# Numerical methods for solving Helmholtz problems

Magdalena Grigoroscuta-Strugaru

ACADÉMIE DE BORDEAUX |2|0|0|9|*P*|*A*|*U*|*U*|3|0|3|8|

**THÈSE DE DOCTORAT**

présentée à

L'Université de Pau et des Pays de l'Adour

École doctorale des sciences et leurs applications - ED 211

par

**Magdalena GRIGOROSCUTA-STRUGARU**

pour obtenir le grade de

**DOCTEUR de l'Université de Pau et des Pays de l'Adour**
**Spécialité : Mathématiques Appliquées**

# CONTRIBUTION À LA RÉSOLUTION NUMÉRIQUE DES PROBLÈMES DE HELMHOLTZ

**Soutenue le 18 décembre 2009**

**Après avis de :**

| | | |
|---|---|---|
| M. Abderrahmane BENDALI | Professeur, INSA Toulouse | Rapporteur |
| M. Peter MONK | Professeur, University of Delaware (USA) | Rapporteur |

**Devant la comission d'examen formée des rapporteurs et de :**

| | | |
|---|---|---|
| M. Mohamed AMARA | Professeur, Université de Pau et des Pays de l'Adour | Directeur de thèse |
| M. Rabia DJELLOULI | Professeur, California State University Northridge (USA) | Co-directeur de thèse |
| M. Henri CALANDRA | Ingénieur de Recherche Expert, centre CSTJF TOTAL | Co-directeur de thèse |
| Mme. Hélène BARUCQ | Directrice de Recherche INRIA, INRIA Bordeaux Sud-Ouest | Présidente du jury |
| M. David PARDO | Professeur, Basque Center for Applied Mathematics (ESPAGNE) | Examinateur |

**Equipe Projet INRIA Magique-3D**, Institut National de Recherche en Informatique et en Automatique (INRIA)

**Laboratoire de Mathématiques Appliquées** Unité Mixte de Recherche CNRS 5142, Université de Pau et des Pays de l'Adour (UPPA)

## Remerciements

Depuis le temps que j'attendais ce moment... D'une part parce que je voulais voir l'aboutissement de ce projet, d'autre part parce que j'avais envie de partager cette fin avec ceux qui m'ont accompagnée, soutenue et encouragée pendant cette aventure unique.

*Merci* tout d'abord à mon directeur de thèse, Mohamed Amara. Vous m'avez fait confiance pour commencer ce voyage dans le monde des éléments finis, de l'analyse numérique et puis, à chaque pas vous m'avez guidée. Vos grandes connaissances m'ont éclairée sur le sujet. Vos conseils, vos idées, vos questions, vos remarques et vos suggestions m'ont énormément apporté et ont fait avancer ce travail. Votre gentillesse et votre bonne humeur m'ont beaucoup aidée surtout dans mes moments de doutes. Pour tout cela, je tiens à vous faire part de ma profonde reconnaîssance !
C'est (enfin !) le moment de dire *Merci* à mon codirecteur de thèse, Rabia Djellouli. Je tiens à vous remercier en premier lieu pour comment vous avez pris le temps de discuter avec moi afin de me faire comprendre ce que c'est la recherche, avec ses hauts et ses bas. Ou plutôt avec mes hauts et mes bas... Votre rigueur m'a appris que si quelque chose (code, figure, démonstration, article, présentation) mérite d'être fait, alors il faut le faire de son mieux. *Merci* aussi pour tous les conseils, pour toutes les questions que vous m'avez posées, pour toutes les réponses que vous avez appréciées. *Merci* pour tous les week-ends que vous avez passés à mes côtés pour m'aider à avancer (j'en profite pour remercier également votre famille). Encore *Merci* pour mes supers séjours aux USA. Votre accueil et notre travail aux USA resteront gravés dans ma mémoire. J'ai appris énormément de choses de notre collaboration, tant au niveau professionnel qu'au niveau personnel. *Merci* pour tout !
Cette thèse a eu lieu dans le cadre d'une collaboration avec le groupe TOTAL pour laquelle je remercie vivement Henri Calandra. *Merci* pour votre accueil dans les locaux TOTAL, pour avoir guidé mes premiers pas dans le milieu industriel. *Merci* également pour toutes les suggestions et les questions que vous m'avez posées. Elles m'ont été d'une grande aide dans mon projet de recherche.

Je suis très reconnaissante envers Hélène Barucq pour m'avoir accueillie dans l'EPI INRIA Magique-3D. *Merci*, Hélène pour toutes les opportunités dont j'ai pu bénéficier en faisant partie de cette équipe. Particulièrement *Merci* pour les voyages aux USA qui ont bien fait avancer mon travail et pour tous les congrès auxquels j'ai participé et qui m'ont ouvert la porte vers le monde des chercheurs. En faisant partie de Magique-3D, j'ai appris beaucoup de choses que je n'aurais pas pu apprendre autrement. *Merci* également d'avoir accepté de présider mon jury de thèse.
Je suis très sensible à l'honneur que m'ont fait Peter Monk et Abderrahmane Bendali en acceptant d'être les rapporteurs de mon manuscript de thèse. Vos remarques, vos questions et vos suggestions m'ont été d'une très grande aide et ont sans aucun doute contribué à la qualité de ce travail. *Merci* !
Je n'oublie pas David Pardo qui a accepté de faire partie de mon jury de thèse. *Merci* également pour la proposition de post-doc sur un sujet qui m'intéresse et qui me passionne. C'est une très grande opportunité pour moi, encore *Merci* !

Je parle déjà de l'avenir, mais je voudrais faire quelques pas en arrière et m'arrêter à mes premiers jours en France, à Pau et à la personne qui m'a beaucoup aidée à découvrir la France et la recherche, Daniela Capatina. *Merci* Dana, pour tous tes encouragements, pour tous les moments où tu as été à mon écoute et où tu as guidé mes pas. *Merci* pour tous les moments de détente passés ensemble et pour tout ce que j'ai pu apprendre de toi (maths et bien d'autres encore !). Je tiens à exprimer aussi ma reconnaissance envers David Trujillo. D'abord *Merci* David de m'avoir proposé un sujet de stage de DEA pendant lequel j'ai découvert la mise en œuvre des éléments finis. *Merci* pour tous les conseils et pour toutes les astuces de programmation que j'ai apprises pendant

notre travail et qui m'ont énormément aidée pour la suite. *Merci* également pour tout le soutien concernant les démarches administratives qui ont facilité mon séjour et mon adaptation en France.

Je remercie l'ensemble du personnel du LMA, dirigé d'abord par Mohamed Amara et ensuite par Laurent Bordes. J'ai passé des années inoubliables dans le Laboratoire de Mathématiques Appliquées de Pau et je leur remercie de m'y avoir accueillie. Merci Chantal Blanchard pour votre gentillesse et pour tous les moments d'encouragement et de partage, Marie-Claire Hummel pour votre bonne humeur et pour votre disponibilité et Lina Gonçalves pour votre précieuse aide dans mes recherches documentaires.
Un clin d'œil à Josy Baron : sans que tu sois à Pau, tu m'as soutenue pendant tout ce projet. Ton efficacité, ton savoir et ta patience m'ont bien facilité les missions (et il y en a eu...), ta bonne humeur m'a donné du courage, ta présence le jour de la soutenance m'a touchée. Pour tout cela *Merci* !

Mais qu'auraient été ces années de thèse sans la bonne ambiance du bureau 214 ? On en a vécu des choses et elles resteront pour toujours dans mon cœur. Je remercie avec cette occasion tous les doctorants et docteurs qui ont contribué à ces beaux moments et qui m'ont encouragée, d'une façon ou d'une autre pendant mes moments difficiles. *Merci* Hannen, Ali, Julie, Nour, Véro, Cyril, Anne-Gaëlle, Ronan, Julien, Meriem, Pieyre ! Une pensée particulière pour Florian : *Merci* pour cette belle amitié qui a grandi au fil des années. *Merci* pour la plus belle voiture du monde, tu auras facilité mon chemin vers la recherche. Je te remercie également pour toutes nos discussions philosophiques sur la cuisine française, le foot français, la France en général. Tu es un Français véritable, plus aucun doute ! *Merci* Fabien pour ton amitié inconditionnelle, pour toutes les discussions concernant des idées (voire des événements...), pour tous les moments d'encouragement et pour toutes les rigolades qui m'ont fait beaucoup de bien entre deux codes lancés et trois bugs trouvés. *Merci* particulièrement pour la patience de me dire au moins une fois par semaine de considérer d'abord des cas simples. J'espère avoir appris ma leçon. Je n'oublie pas Agnès : *Merci* d'avoir été en face de moi au bureau 214 et d'avoir profité de cette place pour partager mes moments de joie et de tristesse. Nos discussions sur la terrasse et au RU m'ont enrichie et m'ont appris beaucoup de choses (tu les connaîs !). Entre autre mon français s'est amélioré avec ton aide et puis... on a conclu ensemble que peu importe si c'est 'Salut', 'Hi', 'Hola', il faut suivre ses rêves !
En parlant de collègues de travail, je me dois de remercier mes amis de TOTAL. En particulier je pense à Rached : *Merci* de m'avoir écoutée et aidée lorsque j'en ai eu besoin. *Merci* également pour tous les moments de partage autour d'un café ou d'un repas, ils m'ont beaucoup apporté ! Un clin d'œil à Céline : *Merci* pour ta bonne humeur quotidienne, pour tes conseils et ton soutien, surtout vers la fin de ce projet ! *Merci* aussi pour ta disponibilité et ton aide le jour de ma soutenance de thèse, tu m'auras offert de très jolis souvenirs ! Je pense également à Caro : on a été ensemble dans le bureau 214, en Magique-3D, à Total, pendant nos voyages inoubliables : Londres (la dernière heure a dû te paraître une éternité...), Marseille (ah, ce Tarot) et surtout Cannes (les stars quoi ! ! !), ainsi que pendant nos soirées mémorables avec le GBU. Pour tous ces moments et pour tes délicieux beignets, *Merci* !

Je n'oublie pas Alex & Gaby, Manue & Jérém, Brigitte & Manuel. Vous avez été ma famille en France et vous m'avez entourée d'un amour et d'une amitié rares... *Merci* du fond du coeur ! Vos pensées et vos prières m'ont donné des ailes et du courage. Le jour où je suis arrivée en France vous m'avez accueillie et entourée. Le jour de la soutenance vous avez tous été avec moi, d'une façon ou d'une autre et cela m'a beaucoup touchée. Vous êtes des gens merveilleux, uniques et chers à mon cœur !

J'en arrive maintenant à ceux qui ont été le plus près de moi, tout en étant loin... Maman, tu ne liras jamais ces mots, mais ton cœur, ta voix, ton sourire et ton amour m'ont accompagnée et m'accompagneront toujours. Du fond du cœur *Merci* pour comment tu as coloré ma vie avec la tienne. C'est de toi que j'ai appris l'amour, la foi, la paix, le courage, la force dont j'ai tellement eu besoin pendant ce projet ! Papa, *Merci* d'avoir vécu avec moi cette aventure avec ses moments de bas et de haut. Tu étais en Roumanie, mais tu étais en France aussi. Merci pour tous tes encouragements et pour ton soutien inconditionnel pendant toutes mes études. Je sais que tu seras là pour moi pendant toute ma vie. *Merci* de m'avoir insufflé l'amour pour le savoir et pour la culture, ainsi que la curiosité pour tout ce qui est nouveau. Oh, combien j'en ai eu besoin pendant ma recherche... Merci à tous les deux, je vous dois cette réussite ! Mais je la dois aussi à toi, Daniel ! C'est toi, mon grand frère, qui as guidé mes pas vers les maths, alors que tu ne les aimais vraiment pas... C'est toi qui as été un modèle de rigueur, de passion pour le travail. C'est toi qui m'as donné un conseil si précieux : dès que ta bonne humeur et/ou ta santé sont en danger, fais une pause ! C'est toi qui as su répondre à beaucoup de mes questions informatiques. Ta présence à mes côtés le jour de la soutenance a été mon plus grand cadeau, *Merci* ! Une pensée particulière pour mes grands-parents. Votre gentillesse, votre amour et vos prières sont des piliers pour ma vie et ce depuis mon enfance ! *Merci* !

Je pense aussi à mes chers amis Roumains qui sont, eux aussi, loin, mais si proche de mon cœur, de ma vie. Otilia & Sami, Anca & Iulian, Anca & Petre, Monica & Ben, Eliza & Andrei, *Merci* pour tout votre soutien ! Chaque texto, chaque e-mail, chaque coup de fil, a été un réel cadeau pour moi, un rappel de tous les moments passés ensemble qui resteront à jamais dans mon cœur et un vrai encouragement d'aller plus loin !

Enfin, mais avant tout, *Merci* à mon tendre époux, Ionut. Tu as vécu avec moi cette aventure et tu l'as rendue belle, tout simplement. Tu as été là dans tous mes moments de doute, de larme, de joie, de fatigue, de désespoir, d'espoir, de réussite. Ta patience m'a soutenue quand je n'étais vraiment pas agréable. Ta rigueur m'a aidée à me poser des questions et à ne pas renoncer avant de trouver au moins une réponse. Tes idées m'ont aidée à chasser les bugs. Ta maîtrise de logiciels graphiques et informatiques m'a fait avancer plus vite que je n'aurais su le faire. Ta bonne humeur m'a réconfortée. Ton équilibre m'a aidée à trouver le mien. Ta bonté m'a rempli le cœur de joie et de bonne humeur. Tes petites et grandes attentions m'ont déconnectée alors que j'étais indéconnectable. Ton humour a souvent changé mes larmes de tristesse en larmes de joie et de rire. Ton amour m'a donné des ailes et m'a fait dire, au moins une fois par jour, pendant trois ans : *La vie est très très belle !* Je sais que, à tes côtés, je le dirai tous les jours de ma vie ! Pour tout cela et pour tout ce que tu es, *Merci* !

A Celui qui m'a donné la Vie, à Celui qui m'a gardé en bon état de santé, à Celui qui me montre tous les matins le lever du soleil et tous les soirs son coucher, à Celui qui m'a donné une identité, à Celui qui m'a entourée de tous ces gens merveilleux, à mon Créateur, mon Dieu et mon Père... toute ma reconnaissance !

# Table des matières

# Introduction Générale

# 1. Préambule

La propagation des ondes est un phénomène physique avec d'importantes applications dans l'imagerie sismique, l'imagerie médicale, l'archéologie, la sismologie, le radar, et le sonar. La plupart de ces applications conduisent à ce qu'on appelle des problèmes de scattering, c'est-à-dire des processus où les ondes rencontrent un obstacle et où on se retrouve avec un phénomène de diffraction (scattering). Selon le milieu où les ondes se propagent, on distingue : des ondes acoustiques (propagation dans un milieu fluide), des ondes élastiques (milieu élastique) et des ondes éléctromagnétiques (milieu magnétique). En général, les équations d'ondes sont dépendantes du temps (transitoires).

Les modèles mathématiques pour ces processus physiques sont bien connus. De plus, les problèmes résultants sont, pour la plupart, résolus mathématiquement puisque l'existence et l'unicité des solutions ont déjà été montrées [10, 31, 35, 41, 42]. Par contre, pour la mise en œuvre dans les applications citées ci-dessus, la solution est demandée explicitement ou au moins quantifiée, d'où la difficulté du problème. Les nombreux travaux concernant les méthodes numériques adaptées à ce type de problèmes ont fait avancer la recherche, mais ce domaine reste quand même un grand défi pour le calcul scientifique.

L'équation la plus étudiée parmi celles qui modélisent la propagation des ondes est l'équation des ondes acoustiques. Elle modélise la propagation d'une onde dans un milieu homogène (fluide) $\Omega$ et s'écrit comme une équation hyperbolique du second ordre dont la solution est le champ de pression acoustique que l'on va noter $p = p(t, \boldsymbol{x})$. En désignant par $c$ la vitesse de propagation (que l'on suppose constante), $p$ vérifie :

$$\frac{\partial^2 p}{\partial t^2} - c^2 \Delta p = F \quad \text{dans } \mathbb{R}_+ \times \Omega, \quad \text{avec} \quad \Omega \subset \mathbb{R}^n. \tag{1}$$

Dans certaines applications (comme l'utilisation du radar pour détecter des corps en mouvement), on adopte le régime harmonique et par conséquent $F$ s'écrit sous la forme :

$$F(t, \boldsymbol{x}) = f(\boldsymbol{x})\, e^{-\mathrm{i}\,\omega t}. \tag{2}$$

La solution de l'équation initiale correspond alors à un type particulier d'ondes qui sont qualifiées de monochromatiques. Ce qualificatif vient du fait que ces ondes sont associées à une et une seule fréquence, donc à une et une seule longueur d'onde. Une onde monochromatique définit une solution harmonique de l'équation des ondes. C'est une fonction de la forme :

$$p(t, \boldsymbol{x}) = u(\boldsymbol{x})\, e^{-\mathrm{i}\,\omega t}, \tag{3}$$

où $\omega$ représente la pulsation du signal ainsi émis. Si on cherche la solution de l'équation précédente sous cette forme, on retrouve l'équation de Helmholtz d'inconnue $u$ :

$$-\Delta u - k^2 u = f \quad \text{dans} \quad \Omega, \tag{4}$$

avec $k$ désignant le nombre d'onde, lié à la fréquence $f_0$ de l'onde par $k = \frac{2\pi f_0}{c}$.

La résolution numérique des problèmes de Helmholtz rencontre deux difficultés majeures. La première est liée au caractère non-borné des domaines de résolution. Afin de surmonter cette difficulté, de nombreuses méthodes ont été développées. Le lecteur intéressé pourra consulter les monographies et les articles de synthèse portant sur les trois grandes classes de méthodes : les couches absorbantes,

communément connues sous le nom de PML (Perfectly Matched Layers) [40, 37], les éléments infinis [8, 3] et les conditions absorbantes [30, 25, 20, 19, 33]. La deuxième difficulté majeure et à laquelle nous nous sommes intéressés dans ce travail porte sur l'effet du régime de fréquence caractérisé par le paramètre $ka$ ($k$ est le nombre d'onde et $a$ la dimension caractéristique du domaine). Le nombre d'onde caractérise le comportement oscillatoire de la solution exacte : plus $ka$ est grand, plus la solution exacte oscille. Cette dépendance doit être prise en compte par le modèle numérique considéré, ce qui, pour les éléments finis standards par exemple, devient très cher et même parfois impossible à programmer pour des cas 3D simulant des phénomènes réels.
Les méthodes standards de type éléments finis (FEM) ne sont donc pas adaptées à l'équation de Helmholtz. En effet, il a été montré dans [7, 6] qu'en régime haute fréquence (qui correspond à des $ka$ grands), la précision de l'approximation est sujet d'une pollution numérique. Pour l'élément $P_1$ par exemple, il a été montré dans [7, 29] que pour $k^2h$ assez petit, l'erreur relative en norme $L^2$ (resp. en seminorme $H^1$) est de l'ordre de $k^3h^2$ (resp. $kh$). Si, de plus, $kh$ est assez petit, il a été montré dans [21] que l'erreur en seminorme $H^1$ est, elle aussi, majorée par $k\,(kh)^2$. Cela veut dire que lorsque $ka$ est grand, pour avoir un niveau de précision spécifique, il n'est pas suffisant de maintenir $kh$ constant (ce qui correspond à un nombre constant d'éléments par longueur d'onde) comme il a été longtemps cru, car l'erreur est majorée au mieux par un terme de l'ordre de $k^3h^2$. Par conséquent, augmenter $k$ et approcher la solution avec une précision fixée demande de considérer de plus en plus d'éléments par longueur d'onde. Cela se traduit par la résolution de systèmes de très grandes tailles et la méthode devient trop coûteuse. Il devient donc presque impossible de résoudre avec les méthodes FEM des problèmes 3D issus de situations réelles.

## 2. État de l'Art

Plusieurs des méthodes numériques développées dans le but de surmonter cette difficulté exploitent l'utilisation des ondes planes. D'une part, ce sont des fonctions oscillantes et donc on s'attend à ce qu'elles reproduisent mieux les nombreuses oscillations de la solution à haute fréquence. D'autre part, les ondes planes vérifient l'équation de Helmholtz homogène et par conséquent, l'idée d'approcher la solution avec des ondes planes incorpore des propriétés *a priori* de la solution. La plupart de ces méthodes appartiennent à la classe des méthodes variationnelles multi-échelle (Variational Multiscale, VMS) [27, 28]. L'idée de base de VMS est de décomposer la solution approchée $u_h$ dans un champ polynomial $u^P$ et un champ d'enrichissement $u^E$, de telle sorte que l'espace d'approximation de $u_h$ soit la somme directe entre les espaces discrets correspondant à $u^P$ et $u^E$. Les différentes méthodes issues de VMS sont basées sur différentes interprétations de $u^E$. Les méthodes suivantes nous semblent être les plus prometteuses pour résoudre efficacement les problèmes de Helmholtz.

- *Residual-Free Bubbles (RFB)*
  Cette méthode a été proposée par Franca *et al* dans [17]. L'idée essentielle de cette approche est d'enrichir localement le champ à l'aide des ondes planes. Le calcul de $u^E$ revient à résoudre, au niveau des éléments du maillage, des problèmes de Helmholtz avec des conditions aux limites de type Dirichlet homogène. Ensuite, le champ polynomial est calculé dans un cadre variationnel. Les expérience numériques présentés montrent une supériorité de la méthode sur FEM. Cependant, la méthode est compatible avec des domaines et des maillages rectangulaires seulement car la résolution des problèmes de Helmholtz avec condition de bord Dirichlet homogène au niveau de l'élément est très couteuse pour des maillages non-structurés.
- *Partition of Unity Method (PUM)*
  Cette méthode a été développée par Babuška-Melenk dans [5]. Comme le nom l'indique, PUM est une méthode basée sur une partition de l'unité du domaine. Le champ polynomial classique

est enrichi (par multiplication) avec des ondes planes. La présence des polynômes assure la continuité de la solution et localise l'utilisation des ondes planes. PUM délivre de très bons résultats en régime basse fréquence, contrairement aux régimes moyenne et haute fréquence car la matrice issue de la discrétisation est très mal conditionnée.

- *Generalized Finite Element Method (GFEM)*
  GFEM a été développée par Strouboulis *et al* dans [38] dans le but de corriger le mauvais conditionnement de PUM. La méthode est similaire à PUM dans la façon de combiner le champ polynomial et les ondes planes. Dans PUM cette quantité représente l'approximation complète, ce qui n'est pas le cas dans GFEM, où la solution est approché par le produit en question seulement. Il ne nous semble pas que GFEM ait réussi à traiter efficacement le problème du conditionnement rencontré dans PUM.
- *Weak-Element Method (WEM)*
  Cette méthode développée par Rose dans [36] est la première méthode à avoir utilisé des ondes planes pour l'approximation du champ d'onde. Basée sur une méthode de décomposition du domaine, WEM approche la solution dans chaque maille en utilisant une superposition d'ondes planes, la fonction résultante étant discontinue. Une continuité au sens des valeurs moyennes du champ et de sa dérivée normale est imposée sur les frontières intérieures du maillage. Les conditions de bord sont, elles aussi, imposées sur les valeurs moyennes du champ et de sa dérivée normale. Au niveau algébrique, la méthode revient à résoudre un système linéaire dont la taille dépend du nombre de frontières du maillage et du nombre d'ondes planes approchant la solution au niveau local. C'est donc une méthode qui peut devenir excessivement chère avec l'augmentation de la dimension de la base locale.
- *Least-Squares Method (LSM)*
  Cette méthode a été développée par Monk-Wang dans [34]. C'est une méthode qui s'appuie, elle aussi, sur la décomposition du domaine de calcul en sous-domaines. Pour l'approximation au niveau local deux choix de fonctions sont proposées : les ondes planes et les fonctions de Bessel. Ces deux types de fonctions de forme sont solutions de l'équation de Helmholtz. Dans les deux cas de figures, on introduit un paramètre, dans le but de minimiser le saut de la solution obtenue, ainsi que celui de sa dérivée normale aux interfaces et de reconstituer les conditions aux limites. Cela revient donc à minimiser une fonction au sens des moindres carrés, d'où le nom de la méthode. Cette approche correspond, au niveau algébrique, à la résolution d'un système linéaire, associé à une matrice Hermitienne et définie positive. La taille du système est directement liée au nombre de sous-domaines contenus dans le maillage et à la dimension des espaces locaux d'approximation. La méthode s'avère performante, mais elle risque de devenir assez couteuse en régime haute fréquence notamment en 3D, car augmenter la base locale d'approximation conduit à une augmentation de la taille du système global et donc de l'effort de calcul.
- *Ultra-Weak Variational Formulation (UWVF)*
  Comme le nom l'indique, cette méthode développée par Cessenat-Després dans [9] est construite dans un cadre variationnel. Le nomenclature "ultra faible" vient du fait que le problème variationnel est obtenu après deux intégrations par parties. Comme dans WEM, le champ d'onde est approché dans une base d'ondes planes, résultant dans une fonction discontinue. La continuité aux interfaces est restaurée au sens faible, en résolvant un système dont les inconnues sont définies sur les interfaces du maillage. Cela réduit le coût global de calcul, par rapport à WEM et LSM, mais des problèmes auxiliaires locaux doivent être résolus auparavant. La discrétisation du système obtenu mène, comme dans LSM, à un système linéaire associé à une matrice Hermitienne et définie positive, mais qui souffre d'un mauvais conditionnement.
- *Discontinuous Galerkin Method (DGM)*

La méthode dite DGM, développée par Farhat *et al* dans une série d'articles [13, 14, 15] est une méthode qui provient de DEM (Dicontinuous Enrichment Method) [11, 12], qui est une méthode de type VMS. La particularité de DEM, comparée aux autres techniques (RFB, PUM, GFEM), réside dans son caractère discontinu. Plus précisément, dans chaque maille la solution est approchée par un champ polynomial, auquel on ajoute un champ d'ondes planes qui, lui, est discontinu. Des multiplicateurs de Lagrange sont introduits pour restaurer la continuité de la solution dans un sens faible. Dans [13] il a été montré que dans DEM le poids dans l'approximation est porté par le champ d'enrichissement. Par conséquent, le champ polynomial a été enlevé, résultant dans la méthode DGM (Discontinuous Galerkin Method). Dans DGM la solution est donc approchée dans une base d'ondes planes seulement, ce qui conduit, comme le nom l'indique, à une solution discontinue sur les frontières intérieures du maillage, comme dans WEM, LSM et UWVF. En outre, DGM hérite de DEM la technique de restauration de la continuité au sens faible en utilisant des multiplicateurs de Lagrange. Cette approche conduit à une méthode dont le coût de calcul est donné uniquement par le nombre d'interfaces du maillage et par le nombre des multiplicateurs de Lagrange considérés sur chaque interface. DGM s'avère très performante lors de la comparaison avec FEM ([13, 14, 15]), en réduisant considérablement le nombre de degrés de liberté nécessaires pour obtenir une solution ayant une précision donnée. Les inconvénients de DGM sont principalement liés à la condition *inf-sup* discrète qui se traduit, en pratique, par une condition de compatibilité que les espaces de discrétisation doivent satisfaire. Cette condition restreint le nombre de multiplicateurs de Lagrange par interface, ce qui pose problème pour les éléments d'ordre supérieur, car, à notre connaissance, il n'y a pas de résultat théorique sur comment sélectionner convenablement les multiplicateurs de Lagrange. Le deuxième point faible de DGM est lié aux instabilités numériques observées pour des maillages fins [4]. La source de ces instabilités réside principalement dans le caractère mal-posé des problèmes locaux qui, au niveau algébrique se traduit par des singularités dans les matrices locales. Enfin, l'utilisation de DGM sur des maillages non-structurés entraîne une perte de précision dans l'approximation.

Pour finir cette liste non-exhaustive des méthodes récentes pour la résolution des problèmes de Helmholtz, nous présentons brièvement deux autres méthodes de type stabilisé, basées sur l'approximation polynomiale standard.

- *Galerkin Least Squares Method (GLSM)* and *Galerkin-Gradient Least-Squares Method (GGLSM)*
  GLSM, développée par Harari-Hughes dans [22] s'inscrit, elle aussi, dans la lignée des méthodes dérivées de VMS. C'est une méthode de type stabilisé, dans laquelle un terme contenant le résidu est ajouté dans la formulation variationnelle standard. Ce terme est multiplié par un paramètre de stabilisation, dont la valeur a été déterminée pour des maillages réguliers uniquement en 1D et 2D. Cet inconvénient conduit à des résultats moins précis pour des maillages irréguliers. GGLSM a été développée par Franca *et al* dans [16]. C'est une méthode obtenue en ajoutant dans la formulation variationnelle un terme résiduel contenant des gradients, dans le but de contrôler les dérivées. Dans [23] il a été montré que les performances numériques de GLSM et GGLSM sont comparables pour des maillages structurés et non-structurés et supérieurs à FEM. En régime basse fréquence ($ka < 5$), les résultats obtenus montrent une performance de l'ordre de 1% pour des résolutions qui, en moyenne ne dépassent pas 25 éléments par longueur d'onde. La méthode risque de devenir chère à haute fréquence.

## 3. Objectif et description du travail

Le but de ce travail est la mise au point d'une méthode robuste pour résoudre efficacement les problèmes de Helmholtz. Les performances et les limitations de DGM nous ont poussé à exploiter l'idée de l'approximation du champ d'onde par une superposition d'ondes planes, dans un cadre discontinu, que l'on corrige à l'aide des multiplicateurs de Lagrange. D'autre part, les formulations au niveau continu et les matrices associées au niveau algébrique qui interviennent révèlent des points communs avec LSM. Enfin, dans la manière de poser les problèmes locaux et de choisir les espaces sur lesquels les formulations globales sont écrites, ce travail ressemble aussi à UWVF. Nous proposons donc une méthode qui possède les ingrédients de DGM, LSM et UWVF et qui, par son originalité et ses performances trouvera (espérons-le) sa place parmi les nombreuses méthodes existantes. Il s'agit d'une méthodologie qui, au niveau pratique, revient à résoudre dans un premier temps des problèmes de Helmholtz locaux qui sont indépendants et bien posés. Ensuite, on résout un système global qui nous permet de reconstituer la solution numérique. Le présent manuscript est constitué de quatre parties. Chacune d'elle représente une étape de ce travail de thèse. A noter que dans la partie I on s'est plus penché sur les aspects théoriques d'un problème de scattering (cadre fonctionnel, analyse d'erreur), alors que dans les trois parties suivantes on a plutôt mis l'accent sur la partie algorithmique et numérique de l'étude. Cependant, les résultats théoriques présentés dans la partie I sont facilement applicables aux formulations proposées dans les parties II et III. L'accent mis sur les résultats numériques est justifié par le fait que souvent (et en particulier dans la partie I) les résultats numériques ne confirment pas toujours les attentes motivées par la théorie. Un bon exemple est celui des matrices qui en théorie sont inversibles et qui deviennent numériquement presque singulières.

Dans la partie I on décrit d'abord le cadre fonctionnel et la stratégie de résolution d'une classe de méthodes de type Galerkin discontinu dans lesquelles la continuité aux interfaces est restaurée dans un sens faible, à l'aide des multiplicateurs de Lagrange. Cette stratégie met en évidence les difficultés non-résolues de DGM, toutes liées à des problèmes dont le caractère bien posé n'est pas assuré. Dans un premier temps on a essayé de modifier les problèmes locaux de DGM, ce qui nous a conduit à une formulation plus stable mais moins précise. Afin de traiter les instabilités numériques de DGM, on a proposé une nouvelle approche du système global. Notre approche étant simple (voir peut-être même naïve), on a simplement modifié la manière adoptée dans DGM de restaurer la continuité au sens faible. On a donc réécrit cette condition de façon à ce que la discrétisation du problème global conduise à une matrice Hermitienne et définie positive, sous un hypothèse de compatibilité. Cette nouvelle condition est, à quelques détails près, de type moindres carrés, ce qui représente un point commun avec LSM. Pour la nouvelle méthode nous avons montré la convergence théorique de l'erreur en effectuant l'analyse mathématique de la formulation. La convergence a été confirmée par les résultats numériques dans le cas de l'élément connu comme $R$-4-1. A noter que pour cet élément, la performance de la nouvelle méthode est comparable à DGM. En revanche, la méthode que nous avons construite semble être relativement supérieure à DGM pour les éléments d'ordre supérieur. A titre d'exemple, l'élément $R$-8-3 donne une meilleure précision que DGM. Cependant, les expériences numériques ont montré que la nouvelle méthode restait instable et ceci en dépit des résultats théoriques. Une analyse plus profonde nous a aidé à conclure que la source de ces instabilités résidait dans la résolution des problèmes locaux qui, comme pour DGM, devenaient *numériquement* presque singuliers avec le raffinement du maillage.

Dans la partie II de ce travail on s'est donc penché sur le traitement des problèmes locaux pour rendre leur résolution plus stable. On a donc construit une nouvelle formulation des problèmes de

Helmholtz, que l'on a appelée mDGM (modified Discontinuous Galerkin Method). Plus précisément, on a imposé des conditions de type Robin sur le bord de chaque élément du maillage. Cela nous a permis de construire des problèmes locaux bien posés. Le prix à payer a été la perte de la continuité des multiplicateurs de Lagrange et par conséquent de la continuité de la dérivée normale du champ d'onde. Nous avons restauré la continuité - dans un sens faible - du champ et de sa dérivée normale avec le même type de condition utilisée dans la partie I, ce qui nous a permis de conserver les bonnes propriétés de la matrice globale : Hermitienne et définie positive. Au niveau pratique, cette technique a doublé le nombre d'inconnues du système global à résoudre, mais les résultats obtenus se sont avérés beaucoup plus précis et plus stables que ceux obtenus avec DGM. Par exemple, pour un élément d'ordre inférieur ($R$-4-2) on a montré que mDGM maintenait la stabilité et la précision pour des résolutions aussi fines que 2700 éléments par longueur d'onde, contrairement à DGM, qui explose pour des maillages beaucoup moins fins. Pour des éléments finis d'ordre supérieur (huit ondes planes pour l'approximation au niveau local), la comparaison avec DGM a montré une supériorité nette de mDGM, en terme de précision et de stabilité. En effet, dans la région où DGM est stable, mDGM a réduit l'erreur par un facteur 5. De plus, l'investigation de la stabilité lors du raffinement du maillage a montré que mDGM maintenait une erreur relative de l'ordre de 1% en allant jusqu'à plus de 600 éléments par longueur d'onde, alors que DGM devient complètement instable (une erreur relative de plus 100%) quand la résolution du maillage dépasse 120 éléments par longueur d'onde. En dépit de cette évidente supériorité, on a été surpris par des instabilités numériques que l'on a observées lorsque l'on a utilisé huit ondes planes par élément. Là encore, on a mené une étude approfondie des propriétés des problèmes locaux et on s'est aperçu qu'ils devenaient presque singuliers, contrairement à nos attentes. Les différentes expériences et études théoriques nous ont aidé à conclure que les 8 ondes planes qui se constituaient en fonctions de base devenaient numériquement linéairement dépendantes, ce qui déstabilisait la méthode. La validation de cette hypothèse a été faite en considérant 7 ondes planes par élément, dont la linéaire indépendance s'était avérée moins sensible au raffinement du maillage que celle des 8 ondes considérées précédemment. La précision et la stabilité du nouvel élément (sept ondes planes et deux multiplicateurs de Lagrange) implémenté dans mDGM sont remarquables, comme il sera montré dans la partie II.

La perte de l'indépendance linéaire des fonctions de base s'est constituée en un nouveau défi à relever. En gardant les problèmes locaux construits dans mDGM, on s'est proposé d'adopter une nouvelle formulation variationnelle de ceux-ci pour mieux gérer l'inconvénient de la dépendance linéaire (numérique) des fonctions de base. La partie III propose une nouvelle formulation dérivant de mDGM, mais améliorée, comme le montreront les résultats présentés dans cette partie. Cette supériorité de la nouvelle méthode sur mDGM nous a conduit à l'appeler imDGM (improved modified Discontinuous Galerkin method). La différence majeure entre les deux méthodes réside donc dans la formulation variationnelle adoptée pour la résolution des problèmes locaux. Dans imDGM on a construit des problèmes locaux qui conduisent à des matrices Hermitiennes et définie positives. De plus, ces matrices se sont avérées mieux conditionnées que celle obtenues avec mDGM, ce qui nous a permis d'obtenir des résultats avec un niveau de précision plus élevé. Le caractère Hermitien et défini positif de ces matrices locales nous permettra, si besoin, de les préconditionner plus facilement que les matrices issues de mDGM. A noter que le côut de calcul et l'effort d'implémentation sont identiques pour les deux méthodes, une raison de plus de choisir imDGM car la formulation délivre des meilleurs résultats pour le même prix. La ressemblance de imDGM (et implicitement de mDGM) avec LSM nous a poussé à comparer la précision et la stabilité des résultats avec celles obtenues avec LSM. Par conséquent, dans la partie III on a fait des comparaison entre imDGM et mDGM d'une part et imDGM, LSM et DGM d'autre part. Ces comparaisons concernent la précision et la stabilités des quatre formulations, en régime moyenne et haute fréquence. ImDGM

est indéniable supérieure à DGM en termes de précision et stabilité. En revanche, imDGM et LSM présentent des performances comparables et ce au moins en vu des expériences numériques que l'on a effectuées. La différence entre les deux méthodes pourraient venir du coût de calcul. Il semble que imDGM pourrait s'avérer moins chère que LSM pour les éléments d'ordre supérieur. Cette remarque nécessite une étude plus approfondie pour la vérifier.

Enfin, nous avons fini ce travail dans une note applicative que nous présentons dans la Partie IV. A la suggestion du groupe TOTAL, on a testé et validé la méthode décrite dans la partie I dans le cadre d'une application très simple, mais représentative pour les expériences industrielles. Plus précisément, on a simulé la propagation en temps d'une onde générée par l'explosion d'une source à la surface d'un domaine rectangulaire. On a considéré d'abord un domaine homogène, ensuite un autre constitué de deux couches correspondant à des vitesses différentes. Le but a été de montrer que la méthode proposée délivre des résultats cohérents, correspondant à la réalité. L'expérience en deux couches surtout a dû prouver que le caractère Hermitien et défini positif de la matrice globale n'est pas affecté par les différentes interfaces du domaine (qui correspondent donc à des différentes valeurs de $k$ d'une couche à l'autre). La preuve offerte par le résultat présenté dans la partie IV a encouragé le groupe TOTAL à vouloir utiliser la méthode pour des problèmes inverses. D'une part résoudre l'équation de Helmholtz à la place de l'équation des ondes réduit considérablement les côuts de calcul, car la discrétisation en temps est supprimée. D'autre part, les bonnes propriétés de la matrice globale permettent l'utilisation d'algorithmes robustes et efficaces, comme le gradient conjugué par exemple qui évite le calcul et le stockage de la matrice. Bien évidemment, ce "détail" est extrêmement important pour les applications industrielles. Des tests sur des maillages non-structurés et sur des domaines correspondant à la réalité sont en cours.

# Bibliographie Générale


[1] Amara M., Djellouli R., Farhat C., Convergence analysis of a discontinuous Galerkin method with plane waves and Lagrange multipliers for the solution of Helmholtz problems *SIAM J. Numer. Anal.*, **47**(2), 1038-1066, 2009

[2] Amara M., Barucq H., Bernardini A., Djellouli R., A mixed-hybrid method for solving mid-and high-frequency Helmholtz problems *International Conference on Theoretical and Computational Acoustics Theoretical and Computational Acoustics 2007*, Grèce Heraklion, P. P. Michael Taroudakis (editor), 2008

[3] Astley R. J., Infinite elements for wave problems : a review of current formulations and an assessment of accuracy *Internat. J. Numer. Methhods Eng.*, **49**(7), 951-976, 2000

[4] Amara M., Calandra H., Djellouli R., Grigoroscuta-Strugaru M., A modified discontinuous Galerkin method for Helmholtz problems *Technical Report INRIA No. 7050* 2009, available online at http ://hal.archives-ouvertes.fr/inria-00421584/fr/ ;

[5] Babuška I., Melenk I.J.M., The partition of unity method *Internat. J. Numer. Methhods Eng.*, **40**, 727-758, 1997

[6] Babuška I., Sauter S., Is the Pollution Effect of the FEM Avoidable for the Helmholtz Equation Considering High Wave Numbers ? *SIAM J. Numer. Anal.*, **34**, 2392-2423, 1997

[7] Bayliss A., Goldstein C. I., Turkel E., On accuracy conditions for the numerical computations of waves *J. Comput. Physics*, **59**, 396-404, 1985

[8] Bettess P., Infinite elements *Internat. J. Numer. Methhods Eng.*, **11**, 53-64, 1977

[9] Cessenat O., Despres B., Application of an ultra-weak variational formulation of elliptic PDEs to the two-dimensional Helmholtz problems *SIAM J. Numer. Anal.*, **35**, 255-299, 1998

[10] Colton D., Kress R., Inverse Acoustic and Electromagnetic Scattering Theory *Berlin : Springer* 1992

[11] Farhat C., Harari I., Franca L. P., The discontinuous enrichment method *Comput. Methods Appli. Mech. Eng.*, **190**, 6455-6479, 2001

[12] Farhat C., Harari I., Hetmaniuk U., The discontinuous enrichment method for multiscale analysis *Comput. Methods Appli. Mech. Eng.*, **192**, 3195-3209, 2003

[13] Farhat C., Harari I., Hetmaniuk U., A discontinuous Galerkin method with Lagrange multipliers for the solution of Helmholtz problems in the mid-frequency regime *Comput. Methods Appli. Mech. Eng.*, **192**, 1389-1419, 2003

[14] Farhat C., Wiedemann-Goiran P., Tezaur R., A discontinuous Galerkin method with plane waves and Lagrange multipliers for the solution of short wave exterior Helmholtz problems on unstructured meshes *Wave Motion*, **39**, 307-317, 2004

[15] Farhat C., Tezaur R., Wiedemann-Goiran P., Higher-order extensions of a discontinuous Galerkin method for mid-frequency Helmholtz problems *Internat. J. Numer. Methhods Eng.*, **61**, 1938-1956, 2004

[16] Franca L. P., Dutra do Carmo E. G., The Galerkin gradient least-squares method *Comput. Methods Appli. Mech. Eng.*, **74**, 41-54, 1989

[17] Franca L.P., Farhat C., Macedo A.P., Lesoinne M., Residual-free bubbles for the Helmholtz equation *Internat. J. Numer. Methhods Eng.*, **40**, 4003-4009, 1997

[18] Gillman A., Djellouli R., Amara M., A Mixed Hybrid Formulation Based on Oscillated Finite Element Polynomials for Solving Helmholtz Problems *J. Comput. Appli. Maths*, **204**(2), 515–525, 2007

[19] Givoli D., Numerical Methods for Problems in Infinite Domains. Studies in Applied Mechanics. *Elsevier Scientific Publishing Co., Amsterdam*, **33**, 4003-4009, 1992

[20] Grote M. J., Keller J. B., On nonreflecting boundary conditions *J. Comput. Physics*, **122**, 231-243, 1995

[21] Hadamard J., Lectures on Cauchy's Problem in Linear Partial Differential Equations *Yale University Press, New Haven* 1923

[22] Harari I., Hughes T.J.R., Galerkin/least-squares finite element methods for the reduced wave equation with non-reflecting boundary conditions in unbounded domains *Comput. Methods Appli. Mech. Eng.*, **98**, 411-454, 1992

[23] Harari I., Magoulès F., Numerical investigation of stabilized finite element computation for acoustics *Wave Motion*, **39**, 339-349, 2004

[24] Harari I., Hughes T.J.R., Finite element methods for the Helmholtz equation in an exterior domain : model problems *Comput. Methods Appli. Mech. Eng.*, **87**, 59-96, 1991

[25] Harari I., Hughes T.J.R., Analysis of continuous formulations underlying the computation of time-harmonic acoustics in exterior domains *Comput. Methods Appli. Mech. Eng.*, **97**, 103-124, 1992

[26] Hörmander L., The Analysis of Linear Partial Differential Operator *Springer-Verlag, New York* 1985

[27] Hughes T. J. R., Multiscale phenomena : Green's functions, the Dirichlet-to-Neumann formulation, subgrid scale models, bubbles and the origin of stabilized methods *Comput. Methods Appli. Mech. Eng.*, **127**, 387-401, 1995

[28] Hughes T. J. R., Feijóo G. R., Mazzei L., Quincy J.-B., The variational multiscale method - a paradigm for computational mechanics *Comput. Methods Appli. Mech. Eng.*, **166**, 3-24, 1998

[29] Ihlenburg F., Finite Element Analysis of Acoustic Scattering *Appl. Math. Sci 132, Springer-Verlag, New York* 1998

[30] Keller J. B., Givoli D., Exact non-reflecting boundary conditions *J. Comput. Physics*, **82**, 172-192, 1989

[31] Lax P., Phillips R. S., Scattering Theory *New York : Academic Press* 1989

[32] Magoulès F., Computational Methods for Acoustics Problems *Saxe-Coburg Publications* 2008

[33] Medvinsky M., Turkel E., Hetmaniuk U.,Local absorbing boundary conditions for elliptical shaped bodies *J. Comput. Physics*, **227**, 8254-8267, 2008

[34] Monk P., Wang D.Q., A least-squares method for the Helmholtz equation *Comput. Methods Appli. Mech. Eng.*, **175**, 411-454, 1999

[35] Ramm A. G., Scattering by obstacles *Dotrecht : Reidel* 1986

[36] Rose M.E., Weak element approximations to elliptic differential equations *Numer. Math.*, **24**, 185-204, 1975

[37] Singer I., Turkel E., A perfectly matched layer for the Helmholtz equation in a semi-infinite strip *J. Comput. Physics*, **201**, 439-465, 2004

[38] Strouboulis T., Babuška I., Copps K., The design and analysis of the Generalized Finite Element Method *Comput. Methods Appli. Mech. Eng.*, **181**, 43-69, 2000

[39] Taylor M. E., Partial Differential Equations I : Basic Theory *Springer-Verlag, New York* 1997

[40] Turkel E., Yefet A., Absorbing PML boundary layers for wave-like equations *Appl. Numer. Math.*, **27**, 533-557, 1998

[41] Varadan V. K., Varadan V. V., Elastic Wave Scattering and Propagation *Ann Arbor Science* 1982

[42] Wilcox C. H. Scattering theory for the d'Alembert Equation in Exterior Domains *Lecture Notes in Mathematics 442. Berlin : Springer* 1975

# **Partie I :** A discontinuous Galerkin method for solving Helmholtz problems

# 1. Introduction

The aim of this work is to design a solution methodology for solving efficiently Helmholtz problems in the mid- and high-frequency regime. As stated in the General Introduction of this manuscript, this is a very challenging goal and Engineers and Applied mathematicians around the world have been working on this problem very hard in the past seventy years. The DGM (*d*iscontinuous *G*alerkin *m*ethod) developed by Farhat *et al* in [6, 7, 8] presents very attractive features to achieve this goal. This method is very simple to understand and implement. In addition, the impressive numerical results reported in [6, 7, 8], clearly illustrate its superiority over standard Galerkin high-order finite element method (FEM) and make this method a powerful tool for solving Helmholtz problems in high frequency regime. The main drawback of DGM is its numerical instabilities that occur when refining the mesh. These instabilities deteriorate dramatically (over two orders of magnitude) the relative error (see the numerical results in Part II).
Our goal is to propose a method that preserves the "good" properties and features of DGM without (hopefully) incurring numerical instabilities. For this purpose, we propose to adopt the same methodology as DGM but we rewrite the weak continuity condition for the numerical solution in a way that resembles the least-squares method (LSM) developed by Monk-Wang in [12]. Recall that LSM is also developed using a decomposition of the domain and in a discontinuous framework. In each subdomain the solution is approximated using the solutions of the homogeneous Helmholtz equation (in [12] two such choices are proposed : the Bessel functions and the plane waves). The main difference between the proposed method and LSM is in the functions to which the global bilinear form is applied. This leads to an important difference in the cost of the two methods. More specifically, the size of linear system arising from LSM is directly related to the number of plane waves used for the approximation of the solution at the element level, that is the dimension of the discrete space associated to the primal variable. On the other hand, in the proposed method the size of the global linear system is determined only by the global number of Lagrange multipliers and consequently is related only to the dimension of the dual discrete space. For example in the case of the $R$-8-2b element, the proposed method reduces the size of the global system by a factor four when compared to LSM.
The remainder of the paper is organized as follows. In Section 2, we introduce the model problem and the mathematical framework. Section 3 is devoted to the presentation of the proposed solution methodology. In Section 4, we present the discrete aspects of the formulation. Section 5 is dedicated to the mathematical analysis of the method. We compare in Section 6 the numerical performance of the proposed method to the performance of DGM and LSM. Finally, Section 6 concludes this paper.

# 2. Preliminaries

## 2.1. The mathematical model

We consider the following mathematical model problem :

$$(\mathrm{BVP}) \begin{cases} -\Delta u - k^2 u &= f & \text{in } \Omega, \\ \partial_n u &= \mathrm{i}ku + g_\Sigma & \text{on } \Sigma, \\ \partial_n u &= g_\Gamma & \text{on } \Gamma, \end{cases} \tag{5}$$

where $u$ is the unknown field. $\Omega$ is the computational domain. It is a bounded polygonal shaped domain whose interior (resp. exterior) boundary is $\Gamma$ (resp. $\Sigma$), as illustrated in Fig. 1. $\mathbf{n}$ is the unitary outward normal vector to the boundaries $\Sigma$ and $\Gamma$. $\partial_n$ is the normal derivative. $f$ and $g$ are

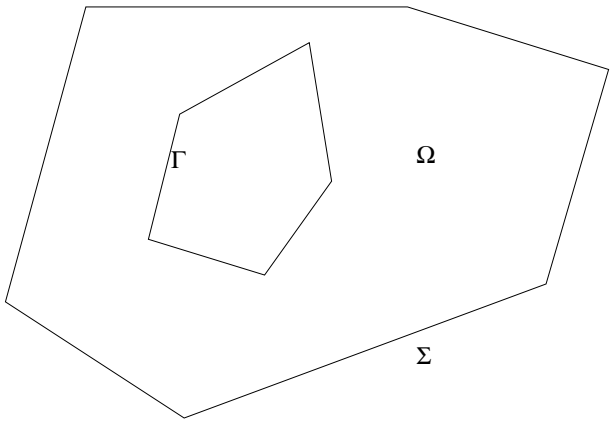


**Fig. 1** – Computational domain in 2D

complex valued functions such that $f \in L^2(\Omega)$, $g_\Sigma \in L^2(\Sigma)$ and $g_\Gamma \in L^2(\Gamma)$. $k$ is a positive number representing the wavenumber.
Note that BVP can be viewed as an acoustic scattering problem formulated in a bounded domain when using absorbing boundary conditions of order 1/2 [2].

## 2.2. The mathematical framework

Let $\tau_h$ be a regular decomposition of $\Omega$ in convex polygonal (triangular- or quadrilateral-) shaped elements $K$. We consider the two following spaces :

$$\forall K \in \tau_h, \quad \mathcal{V}(K) = \left\{ \begin{array}{l} v^K \in H^1(K),\ \Delta v^K + k^2 v^K = 0 \text{ in } K,\ \partial_n^K v^K \in L^2(\partial K), \\ \\ \partial_n^K v^K = \mathrm{i}\, k v^K \text{ on } \partial K \cap \Sigma,\ \partial_n^K v^K = 0 \text{ on } \partial K \cap \Gamma \end{array} \right\} \tag{6}$$

where $\partial_n^K$ represents the derivative with respect to the outward normal $n^K$, and

$$\mathcal{V} = \Big\{ v \in L^2(\Omega)\,;\, \forall K \in \tau_h, v|_K \in \mathcal{V}(K) \Big\}. \tag{7}$$

The spaces $\mathcal{V}(K)$ and $\mathcal{V}$ satisfy the following four properties. Properties 2, 3 and 4 are established in Appendix 1.

**Property 1.** For all $v \in \mathcal{V}$, we have :

$$\Delta v \in L^2(\Omega) \quad \text{iff} \quad \Big( v \in H^1(\Omega) \quad \text{and} \quad \partial_n^K v^K + \partial_n^{K'} v^{K'} = 0 \text{ on } \partial K \cap \partial K', \forall K, K' \in \tau_h. \Big) \tag{8}$$

and

$$\Delta v + k^2 v = 0 \quad \text{in } \Omega \quad \text{iff} \quad v = 0. \tag{9}$$

**Property 2.** Assume $kh$ to be sufficiently small. Then, there is a positive constant $C > 0$ such that for any $K \in \tau_h$ we have :

$$\sqrt{k}|v|_{1,K} + k^2 h_K^{1/2} \left( \|v\|_{0,K} + h_K^{1/2} \|v\|_{0,\partial K} \right) \leq C \|\partial_n v\|_{0,\partial K \cap \mathring{\Omega}}, \quad \forall v \in \mathcal{V}(K), \tag{10}$$

where $\|\cdot\|_{0,K}$ (resp. $|\cdot|_{1,K}$) is the $L^2$-norm (resp. the $H^1$-seminorm). $\|\cdot\|_{0,\partial K}$ is the $L^2$-norm on $\partial K$.

**Property 3.** Assume $kh \leq 1$. Then, there is a positive constant $C > 0$ such that we have :

$$\|\Psi\|_{0,\Omega} \leq C \left( \sum_{e-\text{interior edge}} \frac{1}{k^2|e|} \| [[\partial_n \Psi]] \|_{0,e}^2 + \sum_{e-\text{interior edge}} \frac{1}{|e|} \|[\Psi]\|_{0,e}^2 \right)^{1/2}, \quad \forall \Psi \in \mathcal{V}. \tag{11}$$

The space $\mathcal{M}$ corresponding to the Lagrange multipliers (the dual variable) is given by :

$$\mathcal{M} = \left\{ \mu \in \prod_{K \in \tau_h} L^2(\partial K) \, ; \forall K \in \tau_h, \mu^K = 0 \text{ on } \partial K \cap \partial\Omega, \text{ and } \forall K, K' \in \tau_h, \, \mu^K + \mu^{K'} = 0 \right\}. \tag{12}$$

Last, we denote by $\Phi$ the following application :

$$\begin{array}{rccc} \Phi : & \mathcal{M} & \longrightarrow & \mathcal{V} \\ & \mu & \longmapsto & \Phi(\mu) \end{array} \tag{13}$$

such that we have :

$$\partial_n^K \Phi(\mu) = \mu^K \quad \text{on } \partial K \cap \mathring{\Omega} \quad \forall K \in \tau_h. \tag{14}$$

We prove in Appendix 1. the following property.

**Property 4.** Assume $kh$ to be sufficiently small. Then, $\Phi$ is a linear and well-defined application.

## 3. The solution methodology : The continuous approach

The basic idea of the proposed solution methodology is to determine $u$ (the solution of BVP) under the following form :

$$u = \varphi + \Phi(\lambda), \tag{15}$$

where :
a- $\varphi$ is an element of $L^2(\Omega)$ such that for all $K \in \tau_h$, $\varphi|_K = \varphi^K$ is the unique solution of the following boundary value problem :

$$\left\{ \begin{array}{rll} -\Delta\varphi^K - k^2\varphi^K &= f & \text{in } K, \\ \partial_n \varphi^K &= \mathrm{i}k\varphi^K + g_\Sigma & \text{on } \partial K \cap \Sigma, \\ \partial_n \varphi^K &= g_\Gamma & \text{on } \partial K \cap \Gamma, \\ \partial_n \varphi^K &= 0 & \text{on } \partial K \cap \mathring{\Omega}. \end{array} \right. \tag{16}$$

Note that $\partial_n \varphi$ is continuous across the interior edges, that is :

$$[[\partial_n \varphi]] = \partial_n^K \varphi^K + \partial_n^{K'} \varphi^{K'} = 0 \quad \text{on each interior edge } e. \tag{17}$$

b- The Lagrange multiplier $\lambda \in \mathcal{M}$ is chosen such that

$$[\varphi + \Phi(\lambda)] = 0 \quad \text{on each interior edge } e. \tag{18}$$

We propose to fulfill this requirement in a least-squares sense, that is :

$$\begin{cases} \text{Find } \lambda \in \mathcal{M} \text{ such that :} \\ \displaystyle\sum_{e-\text{interior edge}} \frac{1}{|e|} \int_e [\Phi(\lambda)][\overline{\Phi(\mu)}]\, ds = -\sum_{e-\text{interior edge}} \frac{1}{|e|} \int_e [\varphi][\overline{\Phi(\mu)}]\, ds, \quad \forall \mu \in \mathcal{M}. \end{cases} \tag{19}$$

Observe that this variational formulation uses the functions $\Phi(\mu)$, for all $\mu \in \mathcal{M}$. From the definition (see Eqs. (13)-(14)) of the operator $\Phi$ and Property 4, we deduce that for a fixed $\mu \in \mathcal{M}$, in each element $K \in \tau_h$, the function $\Phi^K(\mu)$ is the unique solution of the following boundary value problem :

$$\begin{cases} -\Delta\Phi^K(\mu) - k^2\Phi^K(\mu) = 0 & \text{in } K, \\ \partial_n\Phi^K(\mu) = \mathrm{i}k\Phi^K(\mu) & \text{on } \partial K \cap \Sigma, \\ \partial_n\Phi^K(\mu) = 0 & \text{on } \partial K \cap \Gamma, \\ \partial_n\Phi^K(\mu) = \mu & \text{on } \partial K \cap \mathring{\Omega}. \\ \Phi^K(\mu) \in H^1(K) \end{cases} \tag{20}$$

Note that the normal derivative of the functions in the space $\mathcal{M}$ is continuous across the interior edges. Hence,

$$[[\partial_n\Phi(\lambda)]] = \partial_n^K\Phi^K(\lambda) + \partial_n^{K'}\Phi^{K'}(\lambda) = 0 \quad \text{on each interior edge } e. \tag{21}$$

**Remark 1.** One can easily deduce that the splitting $\varphi + \Phi(\lambda) = \tilde{u}$ coincides with $u$, solution of BVP, based on the following observations :

- For all $K \in \tau_h$, $\tilde{u}|_K = \tilde{u}^K$ satisfies the Helmholtz equation.
- $[\tilde{u}^K] = 0$ and $\left[[\partial_n\tilde{u}^K]\right] = 0$ on each interior edge $e$.
- $\tilde{u}$ satisfies the boundary conditions on $\Sigma$ and $\Gamma$.

In summary, the proposed solution methodology for determining the solution of BVP under the form $u = \varphi + \Phi(\lambda)$ is a two-step procedure :

**Step 1** For all $\mu \in \mathcal{M}$, we compute $\varphi$ and $\Phi(\mu)$. This is achieved by solving a set of local Helmholtz problems. Note that the two types of boundary value problems to be solved in Step 1 are associated, when written in a variational framework, to the same bilinear form $a_K(\cdot,\cdot)$, that is :

$$a_K(w^K, v^K) = \int_{\partial K} \partial_n^K w^K \overline{v^K}\, ds - \mathrm{i}\, k \int_{\partial K \cap \Sigma} w^K \overline{v^K}\, ds, \quad \forall v^K,\, w^K \in H^1(K). \tag{22}$$

Consequently, we can write the obtained variational formulation for both local boundary value problems in a compact form as follows :

$$\begin{cases} \text{Find } \Psi^K \in H^1(K) \text{ such that :} \\ a_K\left(\Psi^K, v^K\right) = L_K\left(v^K\right), \qquad \forall v^K \in H^1(K) \end{cases} \tag{23}$$

where

$$\Psi^K = \begin{cases} \varphi^K & \text{for (16)}, \\ \Phi^K(\mu) & \text{for (20)}, \end{cases} \quad \forall \mu \in \mathcal{M}.$$

For any $v^K \in H^1(K)$, the right-hand side $L_K(\cdot)$ is given by :

$$L_K\left(v^K\right) = \begin{cases} \displaystyle\int_K f\overline{v^K}\,dx + \int_{\partial K\cap\Sigma} g_\Sigma \overline{v^K}\,ds + \int_{\partial K\cap\Gamma} g_\Gamma \overline{v^K}\,ds & \text{for (16),} \\ \\ \displaystyle\int_{\partial K\cap\mathring{\Omega}} \mu^K \overline{v^K} ds & \text{for (20).} \end{cases} \tag{24}$$

**Step 2** We determine $\lambda \in \mathcal{M}$ by solving the variational problem given by (19).

# 4. The solution methodology : The discrete approach

We describe the discrete aspects and issues of the approach proposed in the previous section. Then, we discuss the similarities and differences between the proposed method, DGM and LSM.

