adaptive hybridized discontinuous Galerkin method for anisotropic nonhomogeneous diffusion. *Computer Methods in Applied Mechanics and Engineering*, 304:118–139, 2016.

[119] M. Sirdey. *Méthode itérative de Trefftz pour la simulation d'ondes électromagnétiques en trois dimensions.* PhD thesis, 2022. URL http://www.theses.fr/2022PAUU3057.

[120] P. Solin, K. Segeth, and I. Dolezel. *Higher-order finite element methods.* Studies in Advanced Mathematics. CRC Press, 2003.

[121] A. Sommerfeld. Die Greensche funktion der schwingungslgleichung. *Jahresbericht der Deutschen Mathematiker-Vereinigung*, 21:309–352, 1912.

[122] L. Thompson and P. Pinsky. Complex wavenumber Fourier analysis of the p-version finite element method. *Computational Mechanics*, 13:255–275, 1994.

[123] E. F. Toro. *Riemann solvers and numerical methods for fluid dynamics: a practical introduction.* Springer Science & Business Media, 2013.

[124] A. Vion and C. Geuzaine. Improved sweeping preconditioners for domain decomposition algorithms applied to time-harmonic Helmholtz and Maxwell problems. *ESAIM: Proceedings and Surveys*, 61:93–111, 2018.

[125] G. B. Whitham. *Linear and nonlinear waves.* John Wiley Sons, Ltd, 1999.

[126] L. C. Wilcox, G. Stadler, C. Burstedde, and O. Ghattas. A high-order discontinuous Galerkin method for wave propagation through coupled elastic–acoustic media. *Journal of Computational Physics*, 229(24):9373–9396, 2010.

[127] J. Zhang and X. Tu. Robust BDDC algorithms for the Brinkman problem with HDG discretizations. *Computer Methods in Applied Mechanics and Engineering*, 400:115548, 2022.

[128] L. Zhu and Y. Du. Pre-asymptotic error analysis of hp-interior penalty discontinuous Galerkin methods for the Helmholtz equation with large wave number. *Computers Mathematics with Applications*, 70(5):917–933, 2015.

[129] O. C. Zienkiewicz, D. W. Kelly, and P. Bettess. The coupling of the finite element method and boundary solution procedures. *International Journal for Numerical Methods in Engineering*, 11(2):355–375, 1977.

List of Figures

List of Tables

List of Symbols and Abbreviations

Common symbols

$\mathbb{R}$	set of real numbers
$\mathbb{C}$	set of complex numbers
$\boldsymbol{x}$	spatial position [m]
t	time [s]
$\boldsymbol{e}_1, \boldsymbol{e}_2, \boldsymbol{e}_3$	cartesian directions
x_1, x_2, x_3	cartesian coordinates [m]
$\imath$	imaginary unit ($= \sqrt{-1}$)
$\boldsymbol{n}$	normal unit vector
$\boldsymbol{\tau}$	tangent unit vector

Symbols for aero/acoustics

p	pressure [Pa]
$\boldsymbol{u}$	velocity [ms^{-1}]
ρ	density [kgm^{-3}]
$\boldsymbol{u}_0$	background flow [ms^{-1}]
ω	angular frequency [rads^{-1}]
c	phase velocity [ms^{-1}]
κ	wavenumber [m^{-1}] ($= \omega/c$)
η	acoustic impendance [$\mathrm{kgm}^{-2}\mathrm{s}^{-1}$]($= \rho c$)

Symbols for numerical discretization

h	mesh size
p	cardinality of local approximation space
K	element
$\mathcal{T}_h$	collection of elements (mesh)
$\mathcal{F}_h$	collection of faces
$\mathcal{V}_h$	approximation space

Operators

^{T}	transpose
†	Hermitian transpose
$\cdot$	scalar product
$\times$	vector product
$\frac{\partial}{\partial x}$	first-order partial derivative with respect to x
$\frac{\partial^n}{\partial x^n}$	nth-order partial derivative with respect to x
∇	nabla

Abbreviations

FEM	finite element method
DG	discontinuous Galerkin
HDG	hybridizable discontinuous Galerkin with numerical traces
CHDG	hybridizzable discontinuous Galerkin with transmission variables
ABC	absorbing boundary condition

