Mathematical and numerical problems of some wave phenomena appearing in magnetic plasmas

Lise-Marie Imbert-Gérard

Laboratoire Jacques-Louis Lions Université Pierre et Marie Curie

Thèse

Présentée à

l'Université Pierre et Marie Curie

Ecole Doctorale de Sciences Mathématiques de Paris Centre

Par Lise-Marie Imbert-Gérard

pour obtenir le grade de

Docteur

Spécialité : Mathématiques

Analyse mathématique et numérique de problèmes d'ondes apparaissant dans les plasmas magnétiques

Directeur de recherche : Bruno Després

Soutenue le 9 septembre 2013 devant le jury composé de :

Devant la commission d'examen formée de :

M. Grégoire Allaire	Pr à l'Ecole polytechnique	examinateur
M. Patrick Ciarlet	Pr à l'ENSTA ParisTech	rapporteur
M. Bruno Després	Pr à l'UPMC	directeur
M. Houssem Haddar	DR à l'INRIA	examinateur
M. Ralph Hiptmair	Pr à l'E.T.H. Zurich	rapporteur
M. Patrick Joly	DR à l'INRIA	examinateur
M. Yvon Maday	Pr à l'UPMC	examinateur

Laboratoire Jacques-Louis Lions
Boîte courrier 187
4 place Jussieu
75 252 Paris cedex 05

UPMC - ED 386
Boîte courrier 290
4 place Jussieu
75 252 Paris cedex 05

Françaises, Français
Belges, Belges,
Poitevins, Poitevines,
Poids lourds, Poids plumes,
Mon Président français de souche, Mon émigré préféré,
Monsieur le Massif Central, au sommet dégarni par la violence du vent de l'histoire,
Mon président, Mon chien,
Mont Saint-Michel, Mon Saint-Bernard,
Public chéri mon amour.

Extraits choisis des Réquisitoires de qui l'on sait.

Et comme me le disait Joe D. hier encore :
A toi.

Par ailleurs, comme disait cet homme qui a changé la vie de tous ceux à qui il a été donné d'assister à l'une de ses représentations :

> *On t'a r'connu,*
> *t'as trop d'racines pour être honnête.*
> *Tu es le polynôme nul.*

Pierre Yves Jamet, professeur O combien enthousiaste et théâtral de mathématiques.

> *[Some] mathematicians [] do not attack difficult problems and settle for more elementary ones (for them), in order to be praised for solving many of these easy problems, instead of trying to open new doors. There should be no shame for failing to open a closed door behind which something very interesting is supposed to be found, but one should be able to show that one made efforts in reasonable directions ; one should also give advice to those who also plan some attempts, explaining what was tried before, and possibly why it did not work.*

Luc Tartar, The General Theory of Homogenization.

Remerciements

A Bruno. Pour tout ce que j'ai appris pendant ma thèse, souvent sans m'en rendre compte. Pour ta façon de faire, de laisser faire et parfois de ne pas faire, qui m'a surprise si souvent. Pour une porte toujours ouverte et des heures passées à échanger devant un tableau. Pour n'avoir jamais cessé de siffloter en arrivant au labo, bravant jusqu'au bout le risque de me voir débarquer sans même attendre le premier café du matin. Pour tous les déplacements et les conférences qui ont égaillé ces trois années, et pour toutes les rencontres qui s'en sont suivies.

Celui de tes concepts qui restera mon favori : "faire au moins une chose par jour". Conseil salutaire s'il en est.

A Stéphane Heuraux, pour les pistes qui nous ont amenés jusqu'ici aujourd'hui.

A Ricardo Weder, pour son regard différent sur un probeème difficile, pour cette collaboration fructueuse, pour ses relectures toujours attentives.

A Peter Monk, pour m'avoir acceuillie à l'Universitè du Delaware. Pour avoir partagé sa grande connaissance de la FVUF (ou UWVF) et son expertise de programmeur. Et pour des idées qui ont germé après cette visite.

A Nicole, Charles et Mamadou, par ordre approximativement croissant de taille. Parce que sans vous, elle aurait été moins bien cette thèse. Tant du point de vue du résultat que du déroulement.

Aux autres mathématiciens dont j'ai croisé la route pendant ces dernières années. Pour leurs conseils, pour leurs cours, pour leur aide ou pour simplement pour le plaisir d'avoir parlé avec eux. Parce que ça donne envie de suivre l'exemple. Grégory Vial, Laurent Boudin, Denis Hartemann, Isabelle Gallagher, Frédéric Coquel, Grégoire Allaire, Frédéric Hecht, Pascal Frey, Simon Masnou, Philippe Saadé, Francisco Sayas, Yvon Maday, Guilaume Sylvand, et tous les autres.

A tous les membres du Laboratoire Jacques-Louis Lions que j'ai croisés pendant ma thèse, parce que c'est aussi grâce à eux que ce laboratoire est si agréable et si stimulant.

A vous qui ne liraient de cette thèse que ces remerciements, et quelques lignes de l'introduction. Il est utile de préciser ici que l'exercice le plus réjouissant d'une thèse, c'est bien l'écriture des remerciements. On s'inspire parfois de certains exemples, évitant de son mieux leurs écueils. Mon choix trouvera son style dans une séance de cinéma qui a marqué les mémoires, à Odéon, "Cool et relax".
Tout cela reste valable pour ceux qui accidentellement auront jeté un coup d'oeil au reste du document, alors qu'il tombait à terre ou servait d'éventail.

Que tous les acteurs de cette petite-histoire-personnelle (PHP) des mathématiques se trouvent ici remerciés à la hauteur parfois vertigineuse, voire rocambolesque, de leur contribution. Et que personne ne prenne ombrage d'un oubli malheureux.

Les prémices de cette épopée se fondent entre un certain Jacques qui verse de l'eau depuis un cône vers un cylindre pour en tirer une loi de volume, jusqu'à cette époque inévitable où les mathématiques deviennent éblouissantes lors de l'entrée en scène d'un professeur théâtral.

Ca commence plus ou moins sérieusement autour d'un panier de cerises, rue de la petite tarasque, par la définition d'un produit scalaire. Un enthousiasme sans fin pour les maths, qui ne souffrira aucune faille pendant des années. Et puis le prof ... Prise de conscience, je commence à investir.

Souvenirs flous, le chiffre 2, accompagné du cardinal 39eme, au numéro 123 d'une rue Saint-Jacques autour de Victor Hugo. Prise de court, je trébuche. Pourtant avec le soutien logistique et moral du N3, des funky pâtes et du C4, je m'acharne au A5. Les seuls encouragements prodigués, néanmoins insensés comme on s'accordât à le dire dès que la décence le permît, furent salutaires.
Still believe in miracles, yst.

Quelques jours en banlieue parisienne, beaucoup de questions et quelques rencontres, pas de concert de L.A., puis c'est le tremplin vers Ker Lann. La rue du Griffon et son tire-bouchon, la rue des portes mordelaises et son BDE. La rue de Lorgeril, les nuits sans fin place de la République, le Qqt. Le mémorable Congo-Mali. Prise de remords, je me re-concentre.

Et puis voilà la place de lices et son tableau noir. Deux cahiers de développements, la BU, même le samedi, la bibliotèque de l'IRMAR, le marché et le bar'antic. Les dimanche-soirs à la colloc' de République, les séances de rugby, le gang de l'option CS et les séances PALOMa rue de Lorgeril(2).

Un Cemracs, il paraît qu'il faut en faire un pendant sa thèse. Maths pendant la jourée, week-ends et soirées à volonté.

Pour les 4 mousquetaires, pour tout ce qu'on a partagé. Pour le toro-piscine, les apéros-sortie-du-bureau, les apéros-sans-fin, le déjeuner quotidien et toutes les soirées, les textos salvateurs et les bonnes blagues, les sushis et les déménagements. Parce qu'on a commencé ensemble, et que c'est pas rien. Mais surtout pour les apéros.

Et puis pour les autres doctorants, ceux d'avant et ceux d'après nous. Pour le français avec accent grec et son rire inimitable, pour le français avec accent chilien et son retard légendaire, pour l'agence de pub pour Taiwan, pour la SMAI, pour les soirées Doc'Up, pour les goûters du GTT, les voisins de bureau et tous les autres.

Plus personnellement, et dans un chaos tant chronologique qu'alphabétique :
pour les morceaux, les vidéos et tout ce que tu regardes, sur l'écran comme dans la vie, et tant pis pour les chemins qui bifurquent,
pour un conseil, pour des conseils, pour des réponses à des questions sans fin,
pour la symbiose des meilleurs apéros du monde, sur les quais ou sur un canapé,
pour le cuistot qui nous régale à chaque fois, que ce soit avec ou sans chocolat,
pour les gorges de samaria et gorges déployées de ces jours-là,
pour les repas de famille du dimanche midi chez les frères,
pour la chapelle des lombards et autres épopées nocturnes,
pour un N'Importe-Quoi, pour tous les autres,
pour les coinches du dimanche et leur Nostalgie,
pour des croissants arrivés d'Ouganda, tôt un matin,
pour le bebop et le wcs, pour les cours-crêperie-caveau, et pour bien plus que ça,
pour la maison de famille, dont mes parents gardent les portes ouvertes,
pour la côte des gardes et la forêt de Meudon,
pour les gens qui vous attendent à l'arrivée,
pour toutes les Embuscades, plus ou moins tendues,
pour le Chili, también,
pour les truffades, les burgers et les sorties ciné,
pour cette soirée inoubliable après le festival d'Avignon,
pour tous les souvenirs d'anniversaires,
pour toi qui es toujours là, quelque soit le pays dans lequel tu vis,
pour une semaine du jour de l'an passée à glander sur mon canapé pour me faire réviser

mes développements tous les soirs,
pour la chronique style et savoir-vivre parisiens,
pour les voyages, mon vieux milou, passés et futurs,
pour les expériences théâtrales, et le cérémonial associé,
pour l'exigence qui pousse loin, et la rando qui tue à Bayonne,
pour les soirées nanto-choletaises,
pour toutes les soirées parisiennes,
pour les week-ends ici ou là,
pour EVN, TAS, et surtout pour SPB,
pour les petites soirées-retrouvailles bi-mensuelles,
pour les balades en forêts parisiennes,
pour les grues nocturnes,
pour les visites de neveux, celles du soir comme celles du lendemain,
pour les divers déménagements,
pour les déjeuners à Chevaleret,
pour les magnums, au miel,
pour toutes les rencontres du Delaware,
pour la presque deuxième maison de la rue du roule, pour sa porte toujours ouverte,
pour ses traditions de tous temps et sa générosité maintenant légendaire.
Et puis pour les règles du jeu que l'on croit connaître.

A la mémoire d'Eric. Pour Blueberry et cet après midi d'avril sur la terrasse.
Ta petite soeur, sur le tard mais pour toujours.

A mes parents et grands-parents, et à tous ceux qui ont travaillé dur pour que les suivants puissent suivre un autre chemin.

Table of contents

Chapter 1

Introduction

Contents

1.1 Le contexte

Ce travail a été initié dans le cadre du développement au sein du laboratoire Jacques-Louis Lions d'un intérêt pour la fusion magnétique. Plus généralement cet intérêt s'inscrit dans le projet mondial ITER [Org] (International Thermonuclear Experimental Reactor), qui vise à démontrer la faisabilité scientifique et technologique de l'énergie de fusion. Le projet est officiellement porté par la Chine, l'Union Européenne, le Japon, la Corée, le Russie et les Etats-Unis, et mobilise par ailleurs des équipes de recherches dans le monde entier. La construction du réacteur de fusion, représenté en figure 1.1, a lieu en ce moment à Cadarache, dans les Bouches-du-Rhône, et a pour but d'obtenir un gain net d'énergie. L'ordre de grandeur envisagé est un facteur 10 entre l'énergie apportée et l'énergie générée par ITER.

La fusion est une réaction nucléaire au cours de laquelle deux noyaux atomiques légers se combinent en un noyau unique. Cette réaction très énergétique se produit au sein d'un plasma, un état de la matière parfois comparé à un gaz à très haute température dans lequel électrons et ions sont dissociés. Il existe plusieurs procédés pour confiner la matière à l'état de plasma. La fusion par confinement magnétique, qui est le procédé mis en œuvre dans le cadre du projet ITER, est basée sur l'utilisation de puissants champs magnétiques. L'étude des ondes dans les plasmas est donc un sujet fondamental, voir [Swa03, Bra98].

Les mécanismes de transport turbulent qui apparaissent dans les plasmas causent d'importantes pertes d'énergie. Les turbulences sont donc non souhaitables dans les réacteurs de fusion nucléaire, puisqu'ils ont pour but de produire de l'énergie. Afin de mieux comprendre la structure de ces turbulences pour un jour pouvoir les contrôler, il est crucial d'étudier la distribution de densité des particules chargées au sein du plasma. Cependant les conditions extrêmes dans lesquelles ont lieu la réaction interdisent tout forme de mesure

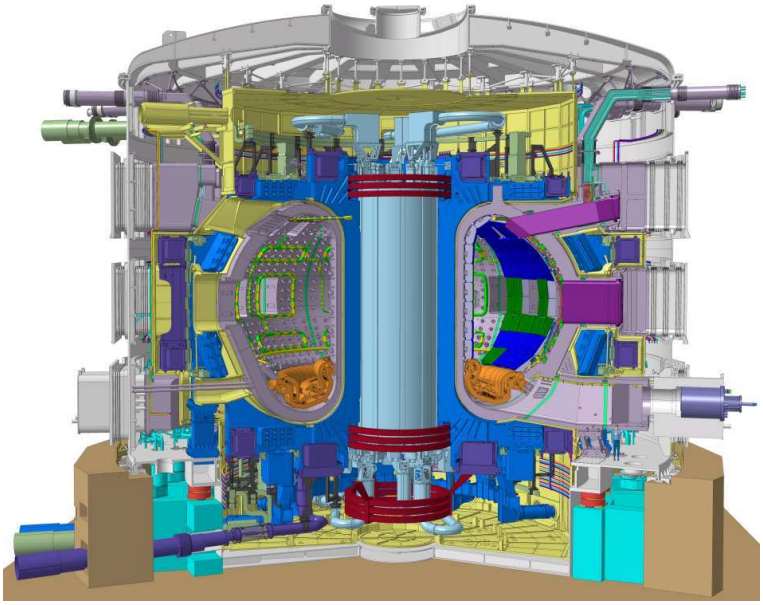


FIGURE 1.1 – Haut de près de trente mètres, lourd de 23 000 tonnes, le tokamak ITER est une machine très complexe. La silhouette du personnage en bleu au pied de la machine nous donne une idée de sa taille. On estime à un million le nombre de composants du tokamak ITER.

intrusive : il n'est envisageable d'introduire aucune sonde à l'intérieur de la chambre magnétique. Ce sont donc des méthodes dites non intrusives qui sont utilisées pour mesurer des caractéristiques des turbulences et de leur dynamique.

Le thème de la réflectométrie pour les plasmas de fusion magnétique intéresse la communauté des physiciens [LDMS96, KGH09, HdSG+11], alors qu'il est encore mal connu par les mathématiciens. Cette méthode de détection de la densité est déjà utilisée en pratique pour sonder les plasmas de fusion. Le réacteur ASDEX (Garching en Allemagne) est par exemple équipé d'un système de réflectométrie micro-onde [MSK+98]. Une onde est envoyée depuis une antenne située sur un mur du réacteur vers le plasma. Comme présenté sur la figure 1.2 elle se propage jusqu'à une certaine profondeur dans le plasma, puis est réfléchie pour une certaine valeur de la densité électronique, dite densité de coupure. Le signal qui est renvoyé vers l'antenne est alors mesuré puis analysé pour calculer la densité à la coupure. Différentes fréquences de l'onde permettent de sonder plusieurs profondeurs, et ainsi de cartographier la densité dans le plasma.

La simulation numérique de la réflectométrie fait l'objet d'une attention particulière, car c'est un problème inverse qu'il faut résoudre afin d'avoir accès à la densité. Il est donc nécessaire de mettre en œuvre des outils de calcul stables, puissants et robustes pour espérer obtenir des résultats pertinents à un prix raisonnable.

Le but de cette thèse était d'étudier certains aspects mathématiques et numériques des équations aux dérivées partielles susceptibles de modéliser la réflectométrie pour les plasmas de fusion, équations qui peuvent également modéliser le chauffage du plasma. Les aspects théoriques liés à l'existence de solutions singulières présentent une complexité surprenante : la mise en place d'un cadre théorique adapté à ces singularités a occupé une place importante dans ce travail de recherche.