## 4.1. The discrete formulation

Let $\mathcal{V}_h$ and $\mathcal{M}_h$ be two finite-dimensional spaces such that :

$$\mathcal{V}_h \subset \mathcal{V} \quad \text{and} \quad \mathcal{M}_h \subset \mathcal{M}. \tag{25}$$

The goal here is to approximate $u$, the solution of BVP, as follows :

$$u_h = \varphi + \Phi_h(\lambda_h), \tag{26}$$

with $\lambda_h \in \mathcal{M}_h$ and $\Phi_h(\lambda_h) \in \mathcal{V}_h$. The previous two steps are then formulated as follows :

**Step 1** We solve the discrete variational problem associated to (23). We will focus on the problems whose unknowns are $\Phi_h(\mu_h)$, for each $\mu_h \in \mathcal{M}_h$. Observe that Step 1 is equivalent to studying the operator :

$$\begin{array}{rccc} \Phi_h : & \mathcal{M}_h & \longrightarrow & \mathcal{V}_h \\ & \mu_h & \longmapsto & \Phi_h(\mu_h) \end{array} \tag{27}$$

such that :

$$\partial_n^K \Phi_h(\mu_h) \cong \mu_h^K \text{ on } \partial K \cap \mathring{\Omega}, \quad \forall K \in \tau_h. \tag{28}$$

In practice, Step 1 requires solving, in each element $K \in \tau_h$, the following variational problem :

$$\begin{cases} \Phi_h^K(\mu_h) \in \mathcal{V}_h(K) \\ \displaystyle\int_{\partial K} \partial_n^K \Phi_h^K(\mu_h)\,\overline{v_h}\,ds - \mathrm{i}\,k \int_{\partial K\cap\Sigma} \Phi_h^K(\mu_h)\,\overline{v_h}\,ds = \int_{\partial K\cap\mathring{\Omega}} \mu_h \overline{v_h}\,ds, \quad \forall v_h \in \mathcal{V}_h(K). \end{cases} \tag{29}$$

**Step 2** We write the global linear system given by (19) in finite dimension, that is :

$$\begin{cases} \text{Find } \lambda_h \in \mathcal{M}_h \text{ such that :} \\ \displaystyle\sum_{e-\text{interior edge}} \frac{1}{|e|} \int_e [\Phi_h(\lambda_h)][\overline{\Phi_h(\mu_h)}]\,ds = -\sum_{e-\text{interior edge}} \frac{1}{|e|} \int_e [\varphi][\overline{\Phi_h(\mu_h)}]\,ds, \quad \forall \mu_h \in \mathcal{M}_h. \end{cases} \tag{30}$$

**Remark 2.** The variational problem (29) suggests that there is no need to require that the elements of $\mathcal{V}_h(K)$ (and even $\mathcal{V}(K)$) satisfy the boundary conditions on $\Sigma$ and $\Gamma$ in the strong sense. In practice, we solve the following variational problem :

$$\begin{cases} \Phi_h^K(\mu_h) \in Y_h(K) \\ \displaystyle\int_{\partial K} \partial_n^K \Phi_h^K(\mu_h)\, \overline{v_h}\, ds - \mathrm{i}\, k \int_{\partial K \cap \Sigma} \Phi_h^K(\mu_h)\, \overline{v_h}\, ds = \int_{\partial K \cap \mathring{\Omega}} \mu_h \overline{v_h}\, ds,\ \forall v_h \in Y_h(K), \end{cases} \tag{31}$$

where $Y_h$ is a finite-dimensional subspace of

$$Y(K) = \left\{ v^K \in H^1(K);\quad \Delta v^K + k^2 v^K = 0 \text{ in } K,\quad \partial_n^K v^K \in L^2(\partial K) \right\}.$$

The choice of using the boundary conditions in the strong sense was driven by the simplicity of the presentation. Note that we can conduct the mathematical analysis and establish the presented results using the boundary conditions in the weak sense.

### 4.2. Some practical issues

The following two remarks are noteworthy :
• The variational problem given by (29) does not guarantee the well-defined character of the application $\Phi_h$. For example, for a fixed $\mu_h$, the existence and uniqueness of $\Phi_h(\mu_h)$ have been proved [1] in the case of the so-called $R$-4-1 element. For higher order elements, the necessary conditions are not obviously verified.

In order to guarantee that $\Phi_h$ is a well defined application, we propose to modify the variational formulation given by (29). More specifically, we consider the following variational problem :

$$\begin{cases} \Phi_h^K(\mu_h) \in \mathcal{V}_h(K) \\ \displaystyle\int_{\partial K \cap \mathring{\Omega}} \partial_n^K \Phi_h^K(\mu_h)\, \partial_n^K \overline{v_h}\, ds = \int_{\partial K \cap \mathring{\Omega}} \mu_h \partial_n^K \overline{v_h}\, ds, \quad \forall v_h \in \mathcal{V}_h(K). \end{cases} \tag{32}$$

**Remark 3.** In practice, using the boundary conditions on $\Sigma$ and $\Gamma$ in a weak sense, we solve the following variational problem :

$$\begin{cases} \Phi_h^K(\mu_h) \in Y_h(K) \\ \displaystyle\int_{\partial K} \left(\partial_n^K \Phi_h^K(\mu_h) - \mathrm{i}\, k \Phi_h^K(\mu_h)\, \mathbb{I}_\Sigma\right) \left(\partial_n^K \overline{v_h} + \mathrm{i}\, k \overline{v_h} \mathbb{I}_\Sigma\right) ds = \int_{\partial K \cap \mathring{\Omega}} \mu_h \partial_n^K \overline{v_h}\, ds, \quad \forall v_h \in Y_h(K), \end{cases} \tag{33}$$

where $\mathbb{I}_\Sigma$ is the indicator function on $\Sigma$.

The variational problem (32) admits a unique solution. Indeed, we first observe that the existence of $\Phi_h(\mu_h)$ is guaranteed by its uniqueness. Then, the homogeneous form of (32) leads to

$$\partial_n^K \Phi_h^K(\mu_h) = 0 \quad \text{on } \partial K. \tag{34}$$

Similarly to the proof of Property 1.3 (see Appendix 1), we conclude that, for $kh$ sufficiently small, $\Phi_h^K(\mu_h) = 0$ in $K$, for all $K \in \tau_h$.

• In the case of the global variational problem given by (19), the existence of a solution is also equivalent to its uniqueness. Let us consider the homogeneous problem associated to (19). We immediately have

$$[\Phi_h(\lambda_h)] = 0 \quad \text{on each interior edge } e. \tag{35}$$

Consequently, $\Phi_h(\lambda_h) \in H^1(\Omega)$, which is not sufficient to conclude $\Phi_h(\lambda_h) = 0$, and thus, even less that $\lambda_h = 0$. In fact, we need to make sure that $\Delta\Phi_h(\lambda_h) \in L^2(\Omega)$ in order to guarantee $\Phi_h(\lambda_h) = 0$. To ensure that $\Delta\Phi_h(\lambda_h) \in L^2(\Omega)$, the function $\Phi_h(\lambda_h)$ must fulfill the following condition :

$$[[\partial_n \Phi_h(\lambda_h)]] = 0 \tag{36}$$

in a *strong* sense.
Therefore, we "enrich" the previous variational formulation as follows :

$$\left\{\begin{array}{l} \text{Find } \lambda_h \in \mathcal{M}_h \text{ such that :} \\ \displaystyle\sum_{e-\text{interior edge}} \frac{1}{|e|}\int_e [\Phi_h(\lambda_h)]\overline{[\Phi_h(\mu_h)]}\,ds + \sum_{e-\text{interior edge}} \frac{1}{k^2|e|}\int_e [[\partial_n\Phi_h(\lambda_h)]]\,\overline{[[\partial_n\Phi_h(\mu_h)]]}\,ds \\ \displaystyle = -\sum_{e-\text{interior edge}} \frac{1}{|e|}\int_e [\varphi]\overline{[\Phi_h(\mu_h)]}\,ds\,, \quad \forall \mu_h \in \mathcal{M}_h. \end{array}\right. \tag{37}$$

First, we observe that the existence of a solution is equivalent to its uniqueness. Then, the homogeneous equation associated to (37), leads to :

$$[\Phi_h(\lambda_h)] = 0 \quad \text{and} \quad [[\partial_n\Phi_h(\lambda_h)]] = 0 \quad \text{on each interior edge } e. \tag{38}$$

Consequently, we have : $\Phi_h(\lambda_h) \in H^1(\Omega)$ and $\Delta\Phi_h(\lambda_h) \in L^2(\Omega)$. Therefore,

$$\Phi_h(\lambda_h) = 0 \quad \text{in} \quad \Omega. \tag{39}$$

However, Eq. (39) does not allow us to conclude that $\lambda_h = 0$ since we only have :

$$\forall K \in \tau_h, \quad \forall v_h \in \mathcal{V}_h(K) \quad \int_{\partial K \cap \mathring{\Omega}} \lambda_h \partial_n \overline{v_h}\,ds = 0. \tag{40}$$

**Remark 4.** At the algebraic level, this means that we deal with a global system whose associated matrix is Hermitian and positive semi-definite. In addition, if for any $\mu_h \in \mathcal{M}_h$ the condition :

$$\left(\forall K \in \tau_h, \quad \forall v_h \in \mathcal{V}_h(K) \quad \int_{\partial K} \mu_h \partial_n \overline{v_h}\,ds = 0\right) \quad \Longrightarrow \quad \mu_h = 0 \tag{41}$$

is fulfilled, than the matrix is positive definite.

Finally, the proposed solution methodology requires :

**Step 1** Solving a set of local variational problems given by (32).

**Step 2** Solving the global variational problem given by (37).

### 4.3. Connection with DGM

Recall that the DGM formulation, developed by Farhat *et al* in [6], consists in formulating BVP using the following mixed-hybrid variational formulation :

$$\left\{\begin{array}{lll} \text{Find } u \in \displaystyle\prod_{K\in\tau_h} H^1(K) \text{ and } \lambda \in \mathcal{M} & \text{such that :} & \\ A(u,v) + B(v,\lambda) = & F(v), & \forall v \in \displaystyle\prod_{K\in\tau_h} H^1(K), \\ B(u,\mu) = & 0, & \forall \mu \in \mathcal{M}, \end{array}\right. \tag{42}$$

where the bilinear forms $A(\cdot,\cdot)$ and $B(\cdot,\cdot)$ are given by :

$$A(w,v) = \sum_{K\in\tau_h} a_K\left(w^K, v^K\right), \quad \forall w, v \in \prod_{K\in\tau_h} H^1(K) \tag{43}$$

and

$$B(v,\mu) = -\sum_{K\in\tau_h} \int_{\partial K} \overline{v^K}\mu^K \, ds, \quad \forall v \in \prod_{K\in\tau_h} H^1(K), \forall \mu \in \mathcal{M}. \tag{44}$$

The right-hand side $F(\cdot)$ is given by :

$$F(v) = \sum_{K\in\tau_h} \left( \int_K f\overline{v^K}\, dx + \int_{\partial K\cap\Sigma} g_\Sigma \overline{v^K}\, ds + \int_{\partial K\cap\Gamma} g_\Gamma \overline{v^K}\, ds \right), \quad \forall v \in \prod_{K\in\tau_h} H^1(K). \tag{45}$$

Similarly to the proposed method, the DGM formulation can also be viewed as a two-step procedure in which the solution $u$ is also given by (15). At the discrete level the local problems are solved using the variational formulation given by (30). The continuity of the numerical solution is restored in the weak sense given by the following global discrete problem :

$$\left\{\begin{array}{l} \text{Find } \lambda_h \in \mathcal{M}_h \text{ such that :} \\ \displaystyle\sum_{e-\text{interior edge}} \frac{1}{|e|} \int_e [\Phi_h(\lambda_h)]\overline{\mu}_h \, ds = -\sum_{e-\text{interior edge}} \frac{1}{|e|} \int_e [\varphi]\overline{\mu}_h \, ds, \quad \forall \mu_h \in \mathcal{M}_h. \end{array}\right. \tag{46}$$

The system given by (46) resembles the one given by (19). The main difference is in the test functions : in DGM the test functions are in $\mathcal{M}_h$, whereas in the proposed method we consider the jumps across interior edges of the functions $\Phi_h(\mu_h)$. This difference does not change the computational cost since both systems have the same size.
Note that the solution of the variational problem given by (46) is not always guaranteed. Let us consider the homogeneous problem associated to (46), since proving the existence of the solution is equivalent to proving its uniqueness. We cannot conclude :

$$\Phi_h(\lambda_h) = 0 \quad \text{and even less} \quad \lambda_h = 0. \tag{47}$$

**Remark 5.** This result is obtained if the following implications hold :

$$\text{If } \forall \mu_h \in \mathcal{M}_h, \sum_{K\in\tau_h} \int_{\partial K} \Phi_h^K(\lambda_h)\, \mu_h^K \, ds = 0, \quad \text{then} \quad \Phi_h(\lambda_h) = 0 \tag{48}$$

and

$$\text{If } \mu_h \in \mathcal{M}_h \text{ and } \left( \forall v_h \in \mathcal{V}_h, \forall K \in \tau_h, \quad \int_{\partial K} \mu_h^K \overline{v_h^K}\, ds = 0 \right), \quad \text{then} \quad \mu_h = 0. \tag{49}$$

**Remark 6.** Note that it is not enough to incorporate the formulation given by (32) into DGM to restore the stability as one could expect. This is illustrated by the following result. We have considered the waveguide problem introduced in [6]. We have performed an experiment in the case where $ka = 1$ and the so-called $R$-8-2b element is used (see Section 6 for the nomenclature). We have measured the relative error with respect to the propagation angles (see Section 6 for more details). The stability of the modified DGM is not satisfactory, as illustrated in Fig. 2.

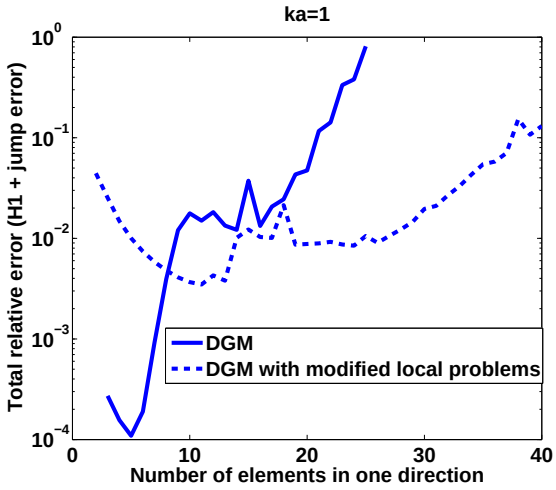


**Fig. 2** – Performance comparison between the DGM($R$-8-2b) formulation given by (31)-(46) and the DGM formulation given by (33)-(46) : $ka = 1$

## 4.4. Connection with LSM

Using the notations adopted in this work, the least squares method (LSM), developed by Monk-Wang in [12], requires solving the following global system :

$$\begin{cases} \text{Find } u_h \in \mathcal{V}_h \text{ such that :} \\ \displaystyle\sum_{e-\text{interior edge}} \frac{k^2}{|e|} \int_e [u_h][\overline{v_h}]\, ds + \sum_{e-\text{interior edge}} \frac{1}{|e|} \int_e [[\partial_n u_h]]\, \overline{[[\partial_n v_h]]}\, ds \\ \displaystyle = - \sum_{e-\text{interior edge}} \frac{1}{|e|} \int_e [\varphi][\overline{v_h}]\, ds\,, \quad \forall v_h \in \mathcal{V}_h. \end{cases} \tag{50}$$

One can easily verify that this problem admits a unique solution. The proposed formulation and LSM differ mainly by the unknowns of the global system. Indeed, the field $u_h$ is the unknown in (50), whereas the unknown in the system (37) is the Lagrange multiplier $\lambda_h$. Consequently, LSM leads to a linear system whose size is of the order of the dimension of $\mathcal{V}_h$, whereas the size of the linear system resulting from (37) depends only on the dimension of $\mathcal{M}_h$. In order to illustrate the computational cost required by both methods, let us consider the $R$-$m$-$n$ element, introduced by

Farhat *et al* in [6]. Here, $R$ stands for rectangular element, $m$ represents the number of plane waves that approximate the solution at the element level, whereas $n$ designates the number of degrees of freedom (dof) per edge. For illustration purpose, consider a square-shaped computational domain to which is applied a rectangular-shaped $p \times p$ mesh. We report in Table 1 the asymptotic size of the solution vector for the proposed method and for LSM in the case of a rectangular-shaped mesh. Note that, similarly to LSM, the proposed formulation can be applied to a triangular-shaped mesh. In Table 2 we report the asymptotic size of the solution vector for the proposed method and for LSM in the case of such a mesh and when considering $m$ plane waves per element and $n$ dofs per edge. Tables 1 and 2 suggest that the proposed method is less expensive in terms of computational

TAB. 1 – Asymptotic size of the solution vector for a $p \times p$ rectangular-shaped uniform mesh.

| Method | Asymptotic size of the solution vector |
|---|---|
| New method | $2p^2\, n$ |
| LSM | $p^2\, m$ |

TAB. 2 – Asymptotic size of the solution vector for a $p \times p$ triangular-shaped uniform mesh.

| Method | Asymptotic size of the solution vector |
|---|---|
| New method | $3p^2\, n$ |
| LSM | $p^2\, m$ |

cost if

$$\dim \mathcal{M}_h \leq \dim \mathcal{V}_h. \tag{51}$$

For a rectangular-shaped mesh, this requirement is fulfilled if $2n \leq m$, whereas for a triangular-shaped mesh the necessary condition is given by $3n \leq m$.

# 5. Mathematical analysis : Error estimates

## 5.1. Announcement of the main results

The following theorem provides a $L^2$-error estimate between the solution $u$ of BVP (see Eq. (5)) and the solution $u_h = \varphi + \Phi_h(\lambda_h)$ where $\lambda_h$ is solution of the problem given by Eq. (37) :

**Theorem 1.** Let $u$ be the solution of BVP (see Eq. (5)), $\lambda \in \mathcal{M}$ be the solution of Eq. (19), $\Phi_h$ be the operator given by Eqs. (27)-(28) and $u_h = \varphi + \Phi_h(\lambda_h)$, where $\lambda_h$ is the solution of the problem given by Eq. (37). Then, there is a positive constant $\hat{C}$ such that :

$$\|u - u_h\|_{0,\Omega} \leq \hat{C} \left( \sum_{K \in \tau_h} \frac{1}{k^4 h_K^3} \left( \|\lambda - \mu_h\|^2_{0,\partial K \cap \mathring{\Omega}} + \|\lambda - \partial_n v_h\|^2_{0,\partial K \cap \mathring{\Omega}} \right) \right)^{1/2}, \ \forall v_h \in \mathcal{V}_h, \ \forall \mu_h \in \widetilde{\mathcal{M}}_h, \tag{52}$$

where the space $\widetilde{\mathcal{M}}_h$ is defined by :

$$\widetilde{\mathcal{M}}_h = \left\{ \mu_h \in \mathcal{M}_h; \int_e \mu_h \, ds = \int_e \lambda \, ds \ \text{ on each interior edge } e \right\}. \tag{53}$$

Let us consider the case where the field is approximated using $m$ plane waves at the element level. We assume $N \in \mathbb{N}$ to be such that $m \geq 2N+1$. Assuming that the needed regularity conditions are verified, we have :

**Theorem 2.** Let $\lambda \in \mathcal{M}$ be the solution of Eq. (19) and $\Phi$ be the application defined by Eqs. (13)-(14). Then, for any $K \in \tau_h$, we have :

$$\|\lambda - \partial_n v_h\|_{0,\partial K \cap \mathring{\Omega}} \leq h_K^{N-1/2} \left( \sum_{l=0}^{N} k^{N+1-l} |\Phi(\lambda)|_{l,K} + |\Phi(\lambda)|_{N+1,K} + h_K |\Phi(\lambda)|_{N+2,K} \right), \quad \forall v_h \in \mathcal{V}_h(K). \tag{54}$$

where the $|\cdot|_{p,\Omega}$ is defined by :

$$|v|_{p,\Omega} = \left( \sum_{K \in \tau_h} |v|_{p,K}^2 \right)^{1/2}. \tag{55}$$

### 5.2. Applications

Let us consider the case where the proposed methodology is equipped with the $\widetilde{R}$-$m$-$n$ element, where $\widetilde{R}$ stands for rectangular-shaped mesh, $m$ denotes the number of plane waves at the element level and $n$ designates the number of dofs per edge for the Lagrange multipliers. We denote by $\mathcal{M}_h$ the set of polynomial functions of degree $n$ defined on the interior edges. We assume $N \in \mathbb{N}$ to be such that $m \geq 2N+1$. Assuming that the needed regularity conditions are verified, the following proposition holds :

**Proposition 1.** Let $\lambda \in \mathcal{M}$ be the solution of Eq. (19) and $\Phi$ be the application defined by Eqs. (13)-(14). Then, for any $K \in \tau_h$, we have :

$$\|\lambda - \mu_h\|_{0,\partial K \cap \mathring{\Omega}} \leq \hat{C} h_K^{n+\frac{3}{2}} |\Phi(\lambda)|_{n+2,K} + \hat{C} h_K^{n+\frac{1}{2}} |\Phi(\lambda)|_{n+1,K}, \quad \forall \mu_h \in \mathcal{M}_h. \tag{56}$$

**Proof of Proposition 1.** From the classical results for the polynomial interpolation [3], we have :

$$\|\lambda - \mu_h\|_{0,\partial K \cap \mathring{\Omega}} \leq \hat{C} h_K^{n+1} |\lambda|_{n+1,\partial K \cap \mathring{\Omega}}. \tag{57}$$

Recall that $\partial_n \Phi(\lambda) = \lambda$ on each interior edge $e$. From the standard regularity results [3] we deduce :

$$\|\lambda - \mu_h\|_{0,\partial K \cap \mathring{\Omega}} \leq \hat{C} h_K^{n+\frac{3}{2}} |\Phi(\lambda)|_{n+2,K} + \hat{C} h_K^{n+\frac{1}{2}} |\Phi(\lambda)|_{n+1,K}, \quad \forall \mu_h \in \mathcal{M}_h, \tag{58}$$

which concludes the proof. ■

It follows from Theorem 1, Theorem 2 and Proposition 1 that :

**Corollary 1.** In the case of the $\widetilde{R}$-$m$-$n$ element, there is a positive constant $\hat{C}$ such that :

$$\begin{aligned}\|u-u_h\|_{0,\Omega} \leq \quad & \frac{\hat{C}}{k^2}\Big[\left(h^n|\Phi(\lambda)|_{n+2,\Omega}+h^{n-1}|\Phi(\lambda)|_{n+1,\Omega}\right)\\ & +h^{N-2}\left(\sum_{l=0}^{N}k^{N+1-l}|\Phi(\lambda)|_{l,\Omega}+|\Phi(\lambda)|_{N+1,\Omega}+h|\Phi(\lambda)|_{N+2,\Omega}\right)\Big].\end{aligned} \tag{59}$$

**Illustrative examples**

- $\widetilde{R}$-4-1 : $m=4$, $n=1$, $N=1$ :

$$\|u-u_h\|_{0,\Omega} \leq \quad \frac{\hat{C}\sqrt{h}}{\sqrt{k}}|\Phi(\lambda)|_{3/2,\Omega}. \tag{60}$$

- $\widetilde{R}$-4-2 : $m=4$, $n=2$, $N=1$ :

$$\begin{aligned}\|u-u_h\|_{0,\Omega} \leq \quad & \hat{C}\Big(\frac{1}{h}\|\Phi(\lambda)\|_{0,\Omega}+\frac{1}{kh}|\Phi(\lambda)|_{1,\Omega}+\frac{1}{k^2h}|\Phi(\lambda)|_{2,\Omega}\\ & +\frac{h}{k^2}|\Phi(\lambda)|_{3,\Omega}+\frac{h^2}{k^2}|\Phi(\lambda)|_{4,\Omega}\Big)..\end{aligned} \tag{61}$$

- $\widetilde{R}$-7-2 : $m=7$, $n=2$, $N=3$ :

$$\begin{aligned}\|u-u_h\|_{0,\Omega} \leq \quad & \hat{C}\Big(k^2h\|\Phi(\lambda)\|_{0,\Omega}+kh|\Phi(\lambda)|_{1,\Omega}+h|\Phi(\lambda)|_{2,\Omega}\\ & +\frac{h}{k}|\Phi(\lambda)|_{3,\Omega}+\frac{h}{k^2}|\Phi(\lambda)|_{4,\Omega}+\frac{h^2}{k^2}|\Phi(\lambda)|_{5,\Omega}\Big).\end{aligned} \tag{62}$$

- $\widetilde{R}$-8-2 : $m=8$, $n=2$, $N=3$ :

$$\begin{aligned}\|u-u_h\|_{0,\Omega} \leq \quad & \hat{C}\Big(k^2h\|\Phi(\lambda)\|_{0,\Omega}+kh|\Phi(\lambda)|_{1,\Omega}+h|\Phi(\lambda)|_{2,\Omega}\\ & +\frac{h}{k}|\Phi(\lambda)|_{3,\Omega}+\frac{h}{k^2}|\Phi(\lambda)|_{4,\Omega}+\frac{h^2}{k^2}|\Phi(\lambda)|_{5,\Omega}\Big).\end{aligned} \tag{63}$$

- $\widetilde{R}$-8-3 : $m=8$, $n=3$, $N=3$ :

$$\begin{aligned}\|u-u_h\|_{0,\Omega} \leq \quad & \hat{C}\Big(k^2h\|\Phi(\lambda)\|_{0,\Omega}+kh|\Phi(\lambda)|_{1,\Omega}+h|\Phi(\lambda)|_{2,\Omega}\\ & +\frac{h}{k}|\Phi(\lambda)|_{3,\Omega}+\frac{h^2}{k^2}|\Phi(\lambda)|_{4,\Omega}+\frac{h^3}{k^2}|\Phi(\lambda)|_{5,\Omega}\Big).\end{aligned} \tag{64}$$

- $\widetilde{R}$-11-3 : $m=11$, $n=3$, $N=5$ :

$$\begin{aligned}\|u-u_h\|_{0,\Omega} \leq \quad & \hat{C}\Big(k^4h^3\|\Phi(\lambda)\|_{0,\Omega}+k^3h^3|\Phi(\lambda)|_{1,\Omega}+k^2h^3|\Phi(\lambda)|_{2,\Omega}\\ & +kh^3|\Phi(\lambda)|_{3,\Omega}+h^3|\Phi(\lambda)|_{4,\Omega}+\frac{h^3}{k}|\Phi(\lambda)|_{5,\Omega}\\ & +\frac{h^3}{k^2}|\Phi(\lambda)|_{6,\Omega}+\frac{h^4}{k^2}|\Phi(\lambda)|_{7,\Omega}\Big).\end{aligned} \tag{65}$$

### 5.3. Proof of the main results

Before proving the results stated in Sections 5.1 and 5.2, we introduce the following notations :
• $a(\cdot,\cdot)$ is the bilinear form on $\mathcal{V}\times\mathcal{V}$ defined by :

$$a(v,w)=\sum_{e-\text{interior edge}}\frac{1}{|e|}\int_e[v][\overline{w}]\,ds+\sum_{e-\text{interior edge}}\frac{1}{k^2|e|}\int_e[[\partial_n v]]\,[[\partial_n\overline{w}]]\;ds,\quad\forall v,w\in\mathcal{V}. \tag{66}$$

• $|||\cdot|||$ is the norm defined by :

$$||||v|||=\left[a\left(v,v\right)\right]^{1/2}\;,\quad\forall v\in\mathcal{V}. \tag{67}$$

Note that the Property 3 ensures that $|||\cdot|||$ is indeed a norm.
• $F(\cdot)$ is a linear form on $\mathcal{V}$ defined by :

$$F(w)=-\sum_{e-\text{interior edge}}\frac{1}{|e|}\int_e[\varphi][\overline{w}]\,ds,\quad\forall w\in\mathcal{V}. \tag{68}$$

Consequently, the variational problem (37) can be written as follows :

$$\left\{\begin{array}{l}\text{Find }\lambda_h\in\mathcal{M}_h\text{ such that :}\\ a\left(\Phi_h\left(\lambda_h\right),\Phi_h\left(\mu_h\right)\right)=F(\Phi_h\left(\mu_h\right)),\quad\forall\mu_h\in\mathcal{M}_h.\end{array}\right. \tag{69}$$

The proof of Theorem 1 is based on the two following technical results that are formulated as lemmas.

**Lemma 1.** Let $\lambda\in\mathcal{M}$ be the solution of Eq. (19), $\Phi$ be the application defined by Eqs. (13)-(14), and $\Phi_h$ the application defined by Eqs. (27)-(28). Then, there is a positive constant $\hat{C}$ such that :

$$\sum_{e-\text{interior edge}}\frac{1}{|e|}\|[\Phi\left(\lambda\right)-\Phi_h\left(\mu_h\right)]\|_{0,e}^2\leq\hat{C}\left(\sum_{K\in\tau_h}\frac{1}{k}\|\lambda-\mu_h\|_{0,\partial K}^2+\frac{1}{k^4h_K^3}\|\mu_h-\partial_n v_h\|_{0,\partial K\cap\mathring{\Omega}}^2\right)^{1/2}, \tag{70}$$

$$\forall v_h\in\mathcal{V}_h\quad\text{and}\quad\forall\mu_h\in\widetilde{\mathcal{M}}_h,$$

where the space $\widetilde{\mathcal{M}}_h$ is defined by Eq. (53).

**Proof of Lemma 1.** Let $\mu_h\in\widetilde{\mathcal{M}}_h$. We set

$$A_h=\sum_{e-\text{interior edge}}\frac{1}{|e|}\|[\Phi\left(\lambda\right)-\Phi_h\left(\mu_h\right)]\|_{0,e}^2. \tag{71}$$

We consider the function $w \in L^2(\Omega)$ such that in each element $K$, $w|_K = w^K$ is the unique solution of the following boundary value problem :

$$\left\{\begin{array}{rll} -\Delta w^K - k^2 w^K &= 0 & \text{in } K, \\ \partial_n^K w^K &= -\mathrm{i}\, k w^K & \text{on } \partial K \cap \Sigma, \\ \partial_n^K w^K &= 0 & \text{on } \partial K \cap \Gamma, \\ \partial_n^K w^K &= \dfrac{1}{|e|}[\Phi(\lambda) - \Phi_h(\mu_h)] & \text{on } e \subset \partial K \cap \mathring{\Omega}. \end{array}\right. \tag{72}$$

Therefore, we have :

$$\begin{aligned} A_h &= \sum_{e-\text{interior edge}} \frac{1}{|e|} \int_e [\Phi(\lambda) - \Phi_h(\mu_h)][\overline{\Phi(\lambda) - \Phi_h(\mu_h)}]\, ds \\ &= -\sum_{K\in\tau_h} \int_{\partial K\cap\mathring{\Omega}} (\Phi(\lambda) - \Phi_h(\mu_h))\, \partial_n \overline{w}\, ds \\ &= -\sum_{K\in\tau_h} \left( \int_{\partial K} (\Phi(\lambda) - \Phi_h(\mu_h))\, \partial_n \overline{w}\ ds - \int_{\partial K\cap\Sigma} (\Phi(\lambda) - \Phi_h(\mu_h))\, \partial_n \overline{w}\, ds \right). \end{aligned} \tag{73}$$

We use the Green's formula and the first boundary condition of the boundary value problem (72). We then obtain :

$$\begin{aligned} A_h &= -\sum_{K\in\tau_h} \left( \int_{\partial K} (\partial_n\Phi(\lambda) - \partial_n\Phi_h(\mu_h))\, \overline{w}\ ds - \mathrm{i}\, k \int_{\partial K\cap\Sigma} (\Phi(\lambda) - \Phi_h(\mu_h))\, \overline{w}\, ds \right) \\ &= -\sum_{K\in\tau_h} \int_{\partial K\cap\mathring{\Omega}} (\partial_n\Phi(\lambda) - \partial_n\Phi_h(\mu_h))\, \overline{w}\, ds. \end{aligned} \tag{74}$$

Using the definition of the operator $\Phi$ (see Eqs. (13)-(14)), we deduce :

$$\int_{\partial K\cap\mathring{\Omega}} (\partial_n\Phi(\lambda) - \partial_n\Phi_h(\mu_h))\, \overline{w}\, ds = \int_{\partial K\cap\mathring{\Omega}} (\lambda - \mu_h)\, \overline{w}\, ds + \int_{\partial K\cap\mathring{\Omega}} (\mu_h - \partial_n\Phi_h(\mu_h))\, \overline{w}\, ds. \tag{75}$$

We are going to estimate each part of the right-hand side of Eq. (75) separately.

- First, we estimate $\left| \int_{\partial K\cap\mathring{\Omega}} (\lambda - \mu_h)\, \overline{w}\, ds \right|$. Recall that $\mu_h$ is an element of $\widetilde{\mathcal{M}}_h$. We have :

$$\begin{aligned} \left| \int_{\partial K\cap\mathring{\Omega}} (\lambda - \mu_h)\, \overline{w}\, ds \right| &= \left| \sum_{e\subset\partial K\cap\mathring{\Omega}} \int_e (\lambda - \mu_h) \left( \overline{w} - \frac{1}{|K|} \int_K \overline{w} \right) ds \right| \\ &\leq \sum_{e\subset\partial K\cap\mathring{\Omega}} \|\lambda - \mu_h\|_{0,e} \left\| w - \frac{1}{|K|} \int_K w \right\|_{0,e} \\ &\leq \hat{C} h_K^{1/2} |w|_{1,K} \|\lambda - \mu_h\|_{0,\partial K\cap\mathring{\Omega}}. \end{aligned} \tag{76}$$

Using Eq. (10) (see Property 2), we deduce that :

$$\begin{aligned} h_K^{1/2}|w|_{1,K} &\leq \frac{\hat{C}h_K^{1/2}}{\sqrt{k}}\|\partial_n w\|_{0,\partial K\cap\mathring{\Omega}} = \frac{\hat{C}h_K^{1/2}}{\sqrt{k}}\left(\sum_{e\subset\partial K\cap\mathring{\Omega}}\frac{1}{|e|^2}\|[\Phi(\lambda)-\Phi_h(\mu_h)]\|_{0,e}^2\right)^{1/2} \\ &\leq \frac{\hat{C}}{\sqrt{k}}\left(\sum_{e\subset\partial K\cap\mathring{\Omega}}\frac{1}{|e|}\|[\Phi(\lambda)-\Phi_h(\mu_h)]\|_{0,e}^2\right)^{1/2}. \end{aligned} \tag{77}$$

From Eqs. (76)-(77), we have :

$$\Big|\int_{\partial K\cap\mathring{\Omega}}(\lambda-\mu_h)\,\overline{w}\,ds\Big| \leq \frac{\hat{C}}{\sqrt{k}}\left(\sum_{e\subset\partial K\cap\mathring{\Omega}}\frac{1}{|e|}\|[\Phi(\lambda)-\Phi_h(\mu_h)]\|_{0,e}^2\right)^{1/2}\|\lambda-\mu_h\|_{0,\partial K\cap\mathring{\Omega}}. \tag{78}$$

- Next, we estimate $\Big|\int_{\partial K\cap\mathring{\Omega}}(\mu_h-\partial_n\Phi_h(\mu_h))\,\overline{w}\,ds\Big|$. We have :

$$\Big|\int_{\partial K\cap\mathring{\Omega}}(\mu_h-\partial_n\Phi_h(\mu_h))\,\overline{w}\,ds\Big| \leq \|\mu_h-\partial_n\Phi_h(\mu_h)\|_{0,\partial K\cap\mathring{\Omega}}\|w\|_{0,\partial K\cap\mathring{\Omega}}. \tag{79}$$

$$\tag{80}$$

Using Eq. (32), we have, for any $v_h\in\mathcal{V}_h(K)$ :

$$\Big|\int_{\partial K\cap\mathring{\Omega}}(\mu_h-\partial_n\Phi_h(\mu_h))\,\overline{w}\,ds\Big| \leq \|\mu_h-\partial_n v_h\|_{0,\partial K\cap\mathring{\Omega}}\|w\|_{0,\partial K\cap\mathring{\Omega}}. \tag{81}$$

From Eq. (10) (see Property 2), we deduce that :

$$\begin{aligned} \|w\|_{0,\partial K\cap\mathring{\Omega}} &\leq \frac{\hat{C}}{k^2h_K}\|\partial_n w\|_{0,\partial K\cap\mathring{\Omega}} \\ &\leq \frac{\hat{C}}{k^2h_K^{3/2}}\left(\sum_{e\subset\partial K\cap\mathring{\Omega}}\frac{1}{|e|}\|[\Phi(\lambda)-\Phi_h(\mu_h)]\|_{0,e}^2\right)^{1/2}. \end{aligned} \tag{82}$$

Using Eqs. (81)-(82), we have :

$$\begin{aligned} \Big|\int_{\partial K\cap\mathring{\Omega}}(\mu_h-\partial_n\Phi_h(\mu_h))\,\overline{w}\,ds\Big| &\leq \frac{\hat{C}}{k^2h_K^{3/2}}\left(\sum_{e\subset\partial K\cap\mathring{\Omega}}\frac{1}{|e|}\|[\Phi(\lambda)-\Phi_h(\mu_h)]\|_{0,e}^2\right)^{1/2} \\ &\|\mu_h-\partial_n v_h\|_{0,\partial K\cap\mathring{\Omega}}. \end{aligned} \tag{83}$$

Last, from Eqs. (78) and (83) we obtain :

$$A_h \leq \hat{C}\sum_{K\in\tau_h}\left(\frac{1}{k}\|\lambda-\mu_h\|_{0,\partial K}^2+\frac{1}{k^4h_K^3}\|\mu_h-\partial_n v_h\|_{0,\partial K\cap\mathring{\Omega}}^2\right),\ \forall v_h\in\mathcal{V}_h, \tag{84}$$

which concludes the proof. ■

**Lemma 2.** Let $\lambda \in \mathcal{M}$ be the solution of Eq. (19), $\Phi$ be the application defined by Eqs. (13)-(14), and $\Phi_h$ be the application defined by Eqs. (27)-(28). Then, there is a positive constant $\hat{C}$ such that :

$$\sum_{e-\text{interior edge}} \frac{1}{|e|} \| \left[ \left[ \partial_n \Phi(\lambda) - \partial_n \Phi_h(\mu_h) \right] \right] \|_{0,e}^2 \leq \hat{C} \sum_{K \in \tau_h} \frac{1}{h_K} \|\mu_h - \partial_n v_h\|_{0,\partial K \cap \mathring{\Omega}}^2, \tag{85}$$

$$\forall v_h \in \mathcal{V}_h \quad \text{and} \quad \forall \mu_h \in \mathcal{M}_h.$$

**Proof of Lemma 2.** Let $\mu_h$ be an element of $\mathcal{M}_h$. We set

$$B_h = \sum_{e-\text{interior edge}} \frac{1}{|e|} \| \left[ \left[ \partial_n \Phi(\lambda) - \partial_n \Phi_h(\mu_h) \right] \right] \|_{0,e}^2. \tag{86}$$

By construction of the operator $\Phi$ (see Eqs. (13)-(14)), we have :

$$\begin{aligned} B_h = \sum_{e-\text{interior edge}} \frac{1}{|e|} \| \left[ \left[ \partial_n \Phi(\lambda) - \partial_n \Phi_h(\mu_h) \right] \right] \|_{0,e}^2 &= \sum_{e-\text{interior edge}} \frac{1}{|e|} \| \left[ \left[ \partial_n \Phi_h(\mu_h) \right] \right] \|_{0,e}^2 \\ &= \sum_{e-\text{interior edge}} \frac{1}{|e|} \| \left[ \left[ \mu_h - \partial_n \Phi_h(\mu_h) \right] \right] \|_{0,e}^2. \end{aligned} \tag{87}$$

Thus,

$$B_h \leq \hat{C} \sum_{K \in \tau_h} \frac{1}{h_K} \|\mu_h - \partial_n \Phi_h(\mu_h)\|_{0,\partial K \cap \mathring{\Omega}}^2. \tag{88}$$

Next, we estimate $\|\mu_h - \partial_n \Phi_h(\mu_h)\|_{0,\partial K \cap \mathring{\Omega}}^2$. From Eq. (32) we have :

$$\|\mu_h - \partial_n \Phi_h(\mu_h)\|_{0,\partial K \cap \mathring{\Omega}} \leq \|\mu_h - \partial_n v_h\|_{0,\partial K \cap \mathring{\Omega}}, \quad \forall v_h \in \mathcal{V}_h. \tag{89}$$

Finally, from Eqs. (88)-(89), we deduce that :

$$B_h \leq \hat{C} \sum_{K \in \tau_h} \frac{1}{h_K} \|\mu_h - \partial_n v_h\|_{0,\partial K \cap \mathring{\Omega}}^2 \, \forall v_h \in \mathcal{V}_h, \tag{90}$$

which concludes the proof. ■

**Proof of Theorem 1.** By construction (see Eqs. (15),(26)), we have :

$$u - u_h = \Phi(\lambda) - \Phi_h(\lambda_h) \tag{91}$$

and therefore (see Eqs. (19),(37)) :

$$a\left(\Phi(\lambda) - \Phi_h(\lambda_h), \Phi_h(\mu_h)\right) = 0, \quad \forall \mu_h \in \mathcal{M}_h. \tag{92}$$

Using Eq. (11) (see Property 3), we obtain :

$$\|\Phi(\lambda) - \Phi_h(\lambda_h)\|_{0,\Omega} \leq \hat{C} |||\Phi(\lambda) - \Phi_h(\lambda_h)|||. \tag{93}$$

In addition, we have :

$$\begin{aligned} a\left(\Phi\left(\lambda\right)-\Phi_h\left(\lambda_h\right),\Phi\left(\lambda\right)-\Phi_h\left(\lambda_h\right)\right) &= a\left(\Phi\left(\lambda\right)-\Phi_h\left(\lambda_h\right),\Phi\left(\lambda\right)-\Phi_h\left(\mu_h\right)\right),\ \forall\mu_h\in\mathcal{M}_h \\ &\leq a\left(\Phi\left(\lambda\right)-\Phi_h\left(\mu_h\right),\Phi\left(\lambda\right)-\Phi_h\left(\mu_h\right)\right),\ \forall\mu_h\in\mathcal{M}_h. \end{aligned} \tag{94}$$

From Eqs. (93)-(94) we deduce :

$$\|\Phi\left(\lambda\right)-\Phi_h\left(\lambda_h\right)\|_{0,\Omega}\leq\hat{C}|||\Phi\left(\lambda\right)-\Phi_h\left(\mu_h\right)|||,\quad\forall\mu_h\in\mathcal{M}_h. \tag{95}$$

Consequently, bounding $\|u-u_h\|_{0,\Omega}$ comes to bounding $|||\Phi\left(\lambda\right)-\Phi_h\left(\mu_h\right)|||$, for all $\mu_h\in\mathcal{M}_h$.
From Lemma 1 and Lemma 2 and using the notations introduced therein, we deduce :

$$\begin{aligned} |||\Phi\left(\lambda\right)-\Phi_h\left(\mu_h\right)||| &\leq \left(\frac{1}{k^2}B_h+A_h\right)^{1/2} \\ &\leq \hat{C}\left(\sum_{K\in\tau_h}\left(\frac{1}{k^4h_K^3}\|\mu_h-\partial_n v_h\|^2_{0,\partial K\cap\mathring{\Omega}}+\frac{1}{k}\|\lambda-\mu_h\|^2_{0,\partial K}\right)\right)^{1/2} \\ &\leq \frac{\hat{C}}{k^2}\left(\left(\sum_{K\in\tau_h}\frac{1}{h_K^3}\|\lambda-\mu_h\|^2_{0,\partial K\cap\mathring{\Omega}}\right)^{1/2}+\left(\sum_{K\in\tau_h}\frac{1}{h_K^3}\|\lambda-\partial_n v_h\|^2_{0,\partial K\cap\mathring{\Omega}}\right)^{1/2}\right), \end{aligned}$$

which proves Theorem 1. ■

**Proof of Theorem 2.** Recall that $\partial_n\Phi\left(\lambda\right)=\lambda$ on each interior edge $e$. Recall also the standard regularity result that holds for sufficiently regular functions [3] :

$$\|\partial_n v\|_{0,\partial K}\leq\hat{C}\left(h_K^{1/2}|v|_{2,K}+\frac{1}{h_K^{1/2}}|v|_{1,K}\right). \tag{96}$$

Theorem 2 is a direct consequence of Eq. (96) and Proposition 2.3 in Appendix 2. ■

## 6. Numerical investigation

We present two sets of experiments for illustration purpose. The first set is obtained when using four plane waves for the approximation at the element level and the second set in the case when eight plane waves are used for approximating the solution locally. In both experiments, we assume $\Omega$ to be an $a\times a$ square-shaped domain. For the numerical results presented in this section we have used discrete spaces containing plane waves. More specifically, we have tested the formulations given by (31)-(30) and (33)-(37) respectively using the spaces proposed in [6].
From now on, we assume $\Omega$ to be an $a\times a$ square-shaped domain. Note that the chosen domain possesses an exterior boundary only. We use a uniform mesh partition of $\Omega$ in rectangular-shaped elements $K$. In the first set of experiments we choose $f$ and $g_\Sigma$ such that the exact solution is given by :

$$u\left(x,y\right)=e^{\mathrm{i}\,kxy}, \tag{97}$$

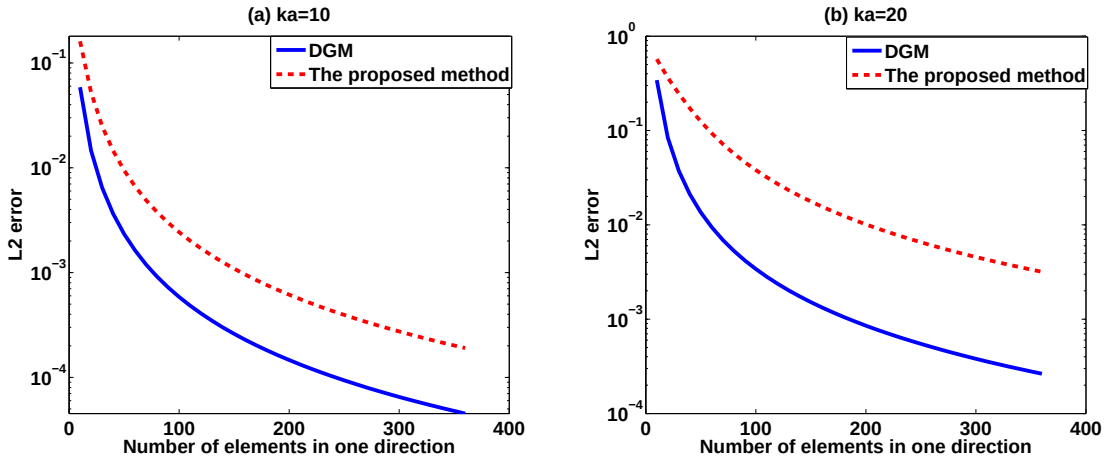


**Fig. 3** – Comparison between DGM and the formulation given by (31)-(30) in the case of the $R$-4-1 element for (a) $ka = 10$ and (b) $ka = 20$.

whereas in the second set of experiments the exact solution is a plane wave propagating in a direction characterized by the vector $\boldsymbol{d} = (\cos\theta, \sin\theta)$, that is :

$$u(x,y) = e^{x\cos\theta + y\sin\theta}. \tag{98}$$

### 6.1. Four plane waves per element

In this paragraph we approximate the solution at the element level using four plane waves that are positioned at :

$$\theta_p = (p-1)\pi/2, \quad 1 \leq p \leq 4. \tag{99}$$

On each interior edge, the Lagrange multiplier is approximated using one constant. This choice corresponds to the so-called $R$-4-1 element, introduced by Farhat *et al* in [6] and discussed from a mathematical point of view in [1]. Note that the discrete spaces that occur in $R$-4-1 element satisfy the compatibility condition given by (49). In addition, in the case of a domain whose elements have the edges parallel to the axis, the normal derivatives of the plane waves propagating in the directions given by (99) are constant on all the interior edges. Therefore, the formulation given by (19) is "sufficient", since the condition $\partial_n^K \Phi_h(\mu_h) \cong \mu_h^K$ on $\partial K \cap \mathring{\Omega}$ is satisfied in this case in the strong sense. These two very important properties of the $R$-4-1 element lead to accurate and stable formulations for both DGM and the method described by (31)-(30).
In Fig. 3 we report the $L^2$-error delivered by DGM and the method described by (31)-(30) for $ka = 10$ and $ka = 20$ respectively. These results show that both methods have the same order of convergence. However, the bounding constant in DGM seems to be smaller than the one appearing in the proposed formulation.

**Remark 7.** We must point out that solving the local problems is a task to be done carefully. The obtained local matrices are ill-conditioned and therefore the method chosen for inverting them is crucial. In order to illustrate the sensitivity of the results to the solving method, we compare in Fig. 4 the $L^2$-error obtained in the previous example when using a standard inversion - that uses the computation of the determinant and of the corresponding minors - and a LU inversion with partial pivot. This result suggests that it is recommended to use a LU factorization. In addition, we have

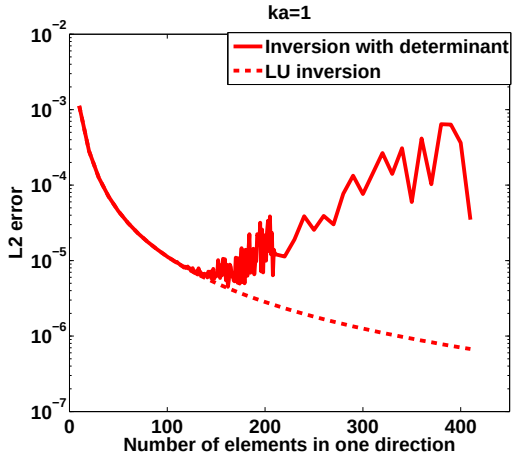


**Fig. 4** – Sensitivity of the relative error to the mesh refinement for the formulation given by (31)-(30) in the case of the $R$-4-1 element when using two methods of inversion

observed during our numerical investigation, that the lack of consistency in the computation of the integrals can also affect considerably the results. More specifically, it is recommended to use the same type of computation (either analytical or numerical) for all the integrals. Note that the local matrices that occur when using higher order elements are worse conditioned, and therefore these considerations become more important than in the case of the $R$-4-1 element.

### 6.2. Eight plane waves per element

We consider here eight plane waves that approximate the solution at the element level. More specifically, the directions of the eight plane waves are given by :

$$\theta_p = (l-1)\pi/4, \quad 1 \leq p \leq 8. \tag{100}$$

We vary $\theta$ (see Eq. (98)) in the interval $[0, 2\pi]$ and for each angle we measure the relative error using the following modified $H^1$-norm :

$$\|v\|_{\widehat{H^1}} = \left( \sum_K \|v\|^2_{H^1(K)} + \sum_{e-\text{interior edge}} \|[v]\|^2_{L^2(e)} \right)^{\frac{1}{2}}, \quad \forall v \in \mathcal{V}. \tag{101}$$

The computation of the normal derivative of a linear combination of these eight basis functions leads, in the case of a rectangular-shaped domain whose subdomains are also rectangular-shaped, to three dofs per edge. More precisely, $\lambda_h$ is approximated on each edge by :

$$\lambda_h = \mu_1 + \mu_2 e^{\mathrm{i}\, k \frac{\sqrt{2}}{2} s} + \mu_3 e^{-\mathrm{i}\, k \frac{\sqrt{2}}{2} s}, \tag{102}$$

where $s$ is the curvilinear abscissa along the edge. Such an approximation of the field and its normal derivative corresponds to the so-called $R$-8-3 element. In Fig. 5 we have represented the relative error for DGM and for the formulation given by (31)-(30). The results depicted herein illustrate the superiority of the proposed formulation over DGM. Indeed, for each propagation angle there is more than two orders of magnitude between the two relative errors.
In [6] it was shown that $R$-8-2b is the most accurate element for DGM when using with eight plane

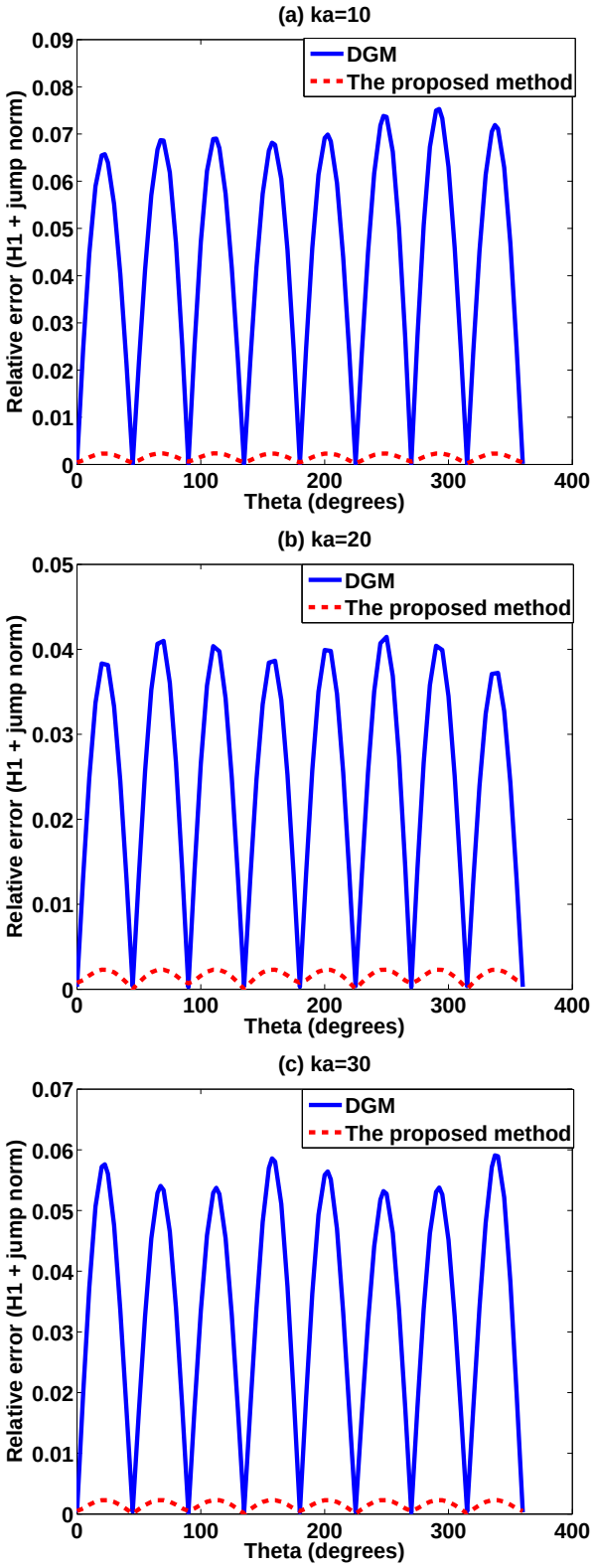


**Fig. 5** – Comparison between DGM and the formulation given by (31)-(30) in the case of the $R$-8-3 element for (a) $ka = 10$, (b) $ka = 20$ and (c) $ka = 30$

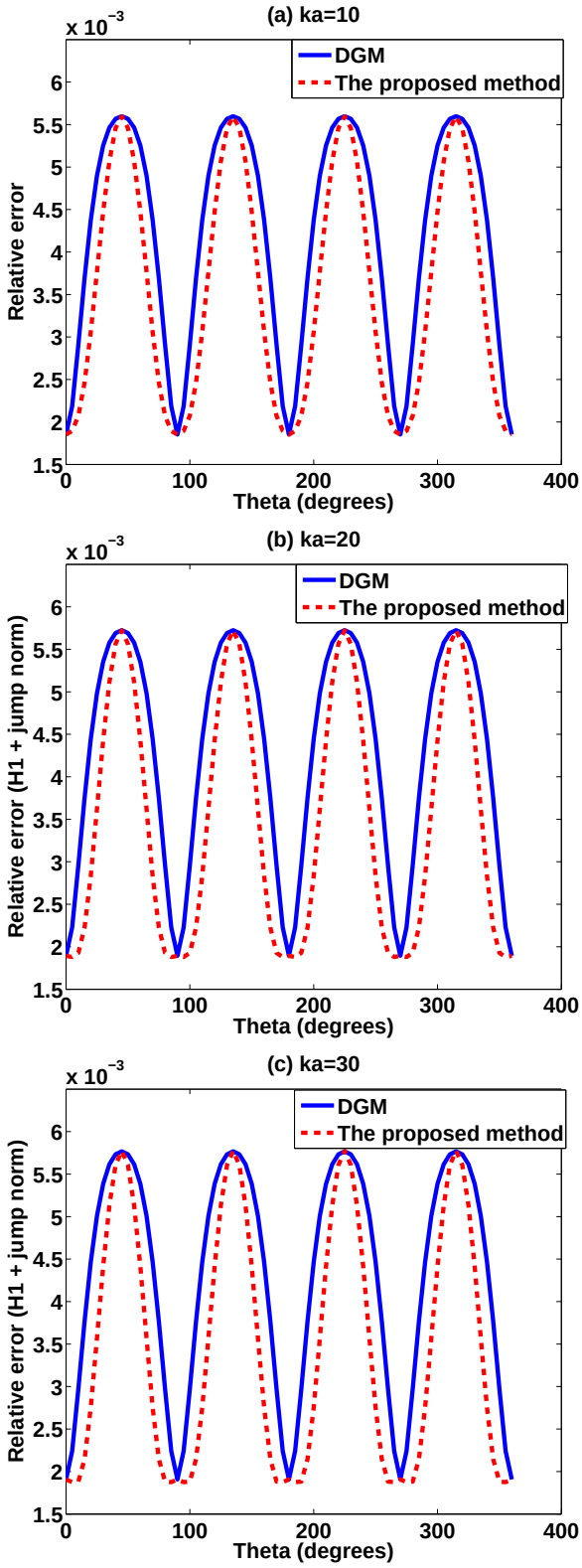


**Fig. 6** – Comparison between DGM and the formulation given by (31)-(30) in the case of the $R$-8-2b element for (a) $ka = 10$, (b) $ka = 20$ and (c) $ka = 30$

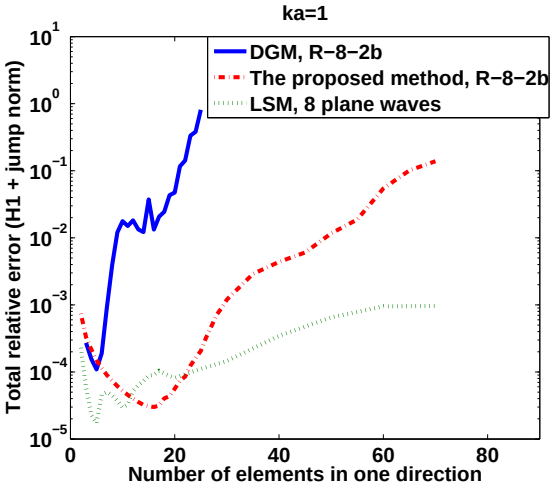


**Fig. 7** – Sensitivity of the total relative error to the mesh refinement in the case of DGM ($R$-8-2b), the formulation given by (33)-(37) ($R$-8-2b), and LSM (8 plane waves) : $ka = 1$

waves at the element level. More specifically, the Lagrange multiplier is approximated, on each interior edge by :

$$\lambda_h = \mu_1 e^{\mathrm{i}\, k \frac{\sqrt{2}}{4} s} + \mu_2 e^{-\mathrm{i}\, k \frac{\sqrt{2}}{4} s}. \tag{103}$$

The comparison of the results delivered by DGM and the formulation given by (31)-(30) are depicted in Fig. 6. These results suggest that for this element the errors are comparable, even though we observe that the proposed formulation performs slightly better.