ECOLE
DOCTORALE
DE MATHEMATIQUES
HADAMARD

Titre: Méthodes d'éléments finis discontinus de type Galerkin avec variables de transmission pour la propagation d'ondes en régime harmonique

Mots clés: Éléments finis, Propagation des ondes, Équation de Helmholtz, Galerkin discontinu, Hybridation, Aéroacoustique

Résumé: L'objectif de cette thèse est de développer des méthodes de Galerkin discontinues (DG) spécifiques aux ondes pour les phénomènes de propagation des ondes. Les systèmes linéaires issus des techniques de discrétisation par éléments finis sont souvent difficiles à résoudre en raison de problèmes de conditionnement liés aux propriétés intrinsèques des problèmes en régime harmonique. Les méthodes DG sont très versatiles et offrent un large cadre pour les formulations qui permettent l'utilisation d'algorithmes itératifs rapides ainsi que des fonctions de base non standard. Cette thèse est motivée par l'idée que l'intégration de la nature oscillante du problème physique dans la méthode de discrétisation devrait conduire à une amélioration des propriétés du système linéaire associé. Spécifiquement, nous considérons une variante de la méthode de Galerkin discontinue hybridable (HDG) avec des flux numériques upwind. Contrairement à l'approche standard, le processus d'hybridation est réalisé en utilisant des variables de transmission dont les définitions sont déterminées par la physique du problème. La méthode résultante, appelée méthode CHDG, présente une amélioration globale de la vitesse de convergence des procédures itératives par rapport aux méthodes standard DG et HDG. Nous étendons la théorie de CHDG, qui a été initialement étudiée pour des problèmes acoustiques, à des problèmes généraux en régime harmonique avec des tests numériques axés sur l'aéroacoustique en milieu homogène. De plus, inspiré par des études récentes sur les méthodes de décomposition de domaine, nous ciblons une autre extension de CHDG qui inclut des variables de transmission d'ordre élevé avec un focus pour la théorie et les tests numériques sur l'acoustique pour les milieux homogènes et hétérogènes. Enfin, avec l'idée de combiner des traitements d'interface adaptés aux ondes avec une discrétisation basée sur les ondes, nous étudions une méthode DG de Trefftz avec des ondes planes propagatives comme fonctions de base appliquée aux problèmes de Helmholtz complétées par des conditions aux limites absorbantes d'ordre élevé. Les méthodes numériques sont étudiées en utilisant des benchmarks numériques bidimensionnels implémentés dans des codes MATLAB dédiés.

Title: Discontinuous Galerkin finite element methods with transmission variables for time-harmonic wave propagation

Keywords: Finite elements, Wave propagation, Helmholtz equation, Discontinuous Galerkin, Hybridization, Aeroacoustics

Abstract: The goal of this thesis is to develop wave-specific discontinuous Galerkin (DG) methods for wave-propagation phenomena. Linear systems arising from finite element discretization techniques are often difficult to solve due to ill-conditioning issues related to intrinsic properties of time-harmonic problems. DG methods are very versatile and offer a wide framework for formulations that allow the use of fast iterative algorithms as well as non-standard basis functions. In addition, part of this thesis is motivated by the intuition that embedding the oscillating nature of the physical problem into the discretization method should lead to an improvement of the properties of the associated linear system. We consider a variant of the hybridizable discontinuous Galerkin (HDG) method with upwind numerical fluxes. Unlike the standard approach, the hybridization process is achieved using transmission variables whose definitions are driven by the physics of the problem. The resulting method, called the CHDG method, delivers an overall improvement of the speed of convergence of iterative procedures when compared to standard DG and HDG methods. We extend the theory of CHDG, which was initially studied for acoustic problems, to general time-harmonic problems with numerical tests focusing on aeroacoustics in homogeneous media. Moreover, inspired by recent studies on domain decomposition methods, we target another extension of CHDG that includes high-order transmission variables with a focus for the theory and the numerical tests on acoustic for both homogeneous and heterogeneous media. Lastly, with the idea of combining wave-tailored interface treatments with a wave-based basis method, we investigate a Trefftz DG method with propagative plane-wave basis functions applied to Helmholtz problems complemented by high-order absorbing boundary conditions. The numerical methods are studied by using two-dimensional numerical benchmarks implemented in dedicated MATLAB codes.

Institut Polytechnique de Paris
91120 Palaiseau, France