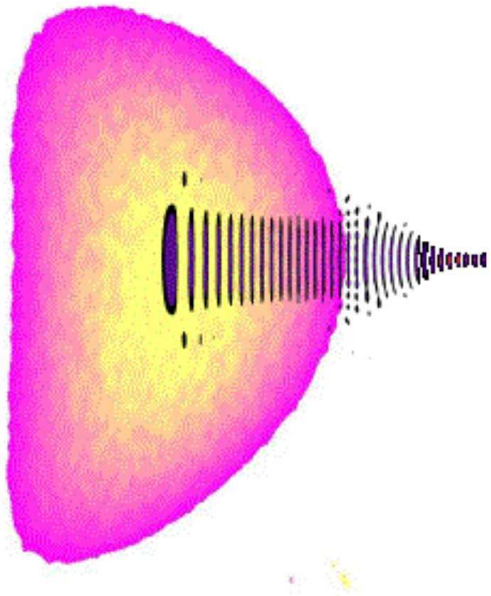


Figure 1.2 – Schéma de principe de la réflectométrie. Coupe du tomkamak. Le plasma qui est représenté en rose est sondé par une onde qui provient de la droite. Fourni par S. Heuraux.

1.2 Le modèle mathématique

L'étude des ondes dans les plasmas est un domaine bien établi, qui se divise principalement en deux grands types de modèles. Les modèles fluides voient les particules chargées comme un écoulement fluide de densié $\rho(t, x)$ tandis que les modèles cinétiques regardent ces particules grâce à leur distribution $f(t, x, v)$.

Dans les deux cas des séries de simplifications sont proposées, mais même avec ces simplifications des modèles établis de longue date ont montré leur pertinence par exemple dans le cadre du chauffage du plasma ou de la génération de courant (current drive). Ces modèles linéarisés considèrent des champs électromagétiques à l'équilibre, perturbés par les ondes. Cela correspond à l'étude d'ondes d'amplitude modérée. La relation entre le courant j et le champ électrique E est décrite par une conductivité matricielle σ, de sorte que localement $j = \sigma E$, si on ne tient compte d'aucune dispersion ni spatiale ni temporelle. Ce tenseur de conductivité éventuellement anisotrope modélise les propriétés électro-magnétiques du milieu.

1.2.1 Le modèle du plasma froid

Les phénomènes physiques prépondérant dans la réflectométrie peuvent être modélisés par un couplage entre les phénomènes électromagnétiques et l'écoulement des particules chargées (c'est un modèle fluide). Dans ce travail les seules particules considérées sont les électrons dont la charge est notée $e < 0$, les ions étant alors considérés comme immobiles en raison de leur forte masse relativement à celle des électrons. Le modèle du plasma froid peut être dérivé avec plusieurs types de particules, voir notamment [Fre07, CW74], mais cette généralization n'entre pas dans le cadre de ce travail.

En électromagnétisme, le champ d'induction électrique D représente la façon dont le champ électrique E influe sur l'organisation des charges électriques dans un matériau

donné, notamment le déplacement des charges et la réorientation des dipôles électriques. D'une manière générale, dans un milieu non homogène comme un plasma cette dépendance varie suivant la position dans le matériau, la fréquence du champ appliqué, la température, et d'autres paramètres. Dans un matériau non linéaire, elle peut impliquer le champ électrique.

Le tenseur de conductivité

L'équation fluide qui décrit la conservation de la masse des électrons est la loi de = en dimension 3. On néglige les collisions ainsi que les mouvements des ions. On note $n_e(\overrightarrow{x})$ la densité des électrons, m_e leur masse, $u_e(\overrightarrow{x},t)$ leur vitesse, e leur charge ; $E(\overrightarrow{x},t)$ désigne le champ électrique. La loi de Newton s'écrit alors :

$$m_e\ n_e \left(\frac{\partial u_e}{\partial t} + (u_e.\nabla)\, u_e\right) = -e\ n_e\,(E + u_e \wedge B)\,.$$

On linéarise autour d'un état d'équilibre :

$$\begin{cases} u_e = & 0 & +u, \\ E = & 0 & +E, \\ B = & B_0 & +\widetilde{B}, \end{cases}$$

où B_0 est le champ magnétique porteur, en supposant $|\tilde{B}|/|B_0| << 1$. D'où en régime harmonique pour une fréquence ω et en négligeant les termes quadratiques :

$$-i\omega m_e u = -e\,(E + u \wedge B_0)\,.$$

En considérant $B_0 = B_0 e_z$ et en introduisant

$$\begin{cases} E_\pm & = \dfrac{1}{\sqrt{2}}(E_x \pm iE_y) \\ (u_e)_\pm & = \dfrac{1}{\sqrt{2}}((u_e)_x \pm i(u_e)_y) \end{cases},$$

ainsi que les pulsations plasma et cyclotron

$$\omega_p^2 = \frac{e^2 n_e(x)}{\varepsilon_0 m_e} \text{ and } \omega_c = \frac{e|B_0|}{m_e}$$

un calcul classique donne

$$\sigma = i\varepsilon_0\omega_p^2 \begin{pmatrix} 1/(\omega+\omega_c) & 0 & 0 \\ 0 & 1/(\omega-\omega_c) & 0 \\ 0 & 0 & 1/\omega \end{pmatrix}$$

dans la base $((e_x + ie_y)/\sqrt{2}, (e_x - ie_y)/\sqrt{2}, e_z)$. Dans cette base le tenseur de permittivité $\varepsilon = \varepsilon_0(I + i/\omega\sigma)$ s'écrit :

$$\varepsilon = \varepsilon_0 \begin{pmatrix} L & 0 & 0 \\ 0 & R & 0 \\ 0 & 0 & P \end{pmatrix} \tag{1.1}$$

avec

$$\begin{cases} R & = 1 - \dfrac{\omega_p^2}{\omega(\omega-\omega_c)} \\ L & = 1 - \dfrac{\omega_p^2}{\omega(\omega+\omega_c)} \\ P & = 1 - \dfrac{\omega_p^2}{\omega^2} \end{cases}$$

La permittivité est donc diagonale dans cette base, pourtant ses vecteurs propres y sont complexes. Exprimés ainsi ils n'ont pas de sens physique. On revient donc à la base initiale (e_x, e_y, e_z) dans laquelle on a

$$\sigma(x) = i\varepsilon_0 \frac{\omega_p^2 \omega}{\omega^2 - \omega_c^2} \begin{pmatrix} 1 & i\frac{\omega_c}{\omega} & 0 \\ -i\frac{\omega_c}{\omega} & 1 & 0 \\ 0 & 0 & 1 - \frac{\omega_c^2}{\omega^2} \end{pmatrix}$$

et on retrouve la formule bien connue par la communauté physique :

$$\varepsilon = \varepsilon_0 \begin{pmatrix} S & -iD & 0 \\ iD & S & 0 \\ 0 & 0 & P \end{pmatrix} \tag{1.2}$$

avec

$$\begin{cases} S &= \dfrac{1}{2}(R+L) = 1 - \dfrac{\omega_p^2}{\omega^2 - \omega_c^2}, \\ D &= \dfrac{1}{2}(R-L) = -\dfrac{\omega_p^2}{\omega} \dfrac{\omega_c}{\omega^2 - \omega_c^2} \end{cases}$$

On remarque que $\varepsilon^* = \varepsilon$.

L'étude de la limite $\omega \to \omega_c$, appelée résonance cyclotron haute si $\omega > \omega_c$ et basse si $\omega < \omega_c$, n'est pas le but de ce travail. On considérera donc systématiquement $\omega \neq \omega_c$.

Une particularité du problème mathématique

Les valeurs propres du tenseur de permittivité diélectrique (1.2) ou (1.1) sont :

$$\begin{cases} \lambda_0 = 1 - \dfrac{\omega_p^2}{\omega^2} \\ \lambda_- = 1 - \dfrac{\omega_p^2}{\omega^2 - \omega_c^2} \left(1 - \dfrac{\omega_c}{\omega}\right), \\ \lambda_+ = 1 - \dfrac{\omega_p^2}{\omega^2 - \omega_c^2} \left(1 + \dfrac{\omega_c}{\omega}\right). \end{cases}$$

Ces trois valeurs dépendent de la variable d'espace à travers leur dépendance en ω_p. De façon générale, on envoie une onde de pulsation ω fixée, supérieure à ω_c. Dans ces conditions, les valeurs propres sont positives si ω_p est petit, ce que l'on suppose vrai près du bord du tokamak, là où l'on place l'antenne. Chacune devient négative lorsque ω_p dépasse une certaine valeur particulière, ce qui se produit au cours de la pénétration au cœur du plasma. Cette transition s'effectue de façon continue, c'est là qu'apparaît la difficulté mathématique : les équations étudiées changent de nature continuement. C'est pour ces valeurs particulières de la pulsation plasma que se produit un phénomène physique fondamental pour cette étude : la coupure.

Un rapide calcul permet de déterminer que les zones de coupure correspondent aux densités telles que :

$$\begin{cases} \omega_p = \omega & \Leftrightarrow \lambda_0 = 0, \\ \omega_p = \omega\sqrt{1 + \frac{\omega_c}{\omega}} & \Leftrightarrow \lambda_- = 0, \\ \omega_p = \omega\sqrt{1 - \frac{\omega_c}{\omega}} & \Leftrightarrow \lambda_+ = 0. \end{cases}$$

Un balayage en fréquence permet donc bien de déplacer les zones de coupure pour différentes valeurs de la densité. C'est le principe même de la réflectométrie qui est ainsi mis en évidence.

Equation d'onde, relation de dispersion, coupure et résonance

Les équations de Maxwell se combinent pour donner l'équation d'onde suivante :

$$\nabla \wedge \nabla \wedge E - \frac{\omega^2}{c^2}\varepsilon E = 0.$$

L'étude de la relation de dispersion consiste alors à identifier les types d'ondes planes solutions de cette équation : soit $e^{i\overrightarrow{\kappa}\overrightarrow{x}}\overrightarrow{E_0}$ une solution de l'équation d'onde, alors nécessairement

$$\overrightarrow{\kappa} \wedge \overrightarrow{\kappa} \wedge \overrightarrow{E_0} - \frac{\omega^2}{c^2}\varepsilon\overrightarrow{E_0} = M_{\overrightarrow{\kappa}}\overrightarrow{E_0} = 0.$$

Une solution non triviale correspond donc nécessairement à un vecteur d'onde $\overrightarrow{\kappa}$ vérifiant $det(M_{\overrightarrow{\kappa}}) = 0$ et cette relation, dénommée relation de dispersion, contient la physique du problème de la propagation des ondes planes pour le modèle du plasma froid. Pour cette raison on s'intéresse à la relation

$$det\left(\kappa_i\kappa_j - \|\overrightarrow{\kappa}\|^2\delta_{ij} - \frac{\omega^2}{c^2}\varepsilon_{ij}\right) = 0.$$

Si l'on s'intéresse aux ondes qui se propagent dans le plan (e_x, e_z), on peut écrire $\overrightarrow{\kappa} = (n\sin\theta, 0, n\cos\theta)$ et ainsi réécrire la relation de dispersion comme une équation du second degré sur n^2 :

$$(S\sin^2\theta + P\cos^2\theta)n^4 - \left(\frac{\omega}{c}\right)^2(RL\sin^2\theta + PS(1+\cos^2\theta))n^2 + \left(\frac{\omega}{c}\right)^4 PRL = 0. \quad (1.3)$$

On peut tout de suite remarquer que dans le cas $\omega_p = 0$, puisqu'alors $S = P = 1$ la relation de dispersion est de nature elliptique : quelle que soit la valeur de θ l'équation est du second degré en n^2 et ses racines $n^2_\pm$ sont donc bornées. En revanche dans le cas $\omega_p \neq 0$ il est possible que que le coefficient dominant $S\sin^2\theta + P\cos^2\theta$ s'annule et que l'équation dégénère en une équation du premier degré pour certaines valeurs de θ. Dans ce cas il est possible que n^2 parte à l'infini pour certaines valeurs de θ. Cela correspond à une structure hyperbolique de la relation de dispersion.

L'équation de dispersion a pour discriminant

$$\left(\frac{\omega}{c}\right)^4\left((RL-PS)^2\sin^4\theta + 4P^2D^2\cos^2\theta\right) > 0.$$

Par conséquent il existe deux racines n^2_+ et n^2_- du polynôme (1.3), qui se factorise sous la forme :

$$P(n^2) = (S\sin^2\theta + P\cos^2\theta)(n^2 - n^2_+)(n^2 - n^2_-).$$

In fine cela montre que deux types d'ondes planes peuvent exister dans un plasma froid : dans un régime propagatif si $n^2 = n^2_\pm > 0$, et dans un régime evanescent si $n^2 = n^2_\pm < 0$.

Le cas de transition $n^2 = 0$ est appelé une coupure : localement le nombre d'onde tend vers zéro. L'autre cas de transition, qui s'effectue pour $n \to \infty$, est appelé résonance : localement le nombre d'onde tend vers l'infini. Ces deux phénomènes, illustrés en figure 1.3, sont spécifiques à l'étude de la propagation des ondes dans les plasmas, c'est leur exploitation qui est à la base du principe de réflectométrie. En effet ces zones de transition interagissent avec les ondes qui proviennent d'une zone propagative. A la résonance l'onde peut transmettre de l'énergie au plasma pour le chauffer, à la coupure l'onde est réfléchie.

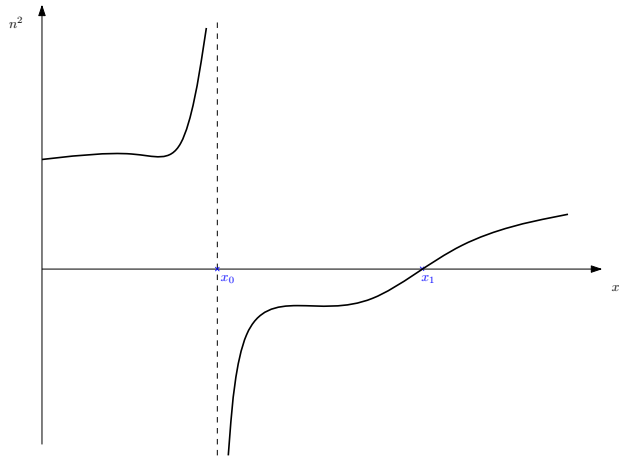


Figure 1.3 – Résonance en $x = x_0$ et coupure en $x = x_1$ dans un même plasma.

C'est ce signal réfléchi qui, mesuré depuis le mur du réacteur, renseigne sur la densité au niveau de la transition.

Coupures et résonances sont bien connues dans la communauté des physiciens dans un cadre plus général que celui de le réflectométrie, et néanmoins peu étudiées par la communauté mathématique. Une vaste littérature sur la physique des plasmas détaille l'étude de cette relation de dispersion, voir [Bra98, Sti92, Swa03]. Dans ce travail on s'intéresse à l'étude des solutions mathématiques dans un régime avec résonance ou coupure, ainsi qu'à la construction d'une classe de méthodes de précision élevée pour la simulation numérique.

Compte tenu de la taille immense du réacteur de fusion ITER, qui mesurera environ six mètres de diamètre interne, dix-neuf mètres de large et onze mètres de haut, les méthodes numériques utilisées aujourd'hui pour la réflectométrie, telles que les différences finies, sont très coûteuses. Elles requièrent en effet un maillage uniforme par essence, or la finesse du maillage est imposée par la taille de l'antenne. Signalons qu'il existe des méthodes d'éléments finis développées par Simon Labrunie [Lab]. Dans le cadre de ce problème modèle on considèrera une autre famille de méthodes numériques.

1.2.2 Modes de propagation

Pour simplifier les notations, dans tout ce document on considérera

$$\nabla \wedge \nabla \wedge E - \varepsilon E = 0, \tag{1.4}$$

$$\varepsilon(x) = \begin{pmatrix} \alpha & i\gamma & 0 \\ -i\gamma & \alpha & 0 \\ 0 & 0 & -\beta \end{pmatrix}. \tag{1.5}$$

Comme le tenseur ε a une structure diagonale par bloc, l'équation peut se décliner sous forme de deux équations appelées modes de propagation.

Le mode dit ordinaire ou mode O correspond au bloc de taille 1 du coefficient. L'équation associée est une équation de Helmholtz, qui est un modèle simple de propagation d'onde dans le cas $\beta < 0$:

$$-\Delta E_z + \beta E_z = 0.$$

Ce mode de propagation présente une coupure associée à la valeur propre λ_0 de ε, c'est-à-dire à la densité telle que $\omega_p = \omega$. Le coefficient β qui est régulier s'annule donc continuement. L'équation de Helmholtz devient localement une fonction d'Airy. Il conviendra donc de prêter une attention particulière à cette zone de coupure.