Last, we have fixed $ka = 1$ and we have investigated the sensitivity of the total relative error - the mean error over all the angles $\theta$ - of DGM and the two formulations given by (31)-(30) and (33)-(37) respectively. The first preliminary tests revealed that (31)-(30) is an unstable formulation. Therefore, we will only present the results obtained with the formulation described by (33)-(37). The results depicted in Fig. 7 illustrate the following :

- The formulation given by (33)-(37) is superior to DGM since the approximation is more accurate and the formulation more stable.
- Numerical instabilities are observed in the proposed method as soon as $h/a < \frac{1}{20}$. These instabilities lead to a loss of almost three orders of magnitude in the accuracy of the solution. We believe that this phenomenon is related to the ill-conditioned matrices that incur when solving the local problems.
- The proposed method is less performant than LSM. The latter method maintains the total relative error to about 0.1% for fine resolutions (about 600 elements per wavelength). On the other hand, note that LSM is four times more expensive since the asymptotic size of the solution vector is $8p^2$, whereas the one of the proposed method is $2p^2$, where $p$ designates the number of elements in one direction (see Table 1).

## 7. Summary and conclusion

We have proposed a new discontinuous Galerkin method for solving Helmholtz problems. The proposed method is midway between DGM and LSM. Similarly to DGM, we use Lagrange multipliers

for restoring the weak continuity of the solution. The proposed method resembles LSM in the manner of writing the continuity condition, that is the least-squares sense. Under a compatibility assumption, this leads, similarly to LSM, to a positive-definite Hermitian matrix. We have established *a priori* error estimates in terms of the wavenumber $k$, the mesh step size $h$ and the number of degrees of freedom in the approximation of both the field and its normal derivative. The numerical experiments revealed that the method is more stable and therefore more accurate than DGM. However, when the mesh size becomes small, numerical instabilities occur, deteriorating the accuracy. We believe that the nearly numerical singularity of the local problems is the source of these instabilities. We plan to address this issue in the next part.

# Bibliography


[1] M. Amara, R. Djellouli, C. Farhat, Convergence analysis of a discontinuous Galerkin method with plane waves and Lagrange multipliers for the solution of Helmholtz problems *SIAM J. Numer. Anal.*, **47**(2), 1038–1066, 2009

[2] X. Antoine, H. Barucq, A. Bendali, Bayliss-Turkel-like Radiation Conditions on Surfaces of Arbitrary Shape. *Journal of Mathematical Analysis and Applications*, **229** 184-211, 1999

[3] Ciarlet P. G., The finite element method for elliptic problems *North Holland, Amsterdam*, 1978

[4] Cessenat O., Després B., Application of an ultra-weak variational formulation of elliptic PDEs to the two-dimensional Helmholtz problems *SIAM J. Numer. Anal.*, **35**, 255-299, 1998

[5] C. Farhat, I. Harari, L. Franca, The discontinuous enrichment method. *Comput. Methods in Appl. Mech. and Engrg.*, **190**, 6455-6479, 2001

[6] C. Farhat, I. Harari, U. Hetmaniuk, A discontinuous Galerkin method with Lagrange multipliers for the solution of Helmholtz problems in the mid-frequency regime. *Comput. Methods in Appl. Mech. and Engrg.*, **192**, 1389-1419, 2003

[7] C. Farhat, P. Wiedemann-Goiran, R. Teazur, A discontinuous Galerkin method with plane waves and Lagrange multipliers for the solution of short-wave exterior Helmholtz problem on unstructured meshes. *Wave Motion*, **39**, 307-317, 2004

[8] Farhat C., Tezaur R., Wiedemann-Goiran P., Higher-order extensions of a discontinuous Galerkin method for mid-frequency Helmholtz problems. *Internat. J. Numer. Methhods Eng.* **61**, 1938-1956, 2004

[9] Lions J. L., Magenes M., Problèmes aux limites non homogènes et applications, *Dunod, Paris*, 1968

[10] Grisvard P., Elliptic Problems in Non Smooth Domains *Pitman, Boston*, 1985

[11] Grisvard P., Singularities in boundary value problems, *Springer-Verlag*, 1992

[12] P. Monk, D.-Q. Wang, A least-squares method for the Helmholtz equation. *Comput. Methods in Appl. Mech. and Engrg.*, **175**(1-2), 121–136, 1999

# Appendix 1

This section is devoted to the proof of Properties 2, 3, 4 in Part I, Section 2.2. The references cited here are listed in the bibliography of Part I. From now on, $\hat{C}$ represents a positive constant which is independent of the mesh partition in elements $K$, of $h$ (the mesh size), as well as of the wavenumber $k$. First, we state the following technical result that we formulate as a lemma :

**Lemma 1.1.** Assume $kh$ to be sufficiently small. Then, there is a positive constant $\hat{C}$ such that we have :

$$|v|_{1,K} \leq \hat{C} h_K^{1/2} \|\partial_n v\|_{0,\partial K}, \quad \forall v \in Y(K) \tag{1}$$

and

$$\|v\|_{0,K} + h_K^{1/2} \|v\|_{0,\partial K} \leq \hat{C} \left( h_K^{3/2} \|\partial_n v\|_{0,\partial K} + \frac{1}{k^2 h_K} \Big| \int_{\partial K} \partial_n v \, ds \Big| \right), \qquad \forall v \in Y(K), \tag{2}$$

where the space $Y(K)$ is given by :

$$Y(K) = \{ v \in H^1(K) ; \quad \Delta v + k^2 v = 0 \text{ in } K \text{ and } \partial_n v \in L^2(\partial K) \}. \tag{3}$$

**Proof of Lemma 1.1.** Let $v \in Y(K)$.
First, we recall the following results :

$$\int_K \nabla v \cdot \nabla \overline{w} \, dx - k^2 \int_K v \overline{w} \, dx = \int_{\partial K} \partial_n v \overline{w} \, ds, \quad \forall w \in H^1(K). \tag{4}$$

which implies :

$$-k^2 \int_K v \, dx = \int_{\partial K} \partial_n v \, ds \tag{5}$$

and

$$|v|_{1,K}^2 = k^2 \|v\|_{0,K} + \int_{\partial K} \partial_n v \overline{v} \, ds. \tag{6}$$

In addition, observe that :

$$\Big\| v - \frac{1}{|K|} \int_K v \, dx \Big\|_{0,K}^2 = \|v\|_{0,K}^2 - \frac{1}{|K|} \int_K v \, dx \int_K \overline{v} \, dx. \tag{7}$$

Recall also that for any $v \in H^1(K)$, we have the following inverse inequalities [3] :

$$\|v\|_{0,\partial K} \leq \hat{C}\left(h_K^{1/2}|v|_{1,K} + \frac{1}{h_K^{1/2}}\|v\|_{0,K}\right) \tag{8}$$

$$\left\|v - \frac{1}{|K|}\int_K v\,dx\right\|_{0,K} \leq \hat{C}h_K|v|_{1,K} \tag{9}$$

$$\left\|v - \frac{1}{|K|}\int_K v\,dx\right\|_{0,\partial K} \leq \hat{C}h_K^{1/2}|v|_{1,K} \tag{10}$$

Next, we prove estimate (1). From (7) we have :

$$k^2\|v\|_{0,K}^2 + \int_{\partial K}\partial_n v\overline{v}\,ds \quad = k^2\left\|v - \frac{1}{|K|}\int_K v\,dx\right\|_{0,K}^2 + \frac{k^2}{|K|}\left|\int_K v\,dx\right|^2 + \int_{\partial K}\overline{v}\partial_n v\,ds \tag{11}$$

Using (5), we deduce that :

$$\begin{aligned} k^2\|v\|_{0,K}^2 + \int_{\partial K}\partial_n v\overline{v}\,ds &= k^2\left\|v - \frac{1}{|K|}\int_K v\,dx\right\|_{0,K}^2 - \frac{1}{|K|}\int_{\partial K}\partial_n v\,ds\int_K \overline{v}\,dx + \int_{\partial K}\overline{v}\partial_n v\,ds \\ &= k^2\left\|v - \frac{1}{|K|}\int_K v\,dx\right\|_{0,K}^2 + \int_{\partial K}\partial_n v\left(\overline{v} - \frac{1}{|K|}\int_K \overline{v}\,dx\right)ds \\ &\leq k^2\left\|v - \frac{1}{|K|}\int_K v\,dx\right\|_{0,K}^2 + \|\partial_n v\|_{0,\partial K}\left\|v - \frac{1}{|K|}\int_K v\,dx\right\|_{0,\partial K} \end{aligned} \tag{12}$$

Therefore, using Eqs. (6), (9) and (10) we have :

$$|v|_{1,K}^2 \leq \hat{C}\left(k^2h_K^2|v|_{1,K}^2 + k_K^{1/2}\|\partial_n v\|_{0,\partial K}|v|_{1,K}\right) \tag{13}$$

Consequently, for $kh$ sufficiently small, we deduce that :

$$|v|_{1,K} \leq \hat{C}h_K^{1/2}\|\partial_n v\|_{0,\partial K} \tag{14}$$

which concludes the proof of estimate 1.

Next, we proof estimate (2). From Eqs. (7) and (5) we have :

$$\begin{aligned} \|v\|_{0,K}^2 &= \left\|v - \frac{1}{|K|}\int_K v\,dx\right\|_{0,K}^2 - \frac{1}{k^4|K|}\int_{\partial K}\partial_n v\,ds\int_{\partial K}\partial_n\overline{v}\,ds \\ &\leq \left\|v - \frac{1}{|K|}\int_K v\,dx\right\|_{0,K}^2 + \frac{1}{k^4|K|}\left|\int_{\partial K}\partial_n v\,ds\right|^2 \end{aligned} \tag{15}$$

Therefore, using Eq. (9) we deduce that :

$$\|v\|_{0,K} \leq \hat{C}h_K|v|_{1,K} + \frac{1}{k^2h_K}\left|\int_{\partial K}\partial_n v\,ds\right| \tag{16}$$

Moreover, using estimate (1), we have :

$$\|v\|_{0,K} \leq \hat{C} h_K^{3/2} \|\partial_n v\|_{0,\partial K} + \frac{1}{k^2 h_K} \Big| \int_{\partial K} \partial_n v \, ds \Big| \tag{17}$$

On the other hand, from Eqs. (8), (16) and estimate (1) we deduce that :

$$\|v\|_{0,\partial K} \leq \hat{C} \left( \frac{1}{h_K^{1/2}} \left( h_K^{3/2} \|\partial_n v\|_{0,\partial K} + \frac{1}{k^2 h_K} \Big| \int_{\partial K} \partial_n v \, ds \Big| \right) + h_K^{1/2} h_K^{1/2} \|\partial_n v\|_{0,\partial K} \right) \tag{18}$$

Consequently, we obtain :

$$\|v\|_{0,\partial K} \leq \hat{C} \left( h_K \|\partial_n v\|_{0,\partial K} + \frac{1}{k^2 h_K^{3/2}} \Big| \int_{\partial K} \partial_n v \, ds \Big| \right) \tag{19}$$

Finally, from Eqs. (17) and (19) we have :

$$\|v\|_{0,K} + h_K^{1/2} \|v\|_{0,\partial K} \leq \hat{C} \left( h_K^{3/2} \|\partial_n v\|_{0,\partial K} + \frac{1}{k^2 h_K} \Big| \int_{\partial K} \partial_n v \, ds \Big| \right), \tag{20}$$

which concludes the proof of estimate (2). ■

**Corollary 1.1.** Assume $kh$ to be sufficiently small. There is a positive constant $\hat{C}$ such that :

$$\|v\|_{0,K} + h_K^{1/2} \|v\|_{0,\partial K} \leq \frac{\hat{C}}{k^2 h_K^{1/2}} \|\partial_n v\|_{0,\partial K} \tag{21}$$

**Proof of Corollary 1.1.** From estimate (2) and using the Cauchy-Schwartz inequality, we obtain :

$$\begin{aligned}
\|v\|_{0,K} + h_K^{1/2} \|v\|_{0,\partial K} &\leq \hat{C} \left( h_K^{3/2} \|\partial_n v\|_{0,\partial K} + \frac{1}{k^2 h_K} \|\partial_n v\|_{0,\partial K} h_K \right) \\
&\leq \hat{C} \left( h_K^{3/2} + \frac{1}{k^2} \right) \|\partial_n v\|_{0,\partial K} \\
&\leq \hat{C} \frac{1}{k^2} \left( \frac{k^2 h_K^2}{h_K^{1/2}} + 1 \right) \|\partial_n v\|_{0,\partial K} \\
&\leq \hat{C} \frac{1}{k^2} \left( \frac{1}{h_K^{1/2}} + 1 \right) \|\partial_n v\|_{0,\partial K} \\
&\leq \frac{\hat{C}}{k^2 h_K^{1/2}} \|\partial_n v\|_{0,\partial K}
\end{aligned} \tag{22}$$

which concludes the proof. ■

**Property 1.1.** Assume $kh$ to be sufficiently small. Then, there is a positive constant $\hat{C}$ such that for any element $K$ of the partition we have :

$$\sqrt{k}|v|_{1,K} + k^2 h_K^{1/2} \left( \|v\|_{0,K} + h_K^{1/2} \|v\|_{0,\partial K} \right) \leq C \|\partial_n v\|_{0,\partial K \cap \mathring{\Omega}}, \quad \forall v \in \mathcal{V}(K), \tag{23}$$

where the space $\mathcal{V}(K)$ is given by :

$$\forall K \in \tau_h, \quad \mathcal{V}(K) = \left\{ \begin{array}{l} v \in H^1(K), \, \Delta v + k^2 v = 0 \text{ in } K, \, \partial_n^K v \in L^2(\partial K), \\ \\ \partial_n^K v = \mathrm{i}\, k v \text{ on } \partial K \cap \Sigma, \, \partial_n^K v = 0 \text{ on } \partial K \cap \Gamma \end{array} \right\} \tag{24}$$

**Proof of Property 1.1.** Let $v \in \mathcal{V}(K)$. We prove Property 1.1. by considering two cases :

**Case 1 :** Assume that the measure of $\partial K \cap \Sigma$ equals 0, that is $|\partial K \cap \Sigma| = 0$. Then, the result is given by Lemma 1.1. and Corollary 1.1.

**Case 2 :** Assume that the measure of $\partial K \cap \Sigma$ is larger than 0, that is $|\partial K \cap \Sigma| > 0$. In this case we have :

$$|v|_{1,K}^2 - k^2 \|v\|_{0,K} - \mathrm{i}\, k \|v\|_{0,\partial K \cap \Sigma}^2 = \int_{\partial K \cap \mathring{\Omega}} \partial_n v \overline{v} \, ds \tag{25}$$

and

$$\|v\|_{0,K} \leq \hat{C} \left( h_K |v|_{1,K} + h_K^{1/2} \|v\|_{0,\partial K \cap \Sigma} \right) \tag{26}$$

From Eqs. (25) and (26) we deduce that :

$$|v|_{1,K}^2 + k\|v\|_{0,\partial K \cap \Sigma}^2 \leq \|\partial_n v\|_{0,\partial K \cap \mathring{\Omega}} \|v\|_{0,\partial K \cap \mathring{\Omega}} + \hat{C} \left( k^2 h_K^2 |v|_{1,K}^2 + k^2 h_K \|v\|_{0,\partial K \cap \Sigma}^2 \right) \tag{27}$$

For $kh_K$ sufficiently small we have :

$$|v|_{1,K}^2 + k\|v\|_{0,\partial K \cap \Sigma}^2 \leq \|\partial_n v\|_{0,\partial K \cap \mathring{\Omega}} \|v\|_{0,\partial K \cap \mathring{\Omega}} \tag{28}$$

Moreover, from Eqs. (8) and(26) we deduce that :

$$\|v\|_{0,\partial K} \leq \hat{C} \left( h_K^{1/2} |v|_{1,K} + \|v\|_{0,\partial K \cap \Sigma} \right) \tag{29}$$

Therefore, using Eq. (29) in Eq. (28) leads to :

$$|v|_{1,K}^2 + k\|v\|_{0,\partial K \cap \Sigma}^2 \quad \leq \hat{C} \left( h_K^{1/2} |v|_{1,K} + \|v\|_{0,\partial K \cap \Sigma} \right) \|\partial_n v\|_{0,\partial K \cap \mathring{\Omega}} \tag{30}$$

and consequently :

$$|v|_{1,K}^2 + k\|v\|_{0,\partial K \cap \Sigma}^2 \leq \frac{\hat{C}}{k} \|\partial_n v\|_{0,\partial K \cap \mathring{\Omega}}^2 \tag{31}$$

Hence,

$$|v|_{1,K} + \sqrt{k}\|v\|_{0,\partial K \cap \Sigma} \leq \frac{\hat{C}}{\sqrt{k}} \|\partial_n v\|_{0,\partial K \cap \mathring{\Omega}} \tag{32}$$

Moreover, combining Eqs. (26) and(29), we obtain :

$$\|v\|_{0,K} + h_K^{1/2} \|v\|_{0,\partial K} \leq \hat{C} \left( h_K |v|_{1,K} + h_K^{1/2} \|v\|_{0,\partial K \cap \Sigma} \right) \tag{33}$$

From Eq. (32) we deduce that :

$$h_K|v|_{1,K} \leq \frac{\hat{C}h_K}{\sqrt{k}}\|\partial_n v\|_{0,\partial K\cap\mathring{\Omega}} \tag{34}$$

and

$$h_K^{1/2}\|v\|_{0,\partial K\cap\Sigma} \leq \frac{\hat{C}h_K^{1/2}}{k}\|\partial_n v\|_{0,\partial K\cap\mathring{\Omega}} \tag{35}$$

Therefore, Eq. (33) becomes :

$$\|v\|_{0,K} + h_K^{1/2}\|v\|_{0,\partial K} \leq \hat{C}\left(\frac{h_K}{\sqrt{k}} + \frac{h_K^{1/2}}{k}\right)\|\partial_n v\|_{0,\partial K\cap\mathring{\Omega}} \tag{36}$$

For $kh_K$ sufficiently small, we have :

$$\begin{aligned} \|v\|_{0,K} + h_K^{1/2}\|v\|_{0,\partial K} &\leq \frac{\hat{C}h_K^{1/2}}{k}\|\partial_n v\|_{0,\partial K\cap\mathring{\Omega}} \\ &\leq \frac{\hat{C}}{k^2 h_K^{1/2}}\|\partial_n v\|_{0,\partial K\cap\mathring{\Omega}} \end{aligned} \tag{37}$$

Estimate (23) is a direct consequence of Eqs. (32) and(37). ■

**Property 1.2.** Assume that $kh \leq 1$. Then, we have :

$$\|\Psi\|_{0,\Omega} \leq \hat{C}\left(\sum_{e-\text{interior edge}} \frac{1}{k^2|e|}\|\left[\left[\partial_n\Psi\right]\right]\|_{0,e}^2 + \sum_{e-\text{interior edge}} \frac{1}{|e|}\|[\Psi]\|_{0,e}^2\right)^{1/2}, \quad \forall\Psi\in\mathcal{V} \tag{38}$$

where the space $\mathcal{V}$ is given by :

$$\mathcal{V} = \left\{v\in L^2(\Omega)\,;\,\forall K\in\tau_h, v|_K\in\mathcal{V}(K)\right\}. \tag{39}$$

**Proof of Property 1.2.** Let $\Psi\in\mathcal{V}$. We consider the following boundary value problem :

$$\left\{\begin{array}{rll} \text{Find } w\in H^1(\Omega) & \text{such that :} & \\ -\Delta w - k^2 w &= \Psi & \text{in } \Omega \\ \partial_n w &= -\mathrm{i}\,kw & \text{on } \Sigma \\ \partial_n w &= 0 & \text{on } \Gamma \end{array}\right. \tag{40}$$

The boundary value problem (40) has a unique solution . In addition, for all the elements $K$ in the mesh, $w$ verifies $\partial_n w\in L^2(\partial K)$ because of the regularity result $w\in H^{5/3}(\Omega)$. This is a standard regularity result for Laplace's operators (reference [10] in the Bibliography of Part I). More specifically, since $\Psi\in L^2(\Omega)$ and $\Omega$ is a polygonal-shaped domain with possible reentrant corners with a measure angle of $\frac{3\pi}{2}$ then, similarly to problem (2.11) given by Lemma 1 (page 1044 in reference [1] in the Bibliography of Part I), the solution is necessarily in $H^{5/3}(\Omega)$ at least.

We also have (see [1]) :

$$\|w\|_{0,\Omega} \leq \frac{\hat{C}}{1+k}\|\Psi\|_{0,\Omega} \tag{41}$$

$$|w|_{1,\Omega} \leq \hat{C}\|\Psi\|_{0,\Omega} \tag{42}$$

$$|w|_{3/2+\varepsilon,\Omega} \leq \hat{C}\,(1+k)^{1/2+\varepsilon}\,\|\Psi\|_{0,\Omega}, \quad 0<\varepsilon\leq\frac{1}{6}. \tag{43}$$

We use also the following inverse inequalities, that we prove in Appendix 3 :

$$\|w\|_{0,e} \leq \hat{C}\left(h_K^{1/2}|w|_{1,K} + \frac{1}{h_K^{1/2}}\|w\|_{0,K}\right) \tag{44}$$

and

$$\|\partial_n w\|_{0,e} \leq \hat{C}\left(h_K^{\varepsilon}|w|_{3/2+\varepsilon,K} + \frac{1}{h_K^{1/2}}|w|_{1,K}\right), \quad 0<\varepsilon\leq\frac{1}{6}. \tag{45}$$

In addition, from Eq. (40) and using Green's formula, we have :

$$\begin{aligned}
\|\Psi\|_{0,\Omega}^2 &= -\int_\Omega \Psi\left(\Delta\overline{w}+k^2\overline{w}\right)\,dx = -\sum_{K\in\tau_h}\int_K \Psi\left(\Delta\overline{w}+k^2\overline{w}\right)\,dx \\
&= \sum_{K\in\tau_h}\left(\int_K \nabla\Psi\cdot\nabla\overline{w}\,dx - k^2\int_K \Psi\overline{w}\,dx - \int_{\partial K}\Psi\partial_n\overline{w}\,ds\right) \\
&= \sum_{K\in\tau_h}\left(-\int_K \overline{w}\left(\Delta\Psi+k^2\Psi\right)\,dx + \int_{\partial K}\overline{w}\partial_n\Psi\,ds - \int_{\partial K}\Psi\partial_n\overline{w}\,ds\right)
\end{aligned} \tag{46}$$

Moreover, in each element $K\in\tau_h$, we have :

$$\left\{\begin{array}{rll} -\Delta\Psi - k^2\Psi &= 0 & \text{in } K \\ \partial_n\Psi &= \mathrm{i}\,k\Psi & \text{on } \partial K\cap\Sigma \\ \partial_n\Psi &= 0 & \text{on } \partial K\cap\Gamma \end{array}\right. \tag{47}$$

Therefore, using Eqs. (46) and (47) we obtain :

$$\|\Psi\|_{0,\Omega}^2 = \sum_{K\in\tau_h}\left(\int_{\partial K\cap\mathring{\Omega}}\overline{w}\partial_n\Psi\,ds + \mathrm{i}\,k\int_{\partial K\cap\Sigma}\overline{w}\Psi\,ds - \int_{\partial K}\Psi\partial_n\overline{w}\,ds\right) \tag{48}$$

In addition, from Eq. (40), we deduce that :

$$\left\{\begin{array}{rll} \partial_n\overline{w} &= \mathrm{i}\,k\overline{w} & \text{on } \partial K\cap\Sigma \\ \partial_n\overline{w} &= 0 & \text{on } \partial K\cap\Gamma \end{array}\right. \tag{49}$$

Therefore, we have :

$$
\begin{aligned}
\|\Psi\|_{0,\Omega}^2 &= \sum_{K\in\tau_h}\left(\int_{\partial K\cap\mathring{\Omega}}\overline{w}\partial_n\Psi\,ds-\int_{\partial K\cap\mathring{\Omega}}\Psi\partial_n\overline{w}\,ds\right)\\
&= \sum_{\substack{e\text{-interior edge}\\ e=\partial K\cap\partial K'}}\int_e\left(\overline{w^K}\partial_n^K\Psi^K+\overline{w^{K'}}\partial_n^{K'}\Psi^{K'}-\Psi^K\partial_n^K\overline{w^K}-\Psi^{K'}\partial_n^{K'}\overline{w^{K'}}\right)\\
&= \sum_{\substack{e\text{-interior edge}\\ e=\partial K\cap\partial K'}}\int_e\overline{w^K}\left(\partial_n^K\Psi^K+\partial_n^{K'}\Psi^{K'}\right)ds-\int_e\partial_n^K\overline{w^K}\left(\Psi^K-\Psi^{K'}\right)ds\\
&\le \sum_{e-\text{interior edge}}\|w\|_{0,e}\|\,[[\partial_n\Psi]]\,\|_{0,e}+\sum_{e-\text{interior edge}}\|\partial_n w\|_{0,e}\|[\Psi]\|_{0,e}\\
&\le \left(\sum_{e-\text{interior edge}}k^2|e|\|w\|_{0,e}^2\right)^{1/2}\left(\sum_{e-\text{interior edge}}\frac{1}{k^2|e|}\|\,[[\partial_n\Psi]]\,\|_{0,e}^2\right)^{1/2}\\
&\quad+\left(\sum_{e-\text{interior edge}}|e|\|\partial_n w\|_{0,e}^2\right)^{1/2}\left(\sum_{e-\text{interior edge}}\frac{1}{|e|}\|[\Psi]\|_{0,e}^2\right)^{1/2}
\end{aligned}
\tag{50}
$$

Moreover, from Eqs. (44)-(45) we deduce that for an edge $e\subset\partial K\cap\mathring{\Omega}$ we have :

$$
|e|^{1/2}\|\partial_n w\|_{0,e}\le\hat{C}\left(h_K^{\varepsilon}|w|_{3/2+\varepsilon,K}+|w|_{1,K}\right),\quad 0<\varepsilon\le\frac{1}{6}
\tag{51}
$$

and

$$
|e|^{1/2}\|w\|_{0,e}\le\hat{C}\left(h_K|w|_{1,K}+\|w\|_{0,K}\right)
\tag{52}
$$

From Eqs.(51)-(52), (41)-(43) and for $kh$ sufficienlty small, we obtain :

$$
\begin{aligned}
\left(\sum_{e-\text{interior edge}}|e|\|w\|_{0,e}^2\right)^{1/2} &\le\hat{C}\left(h|w|_{1,\Omega}+\|w\|_{0,\Omega}\right)\\
&\le\hat{C}\left(h\|\Psi\|_{0,\Omega}+\frac{1}{1+k}\|\Psi\|_{0,\Omega}\right)\\
&\le\hat{C}\frac{h+kh+1}{1+k}\|\Psi\|_{0,\Omega}\\
&\le\hat{C}\frac{1}{1+k}\|\Psi\|_{0,\Omega}
\end{aligned}
\tag{53}
$$

and

$$\begin{aligned}\left(\sum_{e-\text{interior edge}} |e|\|\partial_n w\|_{0,e}^2\right)^{1/2} &\leq \hat{C}\left(h^{1/2+\varepsilon}|w|_{3/2+\varepsilon,\Omega} + |w|_{1,\Omega}\right),\ 0<\varepsilon\leq\tfrac{1}{6}\\ &\leq \hat{C}\left(h^{1/2+\varepsilon}\left(1+k\right)^{1/2+\varepsilon}\|\Psi\|_{0,\Omega} + \|\Psi\|_{0,\Omega}\right),\ 0<\varepsilon\leq\frac{1}{6}\\ &\leq \hat{C}\left(\left(h+kh\right)^{1/2+\varepsilon}\|\Psi\|_{0,\Omega} + \|\Psi\|_{0,\Omega}\right),\quad 0<\varepsilon\leq\frac{1}{6}\\ &\leq \hat{C}\|\Psi\|_{0,\Omega}\end{aligned} \tag{54}$$

Last, from Eqs. (50), (53) and (54) and using $\frac{k}{k+1}<1$, we have :

$$\|\Psi\|_{0,\Omega} \leq \hat{C}\left(\sum_{e-\text{interior edge}} \frac{1}{k^2|e|}\|\left[\left[\partial_n\Psi\right]\right]\|_{0,e}^2 + \sum_{e-\text{interior edge}} \frac{1}{|e|}\|[\Psi]\|_{0,e}^2\right)^{1/2}$$

which concludes the proof. ■

**Property 1.3.** Let $\mathcal{M}$ be the space given by :

$$\mathcal{M} = \left\{\mu\in\prod_{K\in\tau_h} L^2\left(K\right);\, \forall K\in\tau_h, \mu^K = 0 \text{ on } \partial K\cap\partial\Omega, \text{ and } \forall K,K'\in\tau_h,\, \mu^K+\mu^{K'}=0\right\}. \tag{55}$$

We denote by $\Phi$ the following application :

$$\begin{array}{rccc}\Phi: & \mathcal{M} & \longrightarrow & \mathcal{V}\\ & \mu & \longmapsto & \Phi\left(\mu\right)\end{array} \tag{56}$$

such that we have :

$$\partial_n^K\Phi\left(\mu\right) = \mu^K \quad \text{on } \partial K\cap\mathring{\Omega} \quad \forall K\in\tau_h \tag{57}$$

Assume $kh$ to be sufficiently small. Then, the application $\Phi$ is well defined.

**Proof of Property 1.3.** Let $K$ be a fixed element in the mesh and $\mu\in\mathcal{M}$. We consider the following boundary value problem :

$$\left\{\begin{array}{rcll}\text{Find } w\in H^1\left(K\right) & \text{such that :} & & \\ -\Delta w - k^2 w & = 0 & \text{in } K, & \\ \partial_n w & = \mathrm{i}kw & \text{on } \partial K\cap\Sigma & \\ \partial_n w & = 0 & \text{on } \partial K\cap\Gamma & \\ \partial_n w & = \mu & \text{on } \partial K\cap\mathring{\Omega} & \end{array}\right. \tag{58}$$

We have the following equivalent variational formulation :

$$\int_K \nabla w\cdot\nabla\overline{v}dx - k^2\int_K w\overline{v}dx - \mathrm{i}\,k\int_{\partial K\cap\Sigma} w\overline{v}ds = \int_{\partial K\cap\mathring{\Omega}}\mu\overline{v}ds,, \quad \forall v\in H^1\left(K\right). \tag{59}$$

The corresponding bilinear form satisfies the Gärding inequality and consequently, the alternative of Fredholm holds. Therefore, the uniqueness of the solution ensures its existence. To prove the

uniqueness, let us consider the homogeneous associated problem, that is :

$$\left\{\begin{array}{rll} \text{Find } w \in H^1(K) & \text{such that :} & \\ -\Delta w - k^2 w & = 0 & \text{in } K, \\ \partial_n w & = \mathrm{i}kw & \text{on } \partial K \cap \Sigma \\ \partial_n w & = 0 & \text{on } \partial K \cap \Gamma \\ \partial_n w & = 0 & \text{on } \partial K \cap \mathring{\Omega} \end{array}\right. \tag{60}$$

In this case, we have :

$$\int_K \nabla w \cdot \nabla \overline{v} dx - k^2 \int_K w\overline{v} dx - \mathrm{i}\, k \int_{\partial K \cap \Sigma} w\overline{v} ds = 0, \quad \forall v \in H^1(K). \tag{61}$$

We distinguish two different situations :

1. $|\partial K \cap \Sigma| \neq 0$. In this case, by taking $v = w$ in Eq. (61), we deduce that $w = 0$ on $\partial K \cap \Sigma$ and consequently $\partial_n w = 0$ on $\partial K$.
2. $|\partial K \cap \Sigma| = 0$. In this case, we obtain $\partial_n w = 0$ on $\partial K$.

Consequently, we have :

$$\partial_n w = 0 \text{ on } \partial K, \quad \forall K \in \tau_h \tag{62}$$

From the first equation of (60) we deduce that :

$$\int_K \Delta w \, dx = -k^2 \int_K w \, dx \tag{63}$$

In addition, using Green's formula we have :

$$\int_K \Delta w \, dx = \int_{\partial K} \partial_n w \, ds = 0. \tag{64}$$

Combining Eqs. (63)-(64), we obtain :

$$\int_K w \, dx = 0. \tag{65}$$

In this case, we have :

$$\|w\|_{0,K} = \left\| w - \frac{1}{|K|} \int_K w dx \right\|_{0,K} \leq \inf_{a \in \mathbb{C}} \|w - a\|_{0,K} \leq \hat{C} h_K |w|_{1,K} \tag{66}$$

On the other hand, by taking $v = w$ in Eq. (61), we deduce that $w$ satisfies :

$$|w|_{1,K} = k\|w\|_{0,K}. \tag{67}$$

Combining Eqs. (66) and (67) leads to :

$$|w|_{1,K} \leq \hat{C} k h_K |w|_{1,K}. \tag{68}$$

Consequently, for $kh_K$ sufficiently small, we have :

$$|w|_{1,K} = 0 \tag{69}$$

and $w = 0$, which concludes the proof. ■

# Appendix 2

This appendix is devoted to the description of the techniques that allow to prove Theorem 2 in Part I, Section 5.1. From now on, $\hat{C}$ represents a positive constant which is independent of the mesh partition in elements $K$ and $h_K$. First, we introduce the following preliminary notations and properties :
Let $N \geq 0$. We define the space $Z(K)$ as follows :

$$Z(K) = \{v \in H^{N+2}(K); \quad \Delta v + k^2 v = 0 \text{ in } K\}. \tag{1}$$

We introduce also the space $Z_h(K)$ defined by :

$$Z_h(K) = \{v = \sum_{q=1}^{2N+1} \beta_q \phi_q; \quad \beta \in \mathbb{C}^{2N+1}\} \tag{2}$$

with

$$\phi_q = e^{\mathrm{i}\, k \overrightarrow{x} \cdot \overrightarrow{d_q}} \tag{3}$$

and
• $\overrightarrow{d_q} \in \mathbb{R}^2$
• $\|\overrightarrow{d_q}\| = 1$
• $\overrightarrow{d_q} \neq \overrightarrow{d_{q'}} \quad \forall q \neq q'$. Note that $|\phi_q| = 1$ for any $1 \leq q \leq 2N+1$. We introduce $z_q = d_{q_1} + \mathrm{i}\, d_{q_2} \in \mathbb{C}$ with $|z_q| = 1$. The following observations are noteworthy :

1. $\overline{z_q} = \dfrac{1}{z_q}$
2. $\forall m \geq 0$ we have :
   • $(\partial_x + \mathrm{i}\, \partial_y)^m \phi_q(\overrightarrow{x}) = (\mathrm{i}\, k z_q)^m \phi_q(\overrightarrow{x})$.
   • $(\partial_x - \mathrm{i}\, \partial_y)^m \phi_q(\overrightarrow{x}) = (\mathrm{i}\, k \overline{z_q})^m \phi_q(\overrightarrow{x})$.

We introduce also the matrix $\boldsymbol{A} \in \mathbb{C}^{M,M}$, whose entries are given by $\boldsymbol{A}_{mq} = z_q^{m-1}$, $1 \leq m \leq 2N+1 = M$.

**Lemma 2.1.** The matrix $\boldsymbol{A}$ has the following properties :

1. $\boldsymbol{A}$ is invertible.
2. $\|\boldsymbol{A}\|$ is independent of $K$ and of $k$.

**Proof of Lemma 2.1.**

1. Let $\xi \in \mathbb{C}^M$ such that $\boldsymbol{A}\xi = 0$. Then, we have :

$$\sum_{m=1}^{M} \xi_m z_q^{m-1} = 0, \quad 1 \leq q \leq M. \tag{4}$$

We put $P(z) = \sum_{m=1}^{M} \xi_m z_q^{m-1} = 0, \quad 1 \leq q \leq M$. Note that $P(\cdot)$ is a complex polynomial of degree $M-1$. From Eq. (4) we deduce that $(z_q)_{1\leq q\leq M}$ is root of the polynomial $P(\cdot)$. Consequently, for any $z \in \mathbb{C}$ we have $P(z) = 0$ and therefore $\xi = 0$. Therefore $A^T$ is invertible and hence $A$ is invertible.

2. $\boldsymbol{A}$ depends only on $M$ and $\left(\overrightarrow{d_q}\right)_{1\leq q\leq M}$. Consequently, $\|\boldsymbol{A}\|$ is independent of $K$ and of $k$. ■

Let $\overrightarrow{x_G}$ be the barycenter of the element $K$.

**Lemma 2.2.** There is a linear operator $R_N : Z(K) \longrightarrow \mathbb{C}^M$ such that if $v \in Z(K)$ and $\beta = R_N(v)$ then we have :

$$\begin{cases} \sum_{q=1}^{M} \beta_q \phi_q(\overrightarrow{x_G}) = v(\overrightarrow{x_G}) \\ \sum_{q=1}^{M} z_q^m \beta_q \phi_q(\overrightarrow{x_G}) = \frac{1}{(\mathrm{i}\,k)^m} (\partial_x + \mathrm{i}\,\partial_y)^m v(\overrightarrow{x_G}), \quad 1 \leq m \leq N \\ \sum_{q=1}^{M} \frac{1}{z_q^m} \beta_q \phi_q(\overrightarrow{x_G}) = \frac{1}{(\mathrm{i}\,k)^m} (\partial_x - \mathrm{i}\,\partial_y)^m v(\overrightarrow{x_G}), \quad 1 \leq m \leq N \end{cases} \tag{5}$$

**Proof of Lemma 2.2.** Let $v \in Z(K)$ and $b \in \mathbb{C}^M$ defined by :

$$\begin{cases} b_m = \frac{1}{(\mathrm{i}\,k)^{N-m+1}} (\partial_x - \mathrm{i}\,\partial_y)^{N-m+1} v(\overrightarrow{x_G}), \quad 1 \leq m \leq N \\ b_{N+1} = v(\overrightarrow{x_G}) \\ b_{m+N+1} = \frac{1}{(\mathrm{i}\,k)^m} (\partial_x + \mathrm{i}\,\partial_y)^m v(\overrightarrow{x_G}), \quad 1 \leq m \leq N \end{cases} \tag{6}$$

From Lemma 2.1 we know that there is a unique $\xi \in \mathbb{C}^M$ such that :

$$\boldsymbol{A}\xi = b \tag{7}$$

that is

$$\sum_{q=1}^{M} z_q^{m-1} \xi_q = b_m, \quad \forall 1 \leq m \leq 2N+1 \tag{8}$$

For $1 \leq q \leq 2N+1$, we put $\beta_q = \frac{1}{\phi_q(\overrightarrow{x_G})} z_q^N \xi_q$. Then, from Eq. (8) we deduce that :

$$\sum_{q=1}^{M} z_q^{m-1-N} \beta_q \phi_q(\overrightarrow{x_G}) = b_m \quad \text{for } 1 \leq m \leq 2N+1 \tag{9}$$

which leads to :

- $\sum_{q=1}^{M} z_q^{m-1-N}\beta_q\phi_q\left(\overrightarrow{x_G}\right) = b_m \quad$ for $1 \leq m \leq N$, that is

$$\sum_{q=1}^{M} \frac{1}{z_q^m}\beta_q\phi_q\left(\overrightarrow{x_G}\right) = b_{N-m+1} = \frac{1}{(\mathrm{i}\,k)^m}\left(\partial_x - \mathrm{i}\,\partial_y\right)^m v\left(\overrightarrow{x_G}\right) \quad \text{for } 1 \leq m \leq N \tag{10}$$

since $1 \leq m \leq N \Longrightarrow 1 \leq N - m + 1 \leq N$.

- $\sum_{q=1}^{M} \beta_q\phi_q\left(\overrightarrow{x_G}\right) = b_{N+1} = v\left(\overrightarrow{x_G}\right) \quad$ for $m = N + 1$
- $\sum_{q=1}^{M} z_q^{m-1-N}\beta_q\phi_q\left(\overrightarrow{x_G}\right) = b_m \quad$ for $N + 2 \leq m \leq 2N + 1$, that is

$$\sum_{q=1}^{M} z_q^m\beta_q\phi_q\left(\overrightarrow{x_G}\right) = b_{m+N+1} = \frac{1}{(\mathrm{i}\,k)^m}\left(\partial_x + \mathrm{i}\,\partial_y\right)^m v\left(\overrightarrow{x_G}\right) \quad \text{for } 1 \leq m \leq N \tag{11}$$

Consequently, the operator $R_N$ is well defined and it is linear since

$$\beta_q = \frac{z_q^N}{\phi_q\left(\overrightarrow{x_G}\right)}\xi_q \quad \text{for } 1 \leq q \leq M \tag{12}$$

and

$$\xi = \boldsymbol{A}^{-1}b \tag{13}$$

with $b$ linear with respect to $v$. ■

**Corollary 2.1.** There is a linear operator $\Pi_N : Z(K) \longrightarrow Z_h(K)$ such that for any $v \in Z(K)$ we have :

$$\begin{cases} \left(\partial_x + \mathrm{i}\,\partial_y\right)^m \Pi_N v\left(\overrightarrow{x_G}\right) = & \left(\partial_x + \mathrm{i}\,\partial_y\right)^m v\left(\overrightarrow{x_G}\right), \quad 1 \leq m \leq N \\ \left(\partial_x - \mathrm{i}\,\partial_y\right)^m \Pi_N v\left(\overrightarrow{x_G}\right) = & \left(\partial_x - \mathrm{i}\,\partial_y\right)^m v\left(\overrightarrow{x_G}\right), \quad 1 \leq m \leq N \\ \Pi_N v\left(\overrightarrow{x_G}\right) = & v\left(\overrightarrow{x_G}\right) \end{cases} \tag{14}$$

**Proof of Corollary 2.1.** Let $v \in Z(K)$. We put $\beta = R_N(v)$ and we define

$$\Pi_N v\left(\overrightarrow{x}\right) = \sum_{q=1}^{M} \beta_q\phi_q\left(\overrightarrow{x}\right), \quad \forall \overrightarrow{x} \in K \tag{15}$$

We have :

$$\begin{cases} (\partial_x + \mathrm{i}\,\partial_y)^m \Pi_N v(\overrightarrow{x}) = \sum_{q=1}^{M} \beta_q (\mathrm{i}\,k)^m z_q^m \phi_q(\overrightarrow{x}), \quad 0 \le m \le N \\ (\partial_x - \mathrm{i}\,\partial_y)^m \Pi_N v(\overrightarrow{x}) = \sum_{q=1}^{M} \beta_q (\mathrm{i}\,k)^m \frac{1}{z_q^m} \phi_q(\overrightarrow{x}), \quad 0 \le m \le N \end{cases} \tag{16}$$

Hence, for $0 \le m \le N$ from Lemma 2.2 we obtain :

$$\begin{cases} (\partial_x + \mathrm{i}\,\partial_y)^m \Pi_N v(\overrightarrow{x_G}) = (\mathrm{i}\,k)^m \sum_{q=1}^{M} \beta_q z_q^m \phi_q(\overrightarrow{x_G}) = (\partial_x + \mathrm{i}\,\partial_y)^m v(\overrightarrow{x_G}), 0 \le m \le N \\ (\partial_x - \mathrm{i}\,\partial_y)^m \Pi_N v(\overrightarrow{x_G}) = (\mathrm{i}\,k)^m \sum_{q=1}^{M} \frac{\beta_q}{z_q^m} \phi_q(\overrightarrow{x_G}) = (\partial_x - \mathrm{i}\,\partial_y)^m v(\overrightarrow{x_G}),, 0 \le m \le N, \end{cases} \tag{17}$$

which concludes the proof. ■

**Lemma 2.3.** Let $v \in Z(K)$. Then, we have :

$$\partial_x^j \partial_y^l \Pi_N v(\overrightarrow{x_G}) = \partial_x^j \partial_y^l v(\overrightarrow{x_G}), \quad \forall 0 \le j + l \le N \tag{18}$$

**Proof of Lemma 2.3.** We put $w = v - \Pi_N v$. We have $w \in Z(K)$. In addition, from Corollary 2.1 we deduce that :

$$(\partial_x + \mathrm{i}\,\partial_y)^m w(\overrightarrow{x_G}) = (\partial_x - \mathrm{i}\,\partial_y)^m w(\overrightarrow{x_G}) = 0, \quad 0 \le m \le N \tag{19}$$

Observe that for $\overrightarrow{x} \in K$ we have :

$$(\partial_x + \mathrm{i}\,\partial_y)(\partial_x - \mathrm{i}\,\partial_y) w(\overrightarrow{x}) = \Delta w(\overrightarrow{x}) = -k^2 w(\overrightarrow{x}) \tag{20}$$

Hence, for $0 \le j + l \le N$, we have :

$$(\partial_x + \mathrm{i}\,\partial_y)^j (\partial_x - \mathrm{i}\,\partial_y)^l w(\overrightarrow{x}) = \begin{cases} \left(-k^2\right)^j (\partial_x - \mathrm{i}\,\partial_y)^{l-j} & \text{if } j \le l \\ \left(-k^2\right)^l (\partial_x - \mathrm{i}\,\partial_y)^{j-l} & \text{if } l \le j \end{cases} \tag{21}$$

Consequently, using Eq. (19) we have :

$$(\partial_x + \mathrm{i}\,\partial_y)^j (\partial_x - \mathrm{i}\,\partial_y)^l = 0, \quad 0 \le j + l \le N \tag{22}$$

Observe also that for any $\overrightarrow{x} \in K$ we have :

$$\partial_x w(\overrightarrow{x}) = \frac{1}{2}(\partial_x + \mathrm{i}\,\partial_y) w(\overrightarrow{x}) + \frac{1}{2}(\partial_x - \mathrm{i}\,\partial_y) w(\overrightarrow{x}) \tag{23}$$

$$\partial_y w(\overrightarrow{x}) = \frac{1}{2}(\partial_x + \mathrm{i}\,\partial_y) w(\overrightarrow{x}) - \frac{1}{2}(\partial_x - \mathrm{i}\,\partial_y) w(\overrightarrow{x}) \tag{24}$$

This implies that for any $0 \leq j+l \leq N$, $\partial_x^j \partial_y^l w\left(\overrightarrow{x}\right)$ is a linear combination of

$$\left(\partial_x + \mathrm{i}\,\partial_y\right)^{j'} \left(\partial_x - \mathrm{i}\,\partial_y\right)^{l'} w\left(\overrightarrow{x}\right) \quad \text{with } 0 \leq j'+l' \leq j+l. \tag{25}$$

From Eq. (22), we deduce that for $0 \leq j+l \leq N$, $\partial_x^j \partial_y^l w\left(\overrightarrow{x_G}\right) = 0$, which concludes the proof. ■

Next, we define the operator $L_N : H^{N+2}(K) \longrightarrow P_N(K)$, where $P_N(K)$ is the set of polynomials of degree $N$ in $K$ as follows :

$$L_N v\left(\overrightarrow{x}\right) = \sum_{0 \leq j+l \leq N} \frac{1}{j!l!} \left(x - x_G\right)^j \left(y - y_G\right)^l \partial_x^j \partial_y^l v\left(\overrightarrow{x_G}\right), \quad \forall v \in Z(K), \tag{26}$$

that is $L_N v$ is the Taylor expansion of order $N$ of the function $v \in Z(K)$. We immediately have :

$$\forall v \in Z(K) \quad L_N \circ \Pi_N v = L_N v. \tag{27}$$

**Proposition 2.1.** There is a positive constant $\hat{C}$ such that :

$$|\Pi_N v|_{m,K} \leq \hat{C} k^m \left( \sum_{l=0}^{N} \frac{1}{k^l} |v|_{l,k} + \frac{h_K}{k^N} |v|_{N+1,K} + \frac{h_K^2}{k^N} |v|_{N+2,K} \right), \quad \forall v \in Z(K), \forall m \geq 0. \tag{28}$$

**Proof of Proposition 2.1.** Let $v \in Z(K)$. From Corollary 2.1 we have :

$$\Pi_N v = \sum_{q=1}^{M} \beta_q \phi_q \tag{29}$$

where

$$\beta_q = \frac{z_q}{\phi_q\left(\overrightarrow{x_G}\right)} \xi_q \tag{30}$$

and $\xi = \boldsymbol{A}^{-1} b$, with $b$ given by Eqs. (10)-(11). We deduce that :

$$|\Pi_N v|_{m,K} \leq \|\beta\| \left( \sum_{q=1}^{M} |\phi_q|_{m,K}^2 \right)^{1/2}, \quad \forall m \geq 0, \tag{31}$$

where $\|\cdot\|$ is the euclidean norm in $\mathbb{C}^M$. Since $|z_q| = 1$ and $|\phi_q\left(\overrightarrow{x_G}\right)| = 1$, we have :

$$\|\beta\| = \|\xi\| \leq \|\boldsymbol{A}\|^{-1} \|b\| \tag{32}$$

In addition, it is easy to verify that :

$$|\phi_q|_{m,K} \leq \hat{C} k^m h_K \tag{33}$$

Using Eqs. (31)-(33) and Lemma 1.1, we deduce that :

$$|\Pi_N v|_{m,K} \leq \hat{C} k^m h_K \|b\| \tag{34}$$

Moreover, using the definition of $b$ (see Eqs. (10)-(11)) we have :

$$\begin{aligned} \|b\| &= \left( |v\left(\overrightarrow{x_G}\right)|^2 + \sum_{l=1}^{N} \frac{1}{k^{2l}} \left( |\left(\partial_x + \mathrm{i}\,\partial_y\right)^l v\left(\overrightarrow{x_G}\right)|^2 + |\left(\partial_x - \mathrm{i}\,\partial_y\right)^l v\left(\overrightarrow{x_G}\right)|^2 \right) \right)^{1/2} \\ &\leq \hat{C} \left( \sum_{0\leq j+l\leq N} \frac{1}{k^{2(j+l)}} |\partial_x^j \partial_y^j v\left(\overrightarrow{x_G}\right)|^2 \right)^{1/2} \end{aligned} \tag{35}$$

Next, we use the following standard inequality :

$$|\partial_x^j \partial_y^l v\left(\overrightarrow{x_G}\right)| \leq \hat{C} \left( \frac{1}{h_K} \|\partial_x^j \partial_y^l v\|_{0,K} + |\partial_x^j \partial_y^l v|_{1,K} + h_K |\partial_x^j \partial_y^l v|_{2,K} \right) \tag{36}$$

Consequently, Eq. (35) becomes :

$$\begin{aligned} \|b\| &\leq \hat{C} \left( \sum_{0\leq l\leq N} \frac{1}{k^{2l}} \left( \frac{1}{h_K^2} |v|_{l,K}^2 + |v|_{l+1,K}^2 + h_K^2 |v|_{l+2,K}^2 \right) \right)^{1/2} \\ &\leq \hat{C} \left( \sum_{0\leq l\leq N} \frac{1}{k^{2l} h_K^2} |v|_{l,K}^2 + \frac{1}{k^{2N}} |v|_{N+1,K}^2 + \frac{h_K^2}{k^{2N}} |v|_{N+2,K}^2 \right)^{1/2} \end{aligned} \tag{37}$$

Last, we use Eqs. (34)-(37) to conclude the proof. ■

For any $j, l \geq 0$ we put :

$$\mathcal{L}_{j,l}\left(\overrightarrow{x}\right) = \frac{1}{j!l!} \left(x - x_G\right)^j \left(y - y_G\right)^l \tag{38}$$

We have, for any $0 \leq m \leq j + l$ :

$$|\mathcal{L}_{j,l}| \leq \hat{C} h_K^{1+j+l-m}. \tag{39}$$

**Proposition 2.2.** There is a positive constant $\hat{C}$ such that :

$$|L_N v|_{m,K} \leq \hat{C} \sum_{l=0}^{N+2} h_K^{l-m} |v|_{l,K}, \quad \forall v \in Z\left(K\right), \forall 0 \leq m \leq N \tag{40}$$

**Proof of Proposition 2.2.** Let $v \in Z\left(K\right)$. From Eqs. (26) and (38), we deduce that :

$$L_N v = \sum_{0\leq j+l\leq N} \mathcal{L}_{j,l} \partial_x^j \partial_y^l v\left(\overrightarrow{x_G}\right) \tag{41}$$

Using Eq. (39), we have :

$$|L_N v| \leq \hat{C} \sum_{0\leq j+l\leq N} h_K^{j+l+1-m} |\partial_x^j \partial_y^l v\left(\overrightarrow{x_G}\right)| \tag{42}$$

Moreover, from Eq. (36), we obtain :

$$\begin{aligned} |L_N v| &\leq \hat{C} \sum_{0\leq j+l\leq N} \left( h_K^{j+l-m}|v|_{j+l,K} + h_K^{j+l-m+1}|v|_{j+l+1,K} + h_K^{j+l-m+2}|v|_{j+l+2,K} \right) \\ &\leq \hat{C} \sum_{0\leq l\leq N+2} h_K^{l-m}|v|_{l,K} \end{aligned} \tag{43}$$

which concludes the proof. ■

**Lemma 2.4.** There is a positive constant $\hat{C}$ such that :

$$|w - L_N w|_{m,K} \leq \hat{C} h_K^{N+1-m} \left( |w|_{N+1,K} + h_K |w|_{N+2,K} \right), \quad \forall w \in H^{N+2}(K), \forall 0 \leq m \leq N \tag{44}$$

**Proof of Lemma 2.4.** Let $w \in H^{N+2}(K)$. We have :

$$w - L_N w = w - w_h - L_N (w - w_h) \quad \forall w_h \in P_N(K) \tag{45}$$

Let $0 \leq m \leq N$. From Proposition 2.2 we deduce that :

$$|L_N (w - w_h)|_{m,K} \leq \hat{C} \sum_{l=0}^{N+2} h_K^{l-m}|w - w_h|_{l,K} \tag{46}$$

which implies :

$$|w - L_N w|_{m,K} \leq \hat{C} \sum_{l=0}^{N+2} h_K^{l-m}|w - w_h|_{l,K} \quad \forall w_h \in P_N(K) \tag{47}$$

By taking for $w_h$ the Lagrange interpolation we obtain (see [3]) :

$$|w - w_h|_{l,K} \leq \hat{C} h_K^{N+1-l}|w|_{N+1,K} \quad \text{for } 0 \leq l \leq N+1 \tag{48}$$

and

$$|w - w_h|_{N+2,K} = |w|_{N+2,K} \tag{49}$$

Therefore, using Eqs. (48)-(49) in Eq. (47) leads to :

$$\begin{aligned} |w - L_N w|_{m,K} &\leq \hat{C} \sum_{l=0}^{N+1} h_K^{l-m} h_K^{N+1-l}|w|_{N+1,K} + h_K^{N+2-m}|w|_{N+2,K} \\ &\leq \hat{C} h_K^{N+1-m} \left( |w|_{N+1,K} + h_K |w|_{N+2,K} \right) \quad \blacksquare \end{aligned} \tag{50}$$

**Proposition 2.3.** Assume $kh$ to be sufficiently small. Then, there is a positive constant $\hat{C}$ such that :

$$|v - \Pi_N v|_{m,K} \leq \hat{C} h_K^{N+1-m} \left( \sum_{l=0}^{N} k^{N+1-l}|v|_{l,K} + |v|_{N+1,K} + h_K |v|_{N+2,K} \right), \quad \forall v \in Z(K), \forall 0 \leq m \leq N \tag{51}$$

**Proof of Proposition 2.3.** Let $v \in Z(K)$. We have, for any $0 \leq m \leq N$ :

$$\begin{aligned} |v - \Pi_N v|_{m,K} &= |v - L_N v + L_N \circ \Pi_N v - \Pi_N v|_{m,K} \\ &\leq |v - L_N v|_{m,K} + |\Pi_N v - L_N \circ \Pi_N v|_{m,K} \end{aligned} \tag{52}$$

From Lemma 2.4 we have :

$$|v - L_N v|_{m,K} \leq \hat{C} h_K^{N+1-m} \left( |v|_{N+1,K} + h_K |v|_{N+2,K} \right) \tag{53}$$

and

$$|\Pi_N v - L_N \circ \Pi_N v|_{m,K} \leq \hat{C} h_K^{N+1-m} \left( |\Pi_N v|_{N+1,K} + h_K |\Pi_N v|_{N+2,K} \right) \tag{54}$$

We have also from Proposition 2.1 :

$$|\Pi_N v|_{N+1,K} \leq \hat{C} \sum_{l=0}^{N} k^{N+1-l} |v|_{l,k} + k h_K |v|_{N+1,K} + k h_K^2 |v|_{N+2,K} \tag{55}$$

and

$$|\Pi_N v|_{N+2,K} \leq \hat{C} k \sum_{l=0}^{N} k^{N+1-l} |v|_{l,k} + k h_K |v|_{N+1,K} + k h_K^2 |v|_{N+2,K} \tag{56}$$

Combining Eqs. (54)-(56) leads to :

$$|\Pi_N v - L_N \circ \Pi_N v|_{m,K} \leq \hat{C} h_K^{N+1-m} \left( \sum_{l=0}^{N} k^{N+1-l} |v|_{l,K} + k h_K |v|_{N+1,K} + k h_K^2 |v|_{N+2,K} \right) \tag{57}$$

Last, from Eqs. (52), (53) and (57) we obtain :

$$|v - \Pi_N v|_{m,K} \leq \hat{C} h_K^{N+1-m} \left( \sum_{l=0}^{N} k^{N+1-l} |v|_{l,K} + |v|_{N+1,K} + h_K |v|_{N+2,K} \right) \tag{58}$$

which concludes the proof. ■

# Appendix 3

We prove in this section the inverse inequalities given by Eqs. (44)-(45) in Appendix 1. First, we specify the notations and recall some fundamental inverse inequalities.

## Notations and properties

- $K$ is an element of the triangulation $\tau_h$.
- $\hat{K}$ is the element of reference of $\tau_h$.
- The mapping

$$F_K : \begin{array}{l} \hat{K} \longrightarrow K \\ \hat{x} \longmapsto x = F_K(\hat{x}) \end{array} \tag{1}$$

is a diffeomorphism.
- For any function defined on $K$, we set :

$$\hat{v}(\hat{x}) = v \circ F_K(\hat{x}), \quad \hat{x} \in \hat{K}. \tag{2}$$

For $s \in \mathbb{R}^+$, we recall the following basic inverse inequalities (reference [3] in the Bibliography of Part I) :

$$|v|_{s,K} \leq \hat{C} h_K^{1-s} |\hat{v}|_{s,\hat{K}}, \quad \forall v \in H^s(K) \tag{3}$$

and

$$|\hat{v}|_{s,\hat{K}} \leq \hat{C} h_K^{s-1} |v|_{s,K}, \quad \forall v \in H^s(K), \tag{4}$$

where $\hat{C}$ is a positive constant ($\hat{C}$ does not depend on $h_K$). In addition, we have :

$$\|v\|_{0,\partial K} \leq \hat{C} h_K^{1/2} \|\hat{v}\|_{0,\partial\hat{K}}, \quad \forall v \in L^2(\partial K). \tag{5}$$

## The inverse inequalities

First, we prove the estimate (44) in Appendix 1.