Le mode dit extraordinaire ou mode X correspond au bloc 2×2 du coefficient. L'équation associée est beaucoup moins commune. Pour $E = (E_x, E_y)$ elle s'écrit

$$\overrightarrow{\nabla} \wedge \nabla \wedge E - \varepsilon_\perp E = 0,$$

où le rotationnel en dimension deux est défini par

$$\overrightarrow{\nabla} \wedge = \begin{pmatrix} \partial_y \\ -\partial_x \end{pmatrix} \text{ et } \nabla \wedge = \begin{pmatrix} \partial_x \\ \partial_y \end{pmatrix} \wedge .$$

En d'autres termes le champ électrique vérifie

$$\left\{ \begin{array}{rrrl} \partial_y(\partial_x E_y - \partial_y E_x) & -\alpha E_x & -i\gamma E_y & = 0, \\ -\partial_x(\partial_x E_y - \partial_y E_x) & +i\gamma E_x & -\alpha E_y & = 0. \end{array} \right.$$

Ce mode de propagation présente quant à lui deux coupures qui correspondent aux valeurs propres de $\varepsilon_\perp$: une coupure apparaîtra lorsque $\alpha = \pm\gamma$. La seconde particularité correspond au phénomène de résonance décrit plus haut, et c'est elle qui nécessitera dans ce cas une attention particulière. Elle a lieu lorsque $\alpha = 0 \neq \gamma$. Elle est appelée résonance hybride et peut en théorie participer au chauffage des plasmas magnétiques.

1.3 Structure du document

Le point liant de ce travail est ainsi l'étude de l'équation de Maxwell (1.4) pour un tenseur ε défini par (1.5), qui dépend continuement de l'espace.

1.3.1 Principaux résultats obtenus

Avant de s'intéresser à la simulation numérique de problèmes de réflectométrie, il convient de s'assurer que le problème considéré a bien une unique solution. Si le caractère bien posé du problème en mode O résulte d'une décomposition de Fredholm classique, il n'en est rien pour celui du problème en mode X. L'existence de solutions singulières, dites résonantes, est un fait établi pour la communauté physique.

Par un principe d'absorption limite, il a été possible pour la première fois de caractériser une solution singulière dans le cas de coefficients ne dépendant que d'une variable d'espace. La solution limite de l'équation de Maxwell en mode X se compose d'une masse de Dirac sur la composante E_x plus un champ régulier. Ce résultat obtenu en collaboration avec Ricardo Weder et Bruno Després a donné lieu à un article actuellement en cours d'examen. Il est énoncé dans le théoreme 4.1.1.

La mise en place d'une méthode numérique pour des problèmes de réflectométrie requiert comme on l'a vu la prise en charge de coefficients continus. La méthode proposée dans ce travail est basée sur la méthode en ondes planes appelée Ultra-Weak Variational Formulation (UWVF) qui s'appuie sur des ondes planes solutions de l'équation adjointe homogène pour des coefficients constants par morceaux, et qui peut être rapprochée des méthodes de Galerkin discontinues. Adapter cette méthode aux coefficients variables nécessite donc la mise en place de fonctions de base d'un type nouveau, construites sur mesure pour le problème.

Dans le cadre de ce mémoire j'ai proposé, justifié théoriquement et validé numériquement une nouvelle procédure d'approximation en ondes planes généralisées. La procédure explicite présentée ici pour l'équation en mode O dépend d'un paramètre q qui pilote l'ordre de convergence global de la méthode. On obtient donc une méthode d'ordre élevé.

J'ai commencé à aborder l'extension de la méthode pour le mode X. La formule de construction des fonctions de base est établie comme un premier pas vers l'obtention d'une méthode dédiée spécifiquement au mode X. Les premiers calculs avec la méthode UWVF sont présentés à la fin de ce document.

La p convergence qui a été abordée dans la littérature pour la UWVF apparaît dans le résultat d'interpolation obtenu en dimension deux pour les nouvelles fonctions de base.

J'ai développé un code pour la méthode UWVF avec les nouvelles fonctions de base : il permet de calculer une solution approchée pour des problèmes en dimension deux. En mode O le code permet de traiter comme convenu un coefficient qui s'annule : on obtient une approximation satisfaisante d'une onde réfléchie par une coupure, sur un domaine carré pour une cinquantaine de longueurs d'onde. En mode X les premiers résultats numériques sont validés en présence d'une coupure ainsi qu'en présence d'une résonance pour l'approximation d'une solution régulière. Cette validation est faite dans un cas simple qui correspond aux hypothèses de la partie théorique.

En parallèle, j'ai participé à différents projets au Cemracs ainsi que lors de semaines maths-entreprises qui ont donné lieu à des comptes rendus, [DDF+11, WJIG11], mais ne seront pas exposés dans ce mémoire. Les projets math-industrie ont été proposés par A. Fuser de GDF et par P. Saadé de Picviz labs [Saa]. Je me suis également penchée sur quelques problèmes d'estimation numérique de paramètres pour des modèles statisques, principalement pour le modèle de Merton et des extensions modélisant des sauts de loi normale ou double exponentielle.

1.3.2 Plan

Pour plus de clarté le manuscript est divisé en deux parties, la première contenant l'étude mathématique tandis que la seconde comprend l'étude numérique.

Avant de se lancer dans l'étude des équations dissociées en mode O et X, il est intéressant de citer un résultat d'existence et d'unicité pour les équations de Maxwell. C'est l'objet du premier chapitre, qui reprend un résultat classique sous l'hypothèse (plus générale que celle exposée à l'origine) $\Re(\lambda_\varepsilon) > 0$ pour toute valeur propre λ_ε du tenseur de permittivité.

Le second chapitre se focalise sur l'équation en mode X. Un exemple liminaire exhibe une solution singulière, justifiant la mise en place du procédé de régularisation qui permet de se placer dans le cadre classique des espaces de Lebesgue. Une première phase vise à expliciter la solution du problème régularisé en fonction d'une double normalisation, la première au niveau de la résonance et la seconde à l'infini. Suit alors une phase fondamentale d'établissement d'estimations a priori sur la solution du problème régularisé. Ces estimations se veulent uniformes par rapport au paramètre de régularisation afin de rester valables dans la limite $\mu \to 0$. Cette phase se décline en plusieurs étapes : établir des estimations de part et d'autre de la résonance, pour finir par une estimation incluant la zone de résonance.

Le troisième chapitre concerne la méthode numérique. Il en présente le cadre théorique et explicite la procédure de construction des nouvelles fonctions de base, à la fois en dimension un et en dimension deux, pour un paramètre d'approximation q, pour le mode O. Une analyse de convergence est proposée en dimension un. Elle aboutit a une

estimation explicite de l'ordre de convergence en fonction du paramètre q : on prévoit théoriquement une convergence d'ordre $q - 3/2$. La construction des fonctions de base devrait donc permettre d'atteindre l'ordre élevé désiré. Une étude des propriétés d'interpolation des nouvelles fonctions de base est présentée en dimension deux.

Les résultats numériques sont regroupés dans un dernier chapitre. Pour le mode O les ordres théoriques de convergence pour les propriétés d'interpolation sont retrouvés numériquement de façon très satisfaisante. Le code pour la UWVF avec les nouvelles fonctions de base est validé par une série de tests, et appliqué à un premier cas de réflectométrie. Pour le mode X le code est validé dans le cas de l'approximation d'une solution régulière par le problème régularisé.

Chapter 2

Introduction

Contents

2.1 Background

This work initiated with the development of a specific interest for magnetic fusion at Laboratoire Jacques-Louis Lions. It is linked with the worldwide project ITER [Org] (International Thermonuclear Experimental Reactor), that aims at harnessing the energy produced by the fusion reaction. The official project includes China, the European Union, India, Japan, Korea, Russia and the United States. Several research teams around the world are involved as well. The reactor presented in Figure 2.1 is being built in the South of France at Cadarache, and is expected to deliver ten times the power it consumes.

Fusion is a nuclear reaction where hydrogen nuclei collide to form a unique atom. The quantity of released energy is huge. This reaction occurs in a plasma which is a hot electrically charged gas. Several processes can be used to confine plasma. Magnetic confinement fusion, which is based on magnetic forces, is the process chosen for ITER. For this reason studying waves in plasmas is a crucial topic, see [Swa03, Bra98].

Turbulences in the plasma flow are the cause of severe energy loss. So they are not desirable in fusion reactor which aim at producing energy. In order to understand the turbulences, and some day control them, it is important to study the density of charged particles inside the plasma. However because of the high temperature of the plasma it is impossible to probe the plasma from the inside : no probing device could stand the temperature inside the magnetic chamber. The methods used to measure turbulences characteristics and evolution have to be non intrusive.

Reflectometry for magnetic fusion plasmas is a popular field of study within the Physics community [LDMS96, KGH09, HdSG+11], whereas it is hardly known in the Mathematics community. In fact, the density probing method is already used to probe fusion plasma.

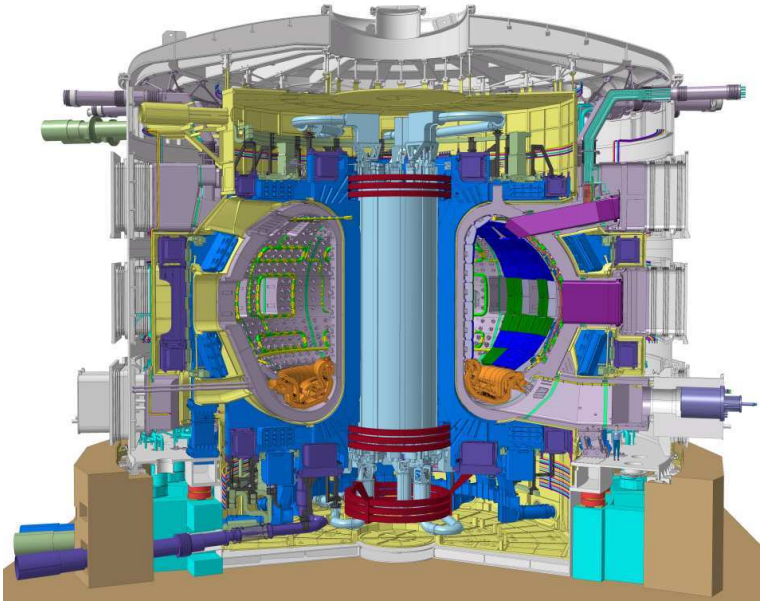


FIGURE 2.1 – The incredibly complex ITER Tokamak will be nearly 30 meters tall, and weigh 23,000 tons. The very small man dressed in blue (bottom right) gives some idea of the machine's scale. The ITER Tokamak is made up of an estimated one million parts. An updated cut-away of the ITER Tokamak, produced by the ITER Design Office in January 2013 [Org].

For instance the reactor ASDEX (Garching, Germany) is equipped with microwave Reflectometry devices [MSK+98]. The principle is the following. A wave is sent toward the plasma from an antenna on the wall of the reactor, see Figure 2.2. It propagates until the electronic density reaches a specific value, called cut-off density. There the signal bounces back toward the antenna where it is measured and analyzed to compute the density at the cut-off. Different frequencies probe the plasma at different depths, providing a map of the plasma density.

Since an inverse problem has to be solved to get the density, efficient numerical simulation methods are required for reflectometry. Specific tools have to be developed and must be stable and powerful enough to get relevant results at a reasonable price.

The goal of this work is then to study some of the mathematical and numerical aspects of the partial differential equations which model reflectometry for fusion plasmas. These equations also model plasma heating. Theoretical aspects concerning existence of singular solutions are surprisingly complex : setting a theoretical framework adapted to these singularities took up a considerable part of this research work.

2.2 The mathematical model

The topic of waves in plasmas is divided into two main types of models. Fluid models consider charged particles as a fluid flow with density $\rho(t, x)$ whereas kinetic models consider their distribution $f(t, x, v)$.

Both types of models rely on a series of simplifications, but some of the models are well known for being relevant for simulating plasma heating or current drive. These linearized models study waves as perturbations of equilibrium electromagnetic fields. This hypothesis

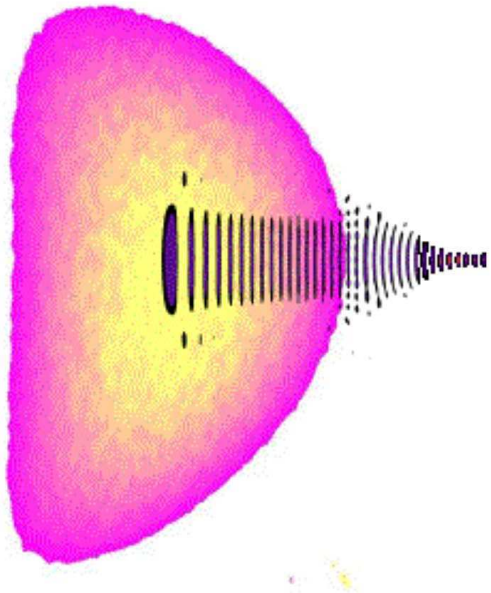


Figure 2.2 – Reflectometry principle. Cut-away of the tokamak. A wave sent from the right is probing the plasma, represented in pink. Courtesy of S. Heuraux.

corresponds to reduced amplitude waves. The expression of the current density j with respect to the electric field E is described by a conductivity matrix σ, $j = \sigma E$ in which there is no space or time dispersion. The conductivity tensor models the electromagnetic properties of the media.

2.2.1 The cold plasma model

The dominating effects in reflectometry can be modeled by a coupling of electromagnetism and charged particles flow. It is a fluid model. In this work the only particles considered are electrons with charge $e < 0$, ions are considered as motionless because of their mass with respect to the electrons mass. The cold plasma model can include several particles, see [Fre07, CW74], but this is not within the scope of this work.

In electromagnetism, the electric induction D represents the way in which the electric field E affects the organization of electric particles in a given media. More generally, in a non homogeneous media such as a plasma it depends on the position, the frequency of the electromagnetic field, the temperature and other parameters.

Conductivity tensor

The fluid equation describing the conservation of electrons mass is Newton's law in dimension 3. Collisions as well as ion motion are neglected. The electrons have a density $n_e(\overrightarrow{x})$, a mass m_e, a velocity $u_e(\overrightarrow{x}, t)$ and a charge e ; $E(\overrightarrow{x}, t)$ refers to the electric field. Newton's law then reads :

$$m_e\, n_e \left(\frac{\partial u_e}{\partial t} + (u_e.\nabla)\, u_e \right) = -e\, n_e\, (E + u_e \wedge B)\,. \tag{2.1}$$

The equilibrium is described by :

$$\left\{\begin{array}{rcl} u_e = & 0 & +u, \\ E = & 0 & +E, \\ B = & B_0 & +\tilde{B}, \end{array}\right. \tag{2.2}$$

where B_0 is the confining magnetic field, and satisfies $|\tilde{B}|/|B_0| << 1$. Neglecting the quadratic terms it yields in the time-harmonic domain, with ω denoting the frequency :

$$-i\omega m_e u = -e\left(E + u \wedge B_0\right). \tag{2.3}$$

The confining magnetic field is such that $B_0 = B_0 e_z$. Defining

$$\left\{\begin{array}{rl} E_\pm & = \dfrac{1}{\sqrt{2}}(E_x \pm iE_y), \\ (u_e)_\pm & = \dfrac{1}{\sqrt{2}}((u_e)_x \pm i(u_e)_y), \end{array}\right.$$

and the plasma and cyclotron frequencies

$$\omega_p^2 = \frac{e^2 n_e(x)}{\varepsilon_0 m_e} \text{ and } \omega_c = \frac{e|B_0|}{m_e}.$$

a classical calculation shows that

$$\sigma = i\varepsilon_0\omega_p^2 \begin{pmatrix} 1/(\omega+\omega_c) & 0 & 0 \\ 0 & 1/(\omega-\omega_c) & 0 \\ 0 & 0 & 1/\omega \end{pmatrix}$$

in the basis $((e_x + ie_y)/\sqrt{2}, (e_x - ie_y)/\sqrt{2}, e_z)$. In this basis the dielectric tensor $\varepsilon = \varepsilon_0(I + i/\omega\sigma)$ reads :

$$\varepsilon = \varepsilon_0 \begin{pmatrix} L & 0 & 0 \\ 0 & R & 0 \\ 0 & 0 & P \end{pmatrix} \tag{2.4}$$

with

$$\left\{\begin{array}{rl} R & = 1 - \dfrac{\omega_p^2}{\omega(\omega-\omega_c)} \\ L & = 1 - \dfrac{\omega_p^2}{\omega(\omega+\omega_c)} \\ P & = 1 - \dfrac{\omega_p^2}{\omega^2} \end{array}\right.$$

The dielectric tensor is diagonal in this basis, nevertheless its eigenvectors there are not real. They have no physical meaning. Going back to the initial basis (e_x, e_y, e_z) one gets

$$\sigma(x) = i\varepsilon_0 \frac{\omega_p^2\omega}{\omega^2 - \omega_c^2} \begin{pmatrix} 1 & i\frac{\omega_c}{\omega} & 0 \\ -i\frac{\omega_c}{\omega} & 1 & 0 \\ 0 & 0 & 1 - \frac{\omega_c^2}{\omega^2} \end{pmatrix}$$

which gives the well known formula :

$$\varepsilon = \varepsilon_0 \begin{pmatrix} S & -iD & 0 \\ iD & S & 0 \\ 0 & 0 & P \end{pmatrix} \tag{2.5}$$

where

$$\begin{cases} S &= \frac{1}{2}(R+L) = 1 - \frac{\omega_p^2}{\omega^2-\omega_c^2}, \\ D &= \frac{1}{2}(R-L) = -\frac{\omega_p^2}{\omega}\frac{\omega_c}{\omega^2-\omega_c^2} \end{cases}$$

Notice that $\varepsilon^* = \varepsilon$.

The behavior at $\omega \to \omega_c$ is called cyclotron resonance. If $\omega < \omega_c$ the case is called a low hybrid resonance. The other case $\omega > \omega_c$ is referred to as the upper hybrid resonance. The assumption $\omega \neq \omega_c$ holds in th whole of this work.

A specificity of the mathematical problem

Eigenvalues of the dielectric tensor (2.5) or (2.4) are :

$$\begin{cases} \lambda_0 = 1 - \dfrac{\omega_p^2}{\omega^2} \\ \lambda_- = 1 - \dfrac{\omega_p^2}{\omega^2-\omega_c^2}\left(1 - \dfrac{\omega_c}{\omega}\right), \\ \lambda_+ = 1 - \dfrac{\omega_p^2}{\omega^2-\omega_c^2}\left(1 + \dfrac{\omega_c}{\omega}\right). \end{cases}$$

These three quantities depend on the space variable through ω_p. Generally speaking, a wave is sent with a given frequency ω, higher than ω_c. In these conditions the eigenvalues are positive for small ω_p which holds close to the wall where the antenna stands and each of the eigenvalues decreases and becomes negative when ω_p passes a threshold. These transitions occur continuously, and this is what gives rise to the mathematical challenge : the type of the equations changes continusouly. The threshold values account for a major physical phenomenon in this study : the cut-off.

A simple calculation shows that the cut-off zones correspond to the densities such that :

$$\begin{cases} \omega_p = \omega & \Leftrightarrow \lambda_0 = 0, \\ \omega_p = \omega\sqrt{1+\frac{\omega_c}{\omega}} & \Leftrightarrow \lambda_- = 0, \\ \omega_p = \omega\sqrt{1-\frac{\omega_c}{\omega}} & \Leftrightarrow \lambda_+ = 0. \end{cases}$$

This illustrates the fact that sweeping through the frequency range moves these zones to different densities and the principle of reflectometry becomes evident.

Wave equation, dispersion relation, cut-off and resonance

Maxwell's equations turn into the following wave equation :

$$\nabla \wedge \nabla \wedge E - \frac{\omega^2}{c^2}\varepsilon E = 0.$$

A plane wave $e^{i\overrightarrow{\kappa}\,\overrightarrow{x}}\overrightarrow{E_0}$ solution to the wave equation has to satisfy

$$\overrightarrow{\kappa} \wedge \overrightarrow{\kappa} \wedge \overrightarrow{E_0} - \frac{\omega^2}{c^2}\varepsilon\overrightarrow{E_0} = M_{\overrightarrow{\kappa}}\overrightarrow{E_0} = 0.$$

So a non trivial solution corresponds to $\overrightarrow{\kappa}$ satisfying $det(M_{\overrightarrow{\kappa}}) = 0$ which is called dispersion relation and describes the physical problem of wave propagation for the cold plasma model :

$$det\left(\kappa_i\kappa_j - \|\overrightarrow{\kappa}\|^2\delta_{ij} - \frac{\omega^2}{c^2}\varepsilon_{ij}\right) = 0.$$

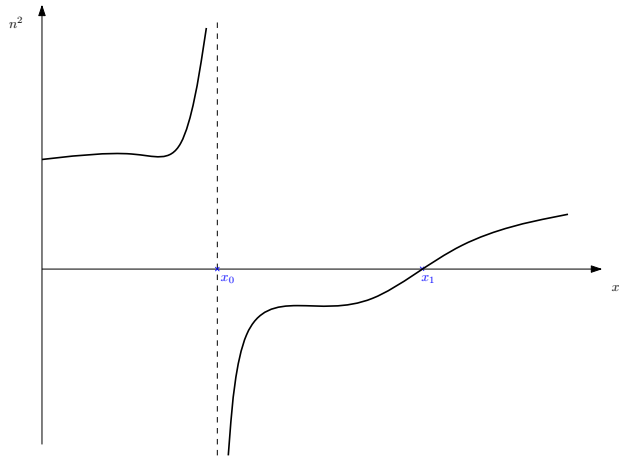


FIGURE 2.3 – Simultaneous resonance at $x = x_0$ and cut-off at $x = x_1$ in a plasma.

Looking for waves propagating in the plane (e_x, e_z), in the form $\overrightarrow{\kappa} = (n\sin\theta, 0, n\cos\theta)$, the dispersion relation turns into a quadratic function with respect to n^2 :

$$(S\sin^2\theta + P\cos^2\theta)n^4 - \left(\frac{\omega}{c}\right)^2 (RL\sin^2\theta + PS(1+\cos^2\theta))n^2 + \left(\frac{\omega}{c}\right)^4 PRL = 0. \quad (2.6)$$

In the case $\omega_p = 0$, it is clear that the dispersion relation is elliptic since $S = P = 1$: the leading term coefficient does not vanish for any value of θ, so that the roots n^2 are bounded. In the case $\omega_p \neq 0$ on the other hand it is possible that this coefficient $S\sin^2\theta + P\cos^2\theta$ vanishes so that the equation turns into a linear function : so for some values of θ, n^2 might go to infinity. This corresponds to a hyperbolic structure in the dispersion relation.

The discriminant of the dispersion relation is

$$\left(\frac{\omega}{c}\right)^4 \left((RL - PS)^2\sin^4\theta + 4P^2D^2\cos^2\theta\right) > 0.$$

As a result the polynomial (2.6) has two roots n_+^2 et n_-^2 and reads :

$$P(n^2) = (S\sin^2\theta + P\cos^2\theta)(n^2 - n_+^2)(n^2 - n_-^2).$$

This shows that two kind of waves can be found : a propagative wave with $n^2 = n_\pm^2 > 0$ and an evanescent wave with $n^2 = n_\pm^2 < 0$.

The transition that occurs at $n^2 = 0$ is called a cut-off : the wave number locally goes to zero. The other possible transition, which occurs at $n^2 \to \infty$, is called a resonance : the wave number locally goes to infinity. These two cases are illustrated in Figure 2.3 and are typical of the study of wave propagation in plasmas. They are actually the basis for reflectometry. The wave coming from a propagating zone interacts with the transition zones. At the resonance the wave can drop energy in the plasma to heat it, at the cut-off the wave is reflected. When this reflected signal is measured at the wall of the reactor it provides information about the density along the cut-off.

Cut-off and resonance are well-known in the Physics community in a wider framework than the one of reflectometry. They are nevertheless not studied that much in the Mathematics community. Several books study the dispersion relation, see for instance [Bra98, Sti92, Swa03]. This work is concerned with the study of mathematical solutions

in a domain containing either a resonance or a cut-off, together with the construction of a new type of numerical methods achieving high precision for numerical simulation.

Because of the huge size of ITER's reactor, that will be around eleven meters high and nineteen meters long, the numerical methods that are currently used - such as finite differences - are very expensive. Indeed, they require a uniform mesh and the size of the mesh must be small enough with respect to the antenna size. Let us note that some finite element methods are developed by Simon Labrunie [Lab]. In the scope of this work another type of numerical methods will be considered.

2.2.2 Propagation modes

For the sake of simplicity, the following notation will hold in the whole document :

$$\nabla \wedge \nabla \wedge E - \varepsilon E = 0, \tag{2.7}$$

$$\varepsilon(x) = \begin{pmatrix} \alpha & i\gamma & 0 \\ -i\gamma & \alpha & 0 \\ 0 & 0 & -\beta \end{pmatrix}. \tag{2.8}$$

As the tensor ε has a block diagonal structure, the equation can be split into two propagation modes.

The Ordinary mode (O mode) corresponds to the block of size 1×1. The associated equation is a Helmholtz equation, which is a basic model for wave propagation is the case $\beta < 0$:

$$-\Delta E_z + \beta E_z = 0.$$

This propagation mode presents a cut-off associated with the eigenvalue λ_0 of ε. The cut-off occurs at a density such that $\omega_p = \omega$. The smooth coefficient β vanishes continuously, and the Helmholtz equation turns locally into an Airy equation. The cut-off zone has to be carefully considered.

The eXtraordinary mode (X mode) corresponds to the block of size 2×2. The associated equation is not as common. For $E = (E_x, E_y)$ it reads

$$\overrightarrow{\nabla} \wedge \nabla \wedge E - \varepsilon_\perp E = 0,$$

where the two dimension *curl* operator is defined by

$$\overrightarrow{\nabla}\wedge = \begin{pmatrix} \partial_y \\ -\partial_x \end{pmatrix} \text{ and } \nabla\wedge = \begin{pmatrix} \partial_x \\ \partial_y \end{pmatrix} \wedge .$$

It means that the electric field satisfies

$$\begin{cases} \partial_y(\partial_x E_y - \partial_y E_x) & -\alpha E_x & -i\gamma E_y & = 0, \\ -\partial_x(\partial_x E_y - \partial_y E_x) & +i\gamma E_x & -\alpha E_y & = 0. \end{cases} \tag{2.9}$$

As for this propagation mode, it presents two cut-offs that correspond to the eigenvalues of the matrix $\varepsilon_\perp$. A cut-off will occur when $\alpha = \pm\gamma$. The second specificity is the resonance previously mentioned, and it requires careful consideration in this case. It happens when $\alpha = 0 \neq \gamma$. It is referred to as hybrid resonance and can theoretically take part in the heating mechanism of magnetic plasmas.

2.3 Structure of the document

The topic of this work is the study of Maxwell's equation (2.7) for a dielectric tensor ε defined in (2.8), which is a smooth function of the space variable.

2.3.1 Major contributions

Before focusing on the numerical simulation for reflectometry problems, it is important to ensure that the problem considered does have a unique solution. The well-posedness for the O mode results from a classical Fredholm decomposition. The X mode problem is more intricate. The existence of singular solutions, called resonant solutions, is accepted in the Physics community.

A limit absorption principle made possible for the first time the definition of a singular solution for a coefficient depending on only one space variable. The limit solution of the K mode Maxwell's equation is the sum of a Dirac mass on the E_x component plus a smooth term. This result was obtained in collaboration with Ricardo Weder and Bruno Després, it led to an article which is actually being reviewed. It is stated in Theorem 4.1.1.

The setting of a numerical method for reflectometry requires as already mentioned the consideration of smooth coefficients. The method proposed in this work is based on the plane wave method called Ultra Weak Variational Formulation (UWVF), which is - in the case of piecewise coefficients - based on plane waves that are solution to the adjoint homogeneous equation, and can be linked with Discontinuous Galerkin (DG) methods. As a result adapting this method to varying coefficients requires the design of new basis functions, tailored to the problem.

In this thesis, I propose, justify theoretically and validate numerically a new approximation procedure based on generalized plane waves. The explicit procedure presented here for the O mode equation depends on a parameter q that steers the global convergence rate of the numerical method. High order convergence is achieved.

I started considering the X mode extension of the method. The design process for the basis function is established, and represents a first step toward a specific method for the X mode problem. The first UWVF computations are displayed at the end of the manuscript.

Some p convergence results can be found in the UWVF literature. I developed a p convergence result regarding the interpolation properties of the new basis functions.

I implemented a code dedicated to the UWVF coupled with the new basis functions : it computes an approximate solution for two dimensional problems. In O mode the code computes as expected a solution in the case of a smooth vanishing coefficient : a satisfying approximation of a wave reflected by a cut-off, on a domain that is fifty wavelengths long. In X mode the first numerical results are validated with a smooth reference solution with a cut-off as well as with a resonance. This validation is performed on a simple case that fits the one considered in the theoretical part.

At the same time, I contributed to several projects at the Cemracs and at the French math-industry weeks, some material is available [DDF+11, WJIG11] but it will not be detailed here. The math-industry projects were proposed by A. Fuser from the French gas company GDF and by P. Saadé from Picviz labs [Saa]. I also worked on a couple of problems aimed at estimating parameters for statistical models, mainly for the Merton model and extended jump models.

2.3.2 Outline

For the sake of clarity, the thesis is divided into two parts, a mathematical study followed by a numerical study.

Before splitting the initial equation into O and X modes, it is interesting to cite an existence and uniqueness result on Maxwell's equation. This is the aim of the first chapter, that describes a classical result under the assumption that $\Re(\lambda_\varepsilon) > 0$ for all eigenvalue λ_ε of the dielectric tensor, that is more general than the original assumption.

The second chapter focuses on the X mode equation. A singular solution is exhibited to justify the need for a regularization process to go back to the classical Lebesgue spaces context. First the solutions of the regularized problem are explicitly built, using a double normalization at the resonance and at infinity. Then comes the crucial stage of establishing a priori estimates on the solution of the regularized problem. These have to be uniform with respect to the regularization parameter μ in order to hold at the limit $\mu \to 0$. It is made of different steps, starting with estimates on each side of the resonance to end with an estimate including the resonance zone.

The third chapter concerns the numerical method. It presents the theoretical setting and explicits the basis functions' design procedure, in dimension one and two, for an approximation parameter q and for the O mode. A convergence analysis is proposed in dimension one. It is concluded by the proof of a theoretical order of convergence explicit with respect to q : the order of convergence is proved to be at least $q - 3/2$. For this reason the basis functions design should lead to the desired high order convergence. A study of interpolation properties of the new basis functions in dimension two follows.

The numerical results are gathered in the last chapter. For the O mode the numerical convergence rates fit the theoretical estimates in a very satisfying way. The UWVF code with the new basis functions is validated on a series of benchmark cases, and applied to a first reflectometry test case. For the X mode the code is validated for the approximation of a regular solution by the regularized problem.

Part I

Theory

Chapter 3

Positive dielectric tensors

Contents

This chapter is inspired by Chapter 4 Variational theory for the cavity problem from Monk [Mon03]. In the introduction of the first chapter there is a remark concerning anisotropic materials : *Although the methods in this book can be applied to anisotropic media, we will not analyze the methods with matrix-valued coefficients. This is mainly due to the difficulty of verifying uniqueness of the solution of Maxwell's equations in this case.*

In the following, we adapt the analysis in the case of a matrix-valued dielectric tensor ε, as presented in (2.5). The analysis of the coercivity of the problem is set on a domain Ω without cavity. The uniqueness result will not be considered, since it is not in the scope of this work, see [Oka02] and the reference given by Monk in [Mon03], namely [Vog91].

The main restriction in this chapter is that the dielectric tensor which must be positive to obtain the coercivity. In some sense it shows that the standard theoretical framework presented here is not adapted to the problem with the cold plasma tensor (2.5).

3.1 Introduction

Consider a Lipschitz domain $\Omega \in \mathbb{R}^3$, and the following problem

$$\begin{cases} \nabla \wedge \nabla \wedge E - \varepsilon E = F & (\Omega), \\ (\nabla \wedge E) \wedge \nu - i\sigma E_T = g & (\partial\Omega), \end{cases} \tag{3.1}$$

where E_T is defined, for ν the outward normal, by

$$E_T = (\nu \wedge E)_{|\partial\Omega} \wedge \nu.$$

Define the classical *curl* space

$$H(curl;\Omega) = \left\{ v \in \left(L^2(\Omega)\right)^3, \nabla \wedge v \in \left(L^2(\Omega)\right)^3 \right\}.$$

A classical variational formulation reads : for all $\phi \in H(curl;\Omega)$

$$\int_\Omega \left[\nabla \wedge E \cdot \nabla \wedge \overline{\phi} - \varepsilon E \cdot \overline{\phi}\right] dV - i\int_{\partial\Omega} \sigma E_T \cdot \overline{\phi}_T dA = \int_\Omega F \cdot \overline{\phi} dV + \int_{\partial\Omega} g \cdot \overline{\phi}_T dA.$$

In order to give a meaning to the different terms appearing in this formulation, a convenient space is

$$X = \left\{ u \in H(curl;\Omega), u_T \in \left(L^2(\partial\Omega)\right)^3 \right\}. \tag{3.2}$$

The parenthesis $(\cdot,\cdot)$ denote the scalar product on $\left(L^2(\Omega)\right)^3$ and the brackets $\langle\cdot,\cdot\rangle$ denote the scalar product on $\left(L^2(\partial\Omega)\right)^3$. The variational problem is then : find $E \in X$ such that

$$(\nabla \wedge E, \nabla \wedge \phi) - (\varepsilon E, \phi) - i\langle\sigma E_T, \phi_T\rangle = (F,\phi) + \langle g, \phi_T\rangle, \tag{3.3}$$

for all $\phi \in X$. The bilinear form a is defined on $X \times X$ as

$$a(u,v) = (\nabla u, \nabla v) - (\varepsilon u, v) - i\langle\sigma u_T, v_T\rangle.$$

3.2 Assumptions on the coefficients

The domain Ω is bounded, simply connected and Lipschitz. The parameters of the problem satisfy :

- $\varepsilon \in (L^\infty(\Omega))^{(3\times3)}$,
- ε is a hermitian matrix which eigenvalues are uniformly strictly positive and uniformly bounded,
- $\sigma \geq c_\sigma > 0$ and $\sigma \in L^\infty(\partial\Omega)$,
- $F \in \left(L^2(\Omega)\right)^3$ and $g \in \left(L^2_t(\partial\Omega)\right)^3$.