**Proposition 3.1.** There is a positive constant $\hat{C}$ ($\hat{C}$ does not depend on $h_K$) such that :

$$\|v\|_{0,\partial K} \leq \hat{C} \left( h_K^{1/2} |v|_{1,K} + \frac{1}{h_K^{1/2}} \|v\|_{0,K} \right), \quad \forall v \in H^1(K). \tag{6}$$

**Proof of Proposition 3.1.** We use the inverse inequality given by (5). We have :

$$\|v\|_{0,\partial K} \leq \hat{C} h_K^{1/2} \|\hat{v}\|_{0,\partial \hat{K}}, \quad \forall v \in H^1(K). \tag{7}$$

Using the trace theorem, we deduce :

$$\|\hat{v}\|_{0,\partial \hat{K}} \leq \hat{C} \|\hat{v}\|_{1,\hat{K}}, \quad \forall v \in H^1(K). \tag{8}$$

From Eqs. (7)-(8) and using the definition of the norm $\|\cdot\|_{1,\hat{K}}$ we have :

$$\|v\|_{0,\partial K} \leq \hat{C} h_K^{1/2} \left( |\hat{v}|_{1,\hat{K}} + \|\hat{v}\|_{0,\hat{K}} \right), \quad \forall v \in H^1(K). \tag{9}$$

Next, we use the inverse inequality given by Eq. (4) with $s = 1$ and $s = 0$. We then obtain :

$$\|v\|_{0,\partial K} \leq \hat{C} h_K^{1/2} \left( |v|_{1,K} + \frac{1}{h_K} \|v\|_{0,K} \right), \quad \forall v \in H^1(K), \tag{10}$$

which concludes the proof. ■

Next, we prove the estimate (45), which is a direct application of the following Proposition.

**Proposition 3.2.** Let $\varepsilon > 0$. Then, there is a positive constant $\hat{C}$ ($\hat{C}$ does not depend on $h_K$) such that :

$$\|v\|_{0,\partial K} \leq \hat{C} \left( h_K^{\varepsilon} |v|_{1/2+\varepsilon,K} + \frac{1}{h_K^{1/2}} \|v\|_{0,K} \right), \quad \forall v \in H^{1/2+\varepsilon}(K). \tag{11}$$

**Proof of Proposition 3.2.** We use the inverse inequality given by Eq. (5). We have :

$$\|v\|_{0,\partial K} \leq \hat{C} h_K^{1/2} \|\hat{v}\|_{0,\partial \hat{K}}, \quad \forall v \in H^{1/2+\varepsilon}(K). \tag{12}$$

On the other hand, using the trace theorem (reference [11] in the Bibliography of Part I), we have :

$$\|\hat{v}\|_{0,\partial \hat{K}} \leq \hat{C} \|\hat{v}\|_{1/2+\varepsilon,\hat{K}}, \quad \forall v \in H^{1/2+\varepsilon}(K). \tag{13}$$

Therefore, it follows from Eqs. (12)-(13) that

$$\|v\|_{0,\partial K} \leq \hat{C} h_K^{1/2} \|\hat{v}\|_{1/2+\varepsilon,\hat{K}}, \quad \forall v \in H^{1/2+\varepsilon}(K). \tag{14}$$

Thus, using the definition of the norm $\|\cdot\|_{1/2+\varepsilon,\hat{K}}$, we have :

$$\|v\|_{0,\partial K} \leq \hat{C} h_K^{1/2} \left( |\hat{v}|_{1/2+\varepsilon,\hat{K}} + \|\hat{v}\|_{0,\hat{K}} \right), \quad \forall v \in H^{1/2+\varepsilon}(K). \tag{15}$$

Next, we apply the inverse inequality given by Eq. (4) with $s = 1/2+\varepsilon$ and $s = 0$, and we substitute into Eq. (15). We then obtain :

$$\|v\|_{0,\partial K} \leq \hat{C} h_K^{1/2} \left( h_K^{\varepsilon-1/2} |v|_{1/2+\varepsilon,K} + \frac{1}{h_K} \|v\|_{0,K} \right), \quad \forall v \in H^{1/2+\varepsilon}(K), \tag{16}$$

which concludes the proof of Proposition 3.2. ■

**Corollary 3.1.** Let $\varepsilon > 0$. Then, there is a positive constant $\hat{C}$ ($\hat{C}$ does not depend on $h_K$) such that :

$$\|\partial_n v\|_{0,\partial K} \leq \hat{C} \left( h_K^{\varepsilon} |v|_{3/2+\varepsilon,K} + \frac{1}{h_K^{1/2}} |v|_{1,K} \right), \quad \forall v \in H^{3/2+\varepsilon}(K). \tag{17}$$

**Proof of Corollary 3.1.** Let $\varepsilon > 0$ and $v \in H^{3/2+\varepsilon}$. Then, $\nabla v \in \left[ H^{1/2+\varepsilon}(K) \right]^2$. In addition, we have :

$$\|\partial_n v\|_{0,\partial K} \leq \|\nabla v\|_{0,\partial K}. \tag{18}$$

We apply the inverse inequality given by Eq. (11) (see Proposition 3.1) to $\nabla v \in \left[ H^{1/2+\varepsilon}(K) \right]^2$. We have :

$$\|\partial_n v\|_{0,\partial K} \leq \hat{C} \left( h_K^{\varepsilon} |\nabla v|_{1/2+\varepsilon,K} + \frac{1}{h_K^{1/2}} \|\nabla v\|_{0,K} \right). \tag{19}$$

This concludes the proof of Corollary 3.1 since

$$|\nabla v|_{1/2+\varepsilon,K} = |v|_{3/2+\varepsilon,K} \quad \text{and} \quad \|\nabla v\|_{0,K} = |v|_{1,K}. \quad \blacksquare \tag{20}$$

# **Partie II :** A modified discontinuous Galerkin method for solving Helmholtz problems

# 1. Introduction

The Helmholtz equation belongs to the classical equations of mathematical physics that are well understood from a mathematical view point. However, the numerical approximation of the solution is still a challenging problem in spite the tremendous progress made during the last years (see, for example, the recent monograph [14] and the references therein). Indeed, the standard finite element method (FEM) is not well suited for solving Helmholtz problems in the mid- and high-frequency regime since highly oscillating solutions are not accurately approximated by piecewise polynomials. This phenomenon, related to the indefiniteness of the Helmholtz operator, is known as the pollution effect [3]. In order to maintain a certain level of accuracy, a mesh refinement is required and/or higher order FEM are used, leading to a prohibitive computational cost for high wavenumbers.

In response to this challenge, alternative techniques for alleviating the pollution effect were proposed. Numerous of these approaches use the plane waves, since they are expected to better approximate highly oscillating waves. Examples of such methods include the weak element method for Helmholtz equation [16], the Galerkin least-squares method [10], the partition of unity method [2], the residual free bubbles method [8], the least-squares method [15], the ultra-weak variational method [4] and recently, the discontinuous Galerkin method (DGM) designed by Farhat *et al* and presented in a series of papers [5, 6, 7]. In the latter method, the solution is approximated at the element mesh level using a superposition of plane waves which results in a discontinuous solution along interior boundaries of the mesh. The continuity is then restored in a weak sense through the use of Lagrange multipliers. The rectangular and quadrilateral elements constructed in [5, 6, 7] clearly outperform the standard Galerkin FEM. For example, for $ka \geq 10$ and for a fixed level of accuracy, the so-called $R$-4-1 element reduces the total number of degrees of freedom (dofs) required by the $Q1$-based finite element discretization for Helmholtz equation by a factor greater or equal to five. Similar results are obtained for the $R$-8-2a and $R$-8-2b elements when compared to the $Q2$ element, and for $Q$-16-4 and $Q$-32-8 when compared to the $Q4$ element. In spite of this impressive performance, the DGM has three important drawbacks. First, the method has to satisfy an *inf-sup* condition which is translated, in practice, as a compatibility requirement : the number of dofs of the Lagrange multiplier (corresponding to the dual variable) and of the field (the primal variable) cannot be chosen arbitrarily. The problem here is that there is no theoretical result on how to satisfy this compatibility requirement, except for the simple case of $R$-4-1 element (see [1]). Hence, for other elements, the existing choices are based on numerical experiments only. The second major issue with the DGM is that it becomes unstable as we refine the mesh. Such instabilities occur because of the singularity of the local problems and, to some extent, to the loss of the linear independence of the plane waves as the step size mesh discretization tends to zero. The latter affects dramatically the stability of the global system due to its ill-conditioning nature. Finally, the DGM exhibits a loss of accuracy for unstructured mesh [6].

We propose a new solution methodology for Helmholtz problems, that falls in the category of discontinuous Galerkin methods. The proposed formulation distinguishes itself from existing procedures by the *well-posed* character of the local problems and by the resulting global system which is associated with a positive semi-definite Hermitian matrix. More specifically, the computation domain is subdivided in quadrilateral- or triangular-shaped elements. The solution is approximated, at the element level, by a superposition of plane waves that are solution of the Helmholtz equation. The continuity of the solution at the interior interfaces of the elements is then enforced by Lagrange multipliers. Unlike the DGM, the proposed method does not require the continuity of the normal derivative. Consequently, Lagrange multipliers are introduced to restore in the weak sense the conti-

nuity of both the field and its normal derivative across interior boundaries of the mesh. Such choice leads to solving (a) local boundary value problems that are well posed in the sense of Hadamard [9] and (b) a global system, whose unknowns are the Lagrange multipliers. The Lagrange multiplier is the solution of a variational problem whose bilinear form can be written into two equivalent expressions. The approximation of both formulations leads to two linear systems corresponding to positive semi-definite matrices. Note that the proposed technique is a two-step procedure where the local problems are first solved and then the Lagrange multipliers are evaluated. This two-step approach allows us to consider equally structured and unstructured meshes with either triangular- or quadrilateral-shaped elements. Since the proposed solution methodology resembles in some aspect the DGM, we will refer to it as mDGM (*m*odified *d*iscontinuous *G*alerkin *m*ethod).

The remainder of the paper is organized as follows. In Section 2, we introduce general notations and the model problem. Section 3 is devoted to the presentation of mDGM. In Section 4, we present the algebraic framework of the formulation. We compare in Section 5 the numerical performance of both methods : DGM and mDGM. The obtained results clearly indicate that mDGM outperforms DGM in terms of stability and accuracy. Finally, Section 6 concludes this paper.

## 2. Preliminaries

In this section, we introduce the model problem and specify the nomenclature and assumptions adopted throughout this paper.

### 2.1. The mathematical model

We consider the following class of waveguide-type problems :

$$(\text{BVP}) \begin{cases} -\Delta u - k^2 u &= f & \text{in } \Omega, \\ \partial_n u &= \mathrm{i}\, k u + g & \text{on } \partial\Omega \end{cases}$$

where $\Omega \subset \mathbb{R}^2$ is an open bounded region, with smooth boundary $\partial\Omega$, $k$ is a positive number representing the wavenumber, $\partial_n$ is the normal derivative and $f$ and $g$ are regular complex valued functions defined respectively on $\Omega$ and $\partial\Omega$. The second equation of BVP is a representation of a class of non-homogeneous Robin boundary conditions, but other types of boundary condition can be considered.
Note that BVP is considered here for its simplicity since it allows us to compute analytically the solution $u$ for a suitable choice of $\Omega$, $f$ and $g$. Such an expression of $u$ is used when assessing the accuracy of mDGM.

### 2.2. Nomenclature and assumptions

In what follows, we consider a regular triangulation $\tau_h$ of $\Omega$ into quadrilateral- or triangular-shaped subdomains $K$ whose boundaries are denoted by $\partial K$. The step size mesh discretization is denoted by $h$. We introduce the space of the primal variable :

$$\mathcal{V} = \left\{ v \in L^2(\Omega)\,;\ v|_K \in H^1(K) \right\},$$

that we equip with the norm :

$$||v||_{\mathcal{V}} = \left( \sum_{K \in \tau_h} ||v^K||^2_{1,K} \right)^{\frac{1}{2}} , \quad \forall\, v \in \mathcal{V},$$

where $||\cdot||_{1,K}$ is the $H^1$-norm on the element $K$. In addition, we introduce $||\cdot||_{0,K}$ and $|\cdot|_{1,K}$ to designate the $L^2$-norm and the $H^1$-seminorm respectively on the element $K$.
Note that $\mathcal{V}$ contains functions that are discontinuous across interior boundaries since their regularity is only $L^2(\Omega)$. For any $v \in \mathcal{V}$, we define the jump across an interior edge $e = \partial K \cap \partial K'$ by :

$$[v] = v^K - v^{K'}.$$

We introduce the space of the dual variable, corresponding here to Lagrange multipliers, by :

$$\mathcal{M} = \left\{ \mu \in \prod_{K \in \tau_h} L^2(\partial K)\,;\ \mu = 0 \text{ on } \partial K \cap \partial\Omega \right\}$$

and we associate to $\mathcal{M}$ the norm given by :

$$||\mu||_{\mathcal{M}} = \left( \sum_{K \in \tau_h} ||\mu^K||^2_{0,\partial K} \right)^{\frac{1}{2}} , \quad \forall\, \mu \in \mathcal{M},$$

where $\mu^K$ designates the restriction of $\mu$ to $\partial K$ : $\mu^K = \mu|_{\partial K}$ and $||\cdot||_{0,\partial K}$ is the $L^2$-norm on $\partial K$. Moreover, for any function $\mu \in \mathcal{M}$, we define the jump across an interior edge $e = \partial K \cap \partial K'$ by :

$$[[\mu]] = \mu^K + \mu^{K'}.$$

## 3. The proposed solution methodology : The continuous approach

The basic idea of mDGM is to evaluate $u$, the solution of BVP, using the following splitting :

$$u = \Phi(\lambda) + \varphi, \tag{1}$$

where $\varphi$ and $\Phi$ are elements of $\mathcal{V}$ and $\lambda \in \mathcal{M}$.
To compute these three quantities, we proceed into two steps :

**Step 1** : For all $K \in \tau_h$ and $\mu \in \mathcal{M}$, we compute $\varphi$ and $\Phi(\mu)$. This is achieved by solving local Helmholtz problems. This step is called the restriction procedure.

**Step 2** : We determine $\lambda \in \mathcal{M}$ by solving a global linear system to ensure the continuity in a weak sense of the solution $u$ given by Eq. (1) and of the normal derivative of $u$. This step is called the optimization procedure.

### 3.1. Step 1 : The restriction procedure

As stated earlier, this step is devoted to the computation of $\varphi$ and $\Phi(\mu)$, for all $\mu \in \mathcal{M}$, by solving locally Helmholtz problems. More specifically, for all $K \in \tau_h$, we compute $\varphi^K$ by solving

the following boundary value problem :

$$\text{(BVP1)} \left\{ \begin{array}{rll} \text{Find } \varphi^K \in H^1(K) & \text{such that :} & \\ -\Delta\varphi^K - k^2\varphi^K & = f & \text{in } K \\ \partial_n\varphi^K & = \mathrm{i}\, k\varphi^K + g & \text{on } \partial K \cap \partial\Omega \\ \partial_n\varphi^K & = \mathrm{i}\, \alpha\varphi^K & \text{on } \partial K \cap \mathring{\Omega} \end{array} \right. .$$

Next, for all $\mu \in \mathcal{M}$ and $K \in \tau_h$, we compute $\Phi\left(\mu^K\right)$ by solving the boundary value problem given by :

$$\text{(BVP2)} \left\{ \begin{array}{rll} \text{Find } \Phi\left(\mu^K\right) \in H^1(K) & \text{such that :} & \\ -\Delta\Phi\left(\mu^K\right) - k^2\Phi\left(\mu^K\right) & = 0 & \text{in } K \\ \partial_n\Phi\left(\mu^K\right) & = \mathrm{i}\, k\Phi\left(\mu^K\right) & \text{on } \partial K \cap \partial\Omega \\ \partial_n\Phi\left(\mu^K\right) & = \mathrm{i}\, \alpha\Phi\left(\mu^K\right) + \mu^K & \text{on } \partial K \cap \mathring{\Omega} \end{array} \right. ,$$

with $\alpha \in \mathbb{R}_+^*$. Note that the presence of $\alpha$ ensures the uniqueness of the solution of BVP 1 and BVP 2, as it will be shown later.

It is easy to verify that the variational formulation of both problems can be expressed in a compact form as follows :

$$\left\{ \begin{array}{l} \text{Find } \Psi^K \in H^1(K) \text{ such that :} \\ a_K\left(\Psi^K, v^K\right) = L_K\left(v^K\right) \qquad \forall v^K \in H^1(K) \end{array} \right. \tag{2}$$

where $a_K(\cdot,\cdot)$ is a bilinear form given by :

$$\begin{array}{rcl} a_K\left(v^K, w^K\right) & = & \displaystyle\int_K \nabla v^K \cdot \nabla\overline{w^K} dx - k^2 \int_K v^K\overline{w^K} dx - \mathrm{i}\,\alpha \int_{\partial K\cap\mathring{\Omega}} v^K\overline{w^K} ds \\ & & \displaystyle -\mathrm{i}\, k \int_{\partial K\cap\partial\Omega} v^K\overline{w^K} ds, \quad \forall\, v^K, w^K \in H^1(K) \end{array} \tag{3}$$

and $\Psi_K$ is either the solution of BVP1 or BVP2, i.e. :

$$\Psi^K = \left\{ \begin{array}{ll} \varphi^K & \text{for BVP1} \\ \Phi\left(\mu^K\right) & \text{for BVP2}, \end{array} \right. \quad \forall\, \mu \in \mathcal{M}.$$

The right-hand side $L_K(\cdot)$ is given by :

$$L_K\left(v^K\right) = \left\{ \begin{array}{ll} \displaystyle\int_K f\overline{v^K} dx + \int_{\partial K\cap\partial\Omega} g\overline{v^K} ds & \text{for BVP1} \\ \displaystyle\int_{\partial K\cap\mathring{\Omega}} \mu^K\overline{v^K} ds & \text{for BVP2} \end{array} \right. , \forall\, v^K \in H^1(K). \tag{4}$$

Consequently, the solving of BVP1 and BVP2 requires solving *one* linear system with multiple right-hand side.
Note that the bilinear form $a_K$ is neither Hermitian, nor symmetric. However, it is easy to check that $a_K(\cdot,\cdot)$ is continuous on $H^1(K) \times H^1(K)$ and satisfies the Gärding inequality in $H^1(K)$ since

$$\Re a_K\left(v^K, v^K\right) + k^2 \|v^K\|_{0,K}^2 = |v|_{1,K}^2, \tag{5}$$

where $\Re$ designates the real part. In addition, we have :

**Proposition 1.** For a fixed $K \in \tau_h$, the variational problem given by Eq. (2) admits a unique solution.

**Proof of Proposition 1.** Let $K$ be a fixed element of $\tau_h$. From Eq. (5), it follows that the bilinear form $a_K(\cdot,\cdot)$ satisfies the Fredholm alternative. Hence, the uniqueness ensures the existence of $\Psi^K \in H^1(K)$, solution of the variational problem given by Eq. (14).
To prove the uniqueness, we consider the homogeneous problem associated to the bilinear form $a_K(\cdot,\cdot)$ and let $w^K$ be its solution. We therefore have :

$$\begin{cases} \text{Find } w^K \in H^1(K) \text{ such that :} \\ a_K\left(w^K, v^K\right) = 0 \qquad\qquad \forall\, v^K \in H^1(K) \end{cases} \tag{6}$$

In particular, for $v^K = w^K$, we have :

$$\alpha \|w^K\|^2_{0,\partial K \cap \mathring{\Omega}} + k\|w^K\|^2_{0,\partial K \cap \partial\Omega} = 0.$$

Since $\alpha > 0$, we must have :

$$w^K = 0 \text{ on } \partial K \quad \text{ and } \quad \partial_n w^K = 0 \text{ on } \partial K.$$

Therefore, using the continuation theorem [12, 17], we deduce that $w^K = 0$ in $K$ and the problem given by Eq. (2) has a unique solution. ■

**Remark 1.** The presence of the Robin condition on $\partial K \cap \mathring{\Omega}$ with $\alpha > 0$ is *crucial* to ensure that 0 is the only solution of the variational problem given by Eq. (6). Indeed, if $\alpha = 0$ and $K$ satisfies $\partial K \cap \partial\Omega = \emptyset$, then the resulting homogeneous Neumann boundary condition on $\partial K$ is not sufficient to guarantee the uniqueness of the solution of the problem given by Eq. (6), since $k^2$ may become an interior eigenvalue.

Next, we define $\varphi$ such that, for all element $K$ in the mesh, the restriction of $\varphi$ to $K$ is $\varphi^K$, the solution of BVP1, i.e. $\varphi|_K = \varphi^K$. Similarly, for all element $K$ and for all $\mu$ in $\mathcal{M}$, we define $\Phi(\mu)$ such that we have $\Phi(\mu)|_K = \Phi\left(\mu^K\right)$, where $\Phi\left(\mu^K\right)$ is the solution of BVP2. Using the definition of $\varphi$ and $\Phi(\mu), \forall \mu \in \mathcal{M}$ we have :

$$\varphi \in \mathcal{V} \quad \text{and} \quad \Phi(\mu) \in \mathcal{V}, \quad \forall \mu \in \mathcal{M} \tag{7}$$

In summary, Step 1 allows us to compute, for all $\mu$ in $\mathcal{M}$ :

$$\varphi + \Phi(\mu) \in \mathcal{V} \tag{8}$$

by solving one variational problem given by Eq. (2) with different right-hand side given by Eq. (4). Step 1 can be viewed, to some extent, as a prediction step.

### 3.2. Step 2 : The optimization procedure

The objective here is to determine $\lambda \in \mathcal{M}$ for which the function given by Eq. (8) is in $H^1(\Omega)$. This requirement can be viewed as a correction stage since we select the best-fit Lagrange multiplier $\lambda$.

The determination of $\lambda$ is accomplished by solving the following global variational problem :

$$(\mathrm{VF}) \begin{cases} \text{Find } \lambda \in \mathcal{M} \text{ such that} \\ A(\lambda, \mu) = F(\mu), \quad \forall\, \mu \in \mathcal{M}, \end{cases} \tag{9}$$

where the bilinear form $A(\cdot,\cdot)$ is given by :

$$\begin{aligned} A(\eta,\mu) = & \sum_{e\text{-interior edge}} \beta_e \int_e [\Phi(\eta)]\, \overline{[\Phi(\mu)]} ds \\ & + \sum_{\substack{e\text{-interior edge} \\ e=\partial K \cap \partial K'}} \gamma_e \int_e \left(\eta^K + \eta^{K'} + \mathrm{i}\,\alpha \left(\Phi\left(\eta^K\right) + \Phi\left(\eta^{K'}\right)\right)\right) \\ & \qquad \left(\overline{\mu^K + \mu^{K'} + \mathrm{i}\,\alpha\left(\Phi\left(\mu^K\right) + \Phi\left(\mu^{K'}\right)\right)}\right) ds \end{aligned} \tag{10}$$

and the linear form $F(\cdot)$ is given by :

$$\begin{aligned} F(\mu) = & - \sum_{e\text{-interior edge}} \beta_e \int_e [\varphi]\, \overline{[\Phi(\mu)]} ds \\ & -\mathrm{i}\,\alpha \sum_{\substack{e\text{-interior edge} \\ e=\partial K \cap \partial K'}} \gamma_e \int_e \left(\varphi^K + \varphi^{K'}\right) \\ & \qquad \left(\overline{\mu^K + \mu^{K'} + \mathrm{i}\,\alpha\left(\Phi\left(\mu^K\right) + \Phi\left(\mu^{K'}\right)\right)}\right) ds. \end{aligned} \tag{11}$$

The parameters $\beta_e$ and $\gamma_e$ are two positive numbers that can be viewed as weight parameters. This problem expresses the continuity in the weak sense of the solution and its normal derivative. Note that the bilinear form $A$ is Hermitian. Consequently, only half of the corresponding matrix will be stored.

**Remark 2.** Unlike the DGM, where only the primal variable $u$ is discontinuous, the mDGM leads to the discontinuity of both variables : the primal variable $u$ and the Lagrange multiplier $\lambda$, the dual variable. Consequently, the normal derivative of $u$ is *discontinuous*.
Alternatively, for numerical approximation purpose, we rewrite Eq. (10) as follows :

$$\begin{aligned} A(\eta,\mu) = & \sum_{e\text{-interior edge}} \beta_e \int_e [\Phi(\eta)]\, \overline{[\Phi(\mu)]} ds \\ & + \sum_{e\text{-interior edge}} \gamma_e \int_e [[\partial_n \Phi(\eta)]]\, \overline{[[\partial_n \Phi(\mu)]]} ds \\ & + \sum_{e \subset \partial\Omega} \omega_e \int_e (\partial_n \Phi(\eta) - \mathrm{i}\,k\Phi(\eta))\, \overline{(\partial_n \Phi(\mu) - \mathrm{i}\,k\Phi(\mu))} ds, \end{aligned} \tag{12}$$

where the weight parameter $\omega_e$ is a positive number. Note that the second integral in Eq. (12) is equal to the second integral in Eq. (10), whereas the third integral in Eq. (12) is in fact equal to 0. Consequently, the right-hand side given by Eq. (11) is modified depending on the use of Eq. (10) or Eq. (12).
The next result states the equivalence between solving BVP and solving the problem arising in the proposed two-step procedure.

**Theorem 1.**
(i) Let $u = \Phi(\lambda) + \varphi$, where for all $K$, $\varphi^K$ is solution of BVP1 and $\Phi\left(\lambda^K\right)$ is solution of BVP2 with $\lambda$ solution of VF. Then $u$ is the unique solution of BVP.
(ii) Conversely, let $u$ be the solution of BVP. For each $K \in \tau_h$, we define $\lambda$ by :

$$\lambda^K = \begin{cases} 0 & \text{on} \quad e \subset \partial K \cap \partial\Omega \\ \partial_n u^K - \mathrm{i}\,\alpha u^K & \text{on} \quad e \subset \partial K \cap \mathring{\Omega} \end{cases} \tag{13}$$

Let $\varphi^K$ be the solution of BVP1 and $\Phi(\lambda^K)$ the solution of BVP2. Then $\lambda$ is solution of VF and $u = \Phi(\lambda) + \varphi$.

**Remark 3.** Eq. (13) indicates a clear distinction between mDGM and DGM, in which $\lambda^K = \partial_n u^K$ and is continuous along the interior edges.

## 4. The proposed solution methodology : The algebraic approach

The implementation of mDGM requires first to introduce two finite-dimensional spaces $\mathcal{V}_h$ and $\mathcal{M}_h$ such that $\mathcal{V}_h \subset \mathcal{V}$ and $\mathcal{M}_h \subset \mathcal{M}$. Similarly to the DGM formulation, we have considered in this paper spaces of plane waves functions. However, other shape functions satisfying the Helmholtz equation can also be considered. Moreover, unlike the DGM, mDGM allows - in principle - to choose the spaces $\mathcal{V}_h$ and $\mathcal{M}_h$ independently.

For any element $K \in \tau_h$, we denote by $\mathcal{V}_h(K)$ (resp. $\mathcal{M}_h(K)$) the set of functions of $\mathcal{V}_h$ (resp. $\mathcal{M}_h$) restricted to $K$ (resp. $\partial K$). Furthermore, $n^K$ (resp. $n^{\lambda^K}$) denotes the dimension of $\mathcal{V}_h(K)$ (resp. $\mathcal{M}_h(K)$). Last, the dimension of $\mathcal{M}_h$, which corresponds to the total number of dofs, is denoted by $n^\lambda$.

We show that when formulated in finite dimensional spaces, the proposed two-step procedure consists in solving linear algebraic systems in each step. Note that in Step 2, the resulting linear system to be solved depends on the approximation of the continuous formulations given by Eq. (10) and Eq. (12) respectively. As stated earlier, Eqs. (10) and (12) are equivalent only at the continuous level. At the discontinuous level, the second and the third equation of BVP2 are satisfied in the weak sense.

### 4.1. Step 1 : The restriction procedure

For an element $K \in \tau_h$ and for any $\mu_h^K \in \mathcal{M}_h(K)$, we denote by $\varphi_h^K \in \mathcal{V}_h(K)$ and $\Phi_h\left(\mu_h^K\right) \in \mathcal{V}_h(K)$ the approximation of $\varphi^K$ and $\Phi\left(\mu_h^K\right)$ respectively. Similarly to the continuous formulation, $\varphi_h$, $\Phi_h(\mu_h)$ and $\mu_h$ are given by : $\varphi_h|_K = \varphi_h^K$, $\Phi_h(\mu_h)|_K = \Phi_h\left(\mu_h^K\right)$ and $\mu_h|_K = \mu_h^K$, for any element $K$ in the mesh.
To compute $\varphi_h$ and $\Phi_h(\mu_h)$, for all $K \in \tau_h$, we set the variational problem given by Eq. (2) in the finite dimensional space $\mathcal{V}_h(K)$, that is :

$$\begin{cases} \text{Find } \Psi_h^K \in \mathcal{V}_h(K) \text{ such that :} \\ a_K\left(\Psi_h^K, v_h^K\right) = L_K\left(v_h^K\right), \quad \forall\, v_h^K \in \mathcal{V}_h(K) \end{cases} \tag{14}$$

where the forms $a_K(\cdot,\cdot)$ and $L_K(\cdot)$ are given by Eq. (3) and Eq. (4) respectively, and

$$\Psi_h^K = \begin{cases} \varphi_h^K & \text{for BVP1} \\ \Phi_h\left(\mu_h^K\right) & \text{for BVP2,} \quad \forall\, \mu_h \in \mathcal{M}_h. \end{cases} \tag{15}$$

Consequently, the variational problem given by Eqs. (14)-(15) can be written in the following matrix form :

$$\left(\mathbf{K}^K - k^2\mathbf{M}^K - \mathrm{i}\,\alpha\mathbf{S}^{\partial K\cap\mathring{\Omega}} - \mathrm{i}\,k\mathbf{S}^{\partial K\cap\partial\Omega}\right)\mathbf{X}^K = \mathrm{rhs}, \tag{16}$$

where $\mathbf{K}^K$ (resp. $\mathbf{M}^K$) is the stiffness (resp. mass) matrix at the element level $K$. $\mathbf{S}^{\partial K\cap\mathring{\Omega}}$ and $\mathbf{S}^{\partial K\cap\partial\Omega}$ are mass-like matrices defined on $\partial K\cap\mathring{\Omega}$ and $\partial K\cap\partial\Omega$ respectively. $\mathbf{X}^K$ is the vector in $\mathbb{C}^{n^K}$ whose components are the values of $\Psi_h^K$ in the basis of $\mathcal{V}_h(K)$.
The linear system given by Eq. (16) possesses the following properties :

- All the entries of the corresponding matrix can be evaluated analytically for plane waves shape functions.
- The linear system admits a unique solution, even when $\partial K\cap\partial\Omega = \emptyset$. Thanks to the positive number $\alpha$ since the presence of the matrix $\mathbf{S}^{\partial K\cap\mathring{\Omega}}$ guarantees the invertibility of the system. Note that this is not the case for the DGM, for which $\mathbf{S}^{\partial K\cap\mathring{\Omega}}$ does not appear, leading to possibly a (weakly) singular system when $\partial K\cap\partial\Omega = \emptyset$.
- The corresponding matrix is neither Hermitian, nor symmetric. This cannot be viewed as a deficiency of the approach since the size of the system is *small* and thus can be solved easily using LU factorization. More specifically, the size of the system is $n^K \times n^K$, where $n^K$ (the number of shape functions at the element level) does not exceed few hundreds.
- For an element $K \in \tau_h$, the number of right-hand side is $n^{\lambda^K} + 1$. We must point out that the obtained problems can be solved in parallel since they are independent from an element $K$ to another.

### 4.2. Step 2 : The optimization procedure

In this step, we set the global problem VF in finite dimension. We have :

$$\begin{cases} \text{Find } \lambda_h \in \mathcal{M}_h \text{ such that :} \\ A_h(\lambda_h,\mu_h) = F_h(\mu_h), \quad \forall\, \mu_h \in \mathcal{M}_h \end{cases} \tag{17}$$

where the forms $A_h(\cdot,\cdot)$ and $F_h(\cdot)$ are obtained from $A(\cdot,\cdot)$ and $F(\cdot)$ respectively by replacing $\varphi$ with $\varphi_h$ and $\Phi(\mu_h)$ with $\Phi_h(\mu_h)$, for $\mu_h \in \mathcal{M}_h$. Hence, solving the variational problem given by Eq. (17) comes to solve the following linear algebraic system :

$$\mathbf{A}\boldsymbol{\Lambda} = \mathbf{b}, \tag{18}$$

where the entries of the matrix $\mathbf{A}$ and of the vector $\mathbf{b}$ depend on the expression of the continuous bilinear form $A$. More specifically, when using the bilinear form given by Eq. (10), the entries of the

matrix $\mathbf{A}$ and of the vector $\mathbf{b}$ are given by :

$$\begin{aligned}\mathbf{A}_{lm} = & \sum_{e\text{-interior edge}} \beta_e \int_e [\Phi_h(\mu_m)] \left[\overline{\Phi_h(\mu_l)}\right] ds \\ & + \sum_{\substack{e\text{-interior edge} \\ e=\partial K\cap\partial K'}} \gamma_e \int_e \left(\mu_m^K + \mu_m^{K'} + \mathrm{i}\,\alpha\left(\Phi_h\left(\mu_m^K\right) + \Phi_h\left(\mu_m^{K'}\right)\right)\right) \\ & \qquad \left(\overline{\mu_l^K + \mu_l^{K'} + \mathrm{i}\,\alpha(\Phi_h\left(\mu_l^K\right) + \Phi_h\left(\mu_l^{K'}\right)}\right) ds\end{aligned} \tag{19}$$

and

$$\begin{aligned}\mathbf{b}_l = & -\sum_{e\text{-interior edge}} \beta_e \int_e [\varphi_h] \left[\overline{\Phi_h(\mu_l)}\right] ds \\ & -\mathrm{i}\,\alpha \sum_{\substack{e\text{-interior edge} \\ e=\partial K\cap\partial K'}} \gamma_e \int_e \left(\varphi_h^K + \varphi_h^{K'}\right) \\ & \qquad \left(\overline{\mu_l^K + \mu_l^{K'} + \mathrm{i}\,\alpha(\Phi_h\left(\mu_l^K\right) + \Phi_h\left(\mu_l^{K'}\right))}\right) ds\end{aligned} \tag{20}$$

for $1 \le l, m \le n^\lambda$.
On the other hand, when the bilinear form $A$ is given by Eq. (12), the entries of the corresponding matrix, denoted by $\widetilde{\mathbf{A}}$, are defined by :

$$\begin{aligned}\widetilde{\mathbf{A}}_{lm} = & \sum_{e\text{-interior edge}} \beta_e \int_e [\Phi_h(\mu_m)]\,\overline{[\Phi_h(\mu_l)]} ds \\ & + \sum_{e\text{-interior edge}} \gamma_e \int_e [[\partial_n\Phi_h(\mu_m)]]\,\overline{[[\partial_n\Phi_h(\mu_l)]]} ds \\ & + \sum_{e\subset\partial\Omega} \omega_e \int_e (\partial_n\Phi_h(\mu_m) - \mathrm{i}\,k\Phi_h(\mu_m))\,\overline{(\partial_n\Phi_h(\mu_l) - \mathrm{i}\,k\Phi_h(\mu_l))} ds,\end{aligned} \tag{21}$$

and consequently, the right-hand side, denoted by $\widetilde{\mathbf{b}}$, is given by :

$$\begin{aligned}\widetilde{\mathbf{b}}_l = & -\sum_{e\text{-interior edge}} \beta_e \int_e [\varphi_h]\,\overline{[\Phi_h(\mu_l)]} ds \\ & - \sum_{e\text{-interior edge}} \gamma_e \int_e [[\partial_n\varphi_h]]\overline{[[\partial_n\Phi_h(\mu_l)]]} ds \\ & - \sum_{e\subset\partial\Omega} \omega_e \int_e (\partial_n\varphi_h - \mathrm{i}\,k\varphi_h - g)\,\overline{(\partial_n\Phi_h(\mu_l) - \mathrm{i}\,k\Phi_h(\mu_l))} ds\end{aligned} \tag{22}$$

for $1 \le l, m \le n^\lambda$.
The unknown $\mathbf{\Lambda}$ is a vector in $\mathbb{C}^{n^\lambda}$ whose components are the values of $\lambda_h$ in the basis of $\mathcal{M}_h$. Hence, from a numerical point of view, two approaches are possible at Step 2 for determining the Lagrange multiplier :

- Approach 1 : solving the linear system given by Eqs. (19)-(20).
- Approach 2 : solving the linear system given by Eqs. (21)-(22).

Note that the matrices $\mathbf{A}$ and $\widetilde{\mathbf{A}}$ are both *Hermitian*. Next, we show that the systems in Approach 1 and 2 are positive semi-definite. Assuming a compatibility condition satisfied, the systems are positive definite.

**Proposition 3.**
(i) The matrix $\mathbf{A}$ is positive semi-definite. Moreover, for $\Phi_h\left(\mu_h^K\right)$ satisfying, in the *strong* sense, the following equations :

$$\forall K \in \tau_h,\, \forall \mu_h \in \mathcal{M}_h, \quad \partial_n \Phi_h\left(\mu_h^K\right) = \mathrm{i}\,\alpha \Phi_h\left(\mu_h^K\right) + \mu_h^K \text{ on } \partial K \cap \mathring{\Omega}, \tag{23}$$

and

$$\forall K \in \tau_h,\, \forall \mu_h \in \mathcal{M}_h, \quad \partial_n \Phi_h\left(\mu_h^K\right) = \mathrm{i}\,k \Phi_h\left(\mu_h^K\right) \text{ on } \partial K \cap \partial\Omega, \tag{24}$$

$\mathbf{A}$ is a positive definite matrix.
(ii) The matrix $\widetilde{\mathbf{A}}$ is positive semi-definite. Moreover, if the spaces $\mathcal{V}_h$ and $\mathcal{M}_h$ satisfy the following condition :

$$\int_{\partial K} \mu_h^K v_h^K \, ds = 0\,, \quad \forall\, v_h^K \in \mathcal{V}_h\left(K\right) \quad \Longrightarrow \mu_h^K = 0, \tag{25}$$

then the matrix $\widetilde{\mathbf{A}}$ is positive definite.

**Proof of Proposition 3.** Let $\{\mu_j\}_{1\leq j\leq n^\lambda}$ be a basis of $\mathcal{M}_h$ and let $\boldsymbol{y} = {}^t[y_1, y_2, ..., y_{n^\lambda}] \in \mathbb{C}^{n^\lambda}$ be an ordinary vector. We set :

$$\eta_h = \sum_{1\leq l\leq n^\lambda} y_l \mu_l. \tag{26}$$

Therefore, it is easy to verify that :

$$\Phi_h\left(\eta_h\right) = \sum_{1\leq l\leq n^\lambda} y_l \Phi_h\left(\mu_l\right). \tag{27}$$

Consequently, in each $K \in \tau_h$, $\Phi_h\left(\eta_h\right)$ satisfies :

$$a_K\left(\Phi_h\left(\eta_h^K\right), v_h^K\right) = \int_{\partial K\cap\mathring{\Omega}} \eta_h^K \overline{v_h^K} ds \quad \forall v_h^K \in \mathcal{V}_h\left(K\right). \tag{28}$$

(i) Using the definition of $\eta_h$, we have :

$$\boldsymbol{y}^* \mathbf{A} \boldsymbol{y} = \sum_{e-\text{interior edge}} \left(\beta_e \| \left[\Phi_h\left(\eta_h\right)\right] \|^2_{L^2(e)} + \gamma_e \| \left[\left[\eta_h + \mathrm{i}\,\alpha \Phi_h\left(\eta_h\right)\right]\right] \|^2_{L^2(e)}\right). \tag{29}$$

Hence,

$$\boldsymbol{y}^* \mathbf{A} \boldsymbol{y} \geq 0, \quad \forall \boldsymbol{y} \in \mathbb{C}^{n^\lambda}$$

that is $\mathbf{A}$ is a positive semi-definite matrix.
Next, we assume that the condition given by Eq. (23) is satisfied. Let $\boldsymbol{y}$ be a vector in $\mathbb{C}^{n^\lambda}$ such

that $\boldsymbol{y}^*\mathbf{A}\boldsymbol{y} = 0$. Then, it follows from Eq. (29) that :

$$\sum_{e-\text{interior edge}} \left(\beta_e \| \left[\Phi_h\left(\eta_h\right)\right] \|^2_{L^2(e)} + \gamma_e \| \left[\left[\eta_h + \mathrm{i}\,\alpha\Phi_h\left(\eta_h\right)\right]\right] \|^2_{L^2(e)}\right) = 0. \tag{30}$$

Since for any interior edge $e$, $\beta_e > 0$ and $\gamma_e > 0$, we have :

$$\| \left[\Phi_h\left(\eta_h\right)\right] \|_{L^2(e)} = 0 \text{ and } \| \left[\left[\eta_h + \mathrm{i}\,\alpha\Phi_h\left(\eta_h\right)\right]\right] \|_{L^2(e)} = 0 \text{ on all interior edges.} \tag{31}$$

Therefore,

$$\left[\Phi_h\left(\eta_h\right)\right] = 0 \quad \text{on all interior edges}$$

and, using Eq. (23), we have also :

$$\left[\left[\partial_n\Phi_h\left(\eta_h\right)\right]\right] = 0 \quad \text{on all interior edges.}$$

Consequently, $\Phi_h\left(\eta_h\right) \in H^1\left(\Omega\right)$ and using Eq. (24), we have :

$$\left\{ \begin{array}{rcll} -\Delta\Phi_h\left(\eta_h\right) - k^2\Phi_h\left(\eta_h\right) & = & 0 & \text{in } \Omega \\ \partial_n\Phi_h\left(\eta_h\right) & = & \mathrm{i}\,k\Phi_h\left(\eta_h\right) & \text{on } \partial\Omega. \end{array} \right. \tag{32}$$

Hence, $\Phi_h\left(\eta_h\right) = 0$ because the boundary value problem given by Eq. (32) admits a unique solution. It follows from Eq. (23) that $\eta_h^K = 0, \forall K \in \tau_h$. Consequently, $\eta_h = 0$ and therefore, $y_l = 0$ for all $1 \leq l \leq n^\lambda$. Thus, $\boldsymbol{y}^*\mathbf{A}\boldsymbol{y} > 0, \forall \boldsymbol{y} \in \mathbb{C}^{n^\lambda} \setminus \{0\}$, that is $\mathbf{A}$ is a positive definite matrix.

(ii) Observe that the matrix $\widetilde{\mathbf{A}}$ satisfies :

$$\begin{aligned} \boldsymbol{y}^*\widetilde{\mathbf{A}}\boldsymbol{y} = & \sum_{e-\text{interior edge}} \left(\beta_e \| \left[\Phi_h\left(\eta_h\right)\right] \|^2_{L^2(e)} + \gamma_e \| \left[\left[\partial_n\Phi_h\left(\eta_h\right)\right]\right] \|^2_{L^2(e)}\right) \\ & + \sum_{e\subset\partial\Omega} \omega_e \| \partial_n\Phi_h\left(\eta_h\right) - \mathrm{i}\,k\Phi_h\left(\eta_h\right) \|^2_{L^2(e)} \geq 0, \end{aligned}$$

that is $\widetilde{\mathbf{A}}$ is a positive semi-definite matrix.
Let $\boldsymbol{y}$ be a vector in $\mathbb{C}^{n^\lambda}$ such that $\boldsymbol{y}^*\widetilde{\mathbf{A}}\boldsymbol{y} = 0$. Following the same reasoning as for (i), we have :

- $\left[\Phi_h\left(\eta_h\right)\right] = 0$ on all interior edges $e$. Hence, $\Phi_h\left(\eta_h\right) \in H^1\left(\Omega\right)$.
- $\left[\left[\partial_n\Phi_h\left(\eta_h\right)\right]\right] = 0$ on all interior edges. Using the fact that $\Delta\Phi_h\left(\eta_h^K\right) \in L^2\left(K\right)$, we deduce that $\Delta\Phi_h\left(\eta_h\right) \in L^2\left(\Omega\right)$.
- $\partial_n\Phi_h\left(\eta_h\right) = \mathrm{i}\,k\Phi_h\left(\eta_h\right)$ on all the boundary edges of the domain.

Consequently, $\Phi_h\left(\eta_h\right) = 0$. Using Eq. (25), we conclude that $y_l = 0$ for all $1 \leq l \leq n^\lambda$, and then $\widetilde{\mathbf{A}}$ is a positive definite matrix. ■

**Remark 4.** Observe that in Approach 2 we obtain $\Phi_h\left(\lambda_h\right) = 0$ without any compatibility condition. From a numerical point of view, this should lead to a more robust approach.

# 5. The proposed solution methodology : Numerical investigation

In order to illustrate and assess the potential of mDGM for solving efficiently Helmholtz problems, we have performed numerical experiments using discrete spaces in which the shape functions are plane waves, as done in DGM [5, 6, 7]. More specifically, $\mathcal{V}_h$ are the spaces introduced in [5]. Once the local space of shape functions, $\mathcal{V}_h(K)$ is chosen, the Lagrange multiplier is approximated on each edge using a subset or all set of shape functions that occur when evaluating $\partial_n v_h^K - \mathrm{i}\,\alpha v_h^K$, for $v_h^K \in \mathcal{V}_h(K)$.
From now on, we suppose that $\Omega$ is an $a \times a$ square domain. We use a uniform partition of $\Omega$ in rectangular-shaped elements $K$. The functions $f$ and $g$ are such that the exact solution $u$ of BVP is a plane wave propagating in a direction $\boldsymbol{d} = (\cos\theta, \sin\theta)$. We vary the propagation angle $\theta$ in the interval $[0, 2\pi)$. In order to compare the results obtained with mDGM to those delivered by DGM, we measure, for each propagating angle $\theta$, the relative error using the following modified $H^1$ norm [5] :

$$\|v\|_{\widehat{H^1}} = \left( \sum_K \|v\|^2_{H^1(K)} + \sum_{e-\text{interior edge}} \|[v]\|^2_{L^2(e)} \right)^{\frac{1}{2}}, \quad \forall v \in \mathcal{V}. \tag{33}$$

Note that Eq. (33) is a modified $H^1$ norm since it takes into account the $H^1$ norm at the element level and the jump of the numerical solution along the interior interfaces of the mesh. We also use the *total* relative error, that is the *mean* value of the relative error obtained when $\theta \in [0, 2\pi)$.
For all numerical experiments, we have set $\alpha = k$, and $\beta_e = 1$ and $\gamma_e = h$ for all the interior edges $e$. The choice of these parameters results from our numerical investigation, given the lack of theoretical guidelines.
We present the results of two classes of numerical experiments : experiments using four plane waves per element, and experiments using eight plane waves per element. All the results are compared to the ones obtained with DGM.

## 5.1. Four plane waves per element

We equip each rectangular element with four plane waves positioned as indicated in Fig. 1.

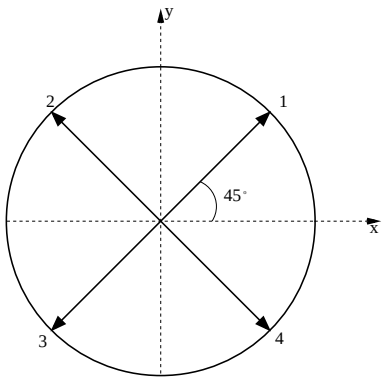


**Fig. 1** – Four plane waves positioned at $\theta_p = \pi/4 + (p-1)\pi/2, \ \forall\ 1 \le p \le 4$

More specifically, for each $K \in \tau_h$, we consider :

$$\mathcal{V}_h(K) = \Big\{ v_h^K = \sum_{1\le p\le 4} e^{\mathrm{i}\, k\boldsymbol{\theta}_p\cdot\boldsymbol{x}} u_p,\ \boldsymbol{\theta}_p = {}^t[\cos\theta_p, \sin\theta_p] \\ \theta_p = \pi/4 + (p-1)\pi/2, 1\le p\le 4,\ u_p \in \mathbb{C}\Big\}.$$

As stated earlier, the choice of the basis of $\mathcal{M}_h$ is related to the computation of $\partial_n v_h^K - \mathrm{i}\, k v_h^K$ on

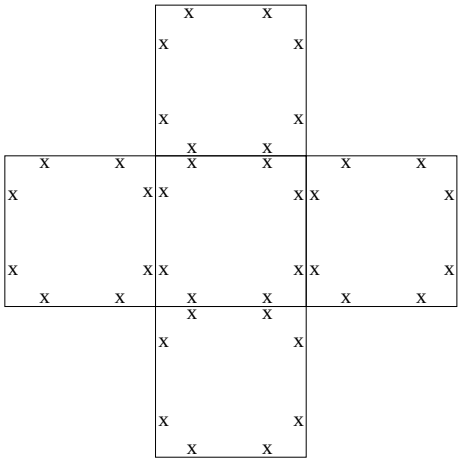

**Fig. 2** – A 40 dofs stencil corresponding to 2 dofs per edge

the edges of the mesh, for $v_h^K \in \mathcal{V}_h(K)$. Observe that, $\forall v_h^K \in \mathcal{V}_h(K)$, on each interior edge $e$ we have :

$$\partial_n v_h - \mathrm{i}\, k v_h = \mu_1 e^{\mathrm{i}\, k\frac{\sqrt{2}}{2}s} + \mu_2 e^{-\mathrm{i}\, k\frac{\sqrt{2}}{2}s}, \tag{34}$$

where $s$ represents the curvilinear abscissa and $\mu_1, \mu_2 \in \mathbb{C}$. Consequently, the Lagrange multiplier is approximated in the following discrete space :

$$\mathcal{M}_h = \Big\{ \mu_h \in \mathcal{M}; \forall K \in \tau_h, \mu_h^K|_e = \mu_1^K e^{\mathrm{i}\, k\frac{\sqrt{2}}{2}x} + \mu_2^K e^{-\mathrm{i}\, k\frac{\sqrt{2}}{2}x} \text{ if } e \parallel \overrightarrow{x}, \\ \mu_h^K|_e = \mu_1^K e^{\mathrm{i}\, k\frac{\sqrt{2}}{2}y} + \mu_2^K e^{-\mathrm{i}\, k\frac{\sqrt{2}}{2}y} \text{ if } e \parallel \overrightarrow{y}, \mu_1, \mu_2 \in \mathbb{C}\Big\}.$$

The spaces $\mathcal{V}_h$ and $\mathcal{M}_h$ defined above correspond to the so-called R-4-2 element in the nomenclature of DGM (see [5]). Note that for the DGM, considering two dofs per edge leads also to a complete approximation of the Lagrange multiplier.
We must point out that, unlike the DGM, the Lagrange multipliers in mDGM are not continuous across interior boundaries. This is why on an edge shared by two elements, the Lagrange multipliers on each side of the edge are different. Consequently, the stencil of the matrix given by Eq. (19) is equal to 40 (see Fig. 2). Note that in the DGM the Lagrange multipliers are equal on both sides of the edge, which leads to a stencil equal to 14.

The first experiments consist in comparing the error delivered by both numerical methods (DGM and mDGM) for different values of $ka$, while maintaining $kh$ constant. More specifically, we consider $ka = 10, 20, 30$ and we choose the step size of the mesh discretization $h/a$ such that $kh = \frac{1}{5}$, which

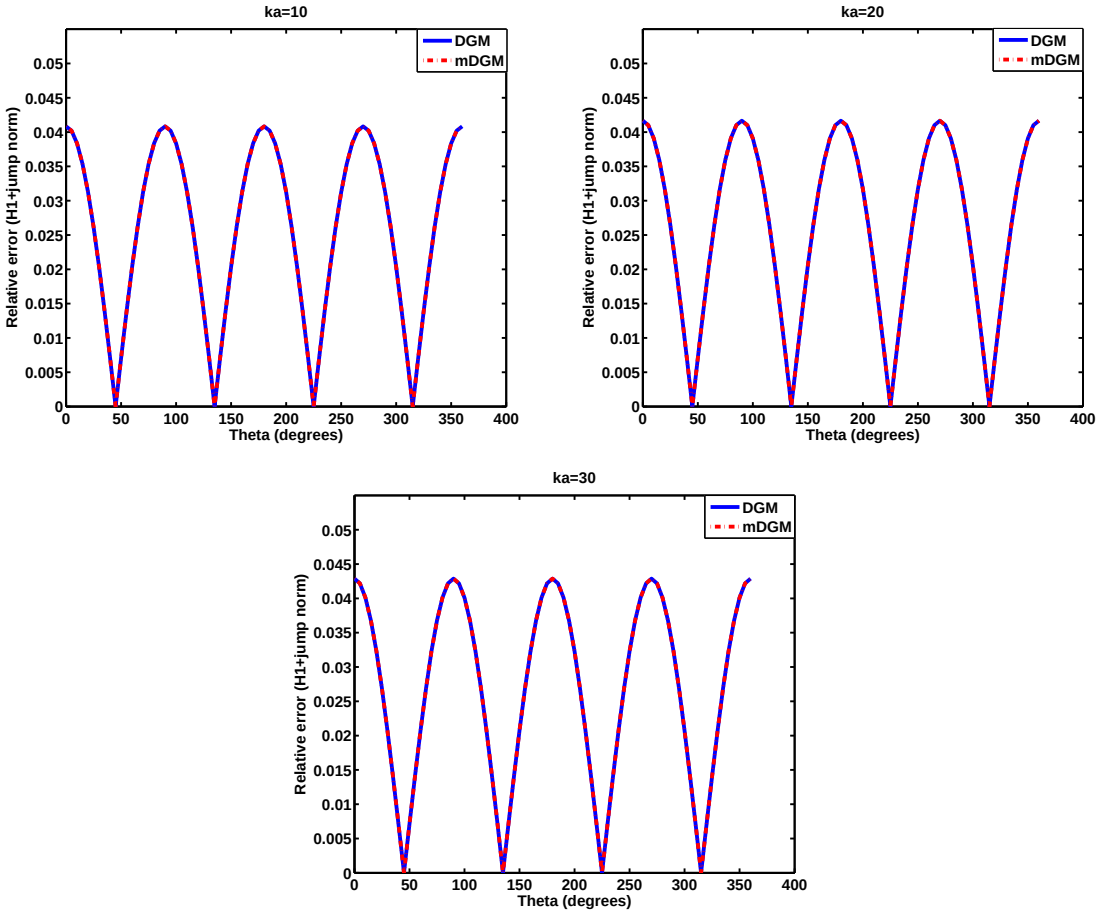


**Fig. 3** – Performance of the two methods for kh=1/5

is about 30 elements per wavelength. The results are depicted in Figs. 3-4 These results indicate the following :

- The two methods deliver results with the same level of accuracy, as indicated in Fig. 3 : both curves are superposed.
- As expected, the relative error is $\pi/2$ periodic (see Fig. 3). On each period $[(l-1)\pi/2, l\pi/2]$ (with $l = 1, 2, 3, 4$), the error is symmetric with respect to the propagation angle $(2l-1)\pi/4$. Moreover, the error is minimal for $\pi/4, 3\pi/4, 5\pi/4, 7\pi/4$ (less than 1%) and maximal for $0, \pi/2, \pi, 3\pi/2$ (about 4%). This is due to the chosen basis, which includes the exact solution when the propagation angle is $(2l-1)\pi/4$, with $l = 1, 2, 3, 4$ and to the fact that the Lagrange multiplier field contains all the functions needed to have a complete approximation.
- Fig. 4 indicates that the R-4-2 element (for both methods) exhibits little pollution : increasing $ka$, while maintaining $kh$ constant, leads to an increase in the relative error which is less than 0.5% at most (see Fig. 4 at angles $\theta = l\pi/2$, with $l = 0, 1, ..., 8$).

Next, we compare the sensitivity of the total relative error (the mean value over the propagation angles) to the mesh size. The result depicted in Fig. 5 is obtained for $ka = 1$. One can observe the following :

- For $h/a > \frac{1}{100}$, the errors delivered by the two methods are comparable. The two curves are on top of each other.

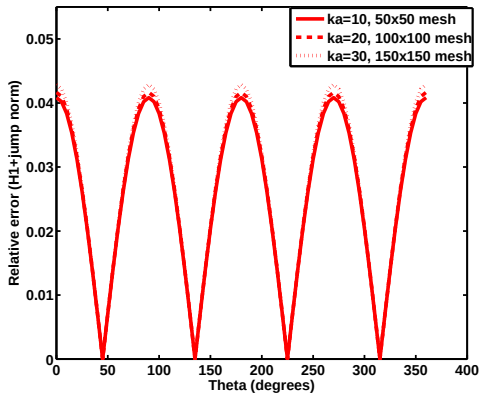


**Fig. 4** – Pollution effect for the R-4-2 element

- For $h/a < \frac{1}{100}$ mDGM outperforms DGM. As we refine the mesh ($h/a < \frac{1}{100}$), DGM becomes unstable. Indeed, there is a dramatic loss in the accuracy of more than one order of magnitude. The error jumps from 0.09% (for $h/a = \frac{1}{100}$)to 1.5% (for $h/a = \frac{1}{190}$). The instability observed in DGM seems to be related to the severe ill conditioning of the local matrices. Observe that mDGM remains stable as we refine the mesh. The last point of the curve was obtained for $h/a = \frac{1}{450}$, the limit of our computing platform. The total relative error for this mesh size is 0.04%.

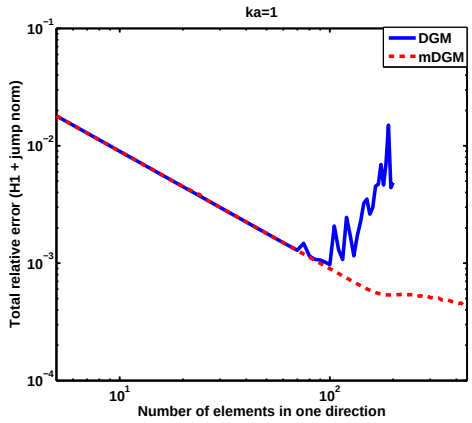


**Fig. 5** – Sensitivity of the total relative error to the mesh refinement : Comparison between DGM (R-4-2) nad mDGM (R-4-2) for ka=1

We must point out that the performance of mDGM in this case is not sensitive to the choice of the approach for solving the linear system in Step 2. Both approaches deliver results with the same level of accuracy.