The space of tangential traces is

$$L^2_t(\partial\Omega) = \left\{ u \in \left(L^2_t(\partial\Omega)\right)^3, \nu \cdot u = 0 \text{ a.e. on } \partial\Omega \right\}.$$

The condition on ε is the matrix generalization of the hypothesis $\Re(\tilde{\varepsilon}) \geq c > 0$ for a dielectric tensor $\varepsilon = \tilde{\varepsilon} Id$ with $\tilde{\varepsilon} \in \mathbb{R}$. This hypothesis is essential to prove some coercivity.

3.3 The space X and the null space of the *curl* operator

This section states classical properties of the X space and the null space of the curl. Only the sketch of the proofs are given for simplicity. All details are to be found in [Mon03]. See also [BBCD97].

Definition 1. The bilinear form $(\cdot,\cdot)_X$ is defined by : for each $(u,v) \in X^2$

$$(u,v)_X = (u,v) + (\nabla \wedge u, \nabla \wedge v) + \langle u_T, v_T\rangle. \tag{3.4}$$

Theorem 3.3.1. The space X defined in (3.2) when equipped with the inner product $(\cdot,\cdot)_X$ is a Hilbert space. The following space is dense in X :

$$\mathcal{X} = \left\{ u | u = \omega_{|\Omega} \text{ for a } \omega \in \mathcal{C}^\infty\left(\mathbb{R}^3\right) \right\}.$$

Note that this result holds for any bounded Lipschitz domain in $\mathbb{R}^3$. *Proof.* First X is well-defined since the tangential trace makes sense for elements of $H(curl;\Omega)$, so that u_T is well-defined.

Then the completeness of X equipped with the inner product (3.4) stems from the study of a Cauchy sequence and the continuity of the trace operator on $H(curl;\Omega)$.

Last the density property is based on two preliminary density results. One of them concerns the space $H_0(curl;\Omega)$ which is defined as the closure of $\left(\mathcal{C}_0^\infty\left(\overline{\Omega}\right)\right)^3$ in $H(curl;\Omega)$. The other one concerns a space adapted to the impedance boundary condition $H_{imp}(curl;\Omega)$ defined by

$$H_{imp}(curl;\Omega) = \left\{u \in H(curl;\Omega), u \wedge \nu \in L_t^2(\partial\Omega)\right\}.$$

□

The following result characterizes the elements of X with vanishing $curl$, that will be of major interest for the Helmholtz decomposition in the next section.

Theorem 3.3.2. Suppose Ω is a simply connected Lipschitz domain and has a boundary consisting of a single connected component. In addition suppose that $u \in X$ is such that $u_T = 0$ on $\partial\Omega$ and $\nabla \wedge u = 0$ in Ω. Then there is a scalar potential $p \in S$ such that $u = \nabla p$, where S is defined by

$$S = \left\{p \in H^1(\Omega), p = 0 \text{ on } \partial\Omega\right\}.$$

Proof. It relies on a more general result for an $(L^2(\Omega))^3$ function on a bounded simply connected Lipschitz domain, that states the equivalence between having a vanishing curl in Ω and being the gradient of an $H^1(\Omega)$ potential. Such a potential is unique up to an additive constant. □

3.4 The Helmholtz decomposition

The version of the Helmholtz decomposition proposed in this section is adapted to the anisotropic case. Its proof relies on the assumption on ε stated in subsection 3.2.

The decomposition itself relies on the coercivity and continuity of a convenient sesquilinear form, identifying a candidate for the $(X_0)^\perp$ component of the decomposition via Lax-Milgram theorem. The compactness properties of X_0 strongly use the decomposition itself.

Lemma 3.1. *The space ∇S is a closed subset of X and one may write*

$$X = X_0 \oplus \nabla S,$$

where

$$X_0 = \{w \in X, (\varepsilon w, \nabla\xi) = 0 \ \forall \xi \in S\}.$$

Proof. The space ∇S is closed in X since S is closed in $H^1(\Omega)$.

To define the decomposition, note $\tilde{a} : X \times X \to \mathbb{C}$ defined by

$$\tilde{a}(u,v) = (\nabla \wedge u, \nabla \wedge v) + (\epsilon u, v) + \langle u_T, v_T\rangle, \ \forall u, v \in X.$$

Thanks to the assumptions on ε the sesquilinear form is coercive and continuous, that is

- There exists a constant $c > 0$ independent of u such that

$$|\tilde{a}(u,u) \geq c\|u\|_X^2 \ \forall u \in X.$$

This comes from the hypothesis on the eigenvalues of ε since one has

$$(\varepsilon u, u) \geq \left(\min_{\lambda \in Sp(\varepsilon)} |\lambda| \right) (u,u) \ \forall u \in X.$$

- There exists a constant $C > 0$ independent of u and v such that

$$|\tilde{a}(u,v) \leq C\|u\|_X \|v\|_X \ \forall u, v \in X.$$

This comes from the boundedness of the diagonalizable matrix ε since one has

$$(\varepsilon u, u) \leq \left(\max_{\lambda \in Sp(\varepsilon)} |\lambda| \right) (u,u) \ \forall u \in X.$$

As a consequence, from Lax-Milgram theorem, for each $u \in X$ there exists a unique function $Pu \in \nabla S$ such that

$$\tilde{a}(Pu, v) = (\varepsilon u, v) \ \forall v \in \nabla S.$$

Then P is a bounded projection from X to ∇S since $Pu = u$ for $u \in \nabla S$. As a result, for each $u \in X$, $u = Pu + (I - P)u$ and $(I - P)u \in X_0$ since for any $\xi \in S$

$$(\varepsilon(I - P)u, \nabla\xi) = \tilde{a}(Pu, \nabla\xi) - \tilde{a}(P^2u, \nabla\xi).$$

It is zero since $Pu \in \nabla S$. □

3.4.1 Compactness properties of X_0

Theorem 3.4.1. Suppose Ω and ε satisfy the conditions given in Section 3.2. Then X_0 is compactly embedded in $\left(L^2(\Omega)\right)^3$.

The proof of this result relies on the Helmholtz decomposition for $\varepsilon = 1$ with

$$X_0^{(1)} = \{w \in X, (w, \nabla\xi) = 0 \ \forall \xi \in S\},$$

and on the compact embedding of $X_0^{(1)}$ in $\left(L^2(\Omega)\right)^3$. This embedding permits to extract a convergent subsequence in $X_0^{(1)}$, which limit will then be decomposed with the non constant ε version of the Helmholtz decomposition. *Proof.* Consider a bounded sequence in X_0, $\{w_n\}_{n=1}^{\infty}$. As announced, each element of the sequence can be decomposed as $w_n = w_n^{(1)} + \nabla p_n^{(1)}$ where $w_n^{(1)} \in X_0^{(1)}$ and $p_n^{(1)} \in S$. Testing this identity against $p_n^{(1)}$ one gets $\|\nabla p_n^{(1)}\|_X \leq \|w_n\|_X$ and as a consequence $\|w_n^{(1)}\|_X \leq \|w_n\|_X$. Thanks to the compact embedding of $X_0^{(1)}$ in $\left(L^2(\Omega)\right)^3$, there is a subsequence still denoted $\{w_n\}_{n=1}^{\infty}$ and an element $w^{(1)} \in X_0^{(1)}$ such that

$$w_n \to w^{(1)} \text{ strongly in } \left(L^2(\Omega)\right)^3. \tag{3.5}$$

Using the Helmholtz decomposition given in lemma 3.1 one gets $w^{(1)} = w^{(\varepsilon)} + \nabla p^{(\varepsilon)}$ for some $w^{(\varepsilon)} \in X_0$ and $p^{(\varepsilon)} \in S$. The last step is to show that $w_n \to w^{(\varepsilon)}$ in $\left(L^2(\Omega)\right)^3$. Since each w_n and $w^{(\varepsilon)}$ are in X_0 then one has

$$\begin{aligned}\left(\varepsilon(w^{(\varepsilon)} - w_n), w^{(\varepsilon)} - w_n\right) &= \left(\varepsilon(w^{(\varepsilon)} - w_n), w^{(\varepsilon)} + \nabla p^{(\varepsilon)} - w_n + \nabla p_n^{(1)}\right) \\ &= \left(\varepsilon(w^{(\varepsilon)} - w_n), w^{(1)} - w_n^{(1)}\right).\end{aligned}$$

Hence $(\min_{\lambda \in Sp(\varepsilon)} |\lambda|)\|w^{(\varepsilon)} - w_n\|_{(L^2(\Omega))^3} \leq \|w^{(1)} - w_n^{(1)}\|_{(L^2(\Omega))^3}$, which right hand side tends to zero from (3.5). It completes the proof thanks to the assumption on the eigenvalues of ε. □ The following result guarantees the coercivity of the *curl-curl* bilinear form.

Corollary 3.2. *Suppose that Ω is a bounded simply connected Lipschitz domain with a connected boundary. In addition suppose that ε satisfies the conditions given in section 3.2. Then there is a constant C such that for every $u \in X_0$*

$$\|u\|_{(L^2(\Omega))^3} \leq \|\nabla \wedge u\|_{(L^2(\Omega))^3} + \|\nu \wedge u\|_{(L^2(\partial\Omega))^3}.$$

Proof. Because of the compact embedding of X_0 in $\left(L^2(\partial\Omega)\right)^3$ and of the continuity of the trace on $H(curl;\Omega)$, it only requires to verify that if $u \in X_0$ satisfies

$$\|\nabla \wedge u\|_{(L^2(\Omega))^3} + \|\nu \wedge u\|_{(L^2(\partial\Omega))^3} = 0,$$

then $u = 0$. From Theorem 3.3.2, since $\nabla \wedge u = 0$, $\nu \wedge u$ and $u \in X_0$, then $u = \nabla p$ for some $p \in S$. Thus $u \in (\nabla S) \cap (\nabla S)^{\perp}$ and so $u = 0$. □

3.5 The variational problem as an operator equation

The whole point of this section is to evidence a Fredholm alternative after splitting the problem thanks to the Helmholtz decomposition.

The decomposition $X = X_0 \oplus \nabla S$ allows to split the initial problem (3.3) into two parts. Any solution of the variational problem (3.3) can be written $E = E_0 + \nabla p$ for some $E_0 \in X_0$ and $p \in S$. Since $\nabla \wedge \nabla p = 0$ and $(\nabla p) \wedge \nu = 0$ on $\partial\Omega$, then

$$(\nabla \wedge E_0, \nabla \wedge \phi) - (\varepsilon(E_0 + \nabla p), \phi) - i\langle \sigma E_{0,T}, \phi_T \rangle = (F, \phi) + \langle g, \phi_T \rangle.$$

Choosing $\phi = \nabla\xi$ one has $\phi \in H_0(curl;\Omega)$ and since $E_0 \in X_0$ then $p \in S$ satisfies

$$-(\varepsilon \nabla p, \nabla \xi) = (F, \nabla \xi), \text{ for all } \xi \in S. \tag{3.6}$$

The solution of this first subproblem is given by the following lemma.

Lemma 3.3. *Assume that ε satisfies the conditions given in Section 3.2. Then there exists a unique solution $\in S$ to* (3.6) *and there is a constant C independent of F such that*

$$\|\nabla p\|_{(L^2(\Omega))^3} \leq C\|F\|_{(L^2(\Omega))^3}.$$

Proof. Define $\tilde{b} : S \times S \to \mathbb{C}$ by $\tilde{b}(p, \xi) = -(\varepsilon \nabla p, \nabla \xi)$. The sesquilinear form $\tilde{b}$ is bounded and coercive from the hypothesis on ε, so that the result stems from Lax-Milgram theorem. □

So considering $p \in S$ is now known, the next problem is to determine $E_0 \in X_0$ such that

$$(\nabla \wedge E_0, \nabla \wedge \phi) - (\varepsilon E_0, \phi) - i\langle \sigma E_{0,T}, \phi_T \rangle = (F, \phi) + \langle g, \phi_T \rangle + (\varepsilon \nabla p, \phi), \tag{3.7}$$

for all $\phi \in X_0$. Define the sesquilinear form $a_+ : X \times X \to \mathbb{C}$ by

$$a_+(u,v) = (\nabla \wedge u, \nabla \wedge v) + (\varepsilon u, v) - i\langle \sigma u_T, v_T \rangle$$

for all $u, v \in X$. The next lemma shows this sesquilinear form is coercive.

Lemma 3.4. *Assume ε and σ satisfy the conditions given in Section 3.2. There is a constant $c > 0$ depending on ε and σ such that*

$$|a_+(u,u)| \geq c\|u\|_X^2 \text{ for all } u \in X.$$

Proof. From the assumptions on ε one can define the square root of the matrix, $\varepsilon^{1/2}$. From the definition of a_+ one has

$$|a_+(u,u)|^2 = \left| \|\nabla \wedge u\|^2_{(L^2(\Omega))^3} + \|\varepsilon^{1/2} u\|^2_{(L^2(\Omega))^3} \right|^2 + \left| \|\sqrt{\sigma} u_T\|^2_{(L^2(\partial\Omega))^3} \right|^2,$$

so that

$$|a_+(u,u)|^2 \geq \|\nabla \wedge u\|^4_{(L^2(\Omega))^3} + \left(\min_{\lambda \in Sp(\varepsilon)} |\lambda| \right) \|u\|^4_{(L^2(\Omega))^3} + \sqrt{c_\sigma} \|u_T\|^4_{(L^2(\partial\Omega))^3},$$

which in turns leads to the desired inequality. □

Define the operator $K : \left(L^2(\Omega)\right)^3 \to \left(L^2(\Omega)\right)^3$ by $K : f \mapsto Kf$ that satisfies

$$a_+(Kf, \phi) = -2 \int_\Omega \varepsilon f \cdot \overline{\phi} dV \text{ for all } \phi \in X_0.$$

Theorem 3.5.1. Assume that ε and σ satisfy the conditions given in Section 3.2. The operator K is a bounded and compact map from $\left(L^2(\Omega)\right)^3$ into $\left(L^2(\Omega)\right)^3$. In addition

$$\|Kf\|_X \leq \|f\|_{(L^2(\Omega))^3}.$$

Proof. Using the Cauchy-Schwarz inequality together with the hypothesis on the coefficients, one gets

$$\begin{aligned} |a_+(u,v)| \quad &\leq \|\nabla \wedge u\|_{(L^2(\Omega))^3} \|\nabla \wedge v\|_{(L^2(\Omega))^3} \\ &+ \left(\max_{\lambda \in Sp(\varepsilon)} |\lambda| \right) \|u\|_{(L^2(\Omega))^3} \|v\|_{(L^2(\Omega))^3} \\ &+ \|\sigma\|_\infty \|u_T\|_{(L^2(\partial\Omega))^3} \|v_T\|_{(L^2(\partial\Omega))^3}. \end{aligned}$$

Hence $|a_+(u,v)| \leq C\|u\|_X \|v\|_X$. Considering the coercivity proved in Lemma 3.4, it shows that a_+ satisfies the conditions of the Lax-Milgram theorem. As a consequence K is well-defined and $\|Kf\|_X \leq C\|f\|_{(L^2(\Omega))^3}$.

The compactness of K directly stems from the above inequality together with the compact embedding of X_0 in $\left(L^2(\Omega)\right)^3$ proved in Theorem 3.4.1. □ A last step leads to a Fredholm alternative.

Applying again Lax-Milgram theorem, one can define $\mathcal{F} \in X_0$ such that

$$a_+(\mathcal{F}, \phi) = (F, \phi) + \langle g, \phi_T \rangle + (\varepsilon \nabla p, \phi) \text{ for all } \phi \in X_0,$$

and see that

$$\|\mathcal{F}\|_X \leq C \left(\|F\|_{(L^2(\Omega))^3} + \|g\|_{(L^2(\partial\Omega))^3} + \|\nabla p\|_{(L^2(\partial\Omega))^3} \right),$$

which from Lemma 3.3 gives

$$\|\mathcal{F}\|_X \leq C\left(\|F\|_{(L^2(\Omega))^3} + \|g\|_{(L^2(\partial\Omega))^3}\right).$$

As a consequence the problem (3.7) is equivalent to finding $E_0 \in \left(L^2(\Omega)\right)^3$ such that

$$(I + K)E_0 = \mathcal{F}, \tag{3.8}$$

and thanks the Fredholm alternative - because K is compact - it is then enough to prove the uniqueness to get the existence. At this point the aim of this section is reached.

Remark that if E_0 stands for a solution of (3.8), then one has

$$\|E_0\|_{(L^2(\Omega))^3} \leq C\|\mathcal{F}\|_{(L^2(\Omega))^3}. \tag{3.9}$$

But since $E_0 = \mathcal{F} - KE_0$, one also has that $E_0 \in X_0$ and thus

$$\|E_0\|_X \leq C\left(\|\mathcal{F}\|_X + \|E_0\|_{(L^2(\Omega))^3}\right),$$

which in turns thanks to (3.9) provides an estimate on E_0 in X_0 with respect to the data of the initial problem

$$\|E_0\|_X \leq C\left(\|F\|_{(L^2(\Omega))^3} + \|g\|_{(L^2(\partial\Omega))^3}\right).$$

Remark 1. In the case $\varepsilon(\overrightarrow{x}) \leq -c < 0$ for a constant $c > 0$ the problem is also well-posed in $H(curl)$. There is no such theory for changing sign ε. See [BCC12] where a different approach, named the T-coercivity, addresses the case of changing sign is the transition is discontinuous.

Chapter 4

Hybrid resonance in planar geometry

Contents

4.1 Introduction

In this chapter, we introduce an approach which is different from the previous chapter, and is inspired by the physics of the problem. Indeed, it is known in plasma physics that Maxwell's equations in the context of a strong background magnetic field may develop singular solutions even for smooth coefficients. This is related to what is called the hybrid resonance [CW74, Fre07, Bra98]. To our knowledge the mathematical analysis of such a phenomenon can not be found in the applied mathematics literature.

Hybrid resonance appears in reflectometry experiments [DHM06, HGP10] and heating devices for fusion plasma [DPS05]. The energy deposit may exceed by far the energy exchange which occurs in Landau damping [Fre07, MV11]. Hybrid resonance is a non damping dissipative phenomenon : a singularity of the solution makes it stronger in some sense than the Landau damping. Indeed Landau damping appears in kinetic models, whereas hybrid resonance appears in a simpler model coupling a fluid with the non-electrostatic part of Maxwell's equations.

Since the mathematical solution is not square integrable, hybrid resonance is also a paradoxical and non standard phenomenon in the context of the mathematical theory of Maxwell's equations, additional references are [DL85, Ces96a, Mon03, Wed91]. The situation can be compared with the mathematical theory of metametarials. In [Wed08a, Wed08b] the electric dielectric and magnetic susceptibility tensors are degenerate -i.e. they have zero eigenvalues- in surfaces, but they remain positive definite. In this case, the solutions are singular, but the problem remains coercive. In [BCZ08, BCC12] the coefficient changes in a discontinuous way from being positive to negative. In this situation coerciveness is lost, but as the absolute value of the coefficient is bounded from below by a positive constant, the solutions are regular. In the case considered here, both difficulties occur at the same time. As the coefficient α (see below) continuously goes from positive to negative values, its absolute value is zero at a point, and as a consequence the problem is not coercive and there are singular solutions.

Our purpose in this chapter is to construct in Theorem 4.1.1 a mathematical solution with a hybrid resonance in planar geometry. It will be done with a limit absorption principle to give a meaning to relevant solutions, and with an extensive use of sharp estimates for singular integral equations. See [JW10] for a recent use of such a method for a completely different problem. An original singular integral equation is attached to the Fourier solution. Introduced in the seminal work of Hilbert [Hil53] and Picard [Pic11], this type of integral equation is referred to as integral equation of the third kind, supplementing the more classical distinction between equations of the first and second kind, see [Vol10, Tri85]. Some references about this type of equations may be found in [BW73, Shu97] for a mathematical analysis, and [Van55, Cas59, FW63] for a theory of particles or plasma physics. The results in this chapter are reminiscent of those of Bart and Warnock [BW73], even if the considered kernel does not satisfy exactly the same hypothesis since it is more singular. Their work stresses the fact that non uniqueness is the rule for such equations. However here the unique solution is obtained thanks to the limit absorption principle which is a physically based selection principle. It is explicitly described as the sum of a singular part plus a

regular part.

One originality of this work lies in the analysis of the properties of this singular equation, since there is no equivalent to it in the classical literature [AS72, Bat53, BM96].

4.1.1 The mathematical model

Introducing the vorticity W, the X mode problem stated in the introduction is expressed as a first order system of three equations

$$\left\{\begin{array}{rrrl} W & +\partial_y E_x & -\partial_x E_y & = 0, \\ \partial_y W & -\alpha E_x & -i\gamma E_y & = 0, \\ -\partial_x W & +i\gamma E_x & -\alpha E_y & = 0. \end{array}\right. \tag{4.1}$$

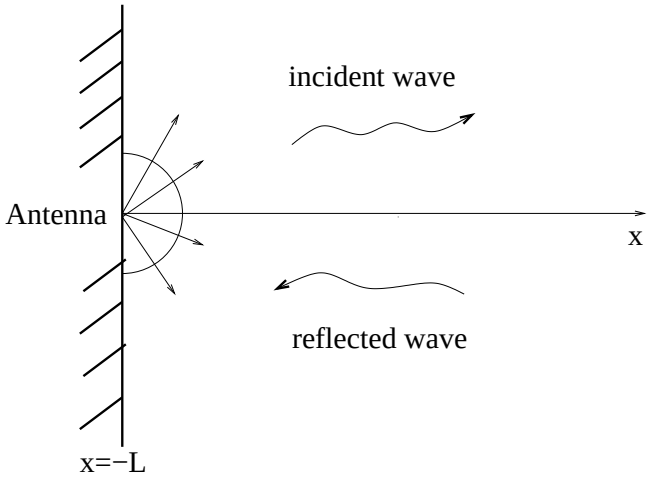


Figure 4.1 – X-mode in planar geometry : the domain. In a real physical device an antenna is on the wall on the left and sends an incident electromagnetic wave through a medium which is assumed infinite for simplicity. The incident wave generates a reflected wave. The antenna will be modeled by the non homogeneous boundary condition (4.2). The medium is filled with a plasma with dielectric tensor given by (2.8).

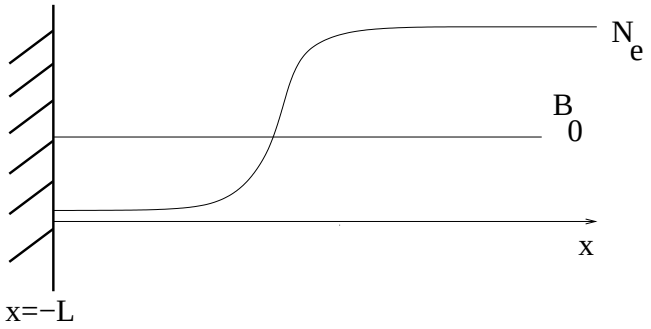


Figure 4.2 – X-mode equations in planar geometry : the physical parameters. The electronic density $x \mapsto N_e(x)$ is low at the boundary, and increases towards a plateau. The background magnetic field B_0 is taken as constant for simplicity.

The simplified 2D domain considered is for some $L > 0$:

$$\Omega = \Big\{(x, y) \in \mathbb{R}^2, \quad -L \leq x, \quad y \in \mathbb{R}\Big\},$$

see Figure 4.1. The X-mode equations (4.1) are combined with a non homogeneous boundary condition

$$W + i\sigma n_x E_y = g \text{ on the left boundary } x = -L, \qquad \sigma > 0, \tag{4.2}$$

which models a given source, typically a radiating antenna meant to heat or to probe the plasma.

4.1.2 Technical hypothesis

Planar geometry will be considered for the sake of simplicity, that is to say the coefficients α and γ only depend on the x variable,

$$\partial_y \alpha = \partial_y \gamma = 0.$$

Other assumptions which correspond to the physical context of idealized reflectometry or heating devices are the following. The diagonal part of the dielectric tensor is dominated by the extra-diagonal part at a finite number o points, that is

$$\alpha(x_i) = 0, \ \alpha'(x_i) \neq 0 \text{ and } \delta(x_i) \neq 0, \quad x_i \in \mathbb{R}, \ i = 1, \ldots, N.$$

For the sake of simplicity suppose here that $N = 1$ and that $x_1 = 0$. Assume that

$$\gamma \in \mathcal{C}^1([-L, \infty[), \qquad \gamma(0) \neq 0. \tag{4.3}$$

The coefficient α satisfies

$$\alpha \in \mathcal{C}^2([-L, \infty[), \qquad \alpha(0) = 0, \quad \alpha'(0) < 0, \tag{H1}$$

and

$$\alpha_- \leq \alpha(x) \leq \alpha_+, \ \forall x \in [-L, \infty[, \qquad \text{and} \quad 0 < r \leq \left|\frac{\alpha(x)}{x}\right|, \ \forall x \in [-L, H] \tag{H2}$$

where $H > 0$. Assume that the coefficients are constant at large scale : there exists γ_∞ and α_∞ so that

$$\gamma(x) = \gamma_\infty \text{ and } \alpha(x) = \alpha_\infty \qquad \forall x \in \mathbb{R} \text{ s.t. } H \leq x < \infty. \tag{H3}$$

Assume that the problem is coercive from H to infinity

$$\alpha_\infty^2 - \gamma_\infty^2 > 0. \tag{H4}$$

These last two assumptions are justified since the electromagnetic wave is strongly absorbed for $x \geq H$.

The coefficients α and γ could be only piecewise smooth and all the theoretical results would remain the same. This justifies the choice of piecewise $\mathcal{C}^\infty$ coefficients for the numerical test cases, see Chapter 6.

An additional condition is defined by

$$4\|\gamma\|_\infty^2 H < r. \tag{H5}$$

It expresses the fact that the length of the transition zone between $x = 0$ and $x = H$ is small with respect to the other parameters of the problem. One can refer to Figures 4.2 and 4.3 for a graphical representation. This hypothesis is physically very reasonable, and will be discussed in Section 4.9.

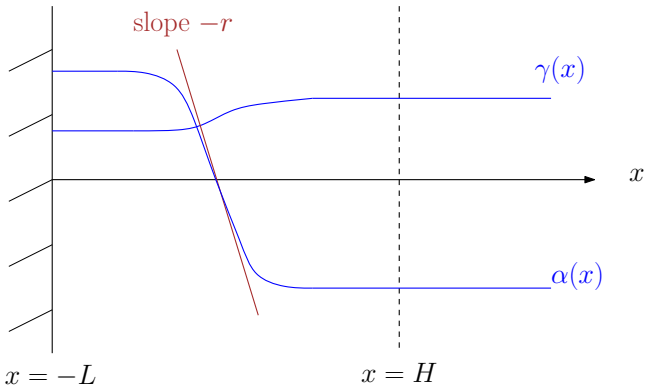


Figure 4.3 – X-mode equations in planar geometry : parameters of the dielectric tensor deduced from the value of the physical parameters described in figure 4.2, assuming that $\omega > \omega_c$. The coefficient α decreases from positive to negative values. It crosses the axis with a slope bounded from below by r. The coefficient γ is positive and bounded, and the two coefficients are constant for $x > H$.

4.1.3 Statement of the main result

Assuming (H1)-(H4), it is known in the Physics community that the problem is highly singular at the origin. The main result of this chapter can be summarized as follows. Denote by $\widehat{g}$ the Fourier transform of g,

$$\widehat{g}(\theta) := \int_{\mathbb{R}} g(y)\, e^{-i\theta y}\, dy.$$

The uniform transversality assumption (H6) is needed, it is a generalization of assumption (H5). See Section 4.8.

Theorem 4.1.1. Assuming (H1)-(H6), for every $g \in L^2(\mathbb{R})$ with $\widehat{g}$ of compact support there exists a solution in the sense of distributions of (4.1) with boundary condition (4.2) that goes to zero at infinity. Moreover, unless the source term g is identically zero, the x component of this solution E_x does not belong to $L^1_{\text{loc}}((-L,\infty)\times\mathbb{R})$. The other components E_y and W are more regular, they belong to $L^2((-L,\infty)\times\mathbb{R})$.

The result relies on a limit absorption principle combined with a specific original integral representation of the solution. The loss of regularity of the electric field is counter intuitive with respect to the standard theory of existence and uniqueness for solutions of time harmonic Maxwell's equations [DL85, Ces96a, Mon03, Wed91]. The essential part of the proof consists in identifying the singularity appearing in the Fourier transform $\widehat{E}_x$. Moreover, the condition (H5) simplifies some parts of the mathematical analysis. The solution is a priori not unique since the limit absorption principle generates two solutions depending on the sign of the regularization.

The singular part of the solution will be presented as the sum of the principal value of the inverse of α, plus a Dirac function, plus a smooth term.

Note that the heating of the plasma (4.90) is directly related to the singular part of the solution.

4.1.4 A wave description of the problem : Phase velocity

The phase velocity measures the velocity of individual Fourier modes. Focus on the dispersion relation in dimension two. A plane wave $(E_x, E_y) = e^{i\overrightarrow{\kappa}\cdot\overrightarrow{x}}\overrightarrow{E_0}$, where $\overrightarrow{E_0} \in \mathbb{C}^2$, is a solution of X-mode equations (4.1) if and only if

$$\left[\begin{pmatrix} \kappa_2^2 & -\kappa_1\,\kappa_2 \\ -\kappa_1\,\kappa_2 & \kappa_1^2 \end{pmatrix} - \begin{pmatrix} \alpha & i\gamma \\ -i\gamma & \alpha \end{pmatrix}\right]\overrightarrow{E_0} = 0, \qquad = (\kappa_1, \kappa_2) \in \mathbb{R}^2.$$

Set $\overrightarrow{\kappa} = n^2(\cos\theta, \sin\theta)$, θ denoting the direction of the wave. The phase velocity $v_\varphi = \frac{\omega}{|n^2|}$ is solution of the eigenvalue problem

$$\begin{pmatrix} \sin\theta^2 - v_\varphi^2\alpha & -\cos\theta\sin\theta - iv_\varphi^2\gamma \\ -\cos\theta\sin\theta + iv_\varphi^2\gamma & \cos\theta^2 - v_\varphi^2\alpha \end{pmatrix} R = 0.$$

The determinant of the matrix is

$$D = v_\varphi^4\left(\alpha^2 - \gamma^2\right) - v_\varphi^2\alpha.$$

Setting $D = 0$ the phase velocity reads

$$v_\varphi^2 = \frac{\alpha}{\alpha^2 - \gamma^2}.$$

Constant coefficients

Consider first that α and γ are constant at least locally. Then the phase velocity v_φ itself is constant as well.

Non constant coefficients

Assume for example that $\alpha = -x$ and that $\gamma = 1$ which is locally compatible with the general assumptions, see Figure 4.3. Figure 4.4 shows the phase velocity as a function of the horizontal space coordinate. When the phase velocity is real we are in a propagating region, and when the phase velocity is pure imaginary the region is non-propagating. One distinguishes two cutoffs where the local phase velocity is infinite

$$\text{Cutoff}: \quad \alpha(x) = \pm\gamma(x)$$

and one resonance where the phase velocity is null

$$\text{Resonance}: \alpha(x) = 0.$$

This structure is characteristic of the hybrid resonance.

4.1.5 Outline

This chapter is organized as follows. Section 4.2 provides an analytic singular solution, that justifies the need for a specific analysis of the problem. The technical tools that are set up to establish this analysis are presented in Section 4.3. In order to display the solutions of the regularized problem, Section 4.4 details a convenient basis of the function space of these solutions, and some continuity estimates are presented in Section 4.5. These estimates are uniform with respect to the regularization process, so that they will still hold at the limit. The process of passing to the limit is described in Section 4.6. A basic numerical validation is then proposed in Section 4.7. The main theorem is proved in Section 4.8. At last the hypothesis (H5) is discussed in Section 4.9.