## 5.2. Eight plane waves per element

We approximate the primal variable using eight plane waves, positioned at :

$$\theta_p = (p-1)\pi/4, \ \forall 1 \leq p \leq 8.$$

This choice corresponds to the following discrete space for the primal variable :

$$\mathcal{V}_h(K) = \Big\{ v_h|^K = \sum_{1\le p\le 8} e^{\mathrm{i}\, k\boldsymbol{\theta}_p\cdot\boldsymbol{x}} u_p,\ \boldsymbol{\theta}_p = {}^t[\cos\theta_p, \sin\theta_p],$$
$$\theta_p = (p-1)\pi/4, 1\le p\le 8,\ u_p\in\mathbb{C}\Big\}.$$

For an element $v_h^K \in \mathcal{V}_h(K)$, the full approximation of $\partial_n v_h^K - \mathrm{i}\, k v_h^K$ leads to five dofs per edge. More specifically in mDGM, the discrete space $\mathcal{M}_h$ corresponding to the full approximation of the Lagrange multiplier is given by :

$$\mathcal{M}_h = \Big\{ \mu_h\in\mathcal{M}; \forall K\in\tau_h, \mu_h^K|_e = \mu_1^K + \mu_2^K e^{\mathrm{i}\, kx} + \mu_3^K e^{-\mathrm{i}\, kx} + \mu_4^K e^{\mathrm{i}\, k\frac{\sqrt{2}}{2}x}$$
$$+\mu_5^K e^{-\mathrm{i}\, k\frac{\sqrt{2}}{2}x} \text{ if } e \parallel \overrightarrow{x},\ \mu_h^K|_e = \mu_1^K + \mu_2^K e^{\mathrm{i}\, ky} + \mu_3^K e^{-\mathrm{i}\, ky}$$
$$+\mu_4^K e^{\mathrm{i}\, k\frac{\sqrt{2}}{2}y} + \mu_5^K e^{-\mathrm{i}\, k\frac{\sqrt{2}}{2}y} \text{ if } e \parallel \overrightarrow{y},\quad \mu_1,\mu_2,\mu_3,\mu_4,\mu_5\in\mathbb{C}\Big\}.$$

Following the nomenclature introduced in [5], such an approximation is called the $R$-8-5 element. Note that mDGM can be implemented using less dofs per edge for the Lagrange multiplier. We recall that in DGM, the maximum number of dofs considered on each edge is three. Indeed, the computation of the normal derivative of the numerical solution leads to the following complete approximation :

$$\lambda_h = \mu_1 + \mu_2 e^{\mathrm{i}\, k\frac{\sqrt{2}}{2}s} + \mu_3 e^{-\mathrm{i}\, k\frac{\sqrt{2}}{2}s},$$

where $s$ represents the curvilinear abscissa. This choice of approximation corresponds to the so-called $R$-8-3 element.
We first present the results obtained in the case of Approach 1, than the ones obtained in the case of Approach 2.

### 5.2.1. Performance assessment in the case of Approach 1

Since the full approximation in DGM requires three dofs per edge, we first compare the performance of mDGM and DGM when using the $R$-8-3 element. The result depicted in Fig. 6 compares the relative error delivered by both methods, as a function of the propagation angle. This result is obtained for $ka = 10$ and $h/a = 1/20$, that is $kh = \frac{1}{2}$, corresponding to about 12 elements per wavelength. It shows a clear superiority of mDGM over DGM. In addition, we have :

- DGM delivers the exact solution for $l\pi/4$, with $l = 0, 1, ..., 8$. Note that in each of these cases, the exact solution is represented by one of the basis functions of the considered element and all the functions obtained when computing the normal derivative are in the Lagrange multiplier field. On the other hand, mDGM computes exactly the solution at angles $\pi/4, 3\pi/4, 5\pi/4, 7\pi/4$ only, as it will be shown later (see Fig. 8). This is not surprising since two dofs are removed from the full approximation of the Lagrange multiplier and therefore, for the plane waves propagating in the directions parallel to the axis, the approximation is not complete.
- The total relative error is about 0.091% for mDGM and is about 5% for DGM. This means that mDGM improves the accuracy by about one and a half order of magnitude.
- Observe that DGM $R$-8-3 is an unstable element. Indeed, the error obtained for $\theta = (2l-1)\pi/8$ should be the same for all $l = 1, 2, ..., 8$ and symmetric with respect to $0, \pi/4, 2\pi/4, ..., 8\pi/4$.

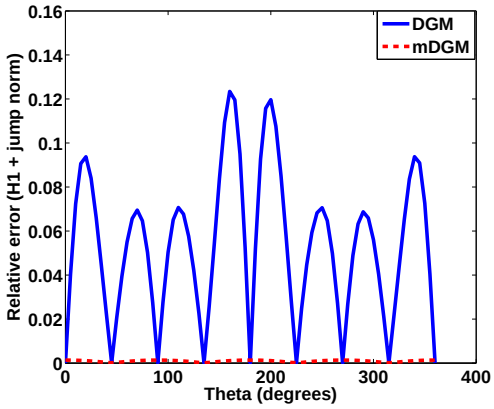


**Fig. 6** – Performance comparison between DGM($R$-8-3) and mDGM ($R$-8-3) for $ka = 10$, $h/a = 1/20$

Fig. 6 shows that these values are not equal.
The instability observed in the DGM approach for $R$-8-3 element is not surprising since this element does not satisfy the numerical inf-sup condition required by DGM [11]. Almost everywhere in the mesh, for an element $K$ there are twelve dofs for the Lagrange multiplier and only eight dofs for the primal variable. A dof must be removed from each edge. For this reason, two discrete spaces were suggested in [5] for the discrete dual variable, leading respectively to the so-called $R$-8-2a and $R$-8-2b elements. Since in the cited paper, the $R$-8-2b element was shown to deliver more accurate results than the $R$-8-2a element, we have compared the two methods when employing the $R$-8-2b element. We recall that the $R$-8-2b element corresponds to the following approximation of the

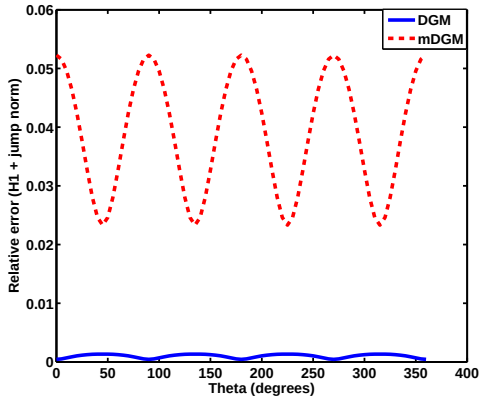


**Fig. 7** – Performance comparison between DGM($R$-8-2b) and mDGM ($R$-8-2b) for $ka = 10$, $h/a = 1/20$

Lagrange multiplier :

$$\lambda_h = \mu_1 e^{\mathrm{i}\, k \frac{\sqrt{2}}{4} s} + \mu_2 e^{-\mathrm{i}\, k \frac{\sqrt{2}}{4} s},$$

where $s$ is the curvilinear abscissa. Similarly to the previous numerical experiment, we have set $ka = 10$ and $h/a = 1/20$, which corresponds to $kh = \frac{1}{2}$. The result depicted in Fig. 7 suggests the following :

- As expected, both methods preserve the symmetry of the error with respect to the propagation angles $\theta = \pi/4, 3\pi/4, 5\pi/4, 7\pi/4$.

- The total relative error obtained with the mDGM is about 5% while the one obtained with DGM is about 0.099%. This superiority of DGM over mDGM is most likely due to the poor approximation of the Lagrange multiplier in the mDGM (three out of five dofs are neglected), compared to the DGM, where only one dof out of three is neglected.

Next, we enrich the approximation of the Lagrange multiplier in mDGM. We use the elements $R$-8-3 and $R$-8-5 and compare mDGM to DGM, equipped with the best element $R$-8-2b. The results reported in Fig. 8 are obtained for $ka = 10$ and $h/a = 1/20$. One can observe the following :

- There is a little improvement in the accuracy of the results delivered by $R$-8-3 and $R$-8-5 mDGM elements over DGM $R$-8-2b element. The total errors obtained with these elements are 0.091% and 0.048% respectively, whereas the one delivered by DGM equipped with the best element, R-8-2b, is 0.099%.

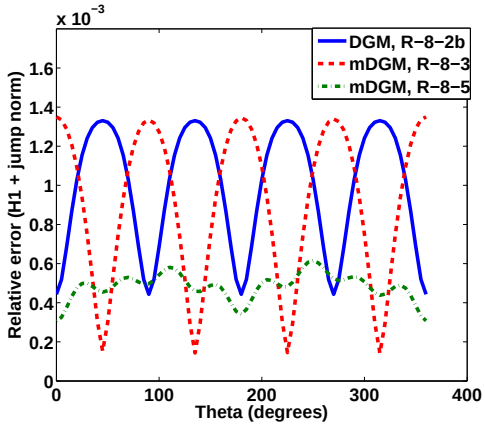


**Fig. 8** – Performance comparison between DGM($R$-8-2b), mDGM($R$-8-3) and mDGM ($R$-8-5) for $ka = 10$, $h/a = 1/20$

- The R-8-5 element seems to be unstable, as illustrated by the loss of the symmetry at the propagation angles $\theta = l\pi/4$, with $l = 0, 1, 2, ..., 7$, corresponding to the basis functions. This numerical instability is also noticeable when we compare $R$-8-5 to $R$-8-3 : for some propagation angles, $R$-8-3 is more accurate than $R$-8-5, which is contrary to what is expected since using five dofs per edge leads to the full approximation of the Lagrange multiplier. We believe that the instability of the $R$-8-5 element is due to the loss of the linear independence of the shape functions, but also to the fact that the matrix of the system is positive semi-definite.

Next, we investigate the sensitivity of the total relative error with respect to $h$, the step size of the mesh discretization. We consider the case of $R$-8-2b element and we set $ka = 1$. We evaluate the total relative error, as well as the smallest eigenvalue of the local system. The results are reported in Table 1.

- The results reveal that in the DGM approach, the error decreases as long as $h/a > \frac{1}{6}$. Then, the error jumps from 0.01% to about 33% for $kh = \frac{1}{23}$. This is not surprising. Indeed, the local systems in the DGM are nearly singular and therefore extremely ill-conditioned when $h$ becomes small, as indicated by the values corresponding to the smallest eigenvalues.
- The smallest error delivered by mDGM, which is about 0.01%, is obtained for $kh = \frac{1}{8}$. Then the error jumps to 37% for $kh = \frac{1}{23}$. This instability is, *a priori*, unexpected and very surprising since we have introduced the Robin-type boundary condition to address specifically this issue, as demonstrated in the case of the $R$-4-2 element. A quick look at Table 1 indicates that the local system corresponding to mDGM becomes nearly singular too. Hence, contrary to our goal, the presence of $\alpha$ seems to be not sufficient to avoid the singularity of the local systems.

TAB. 1 – Dependence with respect to the mesh size of the total relative error and smallest eigenvalue of the local matrix for $R$-8-2b when $ka = 1$

| $h/a$ | DGM | | mDGM | |
|---|---|---|---|---|
| | Total relative error | The smallest eigenvalue | Total relative error | The smallest eigenvalue |
| 1/4 | 0.016% | $5.7 \cdot 10^{-10} + 2.9 \cdot 10^{-18}$i | 0.054% | $5.7 \cdot 10^{-10} - 3.1 \cdot 10^{-11}$i |
| 1/5 | 0.011% | $9.5 \cdot 10^{-11} + 8.0 \cdot 10^{-19}$i | 0.036% | $9.5 \cdot 10^{-11} - 4.1 \cdot 10^{-12}$i |
| 1/6 | 0.019% | $2.2 \cdot 10^{-11} - 2.4 \cdot 10^{-18}$i | 0.025% | $2.2 \cdot 10^{-11} - 7.9 \cdot 10^{-13}$i |
| 1/7 | 0.092% | $6.5 \cdot 10^{-12} + 5.9 \cdot 10^{-19}$i | 0.019% | $6.4 \cdot 10^{-12} - 2.0 \cdot 10^{-13}$i |
| 1/8 | 0.394% | $2.2 \cdot 10^{-12} + 1.8 \cdot 10^{-18}$i | 0.015% | $2.2 \cdot 10^{-12} - 5.9 \cdot 10^{-14}$i |
| 1/9 | 1.206% | $8.6 \cdot 10^{-13} - 1.6 \cdot 10^{-18}$i | 0.017% | $8.6 \cdot 10^{-13} - 2.1 \cdot 10^{-14}$i |
| 1/20 | 4.734% | $1.5 \cdot 10^{-15} - 6.0 \cdot 10^{-21}$i | 3.590% | $1.5 \cdot 10^{-15} + 5.9 \cdot 10^{-17}$i |
| 1/21 | 11.664% | $9.8 \cdot 10^{-16} + 7.0 \cdot 10^{-18}$i | 5.889% | $9.6 \cdot 10^{-16} + 2.4 \cdot 10^{-17}$i |
| 1/23 | 33.435% | $4.7 \cdot 10^{-16} - 3.9 \cdot 10^{-18}$i | 37.378% | $4.7 \cdot 10^{-16} - 6.7 \cdot 10^{-17}$i |

We believe that the singularity of the local system in the mDGM formulation is due to the loss of the linear independence of the shape functions (eight plane waves) as $h$ becomes small, as well as to the non-negativeness nature of the matrix $\mathbf{A}$ of Approach 1.
The loss of the linear independence can be demonstrated as follows : let $K$ be an element of the mesh. For simplicity, we assume $K$ to be the square $[0, h] \times [0, h]$. Then, for a function $w_h^K \in \mathcal{V}_h(K)$, there exist $c_1, c_2, ..., c_8 \in \mathbb{C}$ such that :

$$w_h^K = \sum_{m=1}^{8} c_m e^{\mathrm{i}\, k \boldsymbol{\theta}_m \cdot \boldsymbol{x}}.$$

Assume that :

$$a_K\left(w_h^K, w_h^K\right) = 0. \tag{35}$$

Consequently, $w_h^K = 0$ on $\partial K$, which means :

$$\sum_{m=1}^{8} c_m e^{\mathrm{i}\, k \boldsymbol{\theta}_m \cdot \boldsymbol{x}} = 0 \quad \text{on} \quad \partial K.$$

We write this equality on each of the four edges of $K$. For all $x \in [0, h]$ and $y \in [0, h]$, we have :

$$\left\{\begin{array}{rl}
c_1 + c_5 + (c_2 + c_4)e^{\mathrm{i}\, k \frac{\sqrt{2}}{2} y} + c_3 e^{\mathrm{i}\, k y} + (c_6 + c_8)e^{-\mathrm{i}\, k \frac{\sqrt{2}}{2} y} + c_7 e^{-\mathrm{i}\, k y} & = 0, \\
c_3 + c_7 + (c_2 + c_8)e^{\mathrm{i}\, k \frac{\sqrt{2}}{2} x} + c_1 e^{\mathrm{i}\, k x} + (c_4 + c_6)e^{-\mathrm{i}\, k \frac{\sqrt{2}}{2} x} + c_5 e^{-\mathrm{i}\, k x} & = 0, \\
c_1 e^{\mathrm{i}\, k h} + c_5 e^{-\mathrm{i}\, k h} + (c_2 e^{\mathrm{i}\, k \frac{\sqrt{2}}{2} h} + c_4 e^{-\mathrm{i}\, k \frac{\sqrt{2}}{2} h})e^{\mathrm{i}\, k \frac{\sqrt{2}}{2} y} + c_3 e^{\mathrm{i}\, k y} + & \\
\qquad (c_6 e^{-\mathrm{i}\, k \frac{\sqrt{2}}{2} h} + c_8 e^{\mathrm{i}\, k \frac{\sqrt{2}}{2} h})e^{-\mathrm{i}\, k \frac{\sqrt{2}}{2} y} + c_7 e^{-\mathrm{i}\, k y} & = 0, \\
c_3 e^{\mathrm{i}\, k h} + c_7 e^{-\mathrm{i}\, k h} + (c_2 e^{\mathrm{i}\, k \frac{\sqrt{2}}{2} h} + c_8 e^{-\mathrm{i}\, k \frac{\sqrt{2}}{2} h})e^{\mathrm{i}\, k \frac{\sqrt{2}}{2} x} + c_1 e^{\mathrm{i}\, k x} + & \\
\qquad (c_6 e^{-\mathrm{i}\, k \frac{\sqrt{2}}{2} h} + c_4 e^{\mathrm{i}\, k \frac{\sqrt{2}}{2} h})e^{-\mathrm{i}\, k \frac{\sqrt{2}}{2} x} + c_5 e^{-\mathrm{i}\, k x} & = 0.
\end{array}\right.$$

Therefore, we deduce that :

$$\begin{cases} c_1 = c_3 = c_5 = c_7 = 0 \\ c_4 = -c_2 \\ c_6 = c_2 \\ c_8 = -c_2 \\ c_2 \left( e^{\mathrm{i}\, k \frac{\sqrt{2}}{2} h} - e^{-\mathrm{i}\, k \frac{\sqrt{2}}{2} h} \right) = 0. \end{cases} \tag{36}$$

The problem here is that when $h \to 0$, we have : $e^{\mathrm{i}\, k \frac{\sqrt{2}}{2} h} \to 1$ and $e^{-\mathrm{i}\, k \frac{\sqrt{2}}{2} h} \to 1$ and hence, numerically speaking, it is not necessary to have $c_2 = 0$ in order to have $w_h^K = 0$ on $\partial K$. Consequently, $w_h^K$ may not be equal to 0 when Eq. (35) is satisfied. This computation shows that when $h$ tends to 0, the eight plane waves become linearly dependent, which leads to the singularity in the local matrix.

**Remark 5.** We have performed additional numerical experiments for higher frequencies and observed that both methods become unstable as we refine the mesh. DGM is stable as long as $kh > \frac{1}{6}$, while mDGM seems to remain stable longer ($kh > \frac{1}{9}$). We must point out that the source for the instability is however different. DGM is unstable not only because of the singularity of the local systems, but also due to the loss of the linear independence of the plane waves as $h$ becomes small. We believe that the resulting local system in DGM is also very sensitive to this loss.
On the other hand, mDGM restores the stability of the local problems and leads to a more stable formulation when the shape functions remain linearly independent, as it has been demonstrated in the case of the $R$-4-2 element (see Fig. 5) and in the next experiments. However, the previous numerical results suggest that the loss of the linear independence affects dramatically the stability of mDGM when using Approach 1, that is when the linear system is non-negative only.

The following experiment reveals the behavior of mDGM and DGM when the used shape functions remain linearly independent as we refine the mesh. This experiment consists in approximating the solution, at the element level, using seven plane waves, positioned at :

$$\theta_p = 2(p-1)\pi/7, \quad 1 \leq p \leq 7. \tag{37}$$

We have maintained the same two dofs per edge as in $R$-8-2b (see Section 5.2). Following the nomenclature introduced in [5], we will refer to this element as $R$-7-2. For $ka = 1$, we have compared the sensitivity of the total relative error to the mesh size for the mDGM $R$-8-2b and $R$-7-2 elements. The result depicted in Fig. 9 illustrates the following :

- The error delivered by the $R$-7-2 element decreases as long as $kh > \frac{1}{18}$, unlike the $R$-8-2b element, which delivers the smallest error for $h/a = \frac{1}{9}$. This means that there is a reduction of factor 2 on the mesh size, while maintaining the stability.
- The $R$-7-2 element is shown to be more accurate. For each mesh size, the error delivered by this element is smaller than the one obtained with the $R$-8-2b element. Moreover, the most accurate approximation (about 0.001% for $h/a = \frac{1}{18}$) obtained with $R$-7-2 is one order of magnitude lower than $R$-8-2b element (about 0.01% for $h/a = \frac{1}{9}$).
- Last, note that the $R$-7-2 element remains stable while refining the mesh. For $h/a = \frac{1}{100}$, the total relative error is about 1%, unlike the $R$-8-2b element in which the error jumps from 0.01% (for $h/a = \frac{1}{9}$) to 59% (for $h/a = \frac{1}{24}$).

We have also compared the errors delivered by the mDGM $R$-7-2 element to the ones obtained with the DGM $R$-7-2 element. The result is reported in Fig. 10. The following observations are

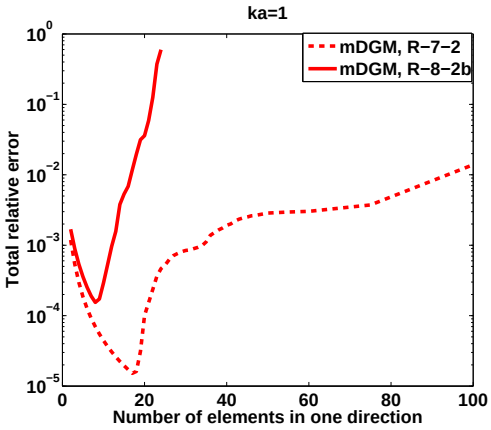


**Fig. 9** – Sensitivity of the total relative error to the mesh refinement : Comparison between mDGM ($R$-8-2b) and mDGM ($R$-8-2b) when $ka = 1$

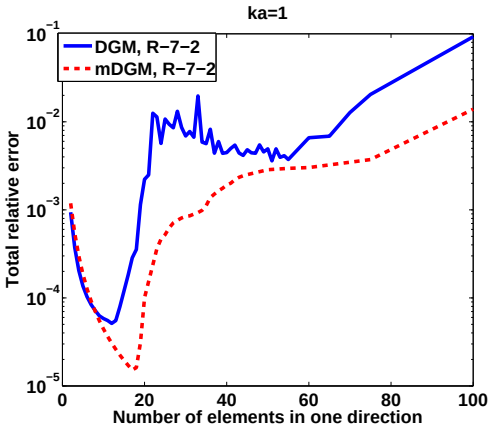


**Fig. 10** – Sensitivity of the total relative error to the mesh refinement : Comparison between mDGM ($R$-7-2) and DGM ($R$-7-2) when $ka = 1$

noteworthy :

- The accuracy of the two methods is comparable for $h/a > \frac{1}{8}$. In this region the two curves are superposed.
- The DGM $R$-7-2 element delivers the most accurate approximation (which is about 0.005%) for $h/a = \frac{1}{12}$. Observe that mDGM becomes unstable slightly later : the smallest error (about 0.001%) is obtained for $h/a = \frac{1}{18}$).
- Although for both methods we observe numerical instabilities as soon as $h/a < \frac{1}{18}$, mDGM is more accurate than DGM. For any mesh size the error delivered by mDGM is smaller than the one obtained with DGM. Moreover, for some mesh sizes, mDGM outperforms DGM by one order of magnitude. We believe that this is due to the local problems which are nearly singular in DGM.

The results depicted in Fig. 9 seem to be surprising since one may expect that approximating the primal variable with eight plane waves leads to more accurate results than when using seven plane waves. Here, it seems that the linear independence when using seven plane waves is less sensitive (remains longer) to the mesh size $h/a$. Consequently, the linear system corresponding to Approach 1 is more stable in this case than when using eight plane waves.

In Table 2 we report the total relative error, as well as the smallest eigenvalue of the local system for the $R$-7-2 element for both DGM and mDGM methods, when $ka = 1$.

TAB. 2 – Dependence with respect to the mesh size of the total relative error and smallest eigenvalue of the local matrix for $R$-7-2 when $ka = 1$

| $h/a$ | DGM | | mDGM | |
|---|---|---|---|---|
| | Total relative error | The smallest eigenvalue | Total relative error | The smallest eigenvalue |
| 1/4 | 0.020% | $5.2 \cdot 10^{-07} - 3.0 \cdot 10^{-17}$i | 0.030% | $5.2 \cdot 10^{-07} - 3.2 \cdot 10^{-08}$i |
| 1/5 | 0.014% | $1.4 \cdot 10^{-07} - 4.8 \cdot 10^{-17}$i | 0.019% | $1.4 \cdot 10^{-07} - 6.6 \cdot 10^{-09}$i |
| 1/6 | 0.010% | $4.6 \cdot 10^{-08} - 8.0 \cdot 10^{-18}$i | 0.012% | $4.6 \cdot 10^{-08} - 1.9 \cdot 10^{-09}$i |
| 1/7 | 0.008% | $1.8 \cdot 10^{-08} + 2.0 \cdot 10^{-17}$i | 0.009% | $1.8 \cdot 10^{-08} - 6.3 \cdot 10^{-10}$i |
| 1/8 | 0.007% | $8.1 \cdot 10^{-09} + 3.9 \cdot 10^{-17}$i | 0.007% | $8.1 \cdot 10^{-09} - 2.5 \cdot 10^{-10}$i |
| 1/9 | 0.006% | $4.0 \cdot 10^{-09} + 2.5 \cdot 10^{-17}$i | 0.005% | $4.0 \cdot 10^{-09} - 1.1 \cdot 10^{-10}$i |
| 1/12 | 0.005% | $7.1 \cdot 10^{-10} + 6.0 \cdot 10^{-18}$i | 0.003% | $7.1 \cdot 10^{-10} - 1.5 \cdot 10^{-11}$i |
| 1/18 | 0.035% | $6.3 \cdot 10^{-11} + 7.7 \cdot 10^{-18}$i | 0.001% | $6.3 \cdot 10^{-11} - 8.5 \cdot 10^{-13}$i |
| 1/20 | 0.222% | $3.3 \cdot 10^{-11} + 3.1 \cdot 10^{-19}$i | 0.010% | $3.3 \cdot 10^{-11} - 4.1 \cdot 10^{-13}$i |
| 1/25 | 1.075% | $8.7 \cdot 10^{-12} - 2.0 \cdot 10^{-18}$i | 0.052% | $8.7 \cdot 10^{-12} + 2.4 \cdot 10^{-17}$i |
| 1/34 | 0.589% | $1.4 \cdot 10^{-12} + 4.4 \cdot 10^{-18}$i | 0.099% | $1.4 \cdot 10^{-12} - 9.9 \cdot 10^{-15}$i |
| 1/50 | 0.493% | $1.4 \cdot 10^{-13} - 1.1 \cdot 10^{-19}$i | 0.285% | $1.4 \cdot 10^{-13} - 6.8 \cdot 10^{-16}$i |
| 1/75 | 2.046% | $1.2 \cdot 10^{-14} + 4.0 \cdot 10^{-18}$i | 0.373% | $1.2 \cdot 10^{-14} - 5.1 \cdot 10^{-17}$i |
| 1/100 | 9.266% | $2.1 \cdot 10^{-15} + 4.5 \cdot 10^{-18}$i | 1.390% | $2.1 \cdot 10^{-15} + 2.5 \cdot 10^{-18}$i |
| 1/150 | 243.2% | $1.9 \cdot 10^{-16} + 1.5 \cdot 10^{-19}$i | 67.75% | $1.8 \cdot 10^{-16} - 8.3 \cdot 10^{-18}$i |

- As in the case of the $R$-8-2b element, the eigenvalues of the local matrices corresponding to each mesh size have the same real part in DGM and mDGM. The Robin-type condition used in mDGM leads to more important imaginary parts in mDGM.
- The values of the smallest eigenvalues and the comparison to the ones reported in Table 1 show that the linearly independence of the seven plane waves is less sensitive to the mesh refinement. Indeed, for example for $h/a = \frac{1}{9}$ the real part of the smallest eigenvalue obtained when using eight plane waves at the element level is $8.64 \cdot 10^{-13}$. This is four orders of magnitude larger than the real part of the smallest eigenvalue of the matrix obtained with seven plane waves $(4.00 \cdot 10^{-9})$.
- A quick comparison of the errors reported in Tables 1 and 2 (see also Fig. 9 for mDGM) shows that in both formulations a better conditioning of the local matrices leads to more accurate approximations. Moreover, in both methods there is a reduction of factor 2 on the mesh size, while maintaining the stability. These two important points are related to the fact that the seven shape functions remain linearly independent as we refine the mesh.
- As it was shown in Fig. 10 and reported in Table 2, the numerical instabilities appear earlier in DGM than in mDGM. We believe that this is due to the local systems which become nearly singular with the mesh refining. In mDGM the observed numerical instabilities are due to the non-negativeness nature of the global matrix corresponding to Approach 1.
- Last, we must point out the fact that the seven shape functions are becoming linearly dependent with the mesh refinement, but more slowly than the eight plane waves corresponding to the $R$-8-2b element. This behavior of the shape functions is predictable when observing the dependence of the smallest eigenvalue of the local matrix with respect to the mesh size.

Consequently, this element will be ultimately unstable.

### 5.2.2. Performance assessment in the case of Approach 2

We have performed several numerical experiments to assess the performance of mDGM when at Step 2 the linear system is given by Approach 2 (see Eqs. (21)-(22)). Due to the Remark 4, it is

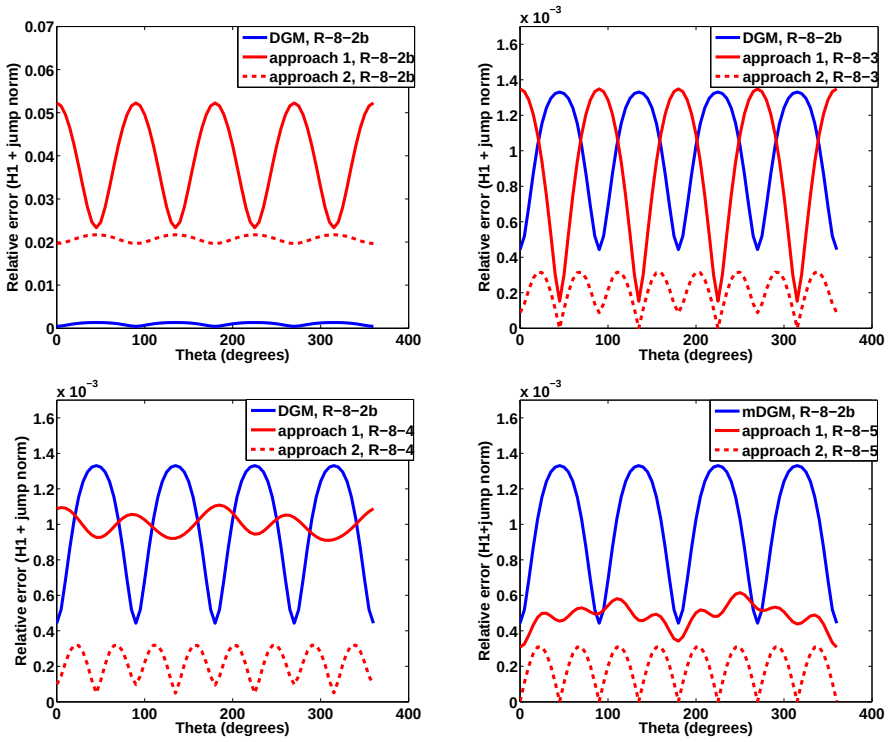


**Fig. 11** – Performance of the three methods, kh=1/2

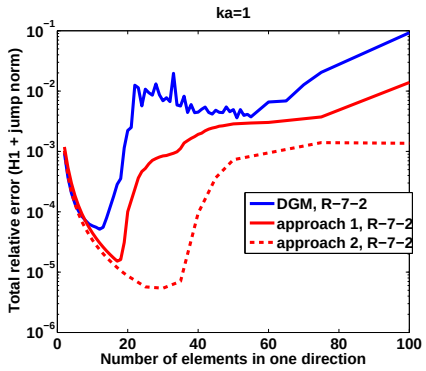


**Fig. 12** – Sensitivity of the total relative error to the mesh refinement : Comparison between the three methods in the case of the R-7-2 element

expected that with the well-posedness of the local boundary value problem in Step 1, the method leads to more stable and thus, more accurate numerical results. The obtained numerical results are depicted in Figs. 11-13. They clearly indicate that Approach 2 of mDGM not only outperforms both Approach 1 of mDGM and DGM, but also delivers more accurate results. The error is reduced by

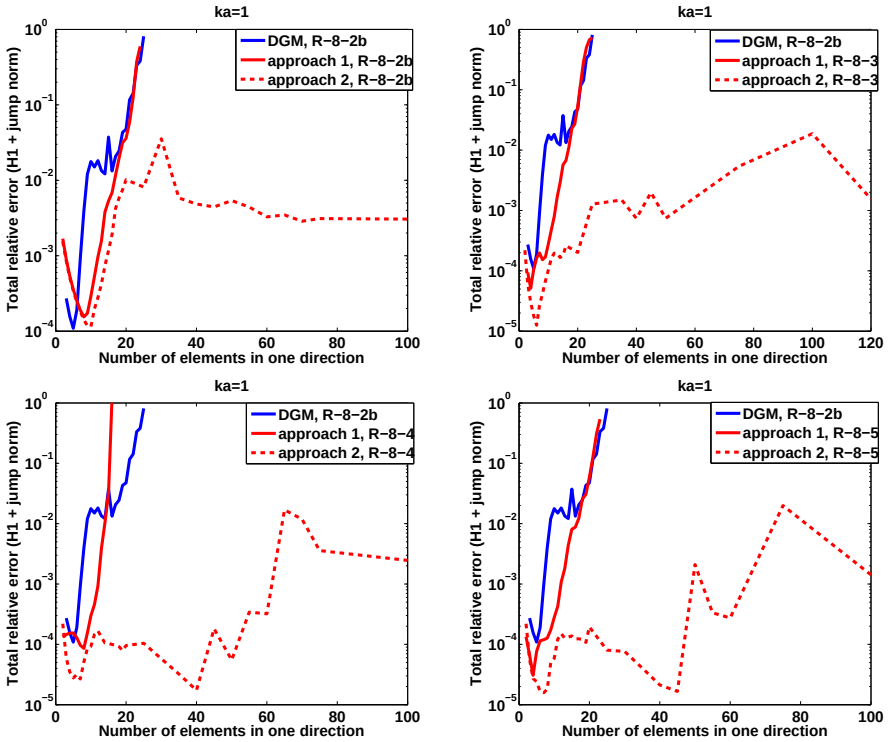


**Fig. 13** – Sensitivity of the total relative error to the mesh refinement : Comparison between the three methods in the case of the R-8-2b element

one to two orders of magnitude depending on the element and the mesh size.

Note that the numerical stability occurs only in the case of mDGM when equipped with Approach 2. Indeed, one can observe that the errors (see dashed curves in Fig. 13) are oscillating as $a/h$ is increasing, but their magnitude remains steadily about 1%, whereas the magnitude of the errors delivered by the two other methods increase to over 100%.

## 6. Summary and conclusion

We have designed a new solution methodology, called mDGM, for Helmholtz problems which is easy to understand and implement. At the element level, we approximate the solution by a superposition of plane waves. Consequently, the obtained solution is discontinuous and Lagrange multipliers are introduced to ensure the continuity in a weak sense. Unlike the DGM, the Lagrange multiplier is also discontinuous, which allows us to consider well-posed local problems. The algebraic approach requires solving local linear systems with multiple right-hand side : the system's size is given by the number of plane waves considered in the local basis. These problems are independent from one element to another and therefore can be solved in parallel. The global system, whose size is the number of total dofs used for approximating the Lagrange multiplier, is positive semi-definite. The numerical results we have presented show that the proposed method is more stable than the DGM. When using Approach 2, mDGM is not only more stable than DGM, but also exhibits a better level of accuracy. More specifically, as indicated by the reported numerical results, mDGM reduces the level of errors by one to two orders of magnitude depending on the mesh size and on the element.

# Bibliography


[1] Amara M., Djellouli R., Farhat C., Convergence analysis of a discontinuous Galerkin method with plane waves and Lagrange multipliers for the solution of Helmholtz problems *SIAM J. Numer. Anal.*, **47**(2), 1038-1066, 2009

[2] Babuška I., Melenk I.J.M., The partition of unity method *Internat. J. Numer. Methhods Eng.*, **40**, 727-758, 1997

[3] Babuška I., Sauter S., Is the Pollution Effect of the FEM Avoidable for the Helmholtz Equation Considering High Wave Numbers ? *SIAM J. Numer. Anal.*, **34**, 2392-2423, 1997

[4] Cessenat O., Després B., Application of an ultra-weak variational formulation of elliptic PDEs to the two-dimensional Helmholtz problems *SIAM J. Numer. Anal.*, **35**, 255-299, 1998

[5] Farhat C., Harari I., Hetmaniuk U., A discontinuous Galerkin method with Lagrange multipliers for the solution of Helmholtz problems in the mid-frequency regime *Comput. Methods Appli. Mech. Eng.*, **192**, 1389-1419, 2003

[6] Farhat C., Wiedemann-Goiran P., Tezaur R., A discontinuous Galerkin method with plane waves and Lagrange multipliers for the solution of short wave exterior Helmholtz problems on unstructured meshes *Wave Motion*, **39**, 307-317, 2004

[7] Farhat C., Tezaur R., Wiedemann-Goiran P., Higher-order extensions of a discontinuous Galerkin method for mid-frequency Helmholtz problems *Internat. J. Numer. Methhods Eng.*, **61**, 1938-1956, 2004

[8] Franca L.P., Farhat C., Macedo A.P., Lesoinne M., Residual-free bubbles for the Helmholtz equation *Internat. J. Numer. Methhods Eng.*, **4O**, 4003-4009, 1997

[9] Hadamard J., Lectures on Cauchy's Problem in Linear Partial Differential Equations *Yale University Press, New Haven* 1923

[10] Harari I., Hughes T.J.R., Galerkin/least-squares finite element methods for the reduced wave equation with non-reflecting boundary conditions in unbounded domains *Comput. Methods Appli. Mech. Eng.*, **98**, 411-454, 1992

[11] Harari I., Hetmaniuk U., Private communication

[12] Hörmander L., The Analysis of Linear Partial Differential Operator *Springer-Verlag, New York* 1985

[13] Ihlenburg F., Finite Element Analysis of Acoustic Scattering *Appl. Math. Sci 132, Springer-Verlag, New York* 1998

[14] Magoulès F., Computational Methods for Acoustics Problems *Saxe-Coburg Publications* 2008

[15] Monk P., Wang D.Q., A least-squares method for the Helmholtz equation *Comput. Methods Appli. Mech. Eng.*, **175**, 411-454, 1999

[16] Rose M.E., Weak element approximations to elliptic differential equations *Numer. Math.*, **24**, 185-204, 1975

[17] Taylor M. E., Partial Differential Equations I : Basic Theory *Springer-Verlag, New York* 1997

# **Partie III :** An improved modified discontinuous Galerkin method for solving Helmholtz problems

# 1. Introduction

Recently, a modified discontinuous Galerkin method (mDGM) with plane wave basis functions and Lagrange multiplier degrees of freedom was proposed for solving efficiently Helmholtz problems in the mid-frequency regime. The method was designed to address the numerical instabilities that arise when using the discontinuous Galerkin method (DGM) designed by Farhat *et al* in [6, 7, 8]. We have observed that these instabilities deteriorate significantly the level of accuracy. For example, when solving waveguide-type problems, using the $R$-8-2b element (eight plane waves per element and two Lagrange multipliers), we have noticed that DGM becomes unstable when $kh \leq \frac{1}{6}$, which corresponds to about 38 elements per wavelength [2]. In addition, in the instability region the error increases by over two orders of magnitude arising from 1% to over 100%. These instabilities are due to the ill-posed character of the local problems as well as to the numerical loss of the linear independence of the shape functions as the mesh size tends to zero.
To address the instability issue that occurs, we have designed recently a method similar to DGM, called mDGM (*m*odified *d*iscontinuous *G*alerkin *m*ethod), in which one needs to solve local problems well posed and then a global problem associated, under a compatibility condition, with a positive definite Hermitian matrix. The numerical results reported in [2] indicate clearly that mDGM outperforms DGM. Nevertheless, mDGM is still sensitive to the numerical loss of the linear independence of the shape functions. Indeed, numerical instabilities still occur with the mesh refinement. However, mDGM retains an acceptable level of accuracy (about 1%) as the mesh size tends to zero.

We propose in this paper to modify the formulation adopted in mDGM. The new formulation has the potential to be more robust and stable and therefore to deliver results with a better level of accuracy with the mesh refinement. Therefore, we will refer to the proposed method as imDGM (*i*mproved *m*odified *d*iscontinuous *G*alerkin *m*ethod). Similarly to mDGM, imDGM is a two-step procedure in which we first solve well-posed local problems and then a global system issued from the continuity condition. The proposed method reformulates the local problems such that at the algebraic level we solve linear systems associated with positive definite Hermitian matrices. The global system to be solved in the second step is the same as in mDGM and corresponds to a positive semi-definite Hermitian matrix.
imDGM resembles in some aspect DGM. In addition, it is close to LSM, as explained below. All the three methods are based on a decomposition of the domain into quadrilateral- or triangular-shaped elements. In imDGM, as well as in DGM and LSM, the solution is approximated, at the element level, by a superposition of plane waves that are solution of Helmholtz equation. This leads to discontinuous functions. Similarly to DGM and unlike LSM, in imDGM the continuity of the solution at the interior interfaces is enforced by Lagrange multipliers. Recall that in DGM the Lagrange multipliers are continuous across the interior interfaces of the mesh and consequently the normal derivative is continuous too. The proposed method does not require the continuity of the normal derivative of the field and therefore the Lagrange multipliers are discontinuous. This allows us to consider well posed local problems, unlike the DGM in which weakly singular systems can occur. As stated earlier, the variational problems proposed in this paper lead to Hermitian positive definite local matrices. Using the solutions of the local problems we write a variational global problem, whose solutions are the Lagrange multipliers. The approximation of the corresponding bilinear form leads to a positive semi-definite Hermitian matrix. In LSM, the discontinuity that arises when approximating locally the solution is corrected by introducing a positive parameter in order to minimize the jumps of both the field and its normal derivative across each interior edge of the mesh. This approach leads to solving a linear system with a Hermitian matrix to obtain the discrete solution. The obtained variational problem and the corresponding linear system are very

similar to the global system that we solve in the second step of imDGM at the continuous and algebraic level respectively. The main difference between the two global problems is the size of the obtained systems, as it will be explained in Section 5.

The remainder of the paper is organized as follows. In Section 2 we introduce general notations and the model problem. Section 3 is devoted to the presentation of imDGM. In Section 4, we present the algebraic framework of the formulation. In Section 5 we compare the computational cost of imDGM to DGM and LSM. Illustrative numerical results comparing the performance of imDGM to mDGM, DGM and LSM are presented in Section 5. Finally, Section 6 concludes this paper.

## 2. Preliminaries

In this section, we introduce the model problem and specify the nomenclature and assumptions adopted throughout this paper.

### 2.1. The mathematical model

We consider the following class of waveguide-type problems :

$$(\text{BVP}) \begin{cases} -\Delta u - k^2 u &= f & \text{in } \Omega, \\ \partial_n u &= \mathrm{i}ku + g & \text{on } \partial\Omega \end{cases}$$

where $\Omega \subset \mathbb{R}^2$ is an open bounded region with a smooth boundary $\partial\Omega$. $k$ is a positive number representing the wavenumber. $\partial_n$ is the normal derivative. $f$ and $g$ are regular complex valued functions defined on the domain $\Omega$ and its boundary $\partial\Omega$ respectively. The second equation of BVP is a representation of a class of non-homogeneous Robin boundary conditions. Other types of boundary condition can be considered.
Note that BVP is considered here for its simplicity since it allows us to compute analytically the solution $u$ for a suitable choice of $\Omega$, $f$ and $g$. An explicit expression of $u$ is crucial for assessing the accuracy of the proposed solution methodology.

### 2.2. Nomenclature and assumptions

In what follows, we consider a regular triangulation $\tau_h$ of $\Omega$ into quadrilateral- or triangular-shaped subdomains $K$ whose boundaries are denoted by $\partial K$. The step size mesh discretization is denoted by $h$. We introduce the local space :

$$\mathcal{V}(K) = \left\{ v^K \in H^1(K),\, \Delta v^K + k^2 v^K = 0 \text{ in } K \right\}, \tag{1}$$

equipped with the following norm :

$$||v^K||_{\mathcal{V}(K)} = \left( ||v^K||^2_{1,K} + ||\partial_n v^K||^2_{0,\partial K} \right)^{\frac{1}{2}}, \tag{2}$$

where $||\cdot||_{m,K}$ is the $H^m$-norm on the element $K$ and $||\cdot||_{0,\partial K}$ is the $L^2$-norm on $\partial K$. In addition, we introduce $|\cdot|_{1,K}$ to designate the $H^1$-seminorm on the element $K$.

Next, we define the space of the primal variable as follows :

$$\mathcal{V} = \left\{ v \in L^2(\Omega);\ v^K = v|_K \in \mathcal{V}(K) \right\} \tag{3}$$

that we equip with the natural norm :

$$||v||_{\mathcal{V}} = \left( \sum_{K \in \tau_h} ||v^K||^2_{\mathcal{V}(K)} \right)^{\frac{1}{2}}, \quad \forall\, v \in \mathcal{V}.$$

Observe that $\mathcal{V}$ contains functions that are discontinuous across interior boundaries since their regularity is only $L^2(\Omega)$. Therefore, for any $v \in \mathcal{V}$, we define the jump across an interior edge $e = \partial K \cap \partial K'$ of two elements $K$ and $K'$ by :

$$[v] = v^K - v^{K'}.$$

We introduce the space of the dual variable, corresponding here to Lagrange multipliers, by :

$$\mathcal{M} = \left\{ \mu \in \prod_{K \in \tau_h} L^2(\partial K);\ \mu = 0 \text{ on } \partial K \cap \partial\Omega \right\}$$

and we associate to $\mathcal{M}$ the norm given by :

$$||\mu||_{\mathcal{M}} = \left( \sum_{K \in \tau_h} ||\mu^K||^2_{0,\partial K} \right)^{\frac{1}{2}}, \quad \forall\, \mu \in \mathcal{M},$$

where $\mu^K$ designates the restriction of $\mu$ to $\partial K$ : $\mu^K = \mu|_{\partial K}$.
For any function $\mu \in \mathcal{M}$, we define the jump across an interior edge $e = \partial K \cap \partial K'$ by :

$$[[\mu]] = \mu^K + \mu^{K'}.$$

## 3. The proposed solution methodology : The continuous approach

Similarly to mDGM [2], the basic idea of the proposed solution methodology, called imDGM, is to evaluate $u$ the solution of BVP using the following splitting :

$$u = \Phi(\lambda) + \varphi, \tag{4}$$

where $\varphi$ and $\Phi$ are elements of $\mathcal{V}$ and $\lambda \in \mathcal{M}$. These quantities are evaluated into two steps :

**Step 1** For all $\mu \in \mathcal{M}$, we compute $\varphi$ and $\Phi(\mu)$. This is achieved by solving a set of local Helmholtz problems.

**Step 2** We determine $\lambda \in \mathcal{M}$ by solving a global linear system to ensure the continuity in a weak sense of $u$ given by (4) and its normal derivative.

### 3.1. Step 1 : The restriction procedure

This step is devoted to the computation of $\varphi$ and $\Phi(\mu)$, for all $\mu \in \mathcal{M}$, by solving local Helmholtz problems. More specifically, for all $K \in \tau_h$, we determine $\varphi^K$ by solving the following boundary

value problem :

$$\text{(BVP1)} \left\{ \begin{array}{rll} \text{Find } \varphi^K \in \mathcal{V}(K) & \text{such that :} & \\ -\Delta\varphi^K - k^2\varphi^K & = f & \text{in } K \\ \partial_n\varphi^K & = \mathrm{i}k\varphi^K + g & \text{on } \partial K \cap \partial\Omega \\ \partial_n\varphi^K & = \mathrm{i}\alpha\varphi^K & \text{on } \partial K \cap \mathring{\Omega} \end{array} \right. ,$$

where $\mathring{\Omega}$ designates the interior domain of $\Omega$.

Next, for all $\mu \in \mathcal{V}$ and $K \in \tau_h$, we compute $\Phi\left(\mu^K\right)$ by solving the boundary value problem given by :

$$\text{(BVP2)} \left\{ \begin{array}{rll} \text{Find } \Phi\left(\mu^K\right) \in \mathcal{V}(K) & \text{such that :} & \\ -\Delta\Phi\left(\mu^K\right) - k^2\Phi\left(\mu^K\right) & = 0 & \text{in } K \\ \partial_n\Phi\left(\mu^K\right) & = \mathrm{i}k\Phi\left(\mu^K\right) & \text{on } \partial K \cap \partial\Omega \\ \partial_n\Phi\left(\mu^K\right) & = \mathrm{i}\alpha\Phi\left(\mu^K\right) + \mu^K & \text{on } \partial K \cap \mathring{\Omega} \end{array} \right. ,$$

with $\alpha \in \mathbb{R}_+^*$. Similarly to the mDGM formulation, we set $\alpha = k$.

We associate to BVP1 and BVP2 a variational formulation that can be expressed in the following compact form :

$$\left\{ \begin{array}{l} \text{Find } \Psi^K \in \mathcal{V}(K) \text{ such that :} \\ a_K\left(\Psi^K, v^K\right) = L_K\left(v^K\right), \qquad \forall v^K \in \mathcal{V}(K) \end{array} \right. , \tag{5}$$

where

$$a_K\left(v^K, w^K\right) = \int_{\partial K} \left(\partial_n v^K - \mathrm{i}\,kv^K\right) \overline{(\partial_n w^K - \mathrm{i}\,kw^K)} ds, \quad \forall\, v^K, w^K \in \mathcal{V}(K) \tag{6}$$

and

$$\Psi^K = \left\{ \begin{array}{ll} \varphi^K + \dfrac{1}{k^2|K|}\displaystyle\int_K f\,dx & \text{for BVP1} \\ \Phi\left(\mu^K\right) & \text{for BVP2}, \quad \forall \mu \in \mathcal{M}. \end{array} \right.$$

For any $v^K \in \mathcal{V}(K)$, the right-hand side $L_K(\cdot)$ is given by :

$$L_K\left(v^K\right) = \left\{ \begin{array}{ll} \displaystyle\int_{\partial K\cap\partial\Omega} \left(g^K - \frac{\mathrm{i}}{k|K|}\int_K f\,dx\right) \overline{(\partial_n v^K - \mathrm{i}\,kv^K)} ds & \\ \displaystyle -\frac{\mathrm{i}}{k|K|}\int_K f\,dx \int_{\partial K\cap\mathring{\Omega}} \overline{(\partial_n v^K - \mathrm{i}\,kv^K)} ds & \text{for BVP1} \\ \displaystyle\int_{\partial K\cap\mathring{\Omega}} \mu^K \overline{(\partial_n v^K - \mathrm{i}\,kv^K)} ds & \text{for BVP2.} \end{array} \right. \tag{7}$$

The following observations are noteworthy :

- Observe that the bilinear form $a_K(\cdot,\cdot)$ given by Eq. (6) is Hermitian. In addition, using Green's formula, one can easily verify that $a_K(\cdot,\cdot)$ can be expressed as follows :

$$a_k\left(v^K, w^K\right) = \int_{\partial K} \left(\partial_n v^K \partial_n \overline{w^K} + k^2 v^K \overline{w^K}\right) ds. \tag{8}$$

Note that we adopt the expression given by Eq. (8) at the discrete level.

- The bilinear form $a_K(\cdot,\cdot)$ given by Eq. (8) is continuous on $\mathcal{V}(K)\times\mathcal{V}(K)$. In addition, for a fixed $K\in\tau_h$, the variational problem given by Eq. (5) admits a unique solution.
- In mDGM, the bilinear form corresponding to the variational formulation of BVP1 and BVP2 is different from Eq. (8). It is given by :

$$b_K\left(v^K,w^K\right)=\int_{\partial K}\left(\partial_n v^K-\mathrm{i}kv^K\right)\overline{w^K}ds,\ \forall v^K,w^K\in H^1(K).$$

Consequently, the matrix corresponding to this bilinear form is neither Hermitian, nor symmetric. In addition, as it was observed in [2], this matrix becomes nearly singular as $h\to 0$ because the plane waves shape functions become (numerically) weakly linearly dependent.

- The variational formulation corresponding to BVP1 is based on the following assumption :

$$f|_K\approx\frac{1}{|K|}\int_K f dx, \tag{9}$$

that is $f$ is assumed to be constant in each element $K$ and its value is its mean value. The hypothesis (9) arises from the fact that we want to rewrite the first equation of BVP1 such that we have a homogeneous Helmholtz equation. Note that when $f=0$, which is the case for the waveguide problems and for most acoustic scattering problems, this hypothesis is always fulfilled. In addition, we have $\Psi^K=\varphi^K$ in the case of BVP1. In fact, when $f=0$, one can impose for $\Phi(\mu)$ the same boundary condition as for $u$, i.e. $\partial_n\Phi(\mu)=\mathrm{i}\,k\Phi(\mu)+g$ on $\partial K\cap\partial\Omega$. Consequently, we only need to solve BVP2. In this case, Step 1 consists in finding a solution $u$ of the form $u=\Phi(\lambda)$.
- Solving BVP1 and BVP2 requires solving *one* linear system with multiple right-hand side.

Next, similarly to mDGM, we define $\varphi$ such that for a given element $K$ of the mesh partition, the restriction of $\varphi$ to $K$ is $\varphi^K$. In addition, for all element $K$ and for all $\mu$ in $\mathcal{M}$, we define $\Phi(\mu)$ such that we have $\Phi(\mu)|_K=\Phi\left(\mu^K\right)$, where $\Phi\left(\mu^K\right)$ is the solution of BVP2.
Hence, solving one variational problem given by Eqs. (5), (8) and with different right-hand side given by Eq. (7) allows us to determine

$$\varphi+\Phi(\mu)\in\mathcal{V},\quad\forall\,\mu\in\mathcal{M}. \tag{10}$$

Step 1 can be viewed, to some extent, as a prediction step since we evaluate $\varphi+\Phi(\mu)$ for all $\mu\in\mathcal{M}$.

### 3.2. Step 2 : The optimization procedure

Similarly to mDGM, the objective here is to determine $\lambda\in\mathcal{M}$ for which the function given by Eq. (10) is in $H^1(\Omega)$. This requirement can be viewed as a correction stage since we select the best-fit Lagrange multiplier $\lambda$. The determination of $\lambda$ is accomplished by solving the following global variational problem :

$$(\mathrm{VF})\begin{cases}\text{Find } \lambda\in\mathcal{M} \text{ such that}\\ A(\lambda,\mu)=F(\mu),\quad\forall\,\mu\in\mathcal{M},\end{cases} \tag{11}$$

where the bilinear form $A(\cdot,\cdot)$ is given by :

$$\begin{aligned} A(\eta,\mu) = & \sum_{e\text{-interior edge}} \beta_e \int_e [\Phi(\eta)]\,\overline{[\Phi(\mu)]}ds \\ & + \sum_{e\text{-interior edge}} \gamma_e \int_e [[\partial_n\Phi(\eta)]]\,\overline{[[\partial_n\Phi(\mu)]]}ds \\ & + \sum_{e\subset\partial\Omega} \omega_e \int_e (\partial_n\Phi(\eta) - \mathrm{i}\,k\Phi(\eta))\,\overline{(\partial_n\Phi(\mu) - \mathrm{i}\,k\Phi(\mu))}ds, \end{aligned} \tag{12}$$

and the linear form $F(\cdot)$ is given by :

$$\begin{aligned} F(\mu) = & -\sum_{e\text{-interior edge}} \beta_e \int_e [\varphi]\,\overline{[\Phi(\mu)]}ds \\ & - \sum_{e\text{-interior edge}} \gamma_e \int_e [[\partial_n\varphi]]\,\overline{[[\partial_n\Phi(\mu)]]}ds \\ & - \sum_{e\subset\partial\Omega} \omega_e \int_e (\partial_n\varphi - \mathrm{i}\,k\varphi - g)\,\overline{(\partial_n\Phi(\mu) - \mathrm{i}\,k\Phi(\mu))}ds, \end{aligned} \tag{13}$$

The parameters $\beta_e$, $\gamma_e$ and $\omega_e$ are three positive numbers that can be viewed as weight parameters. The third integral in Eqs. (12)-(13) is theoretically equal to 0 (see the second boundary condition given by BVP1 and BVP2). However, the presence of this term at the algebraic level leads to a more robust and stable formulation. The variational problem (11) expresses the continuity in the weak sense of both the solution and its normal derivative. Note that the bilinear form $A$ is Hermitian. Consequently, only half of the corresponding matrix will be stored.

**Remark 1.** The expression of the bilinear form $A(\cdot,\cdot)$ is very similar to the one that arises in LSM. The main difference is that the continuity in LSM is imposed directly on the field. Consequently, the bilinear form is applied to functions contained in $\mathcal{V}$, whereas in the proposed method the unknowns are the Lagrange multipliers.

**Remark 2.** Observe that Step 2 is identical for both imDGM and mDGM with a minor difference : the matrix and the right-hand that incur at the algebraic level in imDGM differ from the ones obtained in mDGM because the computation of the solutions of BVP1 and BVP2 are different for the approach adopted in mDGM.

**Remark 3.** When $f = 0$, solving BVP and solving the problem arising in the proposed two-step procedure are equivalent in the following sense :
(i) Let $u = \Phi(\lambda) + \varphi$, where for all $K$, $\varphi^K$ is solution of BVP1, and $\Phi\left(\lambda^K\right)$ is solution of BVP2 with $\lambda$ solution of VF. Then $u$ is the unique solution of BVP.
(ii) Conversely, let $u$ be the solution of BVP. For each $K \in \tau_h$, we define $\lambda$ by :

$$\lambda^K = \begin{cases} 0 & \text{on} \quad e \subset \partial K \cap \partial\Omega \\ \partial_n u^K - \mathrm{i}\,\alpha u^K & \text{on} \quad e \subset \partial K \cap \mathring{\Omega} \end{cases} \tag{14}$$

Let $\varphi^K$ be the solution of BVP1 and $\Phi(\lambda^K)$ the solution of BVP2. Then $\lambda$ is solution of VF and $u = \Phi(\lambda) + \varphi$.

# 4. The proposed solution methodology : The algebraic approach

To derive the discrete version of imDGM, we need to consider first two finite-dimensional spaces $\mathcal{V}_h$ and $\mathcal{M}_h$ such that

$$\mathcal{V}_h \subset \mathcal{V} \quad \text{and} \quad \mathcal{M}_h \subset \mathcal{M}. \tag{15}$$

Similarly to mDGM, $\mathcal{V}_h$ is chosen to be the space of plane wave functions within each element $K$. Other shape functions satisfying Helmholtz equation, such as Bessel functions can also be considered [16]. For $\mathcal{M}_h$ we consider plane waves defined on the edges of the mesh. Note that imDGM allows, in principle, to choose the spaces $\mathcal{V}_h$ and $\mathcal{M}_h$ independently, which is not the case for DGM for which an inf-sup condition must be satisfied.
For any element $K \in \tau_h$, we denote by $\mathcal{V}_h(K)$ (resp. $\mathcal{M}_h(K)$) the set of functions of $\mathcal{V}_h$ (resp. $\mathcal{M}_h$) restricted to $K$ (resp. $\partial K$). Furthermore, $n^K$ (resp. $n^{\lambda^K}$) denotes the dimension of $\mathcal{V}_h(K)$ (resp. $\mathcal{M}_h(K)$). Last, the dimension of $\mathcal{M}_h$, which corresponds to the total number of dofs, is denoted by $n^\lambda$.

We show that when formulated in finite dimensional spaces, the proposed two-step procedure consists in solving linear algebraic systems in each step. In addition, all the local and global systems arising from the discretization are Hermitian and positive definite. This allows the storage of only half the matrix, at both the element and the global level, as well as the use of robust numerical algorithms such as the conjugate gradient scheme for solving the global system.

## 4.1. Step 1 : The restriction procedure

Let $\varphi_h^K$ and $\Phi_h\left(\mu_h^K\right)$ be the approximation of $\varphi^K$ and $\Phi\left(\mu_h^K\right)$ respectively. Similarly to the continuous formulation, we consider $\varphi_h$, $\Phi_h(\mu_h)$ and $\mu_h$ such that : $\varphi_h|_K = \varphi_h^K$, $\Phi_h(\mu_h)|_K = \Phi_h\left(\mu_h^K\right)$ and $\mu_h|_K = \mu_h^K$ for any element $K$ in the mesh.
The goal is to compute $\varphi_h$ and $\Phi_h(\mu_h)$, for all $K \in \tau_h$. To do this, we set the variational problem given by Eqs. (5), (7), (8) in the finite dimensional space $\mathcal{V}_h(K)$, that is :

$$\begin{cases} \text{Find } \Psi_h^K \in \mathcal{V}_h(K) \text{ such that :} \\ a_K\left(\Psi_h^K, v_h^K\right) = L_K\left(v_h^K\right), \quad \forall\, v_h^K \in \mathcal{V}_h(K) \end{cases} \tag{16}$$

where the forms $a_K(\cdot,\cdot)$ and $L_K(\cdot)$ are given by Eq. (8) and Eq. (7) respectively. The function $\Psi_h^K$ is given by :

$$\Psi_h^K = \begin{cases} \varphi_h^K + \frac{1}{k^2|K|}\int_K f dx & \text{for BVP1} \\ \Phi_h\left(\mu_h^K\right) & \text{for BVP2}, \quad \forall\, \mu_h \in \mathcal{M}_h. \end{cases} \tag{17}$$

Consequently, the variational problem (16)-(17) can be written in the following matrix form :

$$\left(\mathbf{P}^{\partial K} + k^2 \mathbf{S}^{\partial K}\right) \mathbf{X}^K = \text{rhs}, \tag{18}$$

where $\mathbf{P}^{\partial K}$ (resp. $\mathbf{S}^{\partial K}$) are stiffness-like (resp. mass-like) matrices defined on $\partial K$. The linear system (18) possesses the following properties :

- The matrix $\mathrm{P}^{\partial K}+k^2\mathbf{S}^{\partial K}$ is Hermitian. In addition, all its entries can be evaluated analytically when plane waves shape functions are used.
- For an element $K \in \tau_h$, the number of right-hand side is $n^{\lambda^K}+1$. We must point out that the obtained problems can be solved in parallel since they are independent from an element $K$ to another $K'$.

In addition, we have :

**Proposition 1.** For a given element $K$ of the mesh, the matrix $\mathbf{B}^K = \mathbf{P}^{\partial K} + k^2\mathbf{S}^{\partial K}$ of the local system given by Eq. (18) is positive definite.

**Proof of Proposition 1.** Let $\left\{\xi_j^K\right\}_{1\leq j\leq n^K}$ be a basis of $\mathcal{V}_h(K)$ and let $\boldsymbol{y} = {}^t[y_1, y_2, ..., y_{n^K}] \in \mathbb{C}^{n^K}$ be an ordinary vector. We set :

$$v_h^K = \sum_{1\leq l\leq n^K} y_l \xi_l^K \in \mathcal{V}_h(K). \tag{19}$$

Then, we have :

$$\boldsymbol{y}^*\mathbf{B}^K\boldsymbol{y} = \|\partial_n v_h\|^2_{L^2(\partial K)} + \|v_h\|^2_{L^2(\partial K)} \geq 0, \tag{20}$$

that is $\mathbf{B}^K$ is a positive semi-definite matrix. Suppose now that $\boldsymbol{y}$ is such that $\boldsymbol{y}^*\mathbf{B}^K\boldsymbol{y} = 0$. Then, it follows from Eq. (20) that :

$$\|\partial_n v_h^K\|^2_{L^2(\partial K)} + \|v_h^K\|^2_{L^2(\partial K)} = 0. \tag{21}$$

Therefore,

$$\partial_n v_h^K = 0 \text{ on } \partial K \quad \text{and} \quad v_h^K = 0 \text{ on } \partial K$$

Using the continuation theorem [13, 18], we deduce that $v_h^K = 0$ in $K$. Consequently, $y_l = 0$ for all $1 \leq l \leq n^K$. Thus, $\boldsymbol{y}^*\mathbf{B}^K\boldsymbol{y} > 0, \forall \boldsymbol{y} \in \mathbb{C}^{n^K} \setminus \{0\}$, that is $\mathbf{B}^K$ is a positive definite matrix. ■

**Remark 4.** Observe that in the case of mDGM the resulting system is neither Hermitian, nor symmetric, but nevertheless invertible. This is the main difference between mDGM and the proposed method imDGM. However, this difference has the potential to provide more stability and robustness for the proposed method.

### 4.2. Step 2 : The optimization procedure

We formulate the global variational problem (VF) in finite dimension. We set :

$$\begin{cases} \text{Find } \lambda_h \in \mathcal{M}_h \text{ such that :} \\ A_h(\lambda_h, \mu_h) = F_h(\mu_h), \quad \forall\, \mu_h \in \mathcal{M}_h \end{cases} \tag{22}$$

where the forms $A_h(\cdot,\cdot)$ and $F_h(\cdot)$ are obtained from $A(\cdot,\cdot)$ and $F(\cdot)$ respectively by replacing $\varphi$ with $\varphi_h$ and $\Phi(\mu_h)$ with $\Phi_h(\mu_h)$, for $\mu_h \in \mathcal{M}_h$. Hence, solving the variational problem (22) comes to solve the following linear algebraic system :

$$\mathbf{A}\boldsymbol{\Lambda} = \mathbf{b}, \tag{23}$$

where the entries of the matrix $\mathbf{A}$ and of the vector $\mathbf{b}$ are given by :

$$\begin{aligned}\mathbf{A}_{lm} = & \sum_{e\text{-interior edge}} \beta_e \int_e [\Phi_h(\mu_m)]\,\overline{[\Phi_h(\mu_l)]}ds \\ & + \sum_{e\text{-interior edge}} \gamma_e \int_e [[\partial_n\Phi_h(\mu_m)]]\,\overline{[[\partial_n\Phi_h(\mu_l)]]}ds \\ & + \sum_{e\subset\partial\Omega} \omega_e \int_e (\partial_n\Phi_h(\mu_m) - \mathrm{i}\,k\Phi_h(\mu_m))\,\overline{(\partial_n\Phi_h(\mu_l) - \mathrm{i}\,k\Phi_h(\mu_l))}ds,\end{aligned} \tag{24}$$

and

$$\begin{aligned}\mathbf{b}_l = & -\sum_{e\text{-interior edge}} \beta_e \int_e [\varphi_h]\,\overline{[\Phi_h(\mu_l)]}ds \\ & - \sum_{e\text{-interior edge}} \gamma_e \int_e [[\partial_n\varphi_h]]\overline{[[\partial_n\Phi_h(\mu_l)]]}ds \\ & - \sum_{e\subset\partial\Omega} \omega_e \int_e (\partial_n\varphi_h - \mathrm{i}\,k\varphi_h - g)\,\overline{(\partial_n\Phi_h(\mu_l) - \mathrm{i}\,k\Phi_h(\mu_l))}ds\end{aligned} \tag{25}$$

The unknown $\boldsymbol{\Lambda}$ is a vector in $\mathbb{C}^{n^\lambda}$ whose components are the values of $\lambda_h$ in the basis of $\mathcal{M}_h$.

**Remark 5.** The linear system (22)-(25) is identical to the one obtained in mDGM. Consequently, the computational complexity (cost and implementation effort) is the same for both methods. Note that the matrix $\mathbf{A}$ is Hermitian. In addition, we have :

**Proposition 2.** The matrix $\mathbf{A}$ is positive semi-definite. Furthermore, if the spaces $\mathcal{V}_h$ and $\mathcal{M}_h$ satisfy the following condition :

$$\int_{\partial K} \mu_h^K \left(\partial_n v_h^K - \mathrm{i}\,k v_h^K\right) ds = 0\,, \quad \forall\, v_h^K \in \mathcal{V}_h(K) \quad \Longrightarrow \mu_h^K = 0, \tag{26}$$

then the matrix $\mathbf{A}$ is positive definite.

**Proof of Proposition 2.** Let $\{\mu_j\}_{1\le j\le n^\lambda}$ be a basis of $\mathcal{M}_h$ and let $\boldsymbol{y} = {}^t[y_1, y_2, ..., y_{n^\lambda}] \in \mathbb{C}^{n^\lambda}$ be an ordinary vector. We set :

$$\eta_h = \sum_{1\le l\le n^\lambda} y_l\mu_l. \tag{27}$$

Therefore, it is easy to verify that :

$$\Phi_h(\eta_h) = \sum_{1\leq l\leq n^\lambda} y_l \Phi_h(\mu_l). \tag{28}$$

Observe that the matrix $\mathbf{A}$ satisfies :

$$\begin{aligned} \boldsymbol{y}^*\mathbf{A}\boldsymbol{y} = & \sum_{e-\text{interior edge}} \left(\beta_e \| [\Phi_h(\eta_h)] \|^2_{L^2(e)} + \gamma_e \| [[\partial_n \Phi_h(\eta_h)]] \|^2_{L^2(e)}\right) \\ & + \sum_{e\subset\partial\Omega} \omega_e \| \partial_n \Phi_h(\eta_h) - \mathrm{i}\, k \Phi_h(\eta_h) \|^2_{L^2(e)} \geq 0, \end{aligned} \tag{29}$$

that is $\mathbf{A}$ is a positive semi-definite matrix.
Assume now that (26) is fulfilled. Let $\boldsymbol{y}$ be a vector in $\mathbb{C}^{n^\lambda}$ such that $\boldsymbol{y}^*\mathbf{A}\boldsymbol{y} = 0$. Since for any interior edge $e$, $\beta_e > 0$ and $\gamma_e > 0$, from Eq. (29) we deduce :

- $[\Phi_h(\eta_h)] = 0$ on all interior edges $e$. Hence, $\Phi_h(\eta_h) \in H^1(\Omega)$.
- $[[\partial_n \Phi_h(\eta_h)]] = 0$ on all interior edges. Using the fact that $\Delta\Phi_h\left(\eta_h^K\right) \in L^2(K)$, we deduce that $\Delta\Phi_h(\eta_h) \in L^2(\Omega)$.
- $\partial_n \Phi_h(\eta_h) = \mathrm{i}\, k \Phi_h(\eta_h)$ on all the boundary edges of the domain.

Therefore, $\Phi_h(\eta_h)$ verifies :

$$\left\{ \begin{array}{rll} -\Delta\Phi_h(\eta_h) - k^2\Phi_h(\eta_h) & = 0 & \text{in } \Omega \\ \partial_n \Phi_h(\mu_h) & = \mathrm{i}k\Phi_h(\mu_h) & \text{on } \partial\Omega \end{array} \right. .$$

This boundary value problem admits a unique solution. Therefore $\Phi_h(\eta_h) = 0$ in $\Omega$. It follows from Eq. (16) and Eq. (26) that $\eta_h^K = 0, \forall K \in \tau_h$. Consequently, $\eta_h = 0$ and therefore, $y_l = 0$ for all $1 \leq l \leq n^\lambda$. Thus, $\boldsymbol{y}^*\mathbf{A}\boldsymbol{y} > 0, \forall \boldsymbol{y} \in \mathbb{C}^{n^\lambda} \setminus \{0\}$, that is $\mathbf{A}$ is a positive definite matrix. ■

## 5. The proposed solution methodology : Computational cost

Similarly to the DGM formulation designed by Farhat *et al* in [6, 7, 8] and to the mDGM formulation proposed in [2], the computational cost depends mainly on the number of Lagrange multipliers. Indeed, the local problems that incur in Step 1 are small linear systems (see Eq. (18)) and therefore can be solved using $LU$ factorization . In addition, these systems can be solved in parallel since the local problems are independent from one element to another. The increase in the number of the shape functions has a very little effect on the total computational cost. Consequently, the cost of the method is given mainly by the size and the sparsity pattern of the global matrix given by Eq. (24), whose unknowns are the Lagrange multipliers. This size is directly related to the global number of dofs and therefore to the number of interior edges. On the other hand, the cost of the LSM proposed by Monk-Wang in [16] depends on the number of plane waves that approximate the solution at the element level, and therefore to the number of elements in the mesh.
For illustration purpose, consider a square-shaped computational domain to which is applied a rectangular-shaped $n \times n$ mesh. Note that $n$ denotes the number of elements in one direction. Recall that $n^{\lambda^K}$ denotes the number of dofs for an element $K$. Suppose that, at the element level $K$, the

solution is approximated using $m$ plane waves. We report in Table I the asymptotic size of the solution vector and the stencil width for the three methods : imDGM, DGM and LSM. Observe

TAB. 1 – Asymptotic size of the solution vector and stencil width for a $n \times n$ rectangular-shaped uniform mesh.

| Method | Asymptotic size of the solution vector | Stencil width |
|---|---|---|
| imDGM | $(4n^{\lambda^K})n^2$ | $20n^{\lambda^K}$ |
| DGM | $(2n^{\lambda^K})n^2$ | $7n^{\lambda^K}$ |
| LSM | $m\,n^2$ | $5m$ |

that imDGM has twice as many Lagrange multipliers as DGM. Doubling the number of Lagrange multipliers is a key aspect of the proposed approach for ensuring the well-posedness character of the local boundary value problems. The stencil is three times larger in imDGM than in DGM. This increase in the computational cost is the price to pay for obtaining a global positive definite matrix in imDGM.
Note that in imDGM the number of plane waves considered at the element level does not affect the computational cost, which is not the case for LSM. This means that approximating the solution locally with different number of plane waves, while introducing the same number of dofs, leads to the same size and stencil width of the matrix. The computational cost in LSM is smaller than in imDGM when using lower order elements. However, LSM seems to be more expensive for using higher order elements ($m > 4n^{\lambda^K}$), which is in practice what is recommended to use to improve the level of accuracy for higher frequencies [16]. This higher computation cost could be more significant for three dimensional problems.

## 6. The proposed solution methodology : Performance assessment

In order to illustrate and assess the potential of imDGM for solving efficiently Helmholtz problems, we have performed numerical experiments using plane waves shape functions and we have compared the results to those obtained with mDGM, DGM and LSM. The Lagrange multiplier is approximated on each edge using a subset or all set of shape functions that occur when evaluating $\partial_n v_h^K - \mathrm{i}\,k v_h^K$, for $v_h^K \in \mathcal{V}_h(K)$.
From now on, we assume $\Omega$ to be an $a \times a$ square-shaped domain. We use a uniform mesh partition of $\Omega$ in rectangular-shaped elements $K$. The functions $f$ and $g$ are such that the exact solution $u$ of BVP is a plane wave propagating in a direction $\boldsymbol{d} = (\cos\theta, \sin\theta)$ :

$$u(x,y) = e^{x\cos\theta + y\sin\theta}. \tag{30}$$

We vary the propagation angle $\theta$ in the interval $[0, 2\pi)$. In order to compare the results obtained with imDGM to those delivered by mDGM, DGM, and LSM, we measure for each propagating angle $\theta$, the relative error using the following modified $H^1$-norm [6] :

$$\|v\|_{\widehat{H^1}} = \left( \sum_K \|v\|^2_{H^1(K)} + \sum_{e-\text{interior edge}} \|[v]\|^2_{L^2(e)} \right)^{\frac{1}{2}}, \quad \forall v \in \mathcal{V}. \tag{31}$$

Note that Eq. (31) is a modified $H^1$-norm since it takes into account the $H^1$-norm at the element level and the jump of the numerical solution along the interior interfaces of the mesh. We also use

the *total* relative error, that is the *mean* value of the relative error obtained when $\theta \in [0, 2\pi)$.
For all numerical experiments, unless otherwise specified, we have set $\beta_e = k^2$, $\gamma_e = 1$ and $\omega_e = 1$.
We first present the results of the comparison of imDGM to mDGM. Then, we compare the performance of imDGM to LSM and DGM.

## 6.1. Comparison with mDGM

In this section we compare imDGM to mDGM to illustrate the potential benefit of adopting imDGM approach for solving Helmholtz problems. Recall that the computational complexity (computational cost and implementation effort) is identical for both methods.

### Approximation with eight plane waves

We approximate the primal variable, at the element level, using eight plane waves, positioned at :

$$\theta_p = (p-1)\pi/4, \ \forall 1 \leq p \leq 8. \tag{32}$$

This choice corresponds to the following discrete space for the primal variable :

$$\begin{aligned} \mathcal{V}_h(K) \ = \ & \Big\{ v_h|^K = \sum_{1\leq p\leq 8} e^{\mathrm{i}\, k \boldsymbol{\theta}_p \cdot \boldsymbol{x}} u_p, \ \boldsymbol{\theta}_p = {}^t[\cos\theta_p, \sin\theta_p], \\ & \theta_p = (p-1)\pi/4, 1 \leq p \leq 8, \ u_p \in \mathbb{C} \Big\}. \end{aligned} \tag{33}$$

Similarly to mDGM, for an element $v_h^K \in \mathcal{V}_h(K)$, the full approximation of $\partial_n v_h^K - \mathrm{i}\, k v_h^K$ leads to five dofs per edge. More specifically in imDGM, the discrete space $\mathcal{M}_h$ corresponding to the full approximation of the Lagrange multiplier is given by :

$$\begin{aligned} \mathcal{M}_h \ = \ & \Big\{ \mu_h \in \mathcal{M}; \forall K \in \tau_h, \mu_h^K|_e = \mu_1^K + \mu_2^K e^{\mathrm{i}\, kx} + \mu_3^K e^{-\mathrm{i}\, kx} + \mu_4^K e^{\mathrm{i}\, k\frac{\sqrt{2}}{2}x} \\ & + \mu_5^K e^{-\mathrm{i}\, k\frac{\sqrt{2}}{2}x} \text{ if } e \parallel \overrightarrow{x}, \ \mu_h^K|_e = \mu_1^K + \mu_2^K e^{\mathrm{i}\, ky} + \mu_3^K e^{-\mathrm{i}\, ky} \\ & + \mu_4^K e^{\mathrm{i}\, k\frac{\sqrt{2}}{2}y} + \mu_5^K e^{-\mathrm{i}\, k\frac{\sqrt{2}}{2}y} \text{ if } e \parallel \overrightarrow{y}, \quad \mu_1, \mu_2, \mu_3, \mu_4, \mu_5 \in \mathbb{C} \Big\}. \end{aligned} \tag{34}$$

Following the nomenclature introduced in [6], such an approximation is called the $R$-8-5 element. Nevertheless, both imDGM and mDGM can be implemented using less dofs per edge for the Lagrange multiplier. Removing the constant from the Lagrange multiplier field leads to the following approximation of $\lambda$ :

$$\lambda_h = \mu_1 e^{\mathrm{i}\, ks} + \mu_2 e^{-\mathrm{i}\, ks} + \mu_3 e^{\mathrm{i}\, k\frac{\sqrt{2}}{2}s} + \mu_4 e^{-\mathrm{i}\, k\frac{\sqrt{2}}{2}s}, \tag{35}$$

where $s$ is the curvilinear abscissa. We use the acronym $R$-8-4 to designate the element given by Eq. (32) and Eq. (35).
We also consider the $R$-8-3 element, in which the Lagrange multiplier is given, on each interior edge, by :

$$\lambda_h = \mu_1 + \mu_2 e^{\mathrm{i}\, k\frac{\sqrt{2}}{2}s} + \mu_3 e^{-\mathrm{i}\, k\frac{\sqrt{2}}{2}s}. \tag{36}$$

Last, we compare the performance of imDGM and mDGM, when equipped with the so-called $R$-8-2b element [6], which corresponds on each interior edge to the following approximation :

$$\lambda_h = \mu_1 e^{\mathrm{i}\, k \frac{\sqrt{2}}{4} s} + \mu_2 e^{-\mathrm{i}\, k \frac{\sqrt{2}}{4} s}. \tag{37}$$

The first experiment consists in comparing the error delivered by both methods equipped with the $R$-8-2b, $R$-8-3, $R$-8-4 and $R$-8-5 elements for different values of $ka$, while maintaining $kh$ constant. More specifically, we set $ka = 10, 20, 30$ and we fix the step size of the mesh discretization such that $kh = \frac{1}{2}$, which corresponds to about 12 elements per wavelength. The results depicted in Figs. 1-3 suggest the following :

- In the case of the $R$-8-2b element, imDGM clearly outperforms mDGM, as illustrated in Figs. 1(a)-1(c). Indeed, imDGM improves the level of accuracy by - at least - a factor two. In addition, imDGM retains longer the accuracy than mDGM as the frequency is increased. For example, the error delivered by imDGM for $ka = 30$ is less than 10%, whereas the error corresponding to mDGM is about 30%.
- For higher-order elements (corresponding to $R$-8-3, $R$-8-4 and $R$-8-5 elements), imDGM seems to perform slightly better than mDGM. Indeed, the gain in the accuracy level is not really significant (see Figs. 1(d)-1(f) and Fig. 2).
- The results reported in Fig 3 suggest using three Lagrange multipliers with eight plane waves to achieve the best level of accuracy compared to the computational cost. Thus, $R$-8-3 seems to be the best element (accuracy *vs.* computational cost) for imDGM when using eight plane waves.

Next, we compare the stability of both methods. To do this, we investigate the sensitivity of the total relative error with respect to $h/a$, the mesh step size. We also examine the dependence of the smallest eigenvalue of the local system with respect to $h/a$. We perform such an analysis when $ka = 1$ in order to reach over thousand elements per wavelength (an overkilled mesh) for two sets of weight parameters : (1) $\beta_e = k^2$, $\delta_e = 1$ and $\omega_e = 1$ and (2) $\beta_e = k^2$, $\delta_e = h$ and $\omega_e = h$. The results are reported in Figure 4 and Table 2. The following observations are noteworthy :

Tab. 2 – Dependence with respect to the mesh size of the total relative error and smallest eigenvalue of the local matrix for $R$-8-2b and $ka = 1$.

| $h/a$ | imDGM | | mDGM | |
|---|---|---|---|---|
| | Total relative error | The smallest eigenvalue | Total relative error | The smallest eigenvalue |
| 1/5 | 0.03% | $2.2 \cdot 10^{-09} + 2.9 \cdot 10^{-17}\mathrm{i}$ | 0.03% | $9.5 \cdot 10^{-11} - 4.1 \cdot 10^{-12}\mathrm{i}$ |
| 1/10 | 0.009% | $1.7 \cdot 10^{-11} - 1.8 \cdot 10^{-18}\mathrm{i}$ | 0.01% | $3.7 \cdot 10^{-13} - 7.9 \cdot 10^{-15}\mathrm{i}$ |
| 1/15 | 0.004% | $1.0 \cdot 10^{-12} + 1.7 \cdot 10^{-18}\mathrm{i}$ | 0.01% | $1.5 \cdot 10^{-14} - 2.4 \cdot 10^{-16}\mathrm{i}$ |
| 1/20 | 0.01% | $1.4 \cdot 10^{-13} - 1.1 \cdot 10^{-18}\mathrm{i}$ | 1.0% | $1.5 \cdot 10^{-15} + 5.9 \cdot 10^{-17}\mathrm{i}$ |
| 1/25 | 0.03% | $2.8 \cdot 10^{-14} + 1.5 \cdot 10^{-18}\mathrm{i}$ | 0.8% | $2.4 \cdot 10^{-16} + 4.4 \cdot 10^{-17}\mathrm{i}$ |
| 1/40 | 0.3% | $1.0 \cdot 10^{-15} - 7.2 \cdot 10^{-19}\mathrm{i}$ | 0.5% | $1.8 \cdot 10^{-17} + 1.6 \cdot 10^{-17}\mathrm{i}$ |
| 1/50 | 0.3% | $1.7 \cdot 10^{-16} + 2.6 \cdot 10^{-19}\mathrm{i}$ | 0.5% | $-4.4 \cdot 10^{-18} + 6.3 \cdot 10^{-18}\mathrm{i}$ |
| 1/70 | 0.3% | $1.3 \cdot 10^{-17} - 1.7 \cdot 10^{-19}\mathrm{i}$ | 0.3% | $-1.1 \cdot 10^{-18} + 5.4 \cdot 10^{-18}\mathrm{i}$ |
| 1/100 | 0.5% | $-4.5 \cdot 10^{-17} + 7.8 \cdot 10^{-20}\mathrm{i}$ | 0.3% | $3.0 \cdot 10^{-18} + 2.4 \cdot 10^{-17}\mathrm{i}$ |
| 1/180 | 2% | $5.5 \cdot 10^{-18} + 1.5 \cdot 10^{-20}\mathrm{i}$ | 2% | $-7.8 \cdot 10^{-18} - 2.6 \cdot 10^{-18}\mathrm{i}$ |

- A first glance at Fig. 4 indicates that both methods exhibit numerical instabilities. These

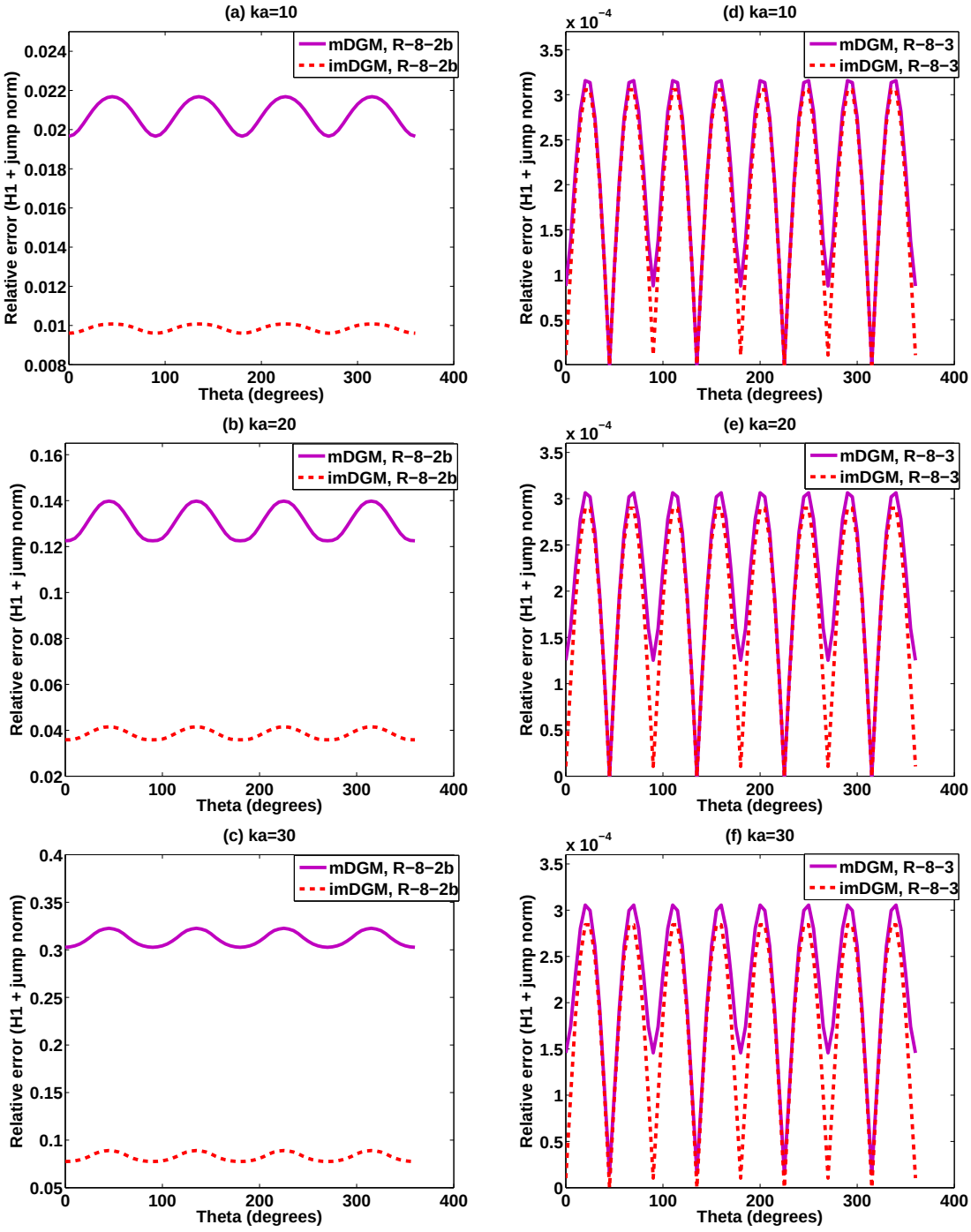


**Fig. 1** – Comparison between mDGM and imDGM in the case of $R$-8-2b element (left) and $R$-8-3 element (right) element for $kh = 1/2$.

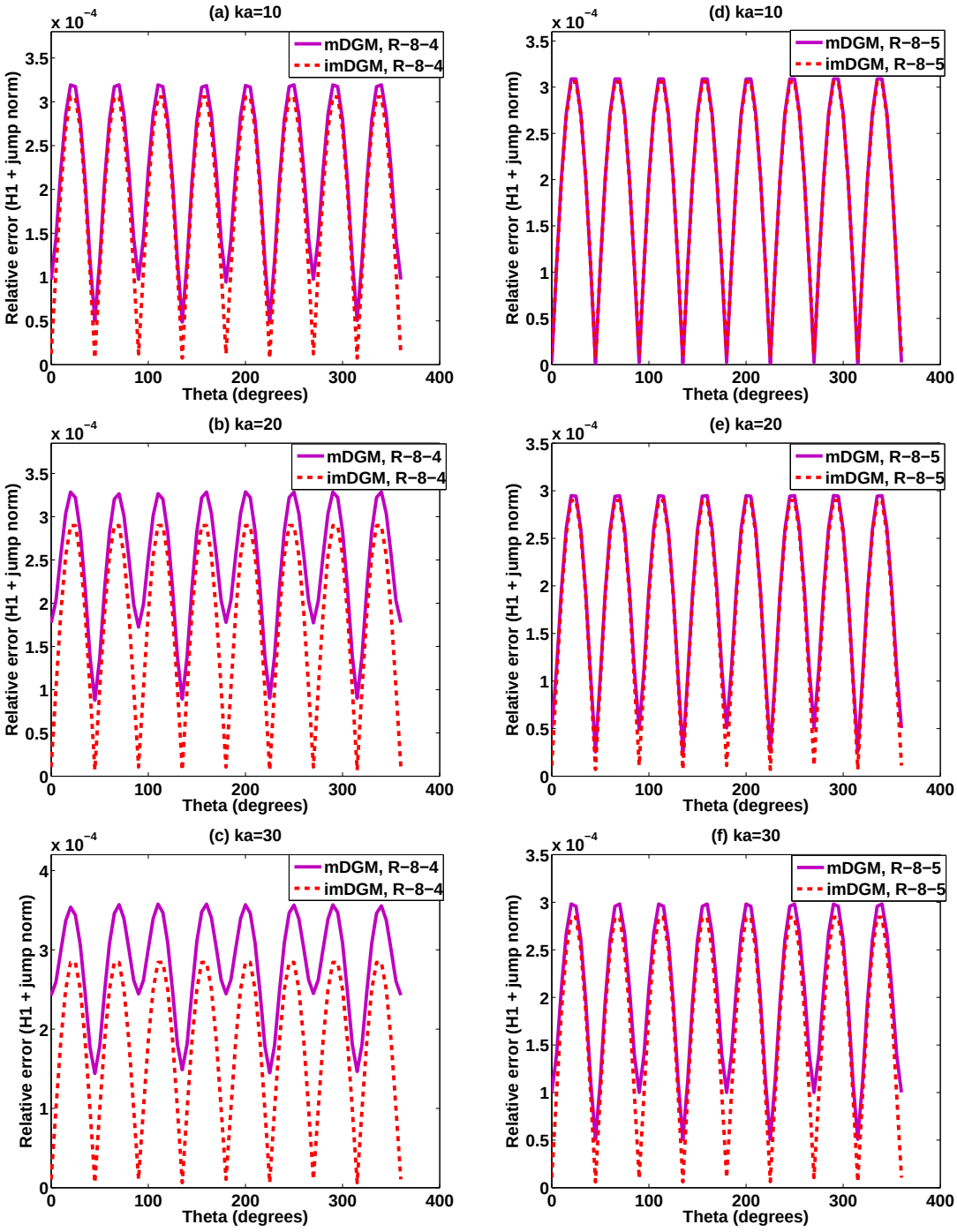


**Fig. 2** – Comparison between mDGM and imDGM in the case of $R$-8-4 element (left) and $R$-8-5 element (right) element for $kh = 1/2$.

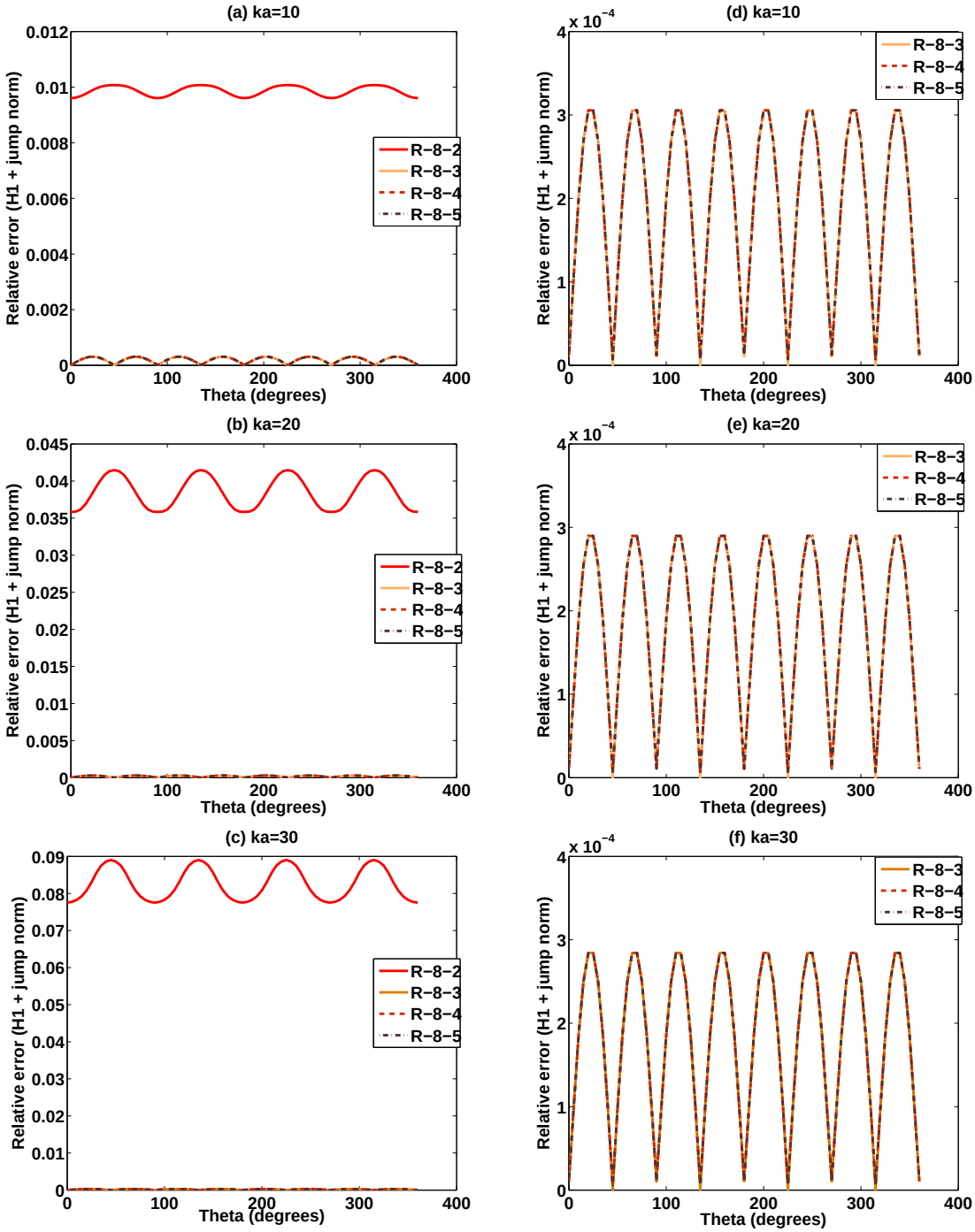


**Fig. 3** – Performance of imDGM for $kh = 1/2$ : Comparison between the elements.

numerical instabilities seem to be related to the local systems that become nearly singular, as illustrated in Table 2 by the values of the smallest eigenvalue. We believe that the eigenvalues tend to zero because of the loss of the linear independence of plane waves shape functions at the element level as the size of the elements becomes very small ($h/a < \frac{1}{40}$ for mDGM and $h/a < \frac{1}{70}$ for imDGM). Table 2 also reveals that the smallest eigenvalue in imDGM tends to zero slower than the one corresponding to mDGM. For example, for $h/a = \frac{1}{40}$ (over 240 elements per wavelength) there is a difference of two orders of magnitude between the two methods. This shows that imDGM remains numerically stable longer than mDGM.

- We note that these numerical instabilities do not deteriorate significantly the accuracy as imDGM retains an acceptable level of accuracy (less than 1%) as $h/a$ tends to zero.
- The presence of these numerical instabilities, in spite of their little effect on the accuracy, suggests to use higher order elements rather then refining the mesh to improve the level of accuracy.
- We have noticed that for $kh > \frac{1}{100}$, imDGM is not sensitive to the choice of the parameters $\beta_e, \delta_e, \omega_e$, whereas for $kh < \frac{1}{100}$, corresponding to over 600 elements per wavelength, the results tend to indicate that the second choice is more appropriate.
- Note that when using overkilled mesh (over thousand elements per wavelength), we have observed isolated values of $h/a$ where the error is over 20%. This phenomenon resembles the behavior of internal resonance-type points. These points occur much more later when adopting the second choice of parameters ($\beta_e = k^2$, $\delta_e = \omega_e = h$). We must point that we have not observed this phenomenon for $kh > \frac{1}{100}$, which is about less than 600 elements per wavelength.

**Approximation with seven plane waves**

In this paragraph, we compare the performance of both methods in the case of the $R$-7-2 element introduced in [2]. This element corresponds to seven plane waves per element, positioned at :

$$\theta_p = 2(p-1)\pi/7, \ \forall 1 \leq p \leq 7. \tag{38}$$

The Lagrange multipliers are given on each side of an interior edge by Eq. (37). Note that it has been observed in [2] that this element is more robust than the $R$-8-3 element because the linear independence of the seven shape functions is less sensitive to the mesh refinement. Therefore, the local problems that result for the $R$-7-2 discretization have a better condition number.

We have performed two sets of experiments. In the first set, we compare the relative error at each propagation angle $\theta$ (see Eq. (30)) delivered by both methods for three frequency values : $ka = 10$, 20, 30 while maintaining $kh = \frac{1}{2}$, which corresponds to about 12 elements per wavelength. The results are depicted in Fig. 5. In the second set of experiments, we analyze the robustness of both methods to the mesh refinement. We set $ka = 1$ and refine the mesh reaching a resolution of about 1200 elements per wavelength. The results are reported in Fig. 6 and Table 3. We also compare the performance of the $R$-8-3 and $R$-7-2 elements for the imDGM formulation. The results are depicted in Fig. 7. These numerical experiments reveal the following :

- imDGM outperforms mDGM, as clearly illustrated in Fig. 5. For example, when $ka = 30$ imDGM reduces the mDGM total relative error by almost a factor 5.
- Fig. 6 indicates that both methods are stable, but imDGM appears to be more robust and therefore more accurate. This is not surprising since the local problems in imDGM are expected

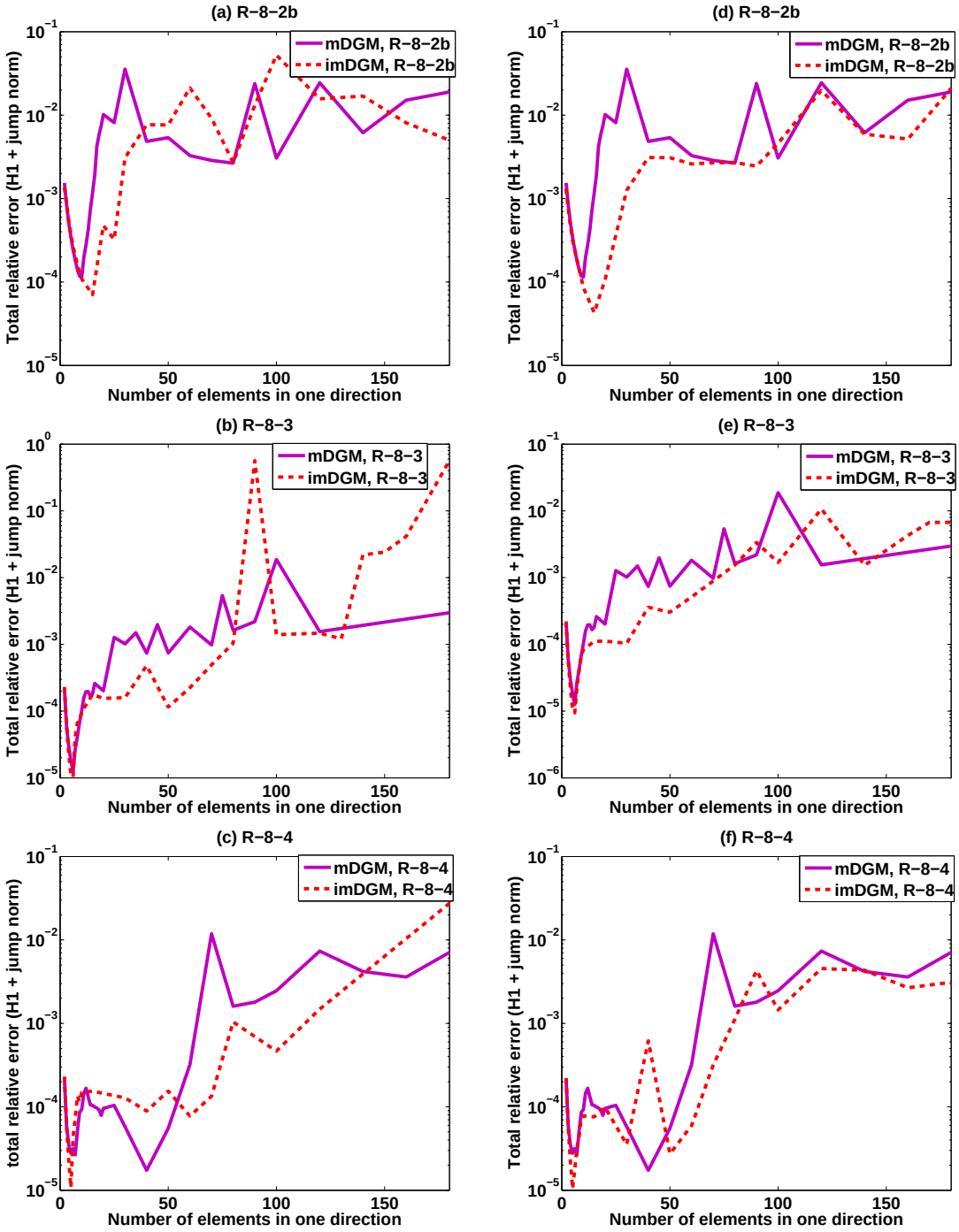


**Fig. 4** – Sensitivity of the accuracy to the mesh refinement : $ka = 1$ and weight parameters values : $\beta_e = k^2$, $\delta_e = \omega_e = 1$ (left), $\beta_e = k^2$, $\delta_e = \omega_e = h$ (right).

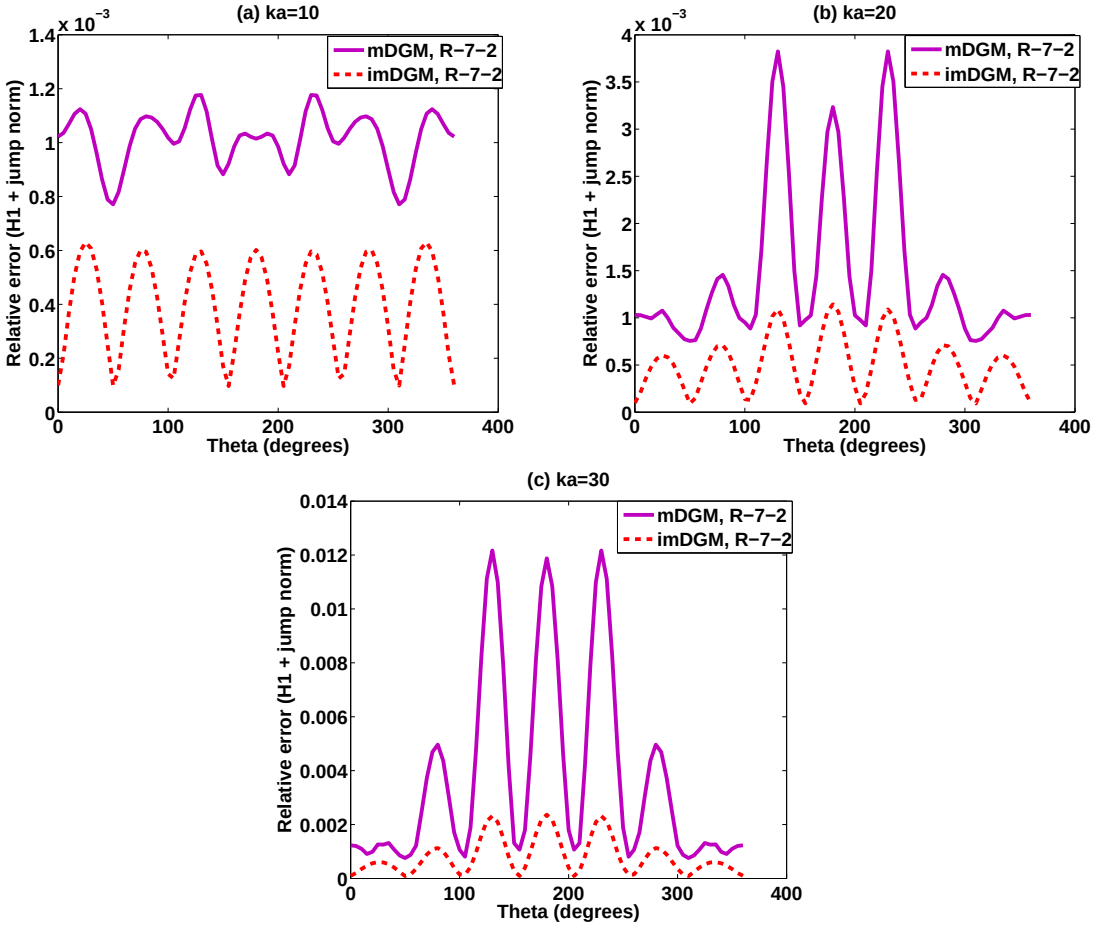


**Fig. 5** – Comparison between mDGM and imDGM in the case of $R$-7-2 element for $kh = 1/2$.

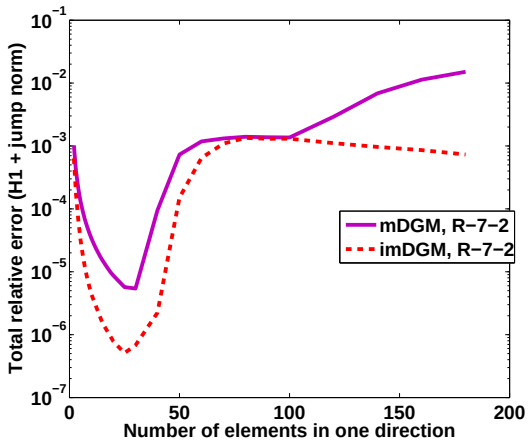


**Fig. 6** – Sensitivity of the total relative error to the mesh refinement : Comparison between mDGM and imDGM in the case of $R$-7-2 element and $ka = 1$.

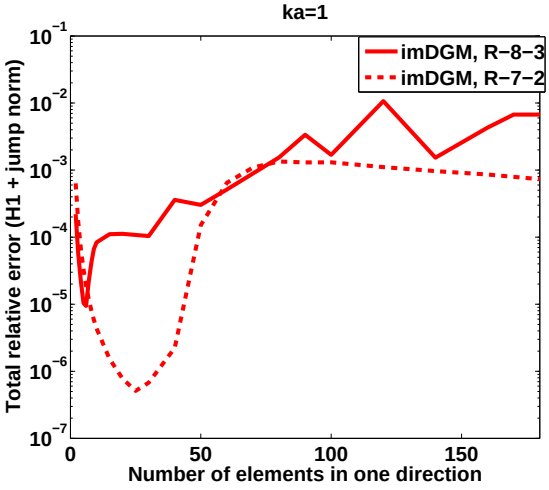


**Fig. 7** – Sensitivity of the total relative error to the mesh refinement : Comparison between imDGM ($R$-8-3) and imDGM ($R$-7-2) for $ka = 1$.

to be better conditioned than in mDGM, as demonstrated by Table 3. Indeed, as $h/a$ tends to zero, there are two orders of magnitude difference in the value of the smallest eigenvalue between the two methods. Table 3 shows that the local problems in mDGM tend to be weakly singular while the ones arising in imDGM remain more stable.

- Fig. 6 also suggests that it is better to use higher elements than refining the mesh. Indeed, for $h/a = \frac{1}{25}$, corresponding to 150 elements per wavelength, the total relative error is about 0.00001%. This is already a very fine mesh considering the value of $ka$ ($ka = 1$). Then, the error deteriorates as we refine the mesh and stagnates at 0.1% for over 600 elements per wavelength. This loss in the accuracy, although not dramatic, might be avoided by using higher order elements rather than refining the mesh.
- Last, Fig. 7 illustrates the superiority of $R$-7-2 over $R$-8-3. Indeed, $R$-7-2 delivers a better level of accuracy than $R$-8-3 because of the robustness of the local systems that arise in $R$-7-2. Recall that $R$-7-2 has a lower computational cost than $R$-8-3 (about 50% less dofs for the Lagrange multiplier).

TAB. 3 – Dependence with respect to the mesh size of the total relative error and smallest eigenvalue of the local matrix for $R$-7-2 and $ka = 1$.

| $h/a$ | imDGM | | mDGM | |
|---|---|---|---|---|
| | Total relative error | The smallest eigenvalue | Total relative error | The smallest eigenvalue |
| 1/5 | 0.003% | $3.1 \cdot 10^{-06} - 4.7 \cdot 10^{-17}$i | 0.01% | $1.4 \cdot 10^{-07} - 6.6 \cdot 10^{-09}$i |
| 1/10 | 0.0004% | $9.7 \cdot 10^{-08} - 5.4 \cdot 10^{-18}$i | 0.003% | $2.1 \cdot 10^{-09} - 5.2 \cdot 10^{-11}$i |
| 1/15 | 0.0002% | $1.3 \cdot 10^{-08} + 2.8 \cdot 10^{-21}$i | 0.0001% | $1.9 \cdot 10^{-10} - 3.0 \cdot 10^{-12}$i |
| 1/20 | 0.00008% | $3.0 \cdot 10^{-09} + 1.2 \cdot 10^{-18}$i | 0.0009% | $3.3 \cdot 10^{-11} - 4.1 \cdot 10^{-13}$i |
| 1/25 | 0.00005% | $1.0 \cdot 10^{-10} - 1.3 \cdot 10^{-19}$i | 0.0006% | $8.7 \cdot 10^{-12} - 8.5 \cdot 10^{-14}$i |
| 1/40 | 0.0002% | $9.5 \cdot 10^{-11} - 1.1 \cdot 10^{-21}$i | 0.009% | $5.2 \cdot 10^{-13} - 3.1 \cdot 10^{-15}$i |
| 1/50 | 0.02% | $3.1 \cdot 10^{-11} + 1.4 \cdot 10^{-19}$i | 0.07% | $1.4 \cdot 10^{-13} - 6.8 \cdot 10^{-16}$i |
| 1/70 | 0.1% | $5.8 \cdot 10^{-12} + 1.4 \cdot 10^{-20}$i | 0.1% | $1.2 \cdot 10^{-14} + 5.1 \cdot 10^{-17}$i |
| 1/100 | 0.1% | $9.7 \cdot 10^{-13} - 1.1 \cdot 10^{-21}$i | 1.1% | $2.1 \cdot 10^{-15} + 2.5 \cdot 10^{-18}$i |
| 1/180 | 0.07% | $5.1 \cdot 10^{-14} - 1.8 \cdot 10^{-20}$i | 1.5% | $1.8 \cdot 10^{-16} - 8.3 \cdot 10^{-18}$i |

## 6.2. Comparison with DGM and LSM

In this section, we compare the performance of imDGM to DGM and LSM. Similarly to the previous numerical experiments (see Section 6.2), we first approximate the field using eight plane waves at the element level. Then, we consider seven plane waves per element since the $R$-7-2 element was shown to be more robust in imDGM. Last, we enrich the local basis by taking eleven shape functions in each element $K$ to illustrate the potential of the method for improving the accuracy level at high frequency, by using higher order elements rather than refining the mesh.

### Approximation with eight plane waves

The first numerical experiments compare the performance of the three methods when the field is approximated, at the element level, using eight plane waves. To do this, we have computed the relative error delivered by the three methods for the propagation angle values $\theta$ in $[0, 2\pi)$ (see Eq. 30), and for different values of $ka$ while maintaining $kh$ constant. More specifically, we have considered $ka = 10$, 20, 30, and we have fixed the step size of the mesh discretization such that $kh = \frac{1}{2}$, which corresponds to about 12 elements per wavelength. The results are depicted in fig. 8. Then, we have performed a second set of experiments with larger frequency values ($ka = 30$, 60, 90, 120) while reducing to 4 the number of elements per wavelength ($kh = \frac{3}{2}$). The results are reported in Fig. 9. Last, we have set $ka = 1$ and investigated the stability of the three methods with respect to the mesh refinement. The results are depicted in Fig. 10. In all these experiments, DGM is equipped with the $R$-8-2b, its best element when using eight plane waves at the element level [6]. Recall that in LSM the only unknown is the field and consequently a single element arises when approximating the solution with $m$ plane waves. For imDGM we have used the $R$-8-3 element since we have observed that this is the best element compared to $R$-8-2b, $R$-8-4 and $R$-8-5 (see Section 6.1). The following observations are noteworthy :

- imDGM outperforms DGM, as illustrated in Figs. 8(a)-8(c). Indeed, imDGM reduces the DGM total relative error by a factor 9.

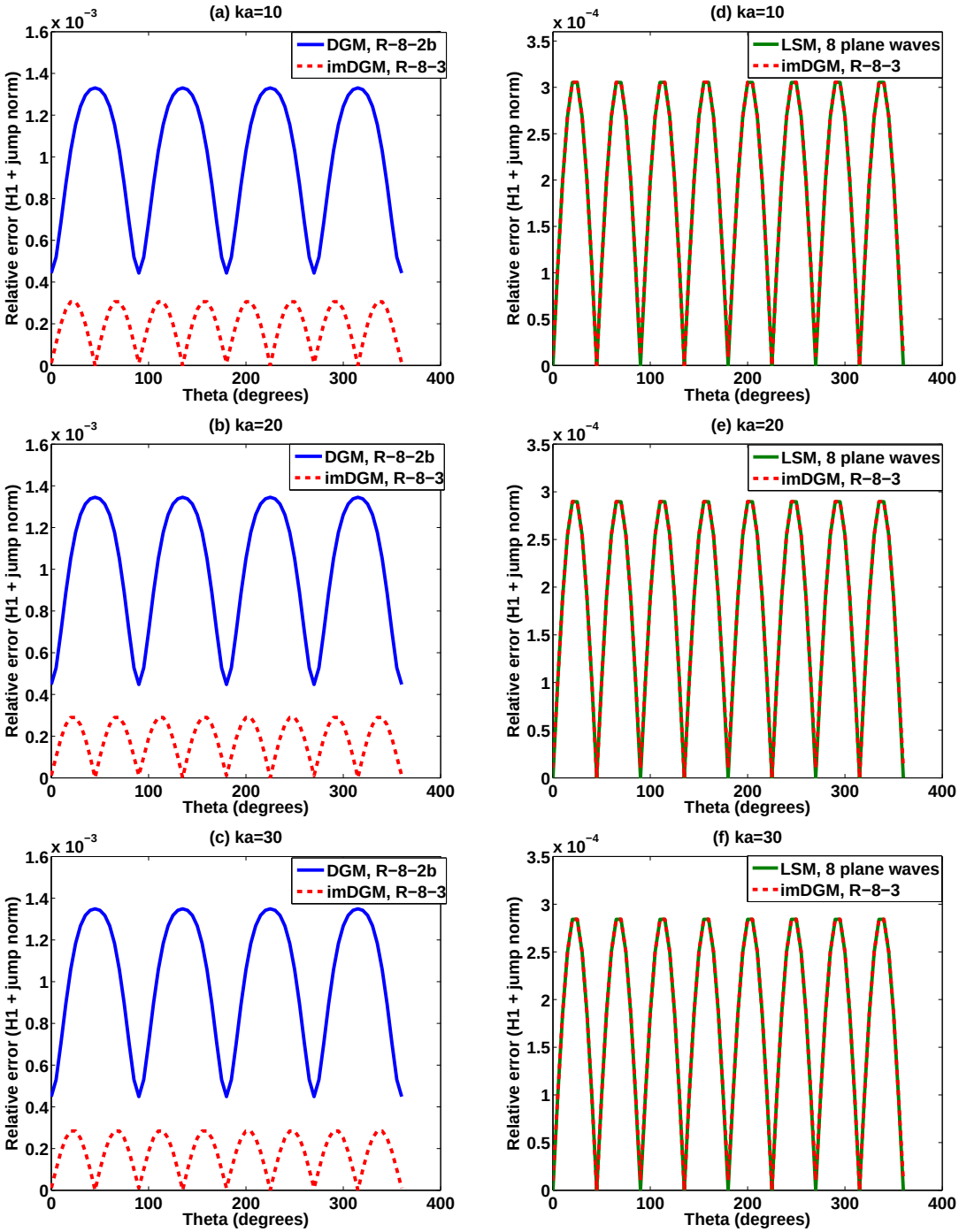


**Fig. 8** – Performance comparison between DGM ($R$-8-2b) and imDGM ($R$-8-3) (left) and LSM (8 plane waves) and imDGM ($R$-8-3) (right) for $kh = 1/2$.

- As indicated in Figs. 8(d)-8(f), imDGM and LSM exhibit a comparable performance.
- The analysis of the pollution effect in the three methods shows the following :
  - For $kh = \frac{1}{2}$, corresponding to about 12 elements per wavelength, the three methods are barely affected by the pollution as the value of $ka$ is increased from 10 to 30 (see Figs. 9(a)-9(c)).
  - For $kh = \frac{3}{2}$, corresponding to 4 elements per wavelength, the pollution effect is more noticeable in imDGM and LSM than in DGM. The error is increased in imDGM and LSM from 0.1% ($ka = 10$) to 8% ($ka = 120$). The latter level of error is still acceptable considering the relatively high frequency regime ($ka = 120$). The superiority of DGM here seems to be related to the fact that the experiments are performed for values of $h/a$ corresponding to the stability region of DGM.
- As already observed in [2] and therefore expected, Fig. 10 shows clearly that DGM is unstable as $kh < \frac{1}{6}$, corresponding to more than 36 elements per wavelength, while both imDGM and LSM remain relatively stable as the mesh is refined. The numerical instabilities do not deteriorate dramatically the level of accuracy. The total relative error remains less than 1% even when using up to 1080 elements per wavelength. This result shows again that instead of refining significantly the mesh (over 1000 elements per wavelength), one may consider using higher order elements to improve the accuracy level.

**Approximation with seven plane waves**

We have performed three types of experiments for the $R$-7-2 element, given by (37)and (38). The first one compares the accuracy of the methods for $ka = 10$, 20, 30 while maintaining the resolution at about 12 elements per wavelength, that is $kh = \frac{1}{2}$ (see Fig. 11). In the second set of experiments, we investigate the pollution effect in the case of 12 elements per wavelength for $ka = 10$, 20, 30 and in the case of 4 elements per wavelength at the frequency regime ranging from $ka = 15$ to $ka = 60$ (see Fig. 12). Last, we have investigated the stability of the three methods for $ka = 1$ and $ka = 20$. The obtained results are reported in Fig. 13.