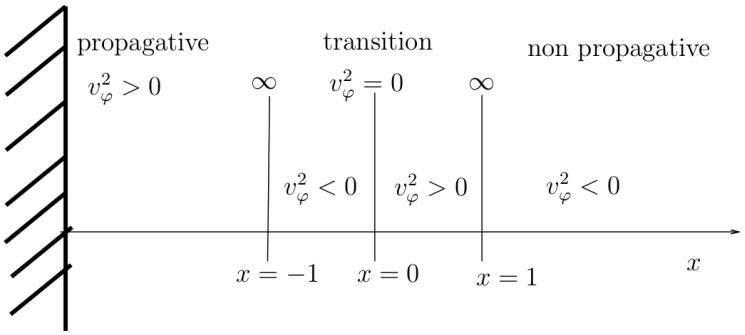


FIGURE 4.4 – Sign of the square of the phase velocity. In this example $\alpha = -x$ and $\gamma = 1$, so that $v_\varphi^2 = \frac{x}{1-x^2}$.

4.2 An example of a singular solution

The aim of this small section is to construct an explicit singular solution.

If the solution is independent of y, which corresponds to normal incidence for plane waves, the system (4.1) is called the Budden problem [CW74]

$$\begin{cases} W - E_y' & = 0, \\ -\alpha E_x - i\gamma E_y & = 0, \\ -W' + i\gamma E_x - \alpha E_y & = 0. \end{cases}$$

After elimination of E_x and W, the y component of the electric field satisfies

$$-E_y'' + \left(\frac{\gamma^2}{\alpha} - \alpha\right) E_y = 0.$$

This equation can be solved analytically in some cases which gives a better understanding of the singularity of the general problem. As pointed out by Olivier Lafitte, a more general theory which refers to this problem as turning points is to be found in [CL55].

Consider that $\alpha = -x$ and γ is solution of $\frac{\gamma^2}{x} - x = \frac{1}{4} - \frac{1}{x}$. The positive solution is

$$\gamma(x) = \sqrt{x^2 - \frac{x}{4} + 1} > 0.$$

The y-component of the electric field is then solution of

$$E_y'' + \left(-\frac{1}{4} + \frac{1}{x}\right) E_y = 0. \tag{4.4}$$

This equation is of Whittaker type [AS72, Bat53]. It is a particular case of the confluent hypergeometric equation, and can also be written in the Kummer form. The general theory shows that the first fundamental solution is regular

$$v(x) = e^{-\frac{x}{2}} x.$$

Indeed $v'(x) = e^{-\frac{x}{2}} \left(1 - \frac{x}{2}\right)$ and $v''(x) = e^{-\frac{x}{2}} \left(-1 + \frac{x}{4}\right)$, so that this function v satisfies

$$v'' + \left(-\frac{1}{4} + \frac{1}{x}\right) v = 0.$$

In order to build a singular solution : consider a second solution w with linear independence with respect to the first one. The linear independence can be imposed by setting the normalized Wronskian relation

$$v(x)w'(x) - v'(x)w(x) = 1.$$

Seeking for a representation $w = vz$, one gets that

$$v^2 z' = 1 \Rightarrow z = \int \frac{dx}{v^2} \Rightarrow w = v \int \frac{dx}{v^2} = x\, e^{-x/2} \int \frac{e^x}{x^2}.$$

Moreover, from formulas 8.212 of [GR65],

$$\int \frac{e^x}{x^2} = -\frac{e^x}{x} + \int \frac{e^x}{x} = -\frac{e^x}{x} + E_i(x),$$

where $E_i(x)$ is the Exponential-integral function. It follows that

$$w(x) = -e^{x/2} + x\, e^{-x/2}\, E_i(x).$$

Furthermore, from formulas 8.214 of [GR65],

$$E_i(x) = \ln|x| + \sum_{j=1}^{\infty} \frac{x^j}{j \cdot j!},$$

and as a consequence

$$w(x) = -1 + Cx + x\, \ln|x| + O(|x|), \quad |x| \to 0.$$

This second function w is bounded, but however non regular at origin. It shows the subtleties associated with the singular Whittaker equation (4.4). Nevertheless the general form of the y-component of the electric field in the case of the Budden problem is bounded

$$E_y = av + bw \Rightarrow E_y \in L^\infty(]-\epsilon, \epsilon[) \quad \forall \epsilon > 0,$$

but the x-component of the electric field is more singular. It is a linear combination of two functions, the first one which is regular and bounded

$$E_x^v(x) = i\frac{\sqrt{x^2 - \frac{x}{4} + 1}}{x} v(x) = ie^{-\frac{x}{2}} \sqrt{x^2 - \frac{x}{4} + 1},$$

and the second one which is singular at the origin since $w(0) = -1$

$$E_x^w(x) = i\frac{\sqrt{x^2 - \frac{x}{4} + 1}}{x} w(x).$$

Since for this second solution $E_x^w \notin L^2(]-\epsilon, \epsilon[)$, the electric field is not a square integrable function in general.

For that reason, it is necessary to develop a different approach to study the general solution of the problem.

4.3 Theoretical tools

This section describes the regularized system that will be studied in order to define a solution to the initial system (4.1), together with a convenient integral representation. Some basic properties of the integral representation follow.

4.3.1 Limit absorption principle and Fourier transform

In order to give a rigorous meaning to the solution at all incidences, a regularized approach is developed. It is based on the limit absorption principle, a standard mathematical principle to give a meaning to ill-posed problems. It can be understood adding a term to the initial Newton law modeling the friction of the electrons on the ions, see [Des] : μ would then be a small collision frequency. Consider a parameter $\mu \neq 0$ (the precise sign will be justified later) and the regularized problem with unknown $(E_x^\mu, E_y^\mu, W^\mu)$

$$\left\{ \begin{array}{cccc} W^\mu & +\partial_y E_x^\mu & -\partial_x E_y^\mu & = 0, \\ \partial_y W^\mu & -(\alpha(x)+i\mu)E_x^\mu & -i\gamma(x)E_y^\mu & = 0, \\ -\partial_x W^\mu & +i\gamma(x)E_x^\mu & -(\alpha(x)+i\mu)E_y^\mu & = 0. \end{array} \right. \tag{4.5}$$

Such a system is well-posed thanks to the result of Chapter 3.

A further simplification consists in Fourier reduction. Since the coefficients do not depend on the y variable, one can perform the usual one dimension reduction. The next system is obtained by applying the Fourier transform to the regularized system (4.5). Denoting the unknowns (U, V, W) it yields

$$\left\{ \begin{array}{cccc} W & +i\theta U & -V' & = 0, \\ i\theta W & -(\alpha(x)+i\mu)U & -i\gamma(x)V & = 0, \\ -W' & +i\gamma(x)U & -(\alpha(x)+i\mu)V & = 0. \end{array} \right. \tag{4.6}$$

Here the notation $'$ refers to the derivative with respect to the x variable.

Note that this last system (4.6) is then a first order system of three equations depending on only one space space variable. Since no derivative of the U component appears, this component is likely to be more singular. Indeed, as was stated in Theorem 4.1.1, the singularity of the solution only concerns the E_x component of the electric field.

4.3.2 A general integral representation

A historical presentation of integral equations was taught by Volterra at the beginning of the twentiest century, see [Vol10]. A more modern and more classical presentation is to be found in [Tri85].

Some definitions are useful to set clearly the next Proposition 4.1.

Definition 2. Denote by (A_μ, B_μ) the two fundamental solutions of the modified equation

$$-u'' - (\alpha(x) + i\mu)u = 0, \tag{4.7}$$

with the usual normalization

$$A_\mu(0) = 1, \quad A'_\mu(0) = 0 \qquad \text{and } B_\mu(0) = 0, \quad B'_\mu(0) = 1. \tag{4.8}$$

The associated kernel k^μ is defined by

$$k^\mu(x,z) = B_\mu(z)A_\mu(x) - B_\mu(x)A_\mu(z). \tag{4.9}$$

Various continuity estimates of A_μ and B_μ are derived in appendix A.1 for the sake of the completeness of this work.

Any couple of independent solutions could be considered in the sequel. However, this choice of normalization at the resonance point will be convenient with respect to the fore coming integral representation properties. Moreover, since it does not depend on the parameter μ it is convenient considering the limit $\mu \to 0$: the functions $A_{\mu=0}$ and $B_{\mu=0}$ are independent solutions of the limit equation $-u'' - \alpha(x)u = 0$.

Definition 3. The operator $\mathcal{D}_z^\theta$ is $i\theta\partial_z - i\gamma(z)\mathrm{Id}$ applied to any function h, that is

$$\mathcal{D}_z^\theta h = i\theta\partial_z h - i\gamma(z)h. \tag{4.10}$$

Definition 4. For some point $G \in \mathbb{R}$, the kernel function is defined by

$$K_1^{\theta,\mu}(x,z;G) = \begin{cases} \dfrac{\mathcal{D}_x^\theta \mathcal{D}_z^\theta k^\mu(x,z)}{\alpha(x)+i\mu}, & \text{for } G \le z \le x \qquad \text{or } x \le z \le G, \\ 0, & \text{in all other cases.} \end{cases} \tag{4.11}$$

Then the kernel sequence is defined by

$$K_{n+1}^{\theta,\mu}(x,z;G) = \int_G^x \frac{\mathcal{D}_x^\theta \mathcal{D}_z^\theta k^\mu(x,t)}{\alpha(x)+i\mu} K_n^{\theta,\mu}(t,z;G)dt. \tag{4.12}$$

The sum is denoted

$$\mathcal{K}^{\theta,\mu}(x,z;G) = \sum_{n=0}^{\infty} K_{n+1}^{\theta,\mu}(x,z;G) \tag{4.13}$$

and define the resolvent kernel.

The integration domain is centered on G, that is

$$\mathrm{supp}\left(K_1^{\theta,\mu}(\cdot,\cdot;G)\right) \subset \left\{(x,y)\in\mathbb{R}^2; \quad G \le z \le x \text{ or } x \le z \le G\right\} \equiv \mathcal{D}_G, \tag{4.14}$$

which yields as well

$$\mathrm{supp}\left(K^{\theta,\mu}(\cdot,\cdot;G)\right) \subset \mathcal{D}_G.$$

Proposition 4.1. *Any triplet (U,V,W) solution of the regularized system (4.6) admits the following integral representation.*

- *Set first an arbitrary reference point $G \in [-L,\infty[$.*
- *The x component of the electric field is solution of the integral equation*

$$U(x) - \int_G^x \frac{\mathcal{D}_x^\theta \mathcal{D}_z^\theta k^\mu(x,z)}{\alpha(x)+i\mu} U(z)dz = \frac{F_G^{\theta,\mu}(x)}{\alpha(x)+i\mu}, \tag{4.15}$$

where the right hand side is defined by

$$F_G^{\theta,\mu}(x) = a_G \mathcal{D}_x^\theta A_\mu(x) + b_G \mathcal{D}_x^\theta B_\mu(x) \tag{4.16}$$

and the kernel is defined thanks to (4.9-4.10). The solution of this kind of integral equation is naturally provided by the resolvent integral formula

$$U(x) = \frac{F_G^{\theta,\mu}(x)}{\alpha(x)+i\mu} + \int_G^x \mathcal{K}^{\theta,\mu}(x,z;G)\frac{F_G^{\theta,\mu}(z)}{\alpha(z)+i\mu}dz \tag{4.17}$$

where the resolvent kernel was defined in (4.12).

- *The y component of the electric field is recovered as*

$$V(x) = a_G A_\mu(x) + b_G B_\mu(x) + \int_G^x \mathcal{D}_z^\theta k^\mu(x,z)U(z)dz, \tag{4.18}$$

and the vorticity is recovered as

$$W(x) = a_G A'_\mu(x) + b_G B'_\mu(x) + \int_G^x \partial_x \mathcal{D}_z^\theta k^\mu(x,z)U(z)dz. \tag{4.19}$$

- *The two complex numbers (a_G, b_G) solve the linear system*

$$\begin{cases} a_G A_\mu(G) + b_G B_\mu(G) = V(G), \\ a_G A'_\mu(G) + b_G B'_\mu(G) = W(G). \end{cases} \tag{4.20}$$

Proof. Eliminating W from the first and third equations of (4.6) gives

$$-V'' - (\alpha + i\mu)V = f \qquad \text{with } f = -i\theta U' - i\gamma U.$$

Since the Wronskian is constant, it follows from the normalization (4.8) that $A_\mu B'_\mu - A'_\mu B_\mu = 1$. Then, from the variation of constants formula,

$$V(x) = a_f A_\mu(x) + b_f B_\mu(x) + \int_G^x f(z) k^\mu(x,z) dz, \quad \forall x \geq -L, \tag{4.21}$$

where a_f and b_f are two integration constants. Now replace f by the corresponding function of U and perform the integration by part

$$\int_G^x U'(z) k^\mu(x,z) dz = U(x)k^\mu(x,x) - U(G)k^\mu(x,G) - \int_G^x U(z)\partial_z k^\mu(x,z) dz.$$

Since $k^\mu(x,x) = 0$ there is a simplification. Therefore (4.21) yields (4.18) with $a_G = a_f + i\theta U(G)B_\mu(G)$ and $b_G = b_f - i\theta U(G)A_\mu(G)$. Next eliminate W from the first and second equations of (4.6) shows that

$$-i\theta V' - \theta^2 U + (\alpha + i\mu)U + i\gamma V = 0. \tag{4.22}$$

The derivative of (4.18) yields

$$V'(x) = a_G A'_\mu(x) + b_G B'_\mu(x) + \int_G^x \partial_x \mathcal{D}_z^\theta k^\mu(x,z) U(z) dz + \mathcal{D}_z^\theta k^\mu(x,x) U(x).$$

Since $\mathcal{D}_z^\theta k^\mu(x,x) = i\theta \left(A_\mu B'_\mu - B_\mu A'_\mu \right)(x) = i\theta$, one gets the identity

$$V'(x) = a_G A'_\mu(x) + b_G B'_\mu(x) + \int_G^x \partial_x \mathcal{D}_z^\theta k^\mu(x,z) U(z) dz + i\theta U(x).$$

The integral equation (4.15) then stems from plugging this expression in (4.22) and performing all simplifications. Finally, one gets the last integral formula (4.19) from $W = -i\theta U + V'$. The linear system (4.20) is obvious from (4.18)-(4.19) at $x = G$. □

Following [Pic11], the equation (4.15) is an integral equation of the third kind in the case $\mu = 0$. In this case the theory is rather incomplete regarding existence and uniqueness [BW73]. However as long as $\mu \neq 0$, the solution based on these integral equations is uniquely defined. Then, the question is to determine the behavior of these solutions when μ goes to 0. Moreover, different choices of G will give different kind of information. A strategy to study of the limit solution $\mu \to 0$ can be the following : *Choose an optimal G, so that a) the integration constants (a_G, b_G) are easy to determine, and b) the resolvent kernel $\mathcal{K}^{\theta,\mu}(\cdot,\cdot;G)$ admits a limit as $\mu \to 0$.*

Considering the form of the right hand side in (4.17), a convenient tool is the Plemelj-Privalov Theorem [Mus92, Pri56]. Unfortunately, a fundamental singularity of the kernel $\mathcal{K}^{\theta,\mu}(\cdot,\cdot;G)$, topic of Subsection 4.3.3, prevents any simple limit procedure. A more convenient technique will be proposed in Section 4.4.

4.3.3 Singularity of the kernels

A fundamental tool in order to pass to the limit in singular integrals is the Plemelj-Privalov Theorem [Mus92, Pri56]. As a reminder, it can be stated as follows.

Theorem 4.3.1. Let $y \mapsto \varphi(y)$ be a function of Holder class $(0 < \alpha < 1)$ on the closed contour $y \in \mathcal{C}$. Then the function

$$y \mapsto s\varphi(y) = \lim_{\mu \to 0^+} \int_{\mathcal{C}} \frac{\varphi(y) - \varphi(t)}{y + i\mu - t} dt$$

is well defined and is also of Holder class $(0 < \alpha < 1)$.

More details are to be found in the appendix.

However, to apply this theorem to pass to the limit $\mu \to 0$ in equation (4.17) it is necessary that the kernel $\mathcal{K}^{\theta,\mu}(x,z)$ be a Hölder continuous function of z for each fixed x. Unfortunately, this regularity is not available in the case studied here.

To illustrate this phenomenon, we study only the first term of the series (4.12) that defines $\mathcal{K}^{\theta,\mu}$, namely

$$\mathcal{K}_1^{\theta,\mu}(x,z) := \frac{\mathcal{D}_x^\theta \mathcal{D}_z^\theta k^\mu(x,z)}{\alpha(x) + i\mu}. \tag{4.23}$$

The study of the singularity depends on the choice of G to be zero or not.

First case : $G \neq 0$

In this case there exists $(0,z) \in \mathcal{D}_G$ with $z \neq 0$. In the limit case $\mu = 0$ one has that $\mathcal{K}_1^{\theta,0}(x,z)$ admits the local expansion :

$$\mathcal{K}_1^{\theta,0}(x,z) \approx \frac{1}{x\ \alpha'(0)}\, \mathcal{D}_x^\theta\, \mathcal{D}_z^\theta\, k^0(x,z).$$

Therefore, $\mathcal{K}_1^{\theta,0}(x,z)$ blows up as $x \to 0$.