- Fig. 11 illustrates the superiority of imDGM and LSM over DGM. Note that this superiority is not really important here since the total relative error delivered by each method is less than 1%.
- As illustrated in Fig. 12, DGM appears to produce less pollution for both meshes. Note that the two considered resolutions correspond to the stability region of DGM. One may improve the accuracy level for both imDGM and LSM by using higher order elements.
- Fig. 13 demonstrates clearly the superiority of imDGM over DGM for both frequencies $ka = 1$ and $ka = 20$. DGM becomes unstable as $kh < \frac{1}{15}$ and the accuracy level deteriorates dramatically reaching more than 20%.
- imDGM and LSM are clearly numerically stable. One can notice a loss of accuracy for $ka = 1$ as $h/a < \frac{1}{40}$ (corresponding to over 240 elements per wavelength). Nevertheless, both methods retain an acceptable level of accuracy even for a resolution corresponding to over 1200 elements per wavelength (the total relative error remains less than 0.1%).
- The results reported in Fig. 13(b) illustrate the superiority of imDGM with $R$-7-2 over DGM with $R$-8-2b. The total relative error delivered by imDGM is one to four orders of magnitude smaller than the error delivered by DGM. Observe a significant loss of the accuracy level (two orders of magnitude) in DGM when the mesh resolution is ranging from 90 to 130 elements per wavelength.
- The curves representing the errors delivered by LSM and imDGM are superposed as long as

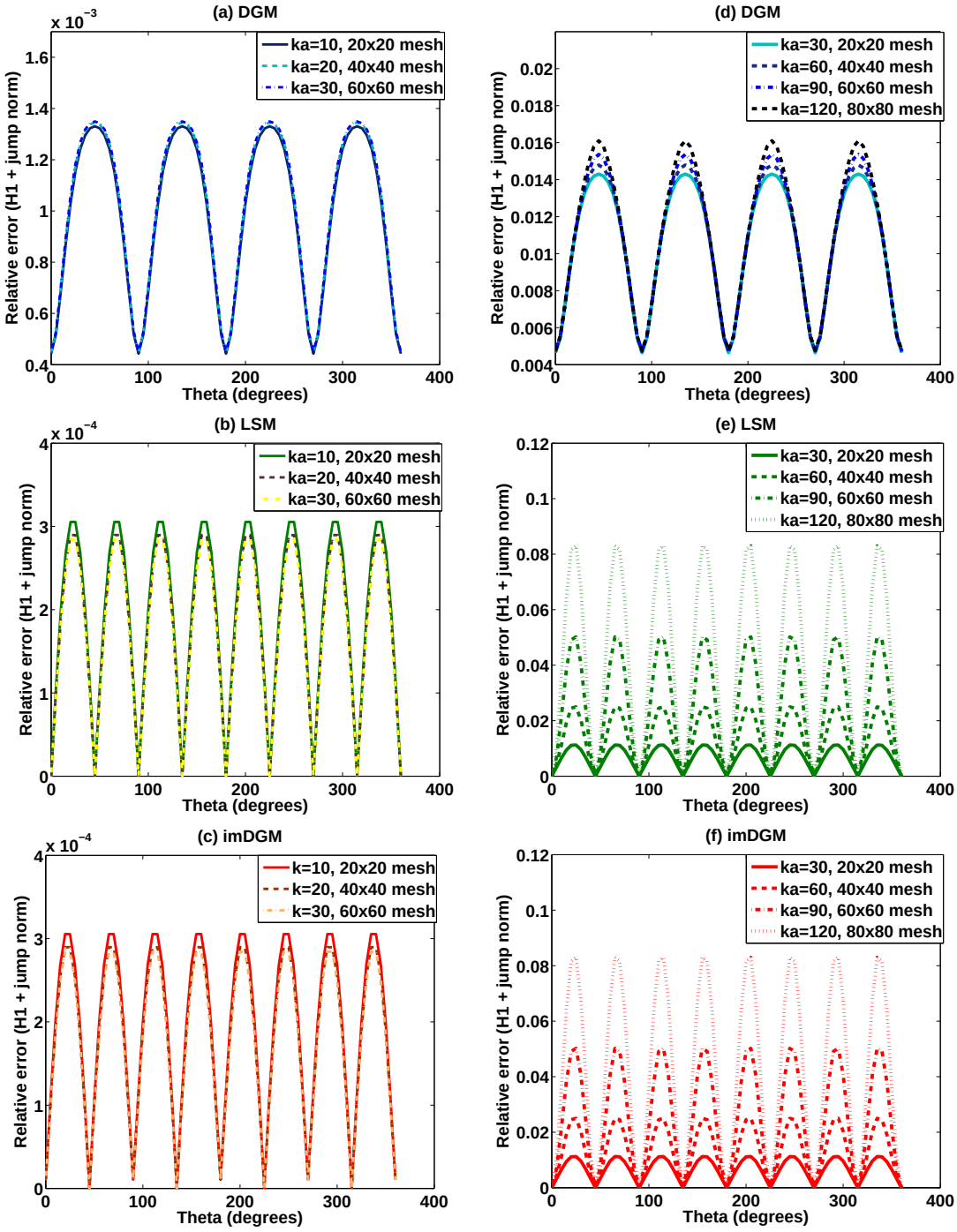


**Fig. 9** – Pollution effect for DGM ($R$-8-2b), LSM (8 plane waves), and imDGM ($R$-8-3) for $kh = 1/2$ (left) and $kh = 3/2$ (right).

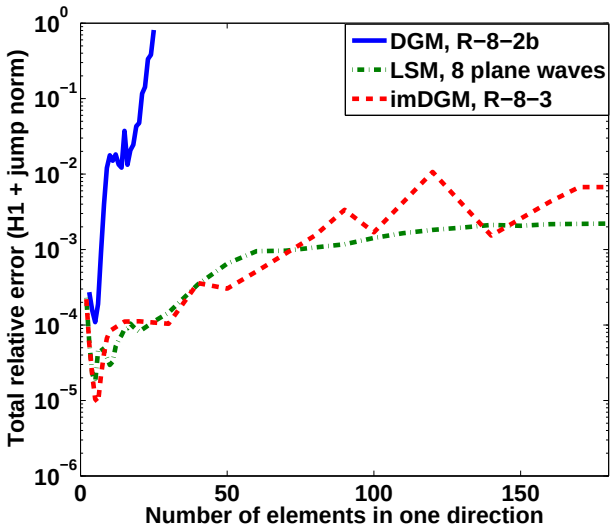


**Fig. 10** – Sensitivity of the total relative error to the mesh refinement : Comparison between DGM ($R$-8-2b), LSM (8 plane waves) and imDGM ($R$-8-3) for $ka = 1$.

$h/a > \frac{1}{400}$. Then, a loss of more than one order of magnitude is observed in LSM. Nevertheless, this accuracy loss is not important since the relative error in LSM remains less than 0.001%.

### Approximation with eleven plane waves

In this section we present preliminary results that illustrate the potential of imDGM for solving efficiently Helmholtz problems in high-frequency regime using high-order elements. We must point out that this numerical investigation is in progress. We expect to include additional numerical results in the near future.
We consider eleven plane waves positioned at :

$$\theta_p = 2(p-1)\pi/11, \ \forall 1 \leq p \leq 11. \tag{39}$$

On each interior edge of an element the Lagrange multiplier is given by (36). We refer to this element by $R$-11-3.
We have performed three sets of experiments. The first experiment compares the accuracy of $R$-7-2 and $R$-11-3 for different values of $ka$ corresponding to a relatively mid-frequency regime and for a fixed resolution. More specifically, we set $ka = 15, 30, 60$ and consider the mesh size $h/a$ such that we have 4 elements per wavelength ($kh = \frac{3}{2}$). The comparison of the relative error for each propagation angle $\theta$ (see Eq. (30) is reported in Fig. 14. Then, we increase the frequency value ($ka = 50$, 100, 200) and reduce to 3 the number of elements per wavelength. The relative errors obtained for each propagation angle $\theta$ are depicted in Fig. 15. In addition, we report in Table 4 the values of the total relative errors delivered by $R$-7-2 and $R$-11-3. Last, we compare the performance of both elements when refining the mesh ($h/a = \frac{1}{10}, \frac{1}{20}, \frac{1}{30}, \frac{1}{40}$) while maintaining fixed the frequency value ($ka = 20$) The results are reported in Table 5. The following observations are noteworthy :

- Fig. 14 clearly indicates that $R$-11-3 outperforms $R$-7-2 in imDGM. In addition, Table 4(a) shows that using $R$-11-3 reduces the total relative error by more than two orders of magnitude for each considered value of $ka$. Observe (Fig. 14(a)) that $R$-11-3 does not exhibit pollution.

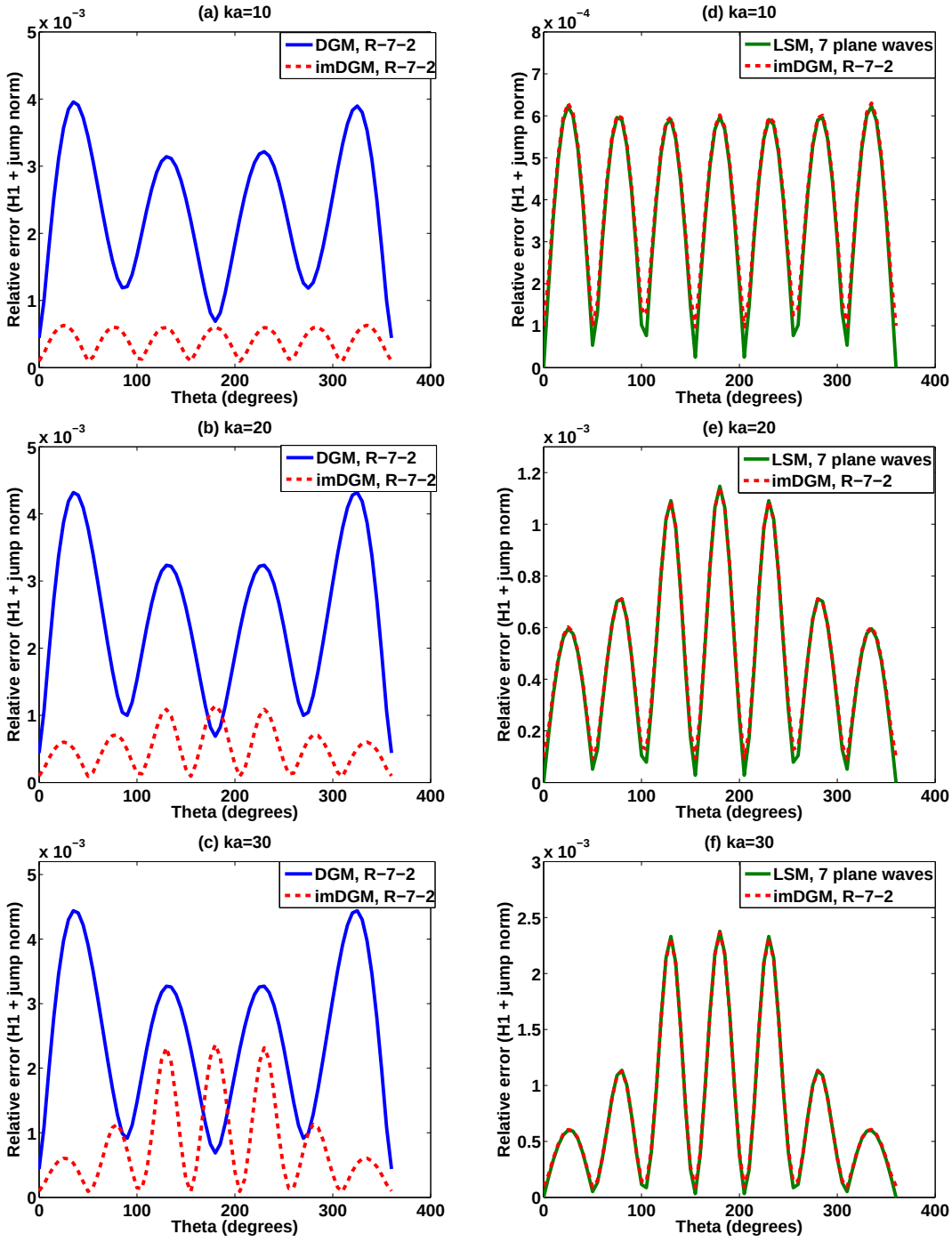


**Fig. 11** – Performance comparison between DGM ($R$-7-2) and imDGM ($R$-7-2) (left) and LSM (7 plane waves) and imDGM ($R$-7-2) (right) for $kh = 1/2$.

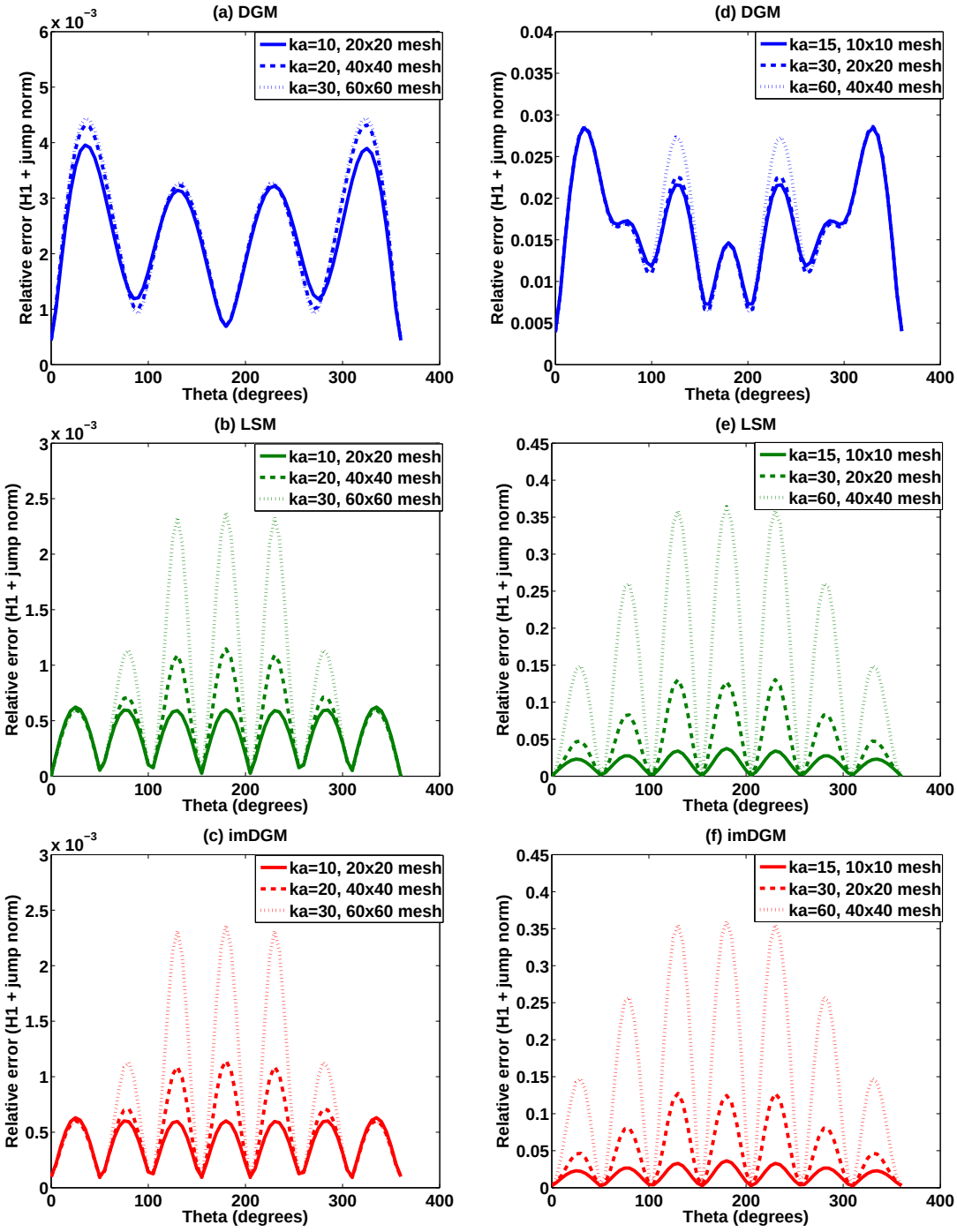


**Fig. 12** – Pollution effect for DGM ($R$-7-2), LSM (7 plane waves), and imDGM ($R$-7-2) for $kh = 1/2$ (left) and $kh = 3/2$ (right).

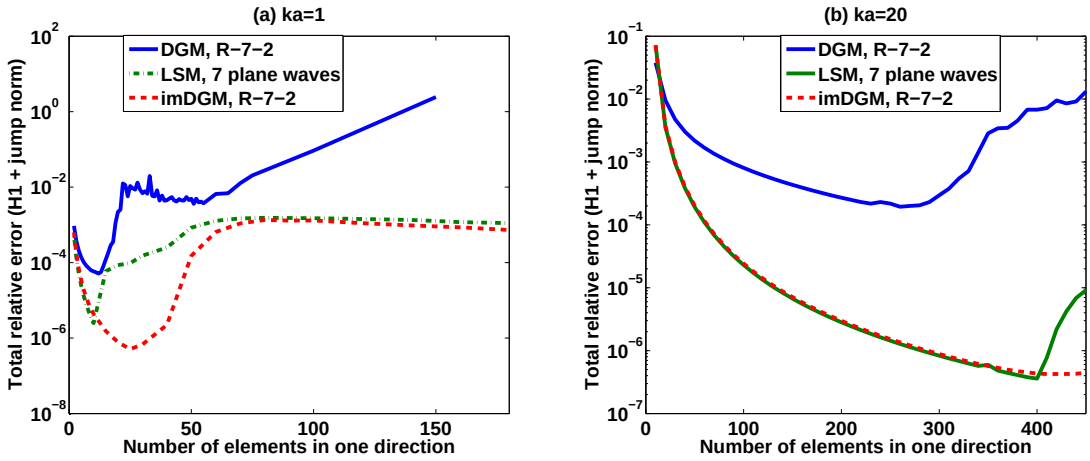


**Fig. 13** – Sensitivity of the total relative error to the mesh refinement : Comparison between DGM ($R$-7-2), LSM (7 plane waves) and imDGM ($R$-7-2) for $ka = 1$ (left) and $ka = 20$ (right).

TAB. 4 – Sensitivity of the total relative error to the frequency : Comparison between imDGM ($R$-7-2) and imDGM ($R$-11-3) for a mesh resolution corresponding to 4 elements per wavelength (left) and 3 elements per wavelength (right).

| $ka$ | $R$-7-2 | $R$-11-3 |
|---|---|---|
| 15 | 1.7% | 0.01% |
| 30 | 4.9% | 0.01% |
| 60 | 15% | 0.01% |

| $ka$ | $R$-7-2 | $R$-11-3 |
|---|---|---|
| 50 | 28% | 0.05% |
| 100 | 51% | 0.07% |
| 200 | 69% | 0.2% |

Indeed, increasing $ka$ from 15 to 60 does not affect the total relative error, which remains at 0.01%.

- As illustrated in Fig. 15 and Table 4(b), $R$-11-3 delivers a better accuracy level than $R$-7-2 in high-frequency regime. Observe that for each value of $ka = 50$ and $ka = 100$, there is again more than two orders of magnitude, whereas for $ka = 200$, $R$-11-3 reduces the error from 69% to 0.2%.
- Last, Table 5 shows that for $ka = 20$, when using the same number of elements per wavelength $R$-11-3 delivers an accuracy which is about two orders of magnitude better than the one obtained with $R$-7-2.

These preliminary experiments show that in mid- and high-frequency regime, it is preferable to use higher-order elements rather than refining the mesh to improve the accuracy. In addition, since the cost of imDGM depends only on the global number of Lagrange multipliers, then using $R$-11-3 instead of $R$-7-2 increases the computational cost by a factor $\frac{1}{2}$ (see Table 1). This additional cost allows however to reduce the level of the total relative error by more than two orders of magnitude.

TAB. 5 – Sensitivity of the total relative error for imDGM ($R$-7-2) and imDGM ($R$-11-3) for $ka = 20$.

| Number of elements per wavelength | $R$-7-2 | $R$-11-3 |
|---|---|---|
| 3 | 7% | 0.04% |
| 6 | 0.4% | 0.002% |
| 9 | 0.1% | 0.0002% |
| 12 | 0.04% | 0.0001% |

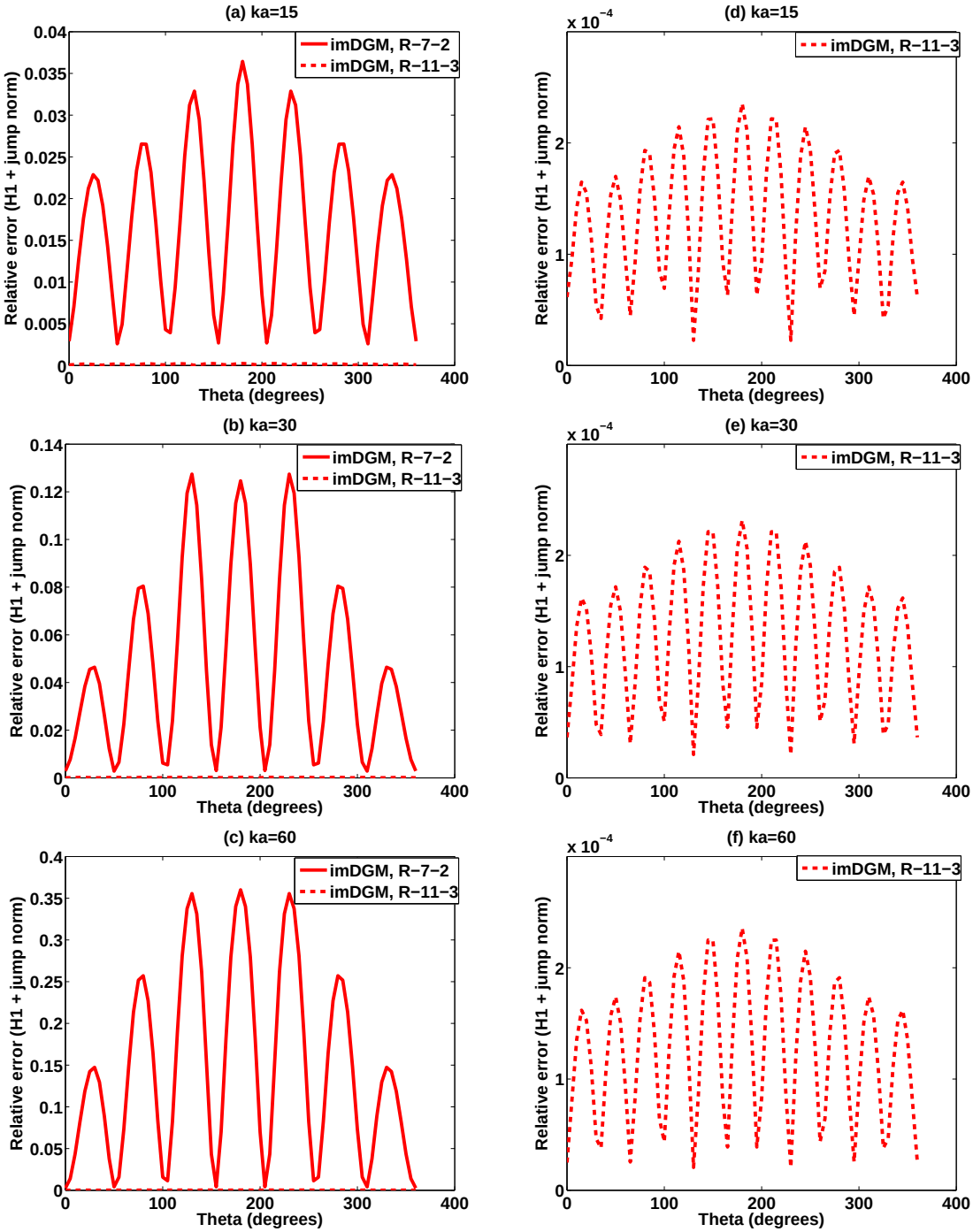


**Fig. 14** – Performance comparison between imDGM ($R$-7-2) and imDGM ($R$-11-3) for $kh = 3/2$ (left). Zoom on imDGM($R$-11-3) for $kh = 3/2$ (right).

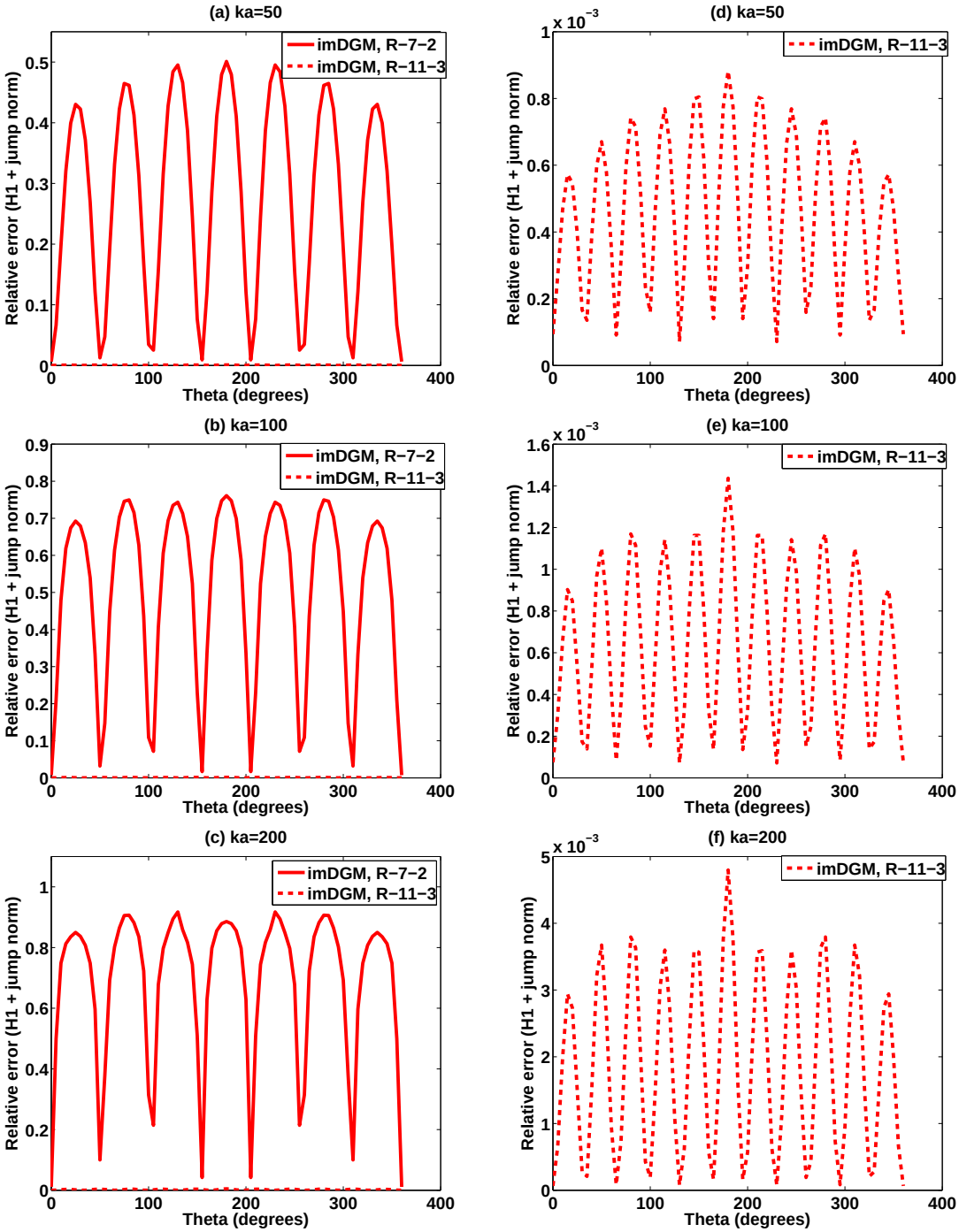


**Fig. 15** – Performance comparison between imDGM ($R$-7-2) and imDGM ($R$-11-3) for $kh = 2$ (left). Zoom on imDGM($R$-11-3) for $kh = 2$ (right).

## 7. Summary and conclusion

We have designed a DG-type method, called imDGM, for solving Helmholtz problems. The method can be viewed "between" DGM formulation, designed by Farhat *et al* in [6, 7, 8] and the LSM formulation, designed by Monk-Wang in [16]. The preliminary numerical results obtained in the case of waveguide are very promising. They show that the method is stable and accurate. For example, for the $R$-7-2 element, the method remains stable for a mesh resolution with over 1000 elements per wavelength. In addition, in the high frequency regime the method delivers results with high level of accuracy . For example, for $ka = 200$ and using only 3 elements per wavelength, imDGM equipped with $R$-11-3 element delivers a total relative error of 0.2%. These preliminary results illustrate the potential of the proposed method for solving efficiently high frequency Helmholtz problems.

# Bibliography


[1] Amara M., Djellouli R., Farhat C., Convergence analysis of a discontinuous Galerkin method with plane waves and Lagrange multipliers for the solution of Helmholtz problems *SIAM J. Numer. Anal.*, **47**(2), 1038-1066, 2009

[2] Amara M., Calandra H., Djellouli R., Grigoroscuta-Strugaru M., A modified discontinuous Galerkin method for Helmholtz problems *Technical Report INRIA No. 7050* 2009, available online at http ://hal.archives-ouvertes.fr/inria-00421584/fr/ ;

[3] Babuška I., Melenk I.J.M., The partition of unity method *Internat. J. Numer. Methhods Eng.*, **40**, 727-758, 1997

[4] Babuška I., Sauter S., Is the Pollution Effect of the FEM Avoidable for the Helmholtz Equation Considering High Wave Numbers ? *SIAM J. Numer. Anal.*, **34**, 2392-2423, 1997

[5] Cessenat O., Després B., Application of an ultra-weak variational formulation of elliptic PDEs to the two-dimensional Helmholtz problems *SIAM J. Numer. Anal.*, **35**, 255-299, 1998

[6] Farhat C., Harari I., Hetmaniuk U., A discontinuous Galerkin method with Lagrange multipliers for the solution of Helmholtz problems in the mid-frequency regime *Comput. Methods Appli. Mech. Eng.*, **192**, 1389-1419, 2003

[7] Farhat C., Wiedemann-Goiran P., Tezaur R., A discontinuous Galerkin method with plane waves and Lagrange multipliers for the solution of short wave exterior Helmholtz problems on unstructured meshes *Wave Motion*, **39**, 307-317, 2004

[8] Farhat C., Tezaur R., Wiedemann-Goiran P., Higher-order extensions of a discontinuous Galerkin method for mid-frequency Helmholtz problems *Internat. J. Numer. Methhods Eng.*, **61**, 1938-1956, 2004

[9] Franca L.P., Farhat C., Macedo A.P., Lesoinne M., Residual-free bubbles for the Helmholtz equation *Internat. J. Numer. Methhods Eng.*, **40**, 4003-4009, 1997

[10] Hadamard J., Lectures on Cauchy's Problem in Linear Partial Differential Equations *Yale University Press, New Haven* 1923

[11] Harari I., Hughes T.J.R., Galerkin/least-squares finite element methods for the reduced wave equation with non-reflecting boundary conditions in unbounded domains *Comput. Methods Appli. Mech. Eng.*, **98**, 411-454, 1992

[12] Harari I., Hetmaniuk U., Private communication

[13] Hörmander L., The Analysis of Linear Partial Differential Operator *Springer-Verlag, New York* 1985

[14] Ihlenburg F., Finite Element Analysis of Acoustic Scattering *Appl. Math. Sci 132, Springer-Verlag, New York* 1998

[15] Magoulès F., Computational Methods for Acoustics Problems *Saxe-Coburg Publications* 2008

[16] Monk P., Wang D.Q., A least-squares method for the Helmholtz equation *Comput. Methods Appli. Mech. Eng.*, **175**, 411-454, 1999

[17] Rose M.E., Weak element approximations to elliptic differential equations *Numer. Math.*, **24**, 185-204, 1975

[18] Taylor M. E., Partial Differential Equations I : Basic Theory *Springer-Verlag, New York* 1997

# **Partie IV :** Application : propagation des ondes en géophysique

# 1. Introduction

Dans le domaine pétrolier, on utilise l'équation des ondes pour produire des images du sous-sol à partir des données qui ont été mesurées lors des campagnes d'acquisition. Les images sont fabriquées à partir de la technique dite de sismique-réflexion. Plus précisément, en générant des onde sismiques artificielles on crée un phénomène de propagation d'ondes dont les réflexions sont enregistrées par des récepteurs placés en surface ou en profondeur. Les ondes réfléchies sont ensuite rétropropagées et en appliquant une condition d'imagerie qui revient à corréler les ondes propagées et rétropropagées, on fabrique une carte de réflecteurs plus ou moins précise. La qualité de l'image est améliorée en itérant le procédé de propagation et rétropropagation. On stoppe l'algorithme une fois que l'image n'évolue pas entre deux itérations. Les logiciels d'imagerie sismique utilisent la solution numérique de l'équation des ondes. La technique d'imagerie est appelée "Reverse Time Migration". Nous renvoyons à la thèse de C. Baldassari [1] pour plus de détails sur cette technique et sa mise en application avec une méthode de type Galerkin discontinu.
Un signal $p$ (l'onde qui se propage) est alors caractérisé par sa valeur en fonction de ses coordonnées $\overrightarrow{x}$ (un vecteur dans un espace à trois dimensions) et du temps $t$ et vérifie l'équation :

$$\frac{\partial^2 p\left(\overrightarrow{x},t\right)}{\partial t^2} - c^2 \Delta p\left(\overrightarrow{x},t\right) = f\left(\overrightarrow{x},t\right), \tag{1}$$

où $f$ est la fonction source et $c$ représente la vitesse de propagation.
La dimension de l'équation (1) peut être réduite à 3, à l'aide de la transformée de Fourier-Laplace. Plus précisément, on a :

$$p\left(\overrightarrow{x},t\right) = \frac{1}{2\pi}\int_{-\infty}^{\infty} P\left(\overrightarrow{x},\omega\right) e^{-\mathrm{i}\omega t} d\omega,$$

avec

$$P\left(\overrightarrow{x},\omega\right) = \int_{-\infty}^{\infty} p\left(\overrightarrow{x},t\right) e^{\mathrm{i}\omega t} dt.$$

On en déduit que $P\left(\overrightarrow{x},\omega\right)$ vérifie l'équation d'Helmholtz :

$$-\Delta P\left(\overrightarrow{x},\omega\right) - k^2\left(\omega\right) P\left(\overrightarrow{x},\omega\right) = F\left(\overrightarrow{x},\omega\right),$$

où $k\left(\omega\right) = \frac{\omega}{c}$ représente le nombre d'onde correspondant à la fréquence angulaire $\omega$ et $F\left(\overrightarrow{x},\omega\right)$ désigne la transformée de Fourier de $f\left(\overrightarrow{x},t\right)$.
La transformée de Fourier permet donc de résoudre l'équation des ondes sous une forme simplifiée (l'équation d'Helmholtz), mais cette simplification demande le calcul de la transformée de Fourier inverse, afin de récupérer la solution en domaine temporel.

Le but de cette partie est de valider le code informatique correspondant à la méthode développée dans la Partie I, pour l'élément $R$-4-1. Le choix de la méthode et de l'élément est justifié par le fait qu'on la considère comme une méthode pas chère et très robuste. Cette propriété est capitale si on tient compte du fait que résoudre l'équation de Helmholtz pour plusieurs fréquences sur un maillage fixe demande une méthode avec une très bonne précision (qui permette donc de retrouver de bonnes approximations avec un nombre réduit d'éléments par longueur d'onde, ce qui est le cas pour les hautes fréquences) et très stable (qui maintienne la précision lorsque la résolution est très fine, correspondant donc à de basses fréquences). De plus, la technique décrite dans la Partie I permet d'utiliser la méthode du gradient conjugué sans avoir à calculer explicitement la matrice du système global. Cette propriété peut s'avérer être un très grand avantage pour les applications industrielles, à condition que l'algorithme converge rapidement. Cette question fait partie de nos

travaux futurs, avec l'étude de la convergence et de la précision obtenues sur des plateformes de calcul parallèle.

Sur un exemple très simple, on se propose donc de regarder la propagation des ondes dans deux types de domaines : l'un avec vitesse constante (ce qui correspond à $k$ constant dans tout le domaine, donc dans tous les sous-domaines), l'autre avec deux couches correspondant à deux vitesses différentes, donc à des $k$ qui peuvent varier d'un élément du maillage à l'autre. Bien évidemment, l'expression de la matrice globale change dans le deuxième cas et l'implémentation de la méthode doit prendre en compte cette possible variation du nombre d'onde d'un sous-domaine à l'autre. Cependant, les caractéristiques de la méthode en question ne changent pas, ce qui permet l'utilisation du schéma du gradient conjugué. On s'intéresse donc aux performances de la méthode en terme de cohérence des résultats (retrouver les réflexions déterminées par le changement de vitesse) et de précision et stabilité dans le sens expliqué précédemment.

## 2. Principes de résolution

Considérons une source ponctuelle de type Ricker qui a pour fréquence centrale $f_0 = 100$ Hz. En temps, son expression est donnée par :

$$f(t) = \left(1 - 2\pi^2 f_0^2 t^2\right) e^{-\pi^2 f_0^2 t^2}. \tag{2}$$

Sur un intervalle de temps égal à $0.1s$ et pour un nombre de pas de temps $nt = 143$, l'évolution du signal émis par la source est donnée par la Fig. 1a.

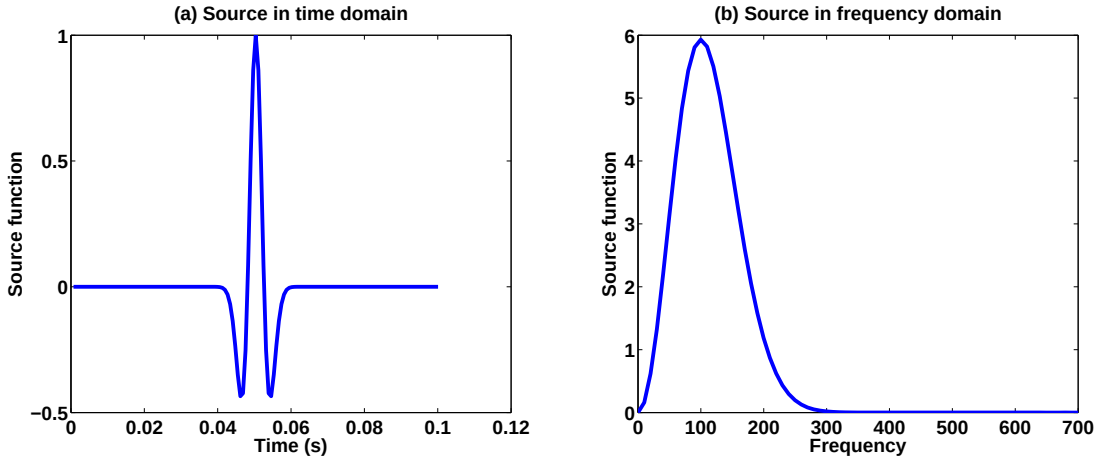


**Fig. 1** – Signal en temps et en fréquence

**Remarque 1.** Le pas de temps $dt = 1/(nt-1)$ correspond ici au rythme (à la cadence) d'échantillonnage. Considérons maintenant l'ensemble de nombres réels $\left\{f^j\right\}_{j=0}^{nt-1}$, où $f^j = f(j \cdot dt)$. Soit $\left\{F^j\right\}_{j=0}^{nt-1}$ la transformée de Fourier discrète de $\left\{f^j\right\}_{j=0}^{nt-1}$. En pratique, les valeurs $\left\{F^j\right\}_{j=0}^{nt-1}$ représentent les amplitudes complexes du signal pour différentes fréquence angulaires, $\omega_j$ avec $j = 0, 1, ..., nt-1$. Plus précisément ces valeurs sont dans l'intervalle $\left[-\frac{\pi}{dt}, \frac{\pi}{dt}\right]$.
Puisque dans l'équation d'Helmholtz le nombre d'onde $k(\omega) = \frac{\omega}{c}$ est élevé au carré, il suffit de résoudre l'équation pour la moitié des valeurs de $\omega_j$. C'est la raison pour laquelle dans la figure 1b

on a représenté le module de l'amplitude complexe du signal en fonction des différentes fréquences $f$, qui varient dans l'intervalle $\left[0, \frac{1}{2dt}\right]$.

**Remarque 2.** Fig. 1b montre que la condition du théorème de Nyquist-Shannon est satisfaite : en effet, le signal occupe une bande de fréquence de 0 à 300 Hz. La valeur maximale est donc inférieure à la moitié de la fréquence d'échantillonnage, qui correspond ici à $\frac{1}{2dt} \cong 700$ Hz.
On est donc amené à résoudre l'équation d'Helmholtz pour différents nombres d'ondes correspondant à $k_j = \frac{\omega_j}{c}$, avec $\omega_j = j\frac{2\pi}{N \cdot dt}$ pour $j = 0, 1, ..., \frac{nt-1}{2}$. Pour chaque fréquence angulaire considérée, la solution obtenue représente le champ d'onde en domaine fréquentiel et est, en général, une fonction complexe.
Considérons maintenant un point du domaine et un récepteur qui enregistre la valeur de la solution en ce point, pour chacune des $\frac{nt-1}{2} + 1$ fréquences angulaires. Afin de reconstituer la solution en temps, via la transformée de Fourier discrète inverse, le vecteur des valeurs en domaine fréquentiel doit être complété avec les conjugués des valeurs obtenues pour $\omega_j$, avec $j = 1, 2, ..., \frac{nt-1}{2}$, de façon à avoir :

$$P(-\omega_j) = \overline{P(\omega_j)}, \forall j = 1, 2, ..., \frac{nt-1}{2}.$$

Cette condition est nécessaire, car on sait que la solution de l'équation des ondes est un signal réel, donc sa transformée de Fourier discrète doit conduire à un vecteur $P$ ayant comme propriété :

$$P_j = \overline{P_{nt-j+1}}, \forall j = 1, 2, ..., \frac{nt-1}{2}.$$

Enfin, pour chaque récepteur, on applique la transformée de Fourier discrète inverse sur l'ensemble des composantes du vecteur associé. On retrouve ainsi, la solution de l'équation des ondes en temps.

**Remarque 3.** Le calcul des transformées de Fourier discrètes directe et inverse peut être fait à l'aide de la transformée de Fourier rapide, directe et respectivement inverse. Pour les expériences présentées ci-après, on a utilisé les routines FFT (Fast Fourier Transform) et IFFT (Inverse Fast Fourier Transform) de Matlab.
En résumé, résoudre l'équation des ondes sous une forme réduite nécessite trois étapes :

1. transformée de Fourier discrète sur l'ensemble des valeurs de la fonction source pour tous les pas de temps $(nt)$ considérés.
2. résolution de $\frac{nt-1}{2} + 1$ équations d'Helmholtz.
3. transformée de Fourier discrète inverse pour chaque maille et ainsi on reconstitue le signal pour chaque pas de temps, dans chaque maille.

## 3. Propagation des ondes en domaine temporel dans un domaine rectangulaire

On considère un domaine rectangulaire de taille 50 m × 100 m. On maille le domaine avec des carrés qui sont tels que $h = 0.5$ m, où $h$ représente le pas de maillage. Dans la première expérience, on suppose que la vitesse de propagation $c$ est constante dans tout le domaine, i.e. $c = 1000$ m/s. Dans la deuxième expérience, on considère qu'à 25 m de profondeur la vitesse de propagation change de façon à ce que l'on ait : $c_1 = c = 1000$ m/s dans la région supérieure et $c_2 = 1500$ m/s dans la région inférieure (voir Fig. 2). Dans les deux expériences, au point situé sur la surface à $x = 25$ m, on fait exploser une source de fréquence centrale $f_0 = 100$ Hz et dont l'expression est donnée par

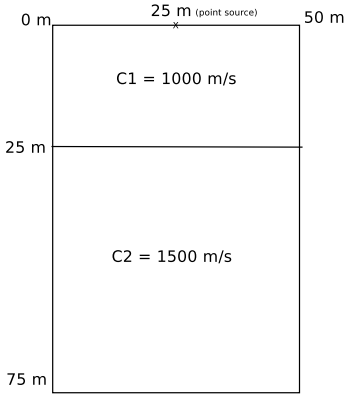


**Fig. 2** – Domaine bicouche

l'équation (2). Pendant 0.1s, avec un pas de temps $dt = 0.0007$s, qui correspond à 143 pas de temps, on observe la propagation des ondes dans le domaine correspondant à chaque expérience. Sur tout le bord du domaine on impose, dans les deux cas, une condition absorbante de type Robin-Fourier (d'ordre 1/2).

Les résultats représentés dans les Fig. 3-7 montrent la propagation du signal émis par la source, à différents instants des deux expériences. A gauche, on a représenté la propagation dans le domaine avec vitesse constante, à droite la propagation en domaine bicouche. Deux figures qui sont situées sur la même ligne correspondent au même pas de temps. Les observations suivantes sont à noter :

- Les Fig. 3-4 montrent que pendant la première partie des deux expériences la propagation des ondes est identique, ce qui n'est pas surprenant, vu que dans la région parcourue pendant cet intervalle de temps, la vitesse est la même.
- L'arrivée de l'onde à l'interface des deux régions génère, comme attendu, une réflexion dans le cas du domaine bicouche. Cette réflexion est visible à partir de la Fig. 5 (droite). Bien entendu, dans le cas du domaine avec vitesse constante, la propagation de l'onde n'est pas perturbée.
- Sur la dernière partie de l'expérience (Fig. 6-7) on voit bien la différence de vitesse entre les deux couches. En effet, l'onde qui se propage dans le domaine bicouche "sort" du domaine avant celle correspondant au domaine de gauche.
- Dans toutes les figures on remarque à l'intérieur du domaine des réflexions au bord. Celles-ci sont dues aux conditions absorbantes (d'ordre 1/2) peu performantes qu'on a choisies pour ce problème modèle. La présence de ces réflexions avant même que l'onde soit générée s'explique par le caractère périodique de la transformée de Fourier (directe et inverse). En d'autres termes, le fait d'avoir choisi le temps de propagation égal à la durée de temps nécessaire à l'onde pour parcourir le domaine homogène conduit à une superposition de phénomènes : d'une part, on a l'onde directe engendrée par l'explosion de la source, d'autre part on a les réflexions du bord correspondant à l'onde qui s'est propagée dans la période précédente.
- Comme indiqué dans la Fig. 1b, la plage des fréquences pour lesquelles on résout l'équation d'Helmholtz est $[0, 700]$ Hz. En tenant compte du pas de maillage et des vitesses considérées, on déduit que la résolution du problème de Helmholtz est effectuée pour des résolutions qui varient entre 3 et 300 points par longueur d'onde. La cohérence des résultats obtenus montrent que l'implémentation de l'élément $R$-4-1 dans la nouvelle formulation conduit à une méthode qui s'avère précise et stable.

La troisième expérience consiste à considérer le même domaine que précédemment, mais en imposant

une différence de vitesse plus grande entre les deux couches afin de mieux observer la réflexion générée par l'interface. De plus, considérer $c_2 = 2500\,m/s$ permet de valider la méthode pour des résolutions encore plus fines (jusqu'à 500 points par longueur d'onde). Cette expérience est représentée dans les Fig. 8-11. Les moments successifs de l'expérience sont à regarder sur chaque ligne de gauche à droite. La différence avec la deuxième expérience réside, comme attendu, dans l'intensité du champ réfléchi et dans la vitesse avec laquelle les ondes sortent du domaine.

Enfin, la dernière expérience consiste à faire propager les ondes engendrées par l'explosion d'un même type de source pour le but de regarder, cette fois-ci, la réflexion déterminée par le passage de l'onde dans un milieu avec vitesse inférieure. Les résultats de l'expérience sont illustrés dans les Fig. 12-14. Contrairement aux résultats de la deuxième expérience, on remarque que le champ d'ondes réfléchies contient d'abord la partie positive du champ. En effet, on observe d'abord un champ représenté par la couleur jaune (qui correspond à des valeurs positives) et ensuite le champ bleu (associé à des valeurs négatives).

## 4. Conclusion et perspectives

Ces résultats préliminaires montrent le potentiel de la méthode pour résoudre efficacement des problèmes issus de la géophysique. Les travaux en cours concernent la résolution de Helmholtz dans des domaines irréguliers, auxquels on associe des maillages non-structurés pour ensuite simuler la propagation des ondes en domaine temporel. On s'intéresse aussi à la comparaison entre la précision des résultats obtenus avec des méthodes directes et des méthodes itératives pour la résolution du système global, tout en prenant en compte le temps de calcul et la mémoire nécessaire. On envisage aussi l'implémentation des éléments finis d'ordre élevé car cela permettra d'avoir des résultats beaucoup plus précis pour un coût de calcul sensiblement plus élevé. Cet objectif nécessitera le passage à la méthode décrite dans la Partie III qui assure la stabilité des éléments d'ordre supérieur. Enfin, le passage sur une plateforme de calcul parallèle réduira le temps de calcul et permettra de tester le code sur des cas issus de la réalité.

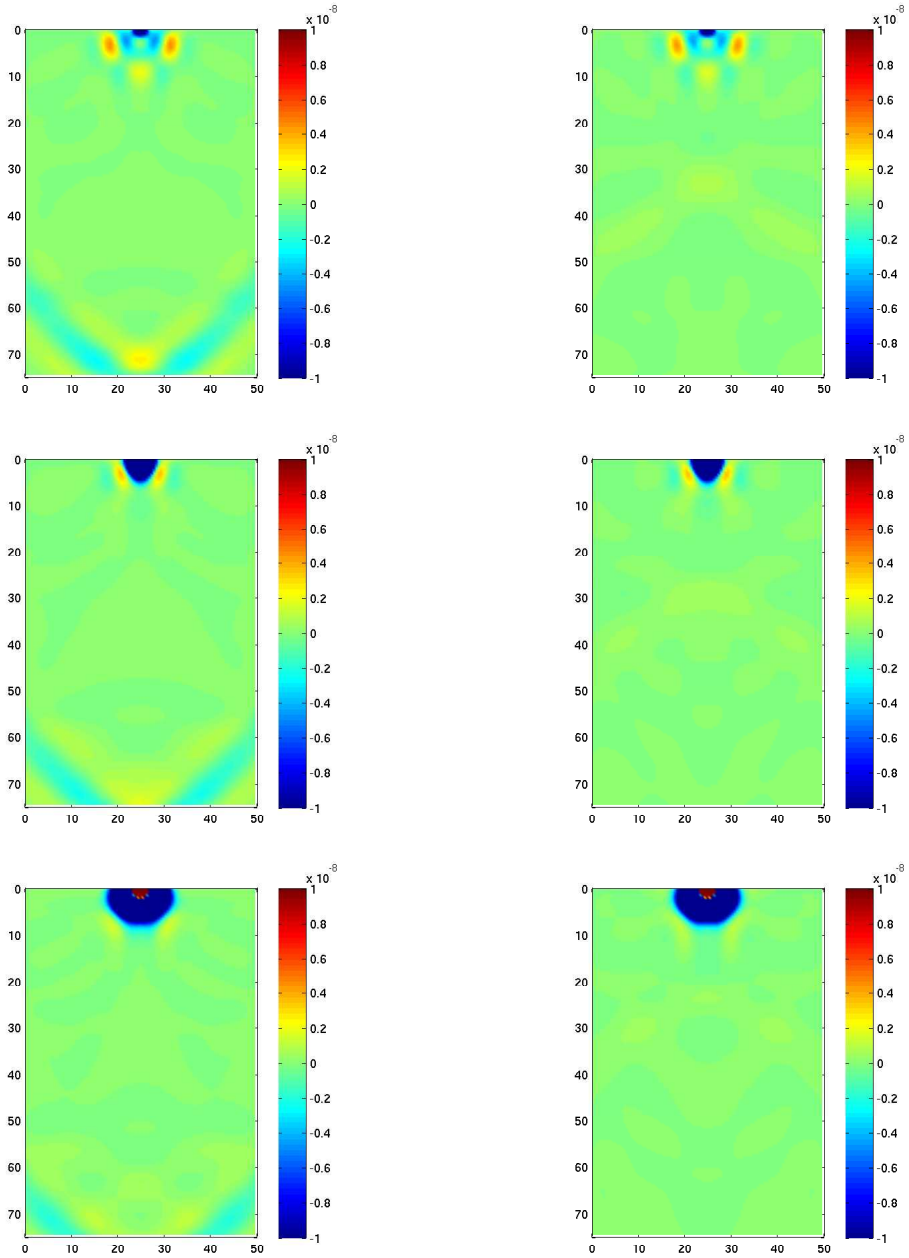


**Fig. 3** – Propagation de l'onde dans un domaine homogène (gauche) et dans un domaine bicouche (droite)

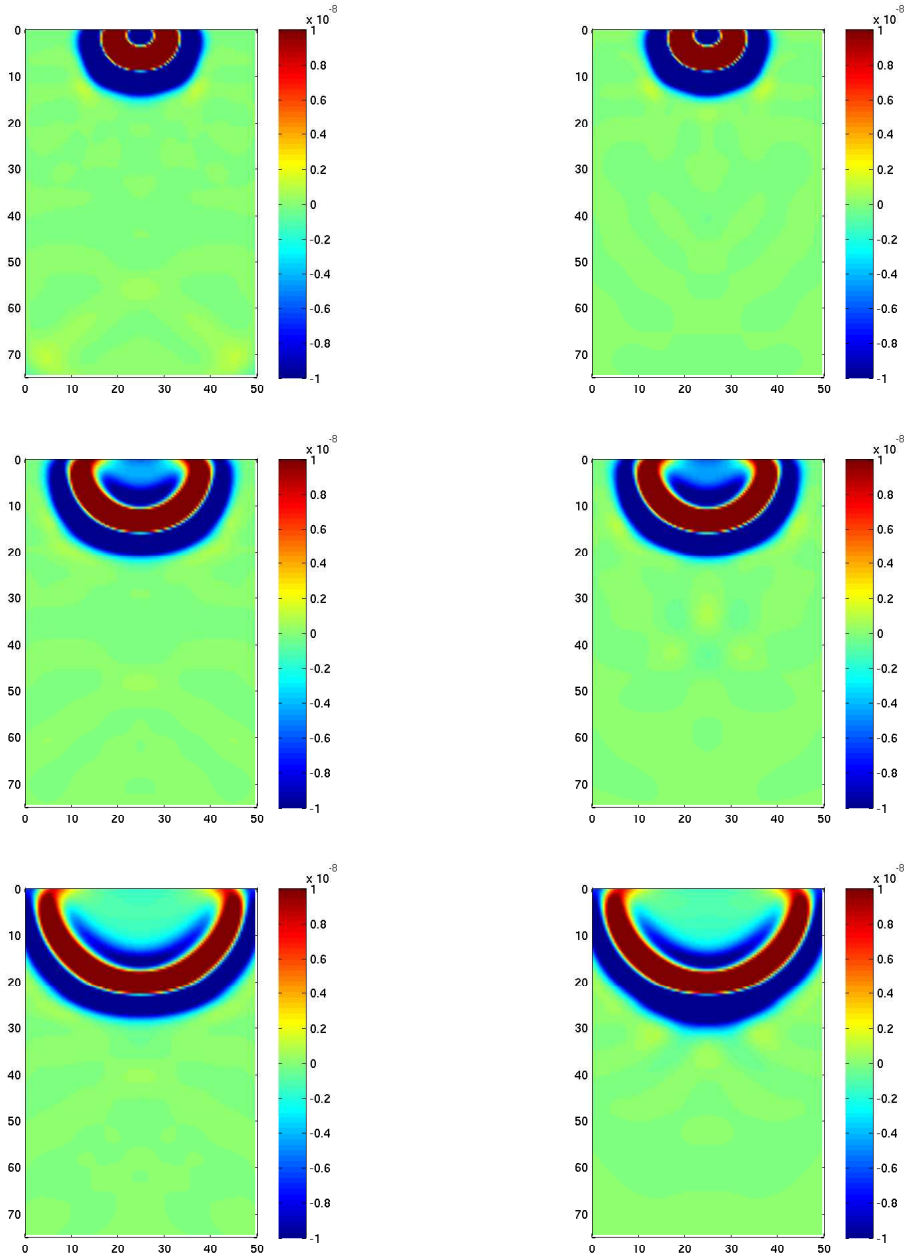


**Fig. 4** – Propagation de l'onde dans un domaine homogène (gauche) et dans un domaine bicouche (droite)

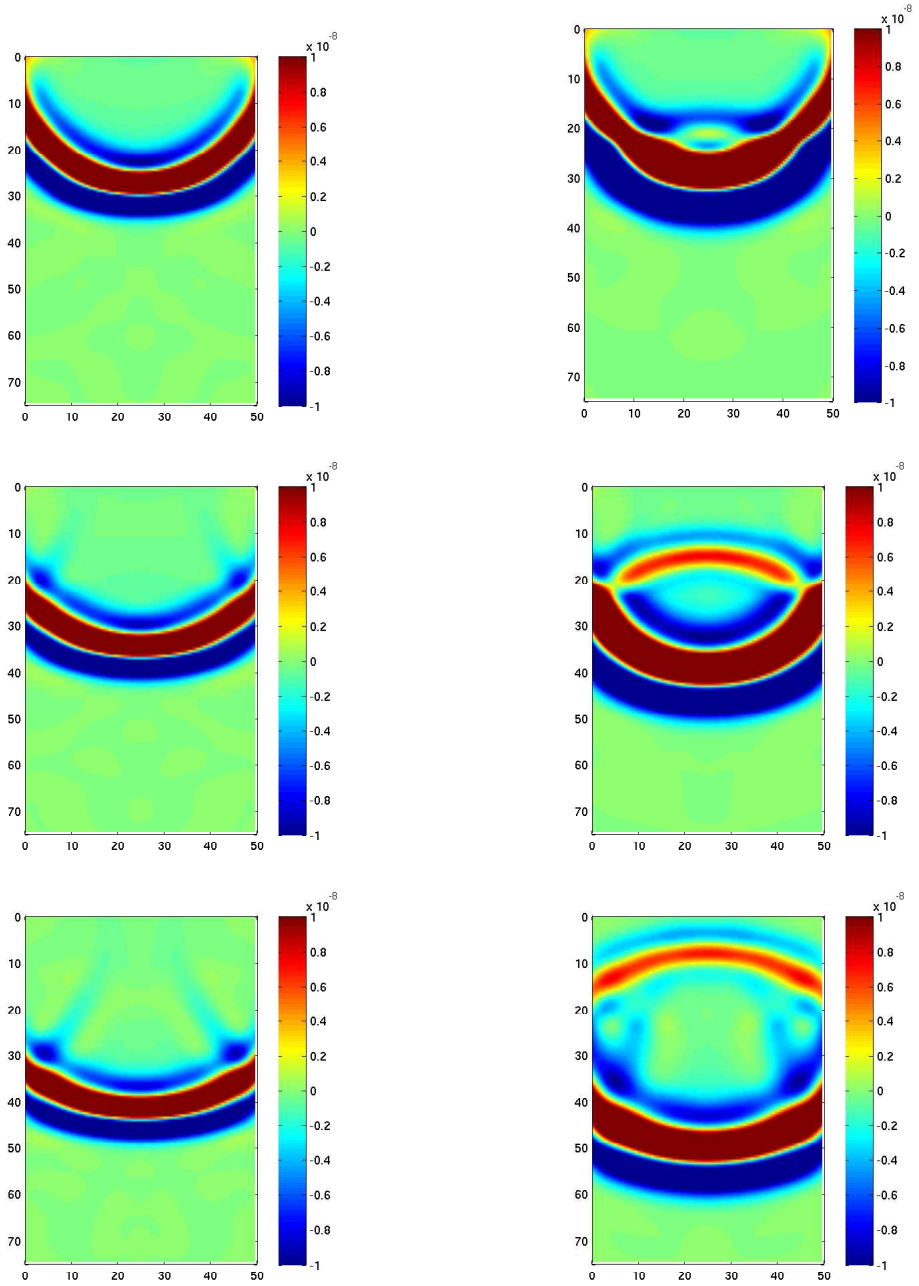


**Fig. 5** – Propagation de l'onde dans un domaine homogène (gauche) et dans un domaine bicouche (droite)

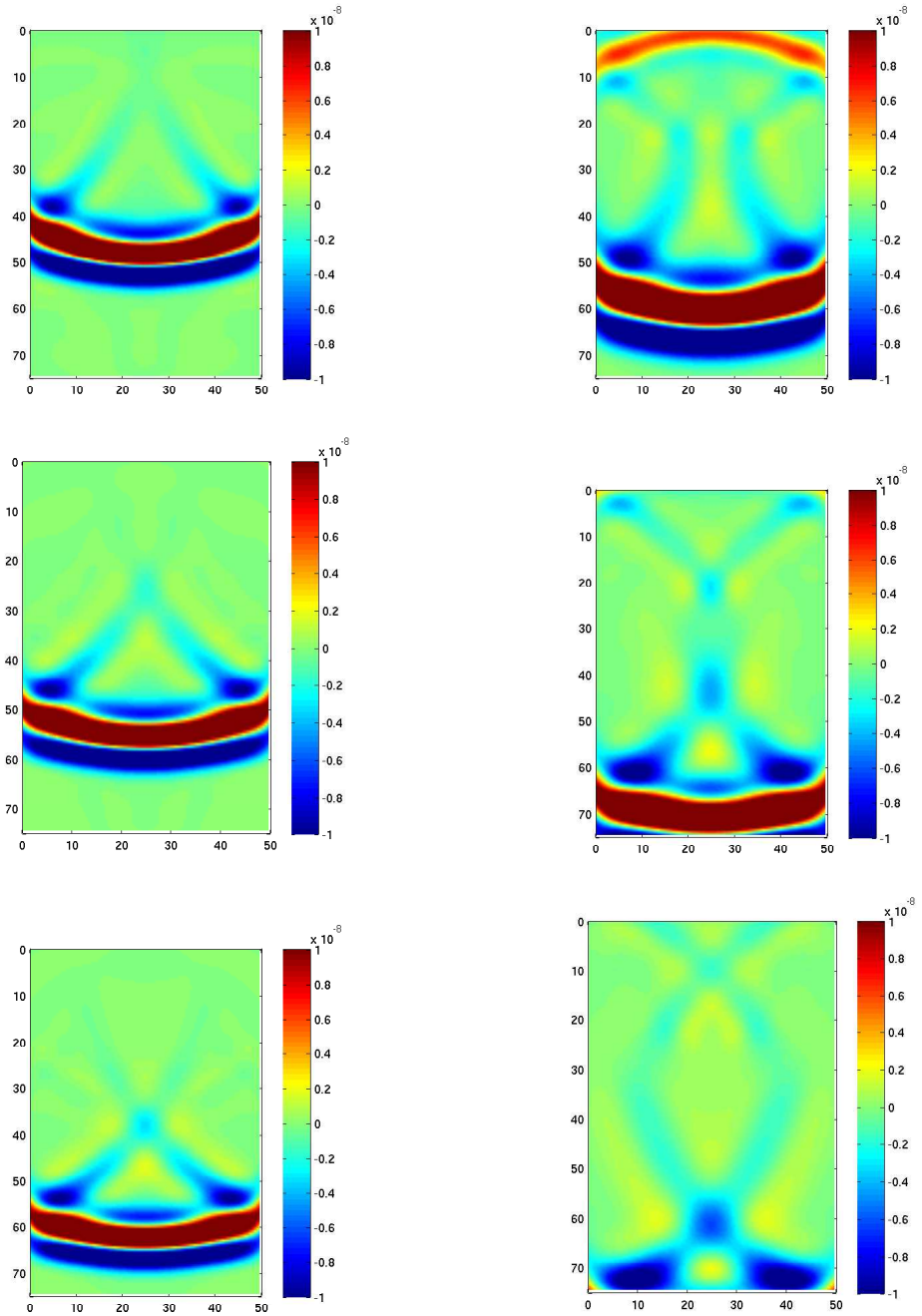


**Fig. 6** – Propagation de l'onde dans un domaine homogène (gauche) et dans un domaine bicouche (droite)

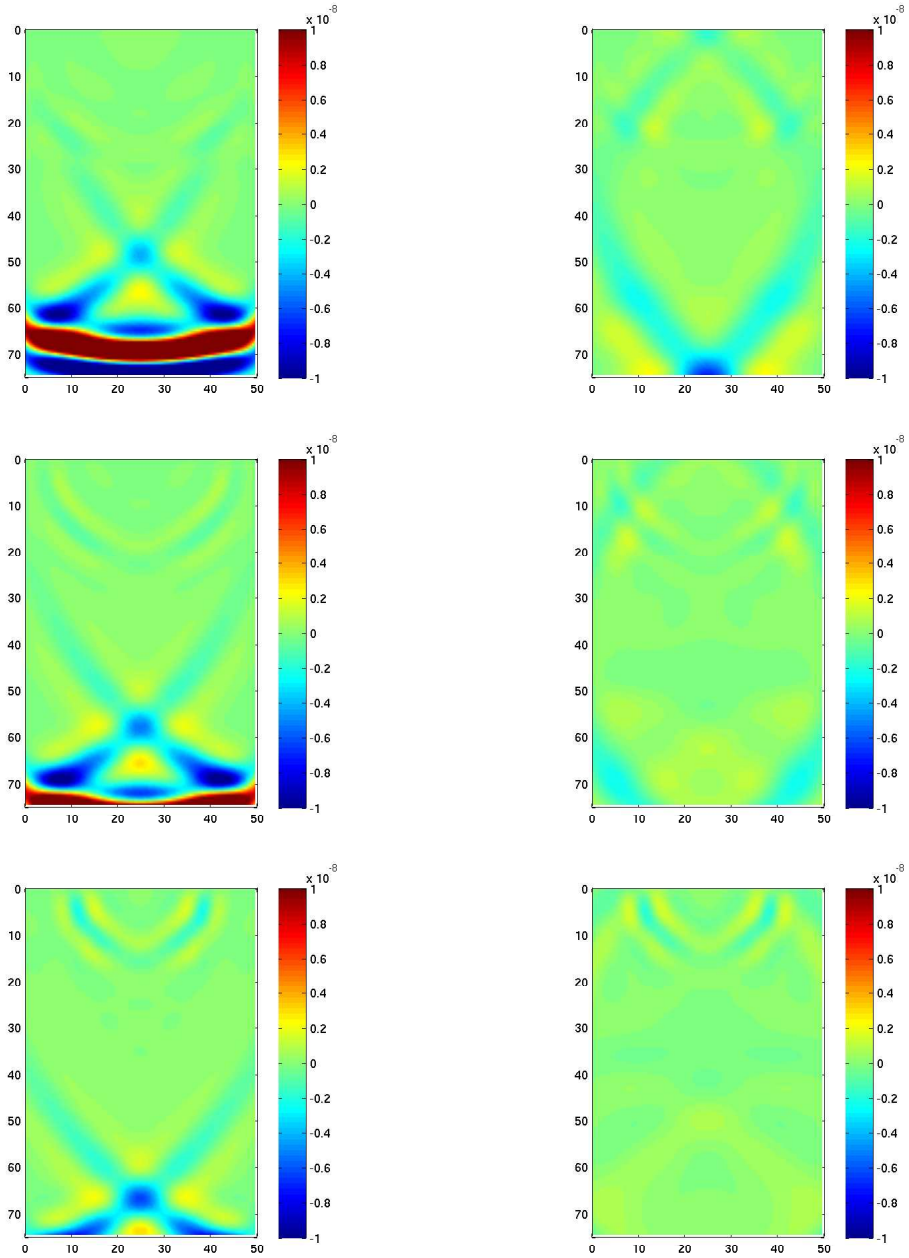


**Fig. 7** – Propagation de l'onde dans un domaine homogène (gauche) et dans un domaine bicouche (droite)

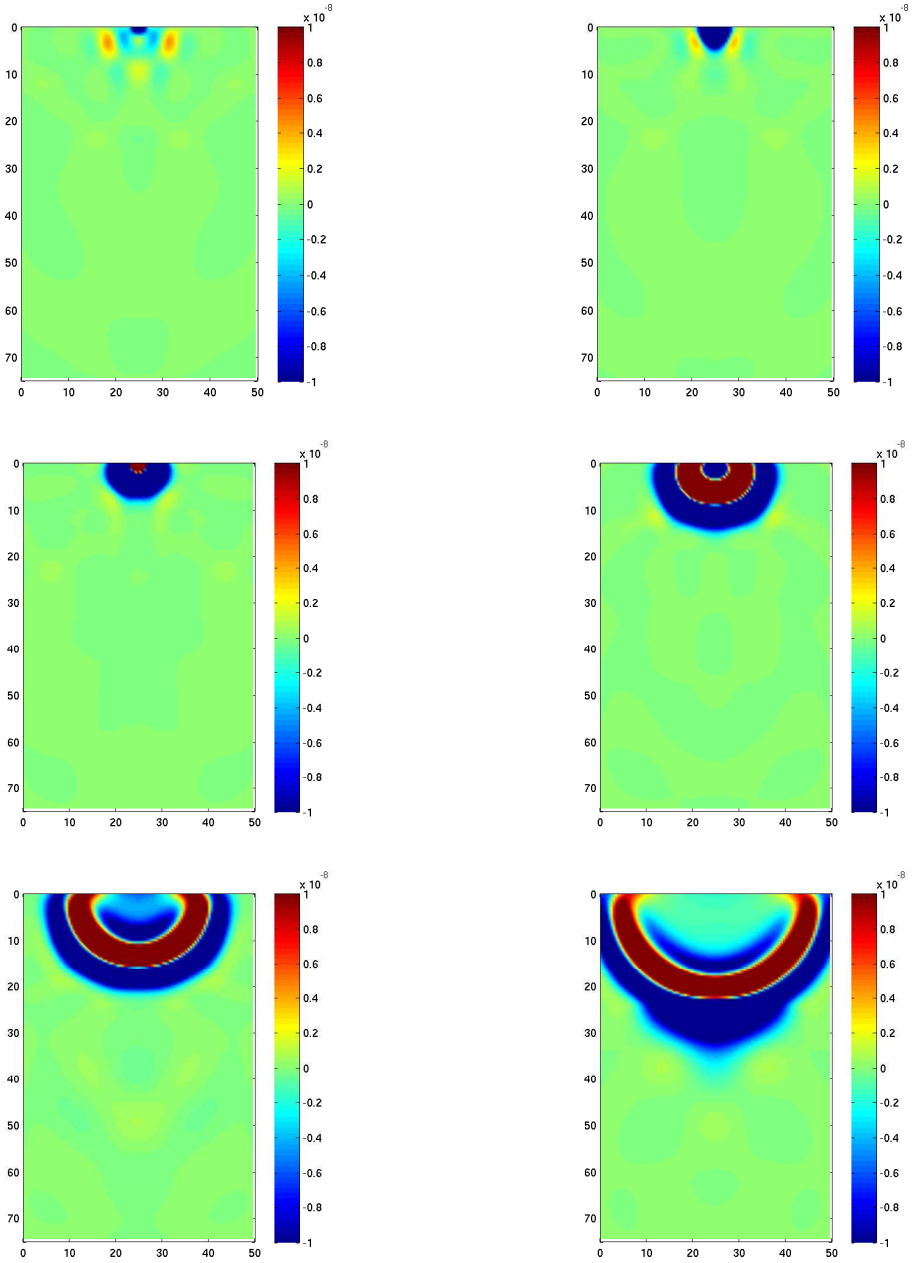

**Fig. 8** – Propagation de l'onde dans un domaine domaine bicouche avec $c_1 = 1000\,m/s$ et $c_2 = 2500\,m/s$

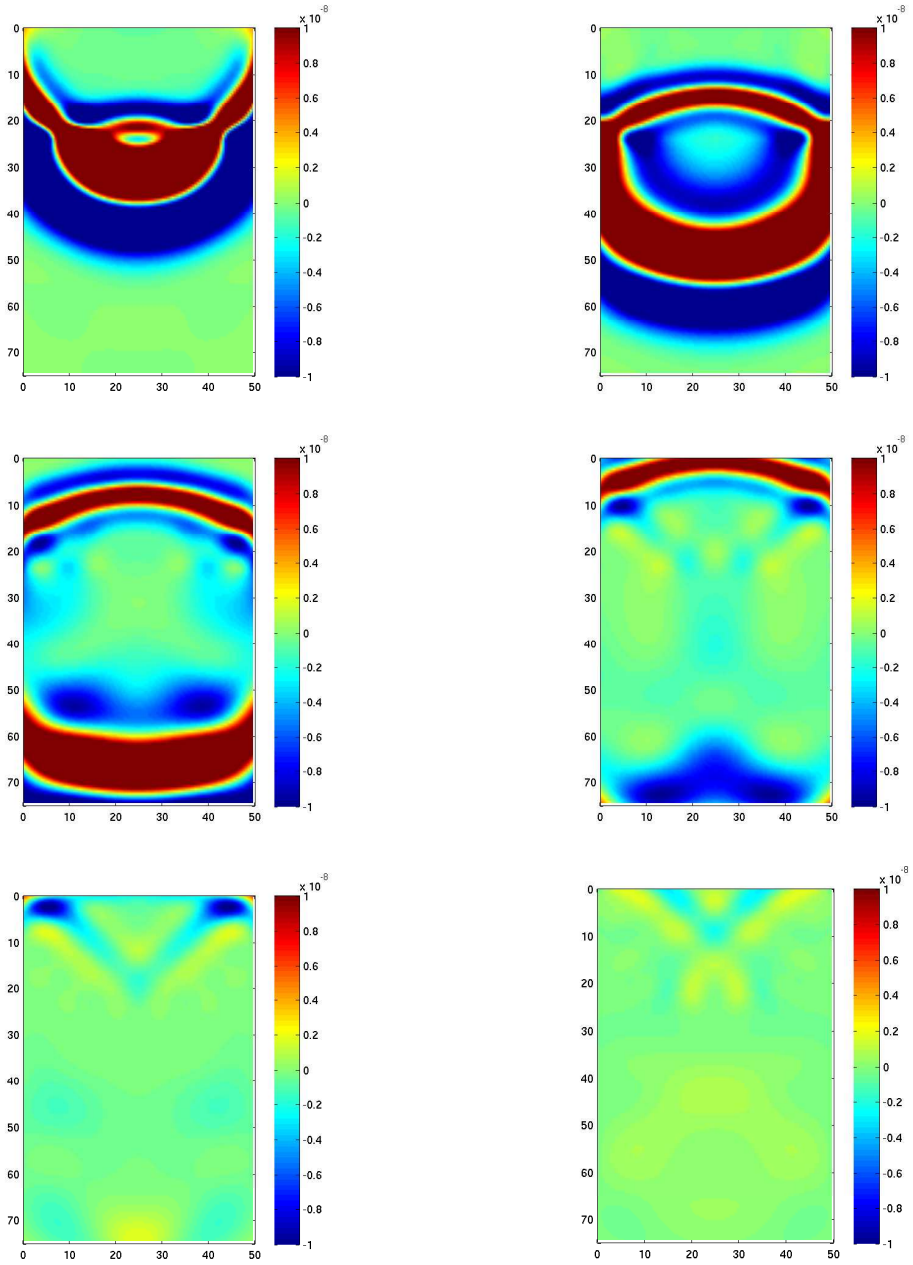

**Fig. 9** – Propagation de l'onde dans un domaine domaine bicouche avec $c_1 = 1000\,m/s$ et $c_2 = 2500\,m/s$

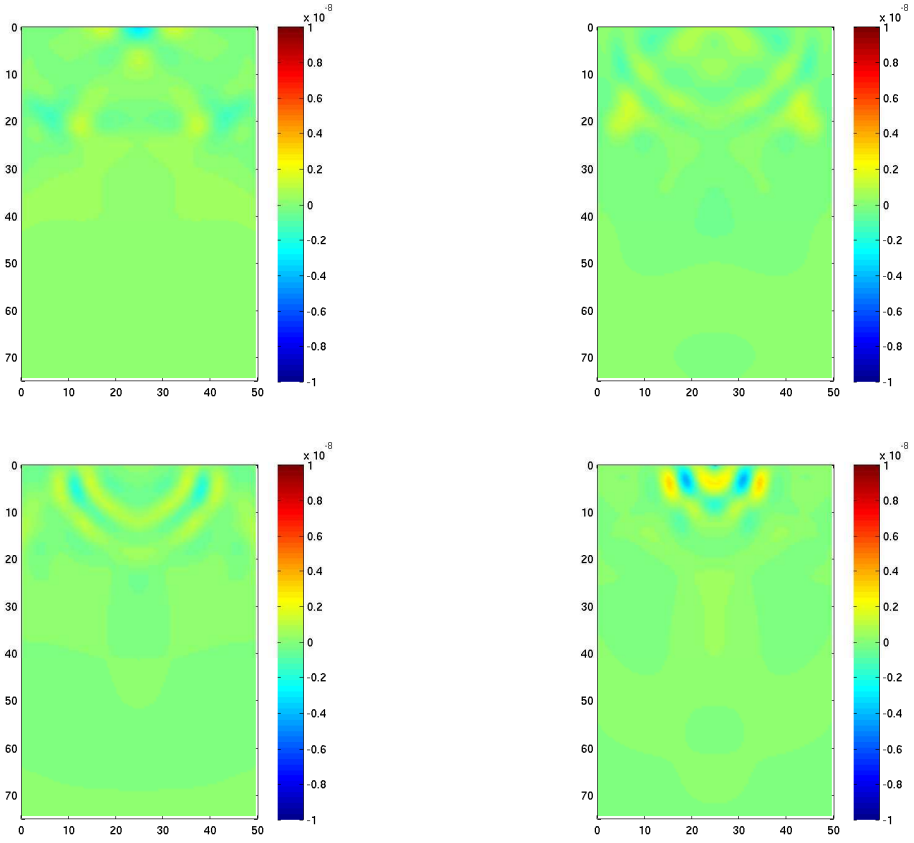


**Fig. 10** – Propagation de l'onde dans un domaine domaine bicouche avec $c_1 = 1000\,m/s$ et $c_2 = 2500\,m/s$

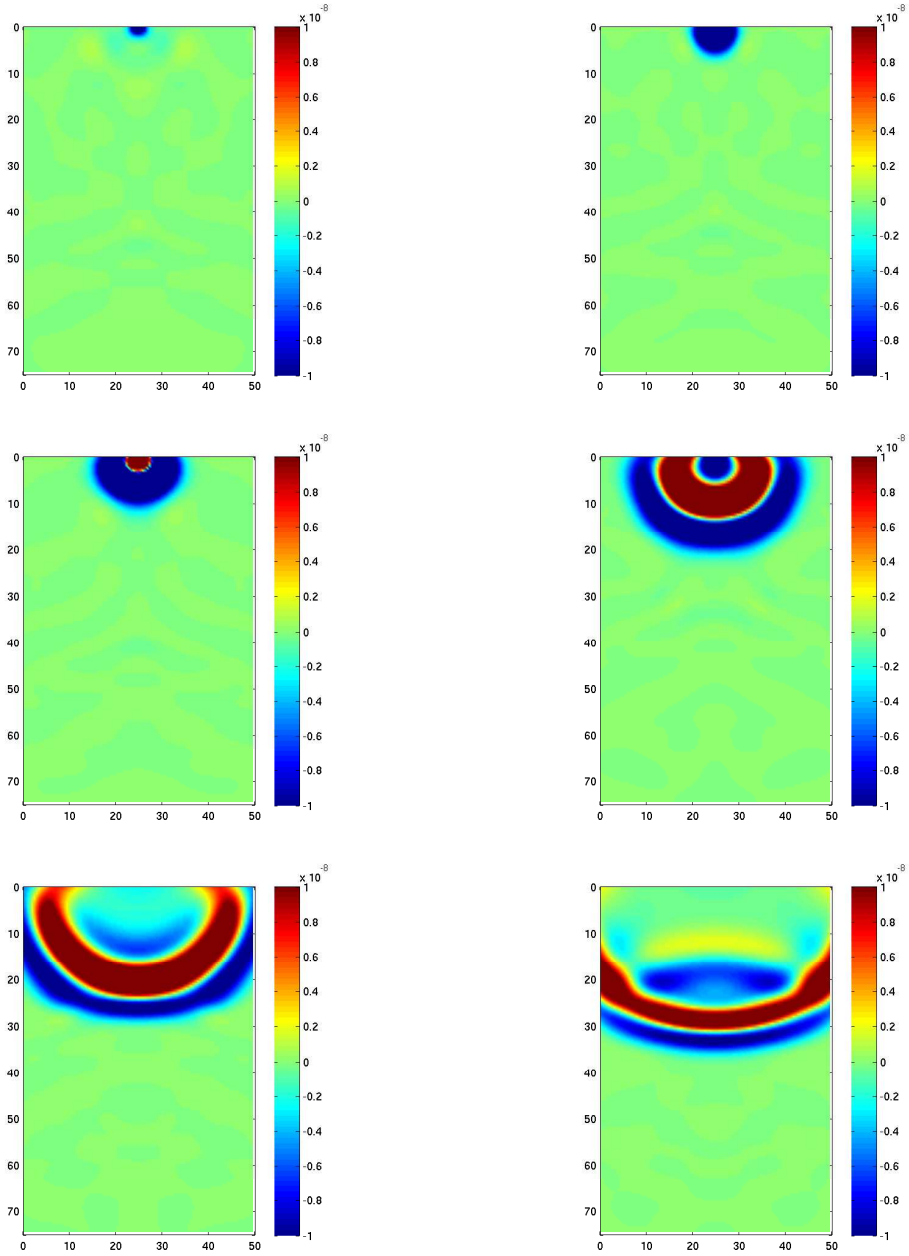

**Fig. 11** – Propagation de l'onde dans un domaine domaine bicouche avec $c_1 = 1500\,m/s$ et $c_2 = 1000\,m/s$

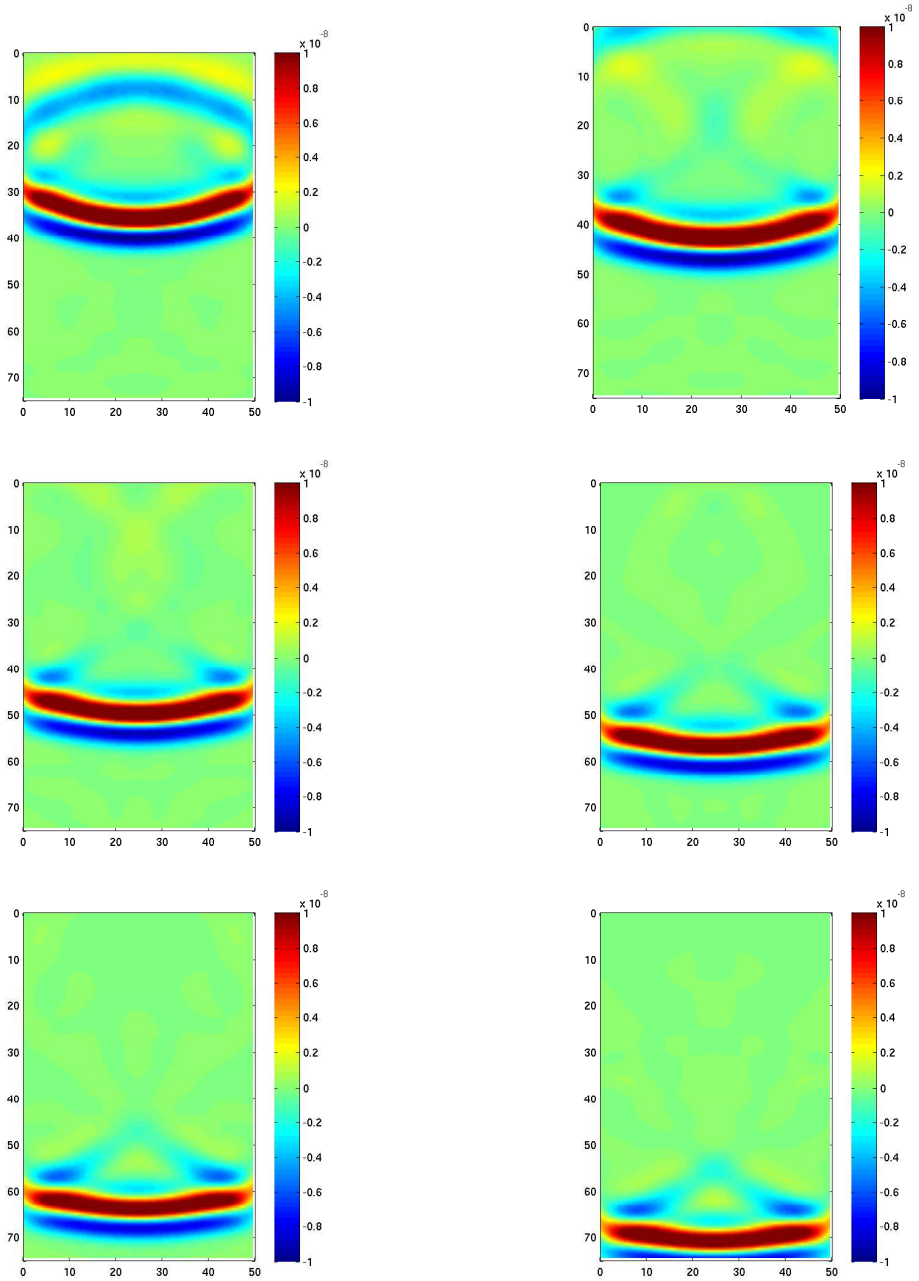

**Fig. 12** – Propagation de l'onde dans un domaine domaine bicouche avec $c_1 = 1500\,m/s$ et $c_2 = 1000\,m/s$

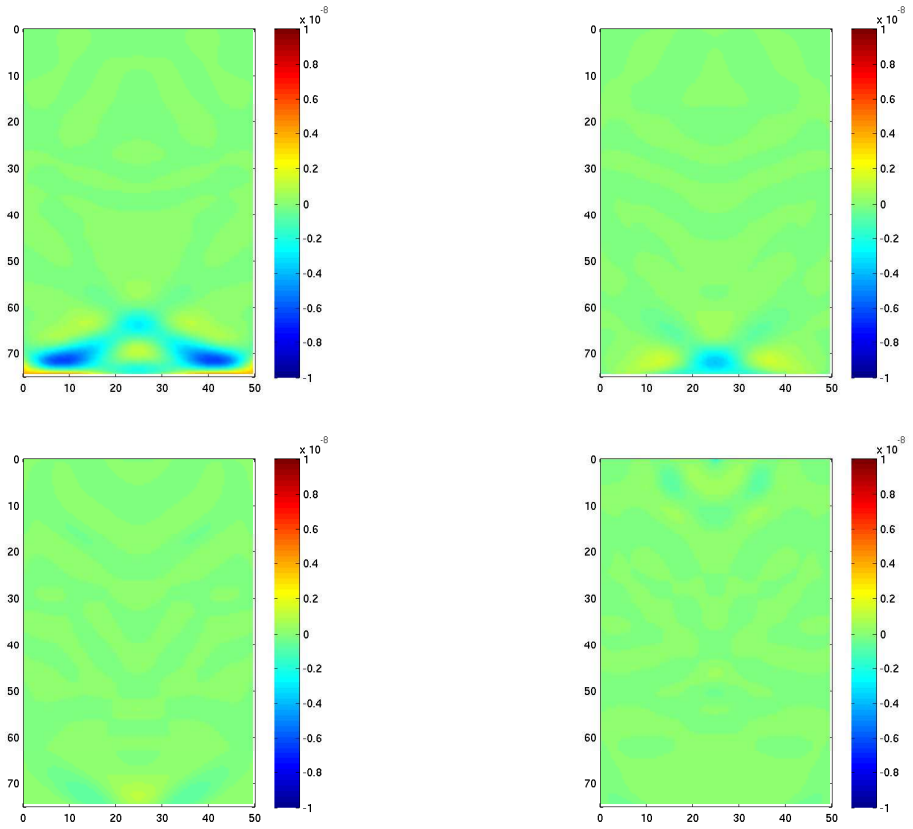


**Fig. 13** – Propagation de l'onde dans un domaine domaine bicouche avec $c_1 = 1500\,m/s$ et $c_2 = 1000\,m/s$

# Bibliographie

[1] C. Baldassari, Modélisation et simulation numérique pour la migration terrestre par équation d'ondes, *Thèse de doctorat*, décembre 2009

[2] J. Rappaz, M. Picasso, Quelques notes en complément du livre "Introduction à l'analyse numérique", Presses Polytechniques Universitaire Romandes (1998), mars 2000

[3] J. Le Roux, La transformée de Fourier et ses applications, EPU, Université de Nice Sophia-Antipolis

# Annexe 4

## 3.1. Transformée de Fourier discrète et transformée de Fourier rapide

Soit une famille de $N$ nombres (réels ou complexes) $\{f_k\}_{k=0}^{N-1}$ et posons $\omega_k = \frac{2k\pi}{N}$. On dira que la famille $\{F_k\}_{k=0}^{N}$ de nombres définis par

$$F_k = \sum_{n=0}^{N-1} f_n e^{-\mathrm{i}\omega_k n},\ k = 0, 1, ..., N-1,$$

est la transformée de Fourier discrète de la famille $\{f_k\}_{k=0}^{N-1}$.

**Remarque 3.1.** Soit $f : \mathbb{R} \to \mathbb{R}$ une fonction continue par morceaux et périodique de période $N$. En posant $\omega_k = \frac{2k\pi}{N}$, on peut écrire la série de Fourier de $f$ :

$$f(x) \sim \sum_{k=-\infty}^{\infty} c_k e^{\mathrm{i}\omega_k x},\ \text{où}\ c_k = \frac{1}{N}\int_0^N f(x)\, e^{-\mathrm{i}\omega_k x} dx.$$

Définissons maintenant les intervalles $I_n = \left[n - \frac{1}{2}, n + \frac{1}{2}\right[$ si $n = 1, 2, 3, ..., N-1$ et $I_0 = \left[0, \frac{1}{2}\right[ \cup \left[N - \frac{1}{2}, N\right]$. En plaçant le segment $[0, N]$ sur un cercle (i.e. en joignant les points 0 et $N$), on obtient une structure périodique de période $N$. La décomposition du domaine $[0, N]$ en sous-intervalles permet d'écrire :

$$c_k = \frac{1}{N}\sum_{j=0}^{N-1}\int_{I_j} f(x)\, e^{-\mathrm{i}\omega_k x} dx.$$

En approchant l'intégrale ci-dessus par la formule de quadrature du rectangle, on a

$$\int_{I_n} f(x)\, e^{-\mathrm{i}\omega_k x} dx \simeq f(n)\, e^{-\mathrm{i}\omega_k n}.$$

Ainsi, on obtient :

$$c_k \simeq \frac{1}{N}\sum_{j=0}^{N-1} f(n)\, e^{-\mathrm{i}\omega_k n}.$$

On conclut donc que si $\{f_k\}_{k=0}^{N-1}$ est la famille de nombres réels définie par $f_k = f(k)$ et si $\{F_k\}_{k=0}^{N-1}$ est la transformée de Fourier discrète de $\{f_k\}_{k=0}^{N-1}$, alors $\frac{1}{N}F_k$ est une approximation du coefficient de Fourier $c_k$ de $f$. En considérant la famille $\{f_k\}_{k=0}^{N-1}$ comme un vecteur $\overrightarrow{f}$ de composantes $f_0, f_1, ..., f_{N-1}$ et la famille $\{F_k\}_{k=0}^{N-1}$ comme un vecteur $\overrightarrow{F}$ de composantes $F_0, F_1, ..., F_{N-1}$, on

déduit facilement que si $A$ est la matrice $N \times N$ donnée par :

$$A = \begin{bmatrix} 1 & 1 & 1 & 1 & \dots & 1 \\ 1 & e^{-\mathrm{i}1\omega_1} & e^{-\mathrm{i}2\omega_1} & e^{-\mathrm{i}3\omega_1} & \dots & e^{-\mathrm{i}(N-1)\omega_1} \\ 1 & e^{-\mathrm{i}1\omega_2} & e^{-\mathrm{i}2\omega_2} & e^{-\mathrm{i}3\omega_2} & \dots & e^{-\mathrm{i}(N-1)\omega_2} \\ 1 & e^{-\mathrm{i}1\omega_3} & e^{-\mathrm{i}2\omega_3} & e^{-\mathrm{i}3\omega_3} & \dots & e^{-\mathrm{i}(N-1)\omega_3} \\ \vdots & & & & & \\ 1 & e^{-\mathrm{i}1\omega_{N-1}} & e^{-\mathrm{i}2\omega_{N-1}} & e^{-\mathrm{i}3\omega_{N-1}} & \dots & e^{-\mathrm{i}(N-1)\omega_{N-1}} \end{bmatrix},$$

alors on a

$$\overrightarrow{F} = A\overrightarrow{f}.$$

**Remarque 3.2.** La matrice $A$ est symétrique : en effet, pour tout $1 \leq j, k \leq N-1$, on a : $j\omega_k = k\omega_j$. De plus, la matrice $A$ vérifie : $AA^* = A^*A = NI$, où $I$ est la matrice $N \times N$ identité. Par conséquent, on obtient :

$$A^{-1} = \frac{1}{N}A^*$$

et puisque $\overrightarrow{f} = A^{-1}\overrightarrow{F}$, on déduit :

$$f_k = \frac{1}{N}\sum_{n=0}^{N-1} F_n e^{\mathrm{i}\omega_k n}.$$

**Définition 3.1.** Connaissant la famille $\{F_k\}_{k=0}^{N-1}$, on dit que la famille $\{f_k\}_{k=0}^{N-1}$ définie par

$$f_k = \frac{1}{N}\sum_{n=0}^{N-1} F_n e^{\mathrm{i}\omega_k n},\ k = 0, 1, ..., N-1$$

est la transformée de Fourier discrète inverse de $\{F_k\}_{k=0}^{N-1}$.
Soit maintenant une famille de nombres $\{f_k\}_{k=0}^{N-1}$ et sa transformée de Fourier discrète $\{F_k\}_{k=0}^{N-1}$. Dans la suite, nous noterons $f_{k+jN} = f_k$ et $F_{k+jN} = F_k$ pour tout $k = 0, 1, 2, ...N-1$ et pour tout $j \in \mathbb{Z}$. Ainsi, nous pouvons considérer $\{f_k\}$ et $\{F_k\}$ comme des suites infinies périodiques de période $N$.

**Définition 3.2.** Soient $\{f_k\}$ et $\{g_k\}$ deux suites de nombre périodiques de période $N$. Nous dirons que $\{h_k\}$ est le **produit de convolution** de $\{f_k\}$ par $\{g_k\}$ et nous noterons :

$$\{h_k\} = \{f_k\} * \{g_k\}$$

si :

$$h_k = \sum_{n=0}^{N-1} f_n g_{k-n}.$$

On vérifie facilement la propriété suivante : si $\{F_k\}$, $\{G_k\}$ et $\{H_k\}$ sont les transformées de Fourier discrètes de $\{f_k\}$, $\{g_k\}$ et $\{h_k\} = \{f_k\} * \{g_k\}$ respectivement, alors on a :

$$H_k = F_k \cdot G_k.$$

Théoriquement, la convolution $\{f_k\} * \{g_k\}$ nécessite $N^2$ opérations. En fait, le théorème de convolution, associé à la transformée de Fourier rapide (FFT=Fast Fourier Transform) nécessite seulement $N \log N$ opérations. La suite de cette section est dédiée à la FFT.

L'idée de base de la FFT est de calculer la transformée de Fourier discrète de deux familles de nombre dont la longueur est la moitié de la famille initiale.
Soit donc un nombre pair de nombres (réels ou complexes) $\{f_k\}_{k=0}^{2M-1}$ et décomposons cette famille de nombres en deux familles $\{f_{2k}\}_{k=0}^{M-1}$ et $\{f_{2k+1}\}_{k=0}^{M-1}$. En renumérotant les éléments de ces deux familles, on pourra écrire :

$$\left\{f_k^{(1)}\right\}_{k=0}^{M-1} \text{ et } \left\{f_k^{(2)}\right\}_{k=0}^{M-1},$$

i.e. $f_k^{(1)} = f_{2k}$ et $f_k^{(2)} = f_{2k+1}$, $k = 0, 1, ..., M-1$. Si $\left\{F_k^{(1)}\right\}_{k=0}^{M-1}$ et $\left\{F_k^{(2)}\right\}_{k=0}^{M-1}$ sont les transformées de Fourier discrètes de ces deux familles, nous avons nécessairement, si $\tilde{\omega}_k = \frac{2\pi}{M}$ et $\omega_k = \frac{2\pi}{2M}$, avec $k = 0, 1, 2, ..., M-1$ :

$$\begin{aligned} F_k^{(1)} &= \sum_{n=0}^{M-1} f_n^{(1)} e^{-\mathrm{i}\tilde{\omega}_k n} = \sum_{n=0}^{M-1} f_{2n} e^{-\mathrm{i}\omega_k 2n}, \\ F_k^{(2)} &= \sum_{n=0}^{M-1} f_n^{(2)} e^{-\mathrm{i}\tilde{\omega}_k n} = \sum_{n=0}^{M-1} f_{2n+1} e^{-\mathrm{i}\omega_k 2n}. \end{aligned}$$

On en déduit que si $\{F_k\}_{k=0}^{2M-1}$ est la transformée de Fourier discrète de $\{f_k\}_{k=0}^{2M-1}$, nous avons pour $k = 0, 1, 2, ..., M-1$ :

$$F_k = \sum_{n=0}^{2M-1} f_n e^{-\mathrm{i}\omega_k n} = F_k^{(1)} + e^{-\mathrm{i}\omega_k} F_k^{(2)}.$$

En tenant compte de $e^{-\mathrm{i}\pi} = e^{\mathrm{i}\pi} = -1$, on obtient pour $k = 0, 1, 2, ..., M-1$ :

$$\begin{aligned} F_{M+k} &= \sum_{n=0}^{M-1} f_{2n} e^{-\mathrm{i}\omega_{M+k} 2n} + \sum_{n=0}^{M-1} f_{2n+1} e^{-\mathrm{i}\omega_{M+k}(2n+1)} \\ &= \sum_{n=0}^{M-1} f_{2n} e^{-\mathrm{i}\omega_k 2n} \cdot \underbrace{e^{-\mathrm{i}\omega_M 2n}}_{1} + \sum_{n=0}^{M-1} f_{2n-1} e^{-\mathrm{i}\omega_k(2n+1)} \cdot \underbrace{e^{-\mathrm{i}\omega_M(2n+1)}}_{-1}. \end{aligned}$$

En résumé, on a le résultat de duplication suivant :

$$\begin{aligned} F_k &= F_k^{(1)} + e^{-\mathrm{i}\omega_k} F_k^{(2)} \\ F_{M+k} &= F_k^{(1)} - e^{-\mathrm{i}\omega_k} F_k^{(2)}, \end{aligned}$$

où $k = 0, 1, 2, ..., M-1$.

Pour construire $\{F_k\}_{k=0}^{2M-1}$, il suffit donc :

1. de construire $\left\{F_k^{(1)}\right\}_{k=0}^{M-1}$ et $\left\{F_k^{(2)}\right\}_{k=0}^{M-1}$ ;
2. d'exécuter les $M$ multiplications $\tilde{F}_k = e^{-\mathrm{i}\omega_k} F_k^{(2)}$, $k = 0, 1, 2, ..., M-1$ et de construire

$F_k = F_k^{(1)} + \tilde{F}_k^{(2)}$ et $F_{k+M} = F_k^{(1)} - \tilde{F}_k^{(2)}$, pour $k = 0, 1, 2, ..., M-1$.

**Nombre de multiplications** La construction de $\{F_k\}_{k=0}^{2M-1}$ à partir de deux familles $\left\{F_k^{(1)}\right\}_{k=0}^{M-1}$ et $\left\{F_k^{(2)}\right\}_{k=0}^{M-1}$ nécessite, comme décrit ci-dessus, $M$ multiplications. Si nous voulons construire les deux familles $\left\{F_k^{(1)}\right\}_{k=0}^{M-1}$ et $\left\{F_k^{(2)}\right\}_{k=0}^{M-1}$ à partir de quatre familles de longueur moitié, on doit exécuter $\frac{M}{2} + \frac{M}{2} = M$ multiplications. En continuant ce raisonnement et en supposant que $N = 2M$ est une puissance de 2, i.e. $N = 2^\gamma$, nous constatons qu'à chaque niveau (il y a $\gamma$ niveaux), on doit exécuter $M$ multiplications, soit en tout $\gamma M$ multiplications. Ainsi, le nombre de multiplications pour construire $\{F_k\}_{k=0}^{N-1}$ récursivement à partir de familles de longueur moitié sera

$$\gamma M = \frac{N}{2} \log_2 N.$$

## 3.2. Reconstruction du signal à temps continu à partir des échantillons

Le traitement numérique des signaux se fait sur des valeurs discrètes : il n'est pas possible de traiter par ordinateur des signaux à temps continu.
Si le signal à analyser ne varie pas trop rapidement et si la cadence d'échantillonnage est suffisamment élevée, on pourra retrouver le signal original à partir du signal échantillonné. Le théorème de H. Nyquist (1928), repris par C. Shannon (1948) traduit cela de manière un peu plus formelle, donnant le lien entre la bande de fréquence occupée par le signal et la cadence d'échantillonnage. Une manière raisonnable de considérer ce problème est de trouver les conditions pour lesquelles il est possible de reconstituer le signal à temps continu à partir des échantillons mémorisés. Ce développement se fait en utilisant une interprétation dans le domaine des fréquences.

### Echantillonnage

Considérons un signal en temps, $f$. Par souci de simplicité, on échantillonne le signal à un rythme régulier. Si on note $T_e$ la période (cadence) d'échantillonnage, on mesurera la valeur du signal à des instants qui sont des multiples de $T_e$. Cette opération peut être formalisée en utilisant une distribution de Dirac :

$$f(nT_e) = \int_{-\infty}^{+\infty} f(t)\, \delta(t - nT_e).$$

On peut interpréter le signal échantillonné comme une séquence d'impulsions de Dirac modulées en amplitude par le signal $f(t)$ :

$$g(t) = \sum_{n=-\infty}^{+\infty} f(nT_e)\, \delta(t - nT_e)$$

ou encore sous la forme d'un produit que l'on peut noter :

$$\begin{aligned} g(t) &= f(t) \times \left( \sum_{n=-\infty}^{+\infty} \delta(t - nT_e) \right) \\ &= f(t)\, s(t). \end{aligned} \tag{3}$$

La représentation (3) est un produit dans le domaine temporel : elle se traduit donc sous la forme d'une convolution dans le domaine des fréquences. La distribution $s(t)$ est un "peigne" d'impulsions de Dirac régulièrement espacées. Elle admet donc une transformée de Fourier, $S(\omega)$, qui, elle aussi est un "peigne" d'impulsions de Dirac régulièrement espacées, l'écart entre les harmoniques étant $\omega_e = 2\pi/T_e$ :

$$S(\omega) = \sum_{k=-\infty}^{+\infty} \delta(\omega - k\omega_e). \tag{4}$$

**Définition 3.3.** La quantité $\omega_e$, définie par :

$$\omega_e = 2\pi/T_e$$

est appelée **fréquence angulaire d'échantillonnage**. On définit aussi la **fréquence d'échantillonnage** $f_e$, par :

$$f_e = \frac{\omega_e}{2\pi}.$$

La transformée de Fourier $G$ de la fonction $g$ s'écrira donc :

$$\begin{aligned} G(\omega) &= \int_{-\infty}^{+\infty} F(\omega - \nu)\, S(\nu)\, d\nu \\ &= \int_{-\infty}^{+\infty} F(\omega - \nu) \left[ \sum_{k=-\infty}^{+\infty} \delta(\nu - k\omega_e) \right] d\nu. \end{aligned}$$

La convolution $G_k(\omega)$ d'une fonction $F(\omega)$ par une impulsion de Dirac décalée en $k\omega_e$ se traduit par une translation de $k\omega_e$ :

$$G_k(\omega) = \int_{-\infty}^{+\infty} F(\omega - \nu)\, \delta(\nu - k\omega_e)\, d\nu = F(\omega - k\omega_e).$$

Pour obtenir $G(\omega)$, on effectue la somme du résultat des convolutions de $X(\omega)$ par les différentes impulsions $\delta(\omega - k\omega_e)$ ; on en déduit donc :

$$G(\omega) = \sum_{k=-\infty}^{+\infty} G_k(\omega) = \sum_{k=-\infty}^{+\infty} F(\omega - k\omega_e). \tag{5}$$

La transformée de Fourier du signal échantillonné s'obtient par addition des reproductions de la transformée de Fourier du signal original identiques en forme, mais décalées les unes des autres de $\omega_e$. C'est donc une fonction périodique dont la période est la fréquence angulaire d'échantillonnage $\omega_e$.

### Réconstruction du signal à temps continu à partir des échantillons

L'opération (5) d'addition des reproductions décalées de la transformée de Fourier n'est pas, en général, une opération réversible : supposons par exemple que le signal en temps continu $f(t)$ est réel et a une composante non-nulle aux fréquences angulaires $\omega_e/2$ et $-\omega_e/2$. On aura :

$$F(-\omega_e/2) = \overline{F(\omega_e/2)}.$$

Lors de l'échantillonnage, on additionnera les répliques décalées de ces composantes, si bien que le résultat sera nécessairement une composante réelle en $\omega_e/2$ en $-\omega_e/2$ : l'information sur la partie imaginaire de $\omega_e/2$ sera perdue.
Pour reconstituer un signal en temps continu à partir de ses échantillons, il faut que le signal continu avant échantillonnage respecte certaines contraintes :

- il ne faut pas qu'une composante à une fréquence $\omega$ du signal échantillonné provienne de plusieurs composantes du signal à temps continu dans l'addition des répliques (5) ;
- pour chaque composante du signal échantillonné, il faut connaître la bande de fréquence de largeur $\omega_e$ dont elle était originaire dans le signal en temps continu.

**Théorème 3.1.** (**Théorème de Nyquist − Shannon**) On peut reconstruire le signal en temps continu à partir de ses échantillons si la bande de fréquence occupée par un signal réel est inférieure à la moitié de la fréquence d'échantillonnage. Alors, les répliques de $F(\omega)$ ne se chevauchent pas et on connait la bande de fréquence initiale du signal $f(t)$.
La fréquence d'échantillonnage doit ainsi être adaptée à la bande de fréquence occupée par le signal.

# Conclusion Générale

# 1. Bilan

Nous nous sommes intéressés dans ce travail au développement et à l'analyse numérique de méthodes numériques capables de résoudre efficacement les problèmes de Helmholtz, notamment en régime haute fréquence. C'est l'objectif principal que nous nous sommes fixé. La méthode dite DGM (*d*iscontinuous *G*alerkin *m*ethod) développée par Farhat *et al* a été la motivation de ce travail de recherche. En effet, en dépit des résultats numériques très impressionnants publiés dans la série d'articles [13,14,15] (références dans la Bibliographie Générale), cette méthode ne nous semble pas en mesure d'atteindre un tel objectif. Ceci est dû essentiellement aux instabilités numériques que nous avons découvertes pour des maillages qui devenaient de plus en plus fin. Nous avons donc décidé de reprendre les idées essentielles de la formulation DGM et de construire une méthode similaire, mais qui soit stable. En d'autres mots : une méthode qui possède les avantages (simplicité de la formulation et implémentation, excellente précision à des coûts raisonnables) de DGM sans ses inconvénients (les instabilités numériques).
Nous avons ainsi construit successivement (dans le temps) trois méthodes qui diffèrent par (a) la manière de traiter la continuité aux interfaces des éléments et (b) la manière de formuler et de résoudre les problèmes locaux. Le tableau 1 récapitule les différences et les similitudes entre les trois méthodes proposées.
Dans la première méthode que nous avons construite, nous avons essayé de résoudre les problèmes d'instabilités numériques qui apparaissent dans DGM, en traitant différemment la façon de restaurer la continuité aux interfaces des éléments. Cette nouvelle approche nous a permis de rendre le système global (correspondant aux multiplicateurs de Lagrange) Hermitien et semi-défini positif. Nous avons aussi analysé mathématiquement les propriétés de cette nouvelle méthode et nous avons réussi à établir des estimations *a priori* explicitant la dépendance de l'erreur en norme $L^2$ par rapport au nombre d'onde $k$, au pas de maillage $h$ et à la fois au nombre d'ondes planes et de multiplicateurs de Lagrange. En revanche, les expériences numériques que l'on a effectuées ont vite montré les limites de cette nouvelle approche puisque les instabilités numériques, bien que moins importantes que dans DGM, sont toujours présentes et gênantes. Ces instabilités sont dues, nous semble-t-il, au fait que les problèmes locaux deviennent presque singuliers lorsque l'on raffine le maillage. Nous avons donc modifié cette méthode en adoptant une nouvelle formulation des problèmes locaux qui permet de les rendre bien posés au sens de Hadamard. C'est ainsi que la deuxième méthode que l'on a appelée mDGM (*m*odified *d*iscontinuous *G*alerkin *m*ethod) est née. Les résultats obtenus pour cette méthode dans le cas du problème de guide d'onde sont nettement meilleurs que l'approche que nous avons proposée au départ. Les instabilités numériques apparaissaient encore, à notre étonnement, lorsque la taille des éléments du maillage devenait très petite. Une étude plus détaillée des problèmes locaux a révélé que les ondes planes utilisées comme fonctions de forme devenaient numériquement linéairement dépendantes à partir d'un certain seuil de pas de maillage. Cette perte de l'indépendance est la source des instabilités numériques car les systèmes locaux deviennent presque singuliers avec le raffinement du maillage. Par conséquent, nous avons reformulé les problèmes locaux de façon à obtenir une formulation plus robuste et moins sensible à la perte de l'indépendance linéaire des fonctions de base. C'est ce qui a donné lieu à la troisième méthode que l'on a appelée imDGM (*i*mproved *m*odified *d*iscontinuous *G*alerkin *m*ethod). Les résultats numériques obtenus montrent clairement que imDGM est stable numériquement même pour des résolutions du maillage qui dépasse 1000 éléments par longueur d'onde. La précision de la méthode est remarquable même dans le cas des maillages très grossiers. Par exemple, pour des problèmes de guide d'onde et pour $ka = 200$, imDGM muni de l'élément $R$-11-3 nécessite seulement 3 éléments par longueur d'onde pour atteindre une précision de l'ordre de 0.2%.
Comme il a déjà été précisé dans le manuscript, imDGM est une méthode qui peut être considérée

comme une approche à mi-chemin entre DGM et LSM (*l*east-*s*quares *m*ethod), développée par Monk-Wang dans [34] (référence dans la Bibliographie Générale). En effet, comme illustré dans le tableau 2, imDGM ne diffère de DGM que par le traitement des problèmes locaux et de la discontinuité aux interfaces des éléments. Cette différence fait que imDGM est nettement supérieure à DGM, à la fois en terme de précision et de stabilité. D'autre part, imDGM diffère de LSM par la présence des multiplicateurs de Lagrange mais ressemble beaucoup à LSM par ce qui est de la résolution du problème global, bien que de taille différente. En effet, la taille du système global de imDGM est liée au nombre de multiplicateurs de Lagrange aux interfaces des éléments, alors que la taille du système à résoudre dans LSM dépend du nombre d'ondes planes (ou de fonctions de forme utilisées) au niveau des éléments. Cette ressemblance est, à notre avis, la raison pour laquelle les performances des deux méthodes sont assez comparables. En revanche, en raison de la différence des tailles des systèmes, imDGM peut s'avérer, nous semble-t-il, moins coûteuse que LSM lorsque l'on utilise des éléments d'ordre élevé et dans le cas des problèmes tridimensionnels à haute fréquence. Cette remarque nécessite cependant une étude plus approfondie pour être validée.

TAB. 1 – Comparaison des trois méthodes proposées : formulation locale pour un élément intérieur et formulation globale.

| Méthode | Partie I | Partie II (mDGM) | Partie III (imDGM) |
|---|---|---|---|
| Formulation locale | $\int_{\partial K} \partial_n v \partial_n \overline{w} = \int_{\partial K} \mu \partial_n \overline{w}$ | $\int_{\partial K} (\partial_n v - \mathrm{i}\, kv)\, \overline{w} = \int_{\partial K} \mu \overline{w}$ | $\int_{\partial K} \left(\partial_n v \partial_n \overline{w} + k^2 v \overline{w}\right) = \int_{\partial K} \mu \left(\partial_n \overline{w} + \mathrm{i}\, k \overline{w}\right)$ |
| Formulation globale | $\sum_{e-\text{interior edge}} \frac{k^2}{\lvert e\rvert} \int_e [\Phi(\lambda)][\overline{\Phi(\mu)}] + \sum_{e-\text{interior edge}} \frac{1}{\lvert e\rvert} \int_e [[\partial_n \Phi(\lambda)]]\, \overline{[[\partial_n \Phi(\mu)]]} = - \sum_{e-\text{interior edge}} \frac{1}{\lvert e\rvert} \int_e [\varphi][\overline{\Phi(\mu)}]$ | | |

TAB. 2 – Comparaison de DGM, LSM et imDGM : formulation locale pour un élément intérieur et formulation globale.

| Méthode | DGM | LSM | imDGM |
|---|---|---|---|
| Formulation locale | $\int_{\partial K} \partial_n v \overline{w} = \int_{\partial K} \mu \overline{w}$ | - | $\int_{\partial K} \left(\partial_n v \partial_n \overline{w} + k^2 v \overline{w}\right) = \int_{\partial K} \mu \left(\partial_n \overline{w} + \mathrm{i}\, k \overline{w}\right)$ |
| Formulation globale | $\sum_{e-\text{interior edge}} \frac{1}{\lvert e\rvert} \int_e [\Phi(\lambda)]\overline{\mu}$ $= - \sum_{e-\text{interior edge}} \frac{1}{\lvert e\rvert} \int_e [\varphi]\overline{\mu}$ | $\sum_{e-\text{interior edge}} \frac{k^2}{\lvert e\rvert} \int_e [u][\overline{w}]$ $+ \sum_{e-\text{interior edge}} \frac{1}{\lvert e\rvert} \int_e [[\partial_n u]]\, \overline{[[\partial_n w]]}$ $= - \sum_{e-\text{interior edge}} \frac{1}{\lvert e\rvert} \int_e [\varphi][\overline{w}]$ | $\sum_{e-\text{interior edge}} \frac{k^2}{\lvert e\rvert} \int_e [\Phi(\lambda)][\overline{\Phi(\mu)}]$ $+ \sum_{e-\text{interior edge}} \frac{1}{\lvert e\rvert} \int_e [[\partial_n \Phi(\lambda)]]\, \overline{[[\partial_n \Phi(\mu)]]}$ $= - \sum_{e-\text{interior edge}} \frac{1}{\lvert e\rvert} \int_e [\varphi][\overline{\Phi(\mu)}]$ |

## 2. Perspectives

Bien entendu, cette étude est loin d'être finie et n'est en fait qu'à son tout début. En effet, il nous semble qu'il serait très intéressant de poursuivre ce travail en accomplissant (à court terme) les tâches suivantes :

- Poursuivre l'analyse de la performance de cette méthode pour les éléments d'ordre élevé. Cette tâche n'est pas difficile à faire. Le code que nous avons mis au point le permet aisément. En effet, la méthode a été implémentée de façon à pouvoir considérer n'importe quel élément sans effort supplémentaire de programmation. Son exécution nécessite seulement du temps.
- Tester d'autres fonctions de forme telles que les fonctions de Bessel ou les polynômes oscillants utilisés dans [2,18] (références dans la Bibliographie Générale).
- Évaluer la performance de la méthode pour les problèmes de scattering et notamment lorsque le maillage est non-structuré afin de mesurer son effet sur la précision.
- Poursuivre l'analyse mathématique de imDGM. Ceci peut être réalisé en utilisant les outils et les techniques présentés dans la partie I.

Les tâches que nous suggérons sont essentielles avant d'appliquer cette méthode aux problèmes tridimensionnels.
A long terme, il serait intéressant d'appliquer cette méthode aux problèmes d'élasto-acoustique et aux problèmes issus de la géophysique. Nous avons déjà entamé un programme de recherche dans cette direction en considérant un problème assez simple de géophysique (voir la partie IV). Bien entendu, (presque) tout reste à faire dans ce domaine.

## 3. Conclusion

Pour conclure, ce travail a abouti au développement et à l'implémentation d'une méthode numérique pour la résolution des problèmes de Helmholtz à haute fréquence qui nous paraît très prometteuse au vu des résultats obtenus assez probants. En outre, ce travail a permis de poser un certain nombre de questions assez intéressantes, nous semble-t-il, qui peuvent constituer un programme de recherche très porteur.

## Contribution à la résolution numérique des problèmes de Helmholtz


### Résumé :

Dans ce travail, nous nous sommes intéressés au développement et à l'analyse numérique de méthodes numériques capables de résoudre efficacement les problèmes de Helmholtz à 2D, notamment en régime moyenne et haute fréquence. La méthode que nous proposons s'inscrit dans la lignée des méthodes de type Galerkin discontinues (DG). Dans chaque élément du maillage, la solution est approchée en utilisant une superposition d'ondes planes. La continuité de la solution aux interfaces est renforcée en utilisant des multiplicateurs de Lagrange. La méthodologie proposée est une procédure en deux étapes : nous résolvons d'abord des problèmes locaux bien posés et ensuite un système global issu de la condition de continuité imposée sur les interfaces. Les plus importantes propriétés de la méthode sont : (a) les problèmes locaux obtenus sont associés à des matrices Hermitiennes et définies positives et (b) le système global, à résoudre dans la deuxième étape, est associé à une matrice Hermitienne et semi-définie positive. Les résultats numériques obtenus montrent la supériorité de la méthode proposée par rapport aux méthodes de type élément fini standard, mais aussi par rapport à d'autres méthodes de type DG, comme par exemple celle développée par Farhat et al (2003).



**Mots clés :** Propagation d'ondes, Éléments finis, Méthodes de Galerkin discontinues, Estimateurs d'erreur *a priori*, Simulation numérique


## Numerical methods for solving Helmholtz problems


### Abstract :

In this work we focus on the design and the analysis of numerical methods for solving efficiently 2D Helmholtz problems in the mid- and high-frequency regime. We propose a new discontinuous Galerkin (DG) method for solving high frequency Helmholtz problems. At the element level, the solution is approximated by a superposition of plane waves. The continuity of the solution at the interior interfaces is enforced weakly with Lagrange multipliers. The proposed formulation can be viewed as a two-step procedure in which we solve well-posed local problems, and then a global system arising from the continuity condition. The main features of the proposed solution methodology are : (a) the resulting local problems are associated with positive definite Hermitian matrices, and (b) the global system to be solved in the second step corresponds to a positive semi-definite Hermitian matrix. The obtained numerical results clearly indicate that the proposed solution methodology outperforms standard finite element methods, as well as existing DG methodologies, such as the method proposed by Farhat et al (2003).



**Keywords :** Wave propagation, Finite elements, Discontinuous Galerkin methods, *A priori* error estimates, Numerical simulation