Second case : $G = 0$

Here is a preliminary result on the kernel k^μ evaluated on the diagonal $\{(x,x) \in \mathbb{R}^2$ such that $x \geq -L\}$.

Proposition 4.2. *One has*

$$(\mathcal{D}_x^\theta \mathcal{D}_z^\theta k^\mu)(x,x) = 0 \quad \forall x \geq -L. \tag{4.24}$$

Proof. Indeed by construction

$$(\mathcal{D}_x^\theta \mathcal{D}_z^\theta k^\mu)(x,x) = -\gamma(x)\gamma(x)k^\mu(x,x)$$
$$+\theta\,\gamma(x)\,((\partial_x k^\mu)(x,x) + (\partial_z k^\mu)(x,x)) - \theta^2(\partial_x\partial_z k^\mu)(x,x).$$

Notice that by definition $k_\mu(x,x) = 0$ for all x so the first contribution vanishes in $(\mathcal{D}_x^\theta \mathcal{D}_z^\theta k^\mu)(x,x)$. One also has that

$$(\partial_x k^\mu)(x,x) + (\partial_z k^\mu)(x,x)$$
$$= B_\mu(x)A'_\mu(x) - B'_\mu(x)A_\mu(x) + B'_\mu(x)A_\mu(x) - B_\mu(x)A'_\mu(x) = 0,$$

so, the second contribution vanishes as well. Furthermore,

$$(\partial_x\partial_z k^\mu)(x,x) = B'_\mu(x)\, A'_\mu(x) - B'_\mu(x)\, A'_\mu(x) = 0.$$

This completes the proof of identity (4.24). □

Proposition 4.3. *The limit kernel*

$$\frac{\mathcal{D}_x^\theta \mathcal{D}_z^\theta k^{\mu=0}(x,z)}{\alpha(x)}$$

belongs to $L^\infty\left(\mathcal{D}_0 \cap \{x \in [-L,H] \setminus \{\hat{\text{A}}\ 0\hat{\text{A}}\ \}\}\right)$.

Proof. A first order Taylor expansion of $\mathcal{D}_x^\theta \mathcal{D}_z^\theta k^\mu$ around 0 yields

$$\mathcal{D}_x^\theta \mathcal{D}_z^\theta k^\mu(x,z) = \alpha_\mu x + \beta_\mu z + O(|x|^2 + |z|^2).$$

Notice that (4.24) implies $\beta_\mu = -\alpha_\mu$. The coefficient α_μ is easily computed using

$$(\mathcal{D}_x^\theta \mathcal{D}_z^\theta k^\mu)(x,0) = \mathcal{D}_x^\theta A_\mu(x) \mathcal{D}_z^\theta B_\mu(0) - \mathcal{D}_x^\theta B_\mu(x) \mathcal{D}_z^\theta A_\mu(0)$$

and the definition (4.7-4.8). One gets that

$$\begin{cases} \mathcal{D}_x^\theta A_\mu(x) = -i\gamma(0) - i\gamma'(0)x + \theta\mu x + O(x^2), \\ \mathcal{D}_x^\theta B_\mu(x) = i\theta - i\gamma(0)x + O(x^2). \end{cases}$$

So

$$\begin{aligned} (\mathcal{D}_x^\theta \mathcal{D}_z^\theta k^\mu)(x,0) &= \left(-i\gamma(0) - i\gamma'(0)x + \theta\mu x + O(x^2)\right) i\theta \\ &\quad - \left(i\theta - i\gamma(0)x + O(x^2)\right)\left(-i\gamma(0)\right) \\ &= \left(\gamma(0)^2 + \theta\gamma'(0) + i\theta^2\mu\right) x + O(x^2). \end{aligned}$$

This coefficient α_μ being constant, one obtains that

$$\varphi_x(z) := \frac{\mathcal{D}_x^\theta \mathcal{D}_z^\theta k^{\mu=0}(x,z)}{\alpha(x)} = \frac{\left(\gamma(0)^2 + \theta\gamma'(0)\right)(x-z) + O(|x|^2 + |z|^2)}{\alpha(x)}. \tag{4.25}$$

This expansion is valid for $(x,z) \in \mathcal{D}_0$,see (4.14) for its definition : in this case $|x-z| \le |x|$ and $|z| \le |x|$. Moreover, since $\alpha(x) = x(\alpha'(0) + O(1))$ we obtain that

$$|\varphi_x(z)| \le \frac{\left|\gamma(0)^2 + \theta\gamma'(0)\right|}{\sqrt{\alpha'(0)^2}} + O(|x|).$$

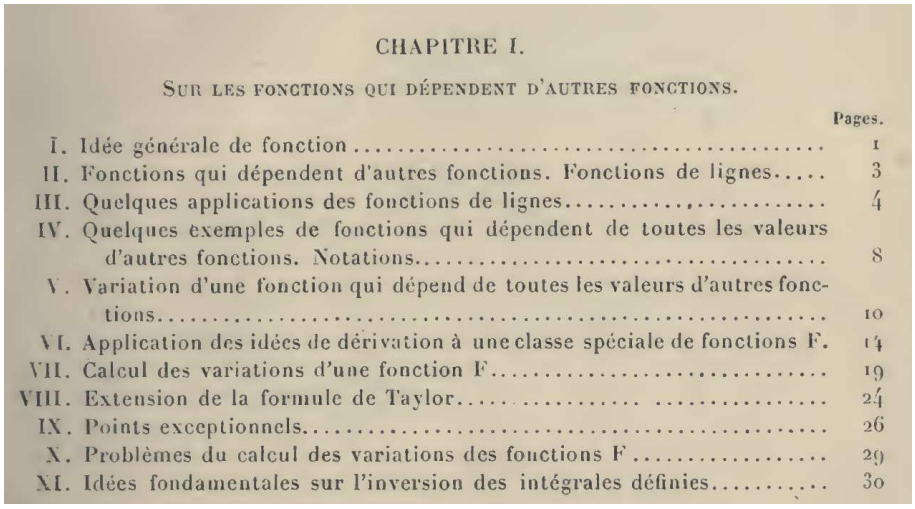
Since there is no such difficulty for x away from 0, this inequality ends the proof of the proposition. □

Remark 2. A similar property holds for the kernel $\mathcal{K}_1^{\theta,\mu}(x,z) = \frac{\mathcal{D}_x^\theta \mathcal{D}_z^\theta k^\mu(x,z)}{\alpha(x)+i\mu}$ which belongs to $L^\infty\left(\mathcal{D}_0 \cap \{x \in [-L,H]\}\right)$ for all θ and uniformly for $\mu \in [-1,1] \setminus \{0\}$, that is

$$\left\| \frac{\mathcal{D}_x^\theta \mathcal{D}_z^\theta k^\mu(x,z)}{\alpha(x) + i\mu} \right\|_{L^\infty(\mathcal{D}_0 \cap \{x \in [-L,H]\})} \le C_\theta, \quad \forall \mu \in [-1,1] \setminus \{0\}. \tag{4.26}$$

Such an estimate is sufficient to control some L^∞ bounds of the series that defines the iterated kernel $\mathcal{K}^{\theta,\mu}(x,z;0)$:

$$\begin{aligned} \left|K_{n+1}^{\theta,\mu}(x,z;0)\right| &= \left| \int_0^x \frac{\mathcal{D}_x^\theta \mathcal{D}_z^\theta k^\mu(x,t)}{\alpha(x)+i\mu} K_n^{\theta,\mu}(t,z;0) dt \right|, \\ &\le C_\theta^{n+1} \underbrace{\int_{0<x_1<\cdots<x_n<x} \prod_{1\le i\le n} dx_i}_{\frac{x^n}{n!}}, \end{aligned}$$



CHAPITRE I.

SUR LES FONCTIONS QUI DÉPENDENT D'AUTRES FONCTIONS.

FIGURE 4.5 – Abstract of the Table of Contents from Volterra's *Leçons sur les équations intégrales et les équations intégro-différentielles*, [Vol10].

so that

$$\left|\mathcal{K}^{\theta,\mu}(x,z;0)\right| = \left|\sum_{n=0}^{\infty} K_{n+1}^{\theta,\mu}(x,z;0)\right| \leq C_\theta \left(e^{C_\theta H} - 1.\right)$$

However, L^∞ bounds are not sufficient to show that $\mathcal{K}^{\theta,\mu}(x,z;0)$ is of Hölder class in z in the vicinity of $x = 0$: that is, one cannot pass to the limit using the Plemelj-Privalov Theorem for all values of the parameters involved in (4.12, 4.17). This is why another approach is developed hereafter to give a meaning to the limit value.

Remark 3. Equation (4.15) is a classic second kind Volterra equation, that reads

$$f(x) = u(x) + \int_0^x K(x,\xi)u(\xi)d\xi.$$

Volterra himself considered continuous kernels K, see his notes from 1910 [Vol10]. He discussed the notion of *functions that depend on other functions* - see Figure 4.5 - and at that time could only use Riemann integrals. But the classical framework for such integral equations described for instance in [Tri85] is to consider L^2 kernels and functions - see Figure 4.6. A further result shows that if f belongs to L^p and K belongs to L^∞, then u belongs to L^p as well. An appropriate version of Gronwall's inequality to be found in [Bee69] indeed shows that

$$|u(x)| \leq |f(x)| + \int_0^x |S(x,\xi)|\ |f(\xi)|d\xi,$$

where S stands for the resolvent kernel of K, then the result yields from Holder's inequality.

4.4 The space $\mathbb{X}^{\theta,\mu}$ $(\mu \neq 0)$

The solutions of the integral equations described in Proposition 4.1 belong to a vectorial space of dimension two : see also (4.29). The aim of this section is to define a basis of this vectorial space. To define the two convenient basis functions, two aspects will be considered :

1·4. L_2-Kernels and Functions

As we shall see, the method of successive approximations can be applied to a large number of integral equations, not only those of the Volterra type. Thus, it is advantageous to study its convergence under conditions for the kernel $K(x,y)$ and the function $f(x)$ which are not too restrictive. Although these conditions will be quite weak, they will be suitable for the application of this method (but not only this method) to Fredholm integral equations.

By using the well-known *Schwarz inequality*†

$$\left[\int_a^b f(x)\,g(x)\,dx\right]^2 \leqslant \int_a^b f^2(x)\,dx \int_a^b g^2(x)\,dx, \tag{1}$$

which will be an important tool in the foregoing theory, we can avoid the customary hypothesis that $K(x,y)$ is continuous (and consequently bounded) by placing it in the L_2-space. Namely, we

shall suppose that the kernel $K(x,y)$ is *quadratically integrable* in the square $(0 \leqslant x \leqslant h,\ 0 \leqslant y \leqslant h)$, where h is a positive constant, i.e. that the integral

$$\| K \|^2 = \int_0^h \int_0^h K^2(x,y)\,dx\,dy \leqslant N^2 \tag{2}$$

exists (at least in the Lebesgue sense†) and is less than a certain constant N^2. Such a kernel will be called an *L_2-kernel* and $\| K \|$ its *norm*.

Similarly, we shall suppose that the given function $f(x)$ of our integral equations is always an L_2-function, i.e. that its norm $\| f \|$, given by

$$\| f \|^2 = \int_0^h f^2(x)\,dx, \tag{3}$$

exists and is finite.

FIGURE 4.6 – Abstract from Tricomi's book *Integral equations*, [Tri85].

- the behavior at infinity $x = \infty$,
- the behavior at the origin $x = 0$.

For the sake of simplicity, the parameter μ is restricted to $0 < \mu \leq 1$ without loss of generality. The extension to negative values of μ will be considered in Section 4.6.4. Define the vectorial space of all solutions of the X-mode equations

$$\mathbb{X}^{\theta,\mu} = \{x \mapsto (U(x), V(x), W(x)), \text{ for all solutions of the system (4.6)}\}. \tag{4.27}$$

One may also use the notation

$$\mathbf{U}^{\theta,\mu} = (U^{\theta,\mu}, V^{\theta,\mu}, W^{\theta,\mu}) \in \mathbb{X}^{\theta,\mu}.$$

The fact that $\dim \mathbb{X}^{\theta,\mu} = 2$ is also evident considering the right hand side of the integral equation (4.15).

Definition 5. Define the matrix $A^{\theta,\mu}$ by

$$A^{\theta,\mu}(x) = \begin{pmatrix} \frac{\theta\gamma(x)}{\alpha(x)+i\mu} & 1 - \frac{\theta^2}{\alpha(x)+i\mu} \\ \frac{\gamma(x)^2}{\alpha(x)+i\mu} - \alpha(x) - i\mu & -\frac{\theta\gamma(x)}{\alpha(x)+i\mu} \end{pmatrix}. \tag{4.28}$$

By elimination of $U^{\theta,\mu}$ in (4.6), one gets a system of two coupled ordinary differential equations

$$\frac{d}{dx}\begin{pmatrix} V^{\theta,\mu} \\ W^{\theta,\mu} \end{pmatrix} = A^{\theta,\mu}(x)\begin{pmatrix} V^{\theta,\mu} \\ W^{\theta,\mu} \end{pmatrix}. \tag{4.29}$$

In the case $\mu \neq 0$ the matrix is non singular for all x, which gives a meaning to the regularized problem. One notices the matrix is singular for $\mu = 0$.

The Wronskian of two solutions will be used to ensure their linear independence.

Lemma 4.4. *Take two solutions* $\left(V^{\theta,\mu}, W^{\theta,\mu}\right)$ *and* $\left(\widetilde{V}^{\theta,\mu}, \widetilde{W}^{\theta,\mu}\right)$ *of (4.29). Define the Wronskian*

$$\mathcal{W}(x) = V^{\theta,\mu}(x)\widetilde{W}^{\theta,\mu}(x) - W^{\theta,\mu}(x)\widetilde{V}^{\theta,\mu}(x). \tag{4.30}$$

Then the Wronskian is constant

$$\mathcal{W}(x) = \mathcal{W}(0), \qquad \forall x.$$

Proof. The system (4.29) main be rewritten as

$$\frac{d}{dx}\begin{pmatrix} V \\ W \end{pmatrix} = \begin{pmatrix} a & b \\ c & -a \end{pmatrix}\begin{pmatrix} V \\ W \end{pmatrix}.$$

Therefore

$$\frac{d}{dx}\mathcal{W} = \frac{d}{dx}\left(V(x)\widetilde{W}(x) - W(x)\widetilde{V}(x)\right)$$

$$= (aV + bW)\,\widetilde{W} + V\left(c\widetilde{V} - a\widetilde{W}\right) - (cV - aW)\,\widetilde{V} - W\left(a\widetilde{V} + b\widetilde{W}\right) = 0$$

since all terms cancel each other. □

4.4.1 Behavior at infinity

Thanks to hypothesis (H3), for $x \geq H$ the model is simplified. In fact, it corresponds to a system as in (4.29) with constant coefficients, which matrix will be denoted $A_\infty^{\theta,\mu}$. As a consequence the solution will be explicitly described for $x \geq H$.

Proposition 4.5. *The matrix $A_\infty^{\theta,\mu}$ has two distinct eigenvalues. The first eigenvalue $\lambda^{\theta,\mu}$ has a positive real part. The second eigenvalue is $-\lambda^{\theta,\mu}$.*

Proof. The eigenvalues are solution to the characteristic equation

$$\lambda^2 - \mathrm{tr}(A_\infty^{\theta,\mu})\lambda + \det(A_\infty^{\theta,\mu}) = 0$$

where $\mathrm{tr}(A_\infty^{\theta,\mu}) = 0$ and

$$\det(A_\infty^{\theta,\mu}) = \alpha_\infty + i\mu - \theta^2 - \frac{\gamma_\infty^2}{\alpha_\infty + i\mu}.$$

The real part is

$$\mathrm{real}\left(\det(A_\infty^{\theta,\mu})\right) = \alpha_\infty - \theta^2 - \frac{\gamma_\infty \alpha_\infty}{\alpha_\infty^2 + \mu^2} = \alpha_\infty \left(1 - \frac{\gamma_\infty^2}{\alpha_\infty^2 + \mu^2}\right) - \theta^2$$

and is therefore negative due to the coercivity assumption (H4). So the usual square root $\lambda^{\theta,\mu} = \sqrt{-\det(A_\infty^{\theta,\mu})}$ has a positive real part. The other one has a negative real part. □

As a consequence any $\mathbf{U} \in \mathbb{X}^{\theta,\mu}$ is at large scale a linear combination of an exponential increasing function and an exponential decreasing function

$$\mathbf{U}(x) = c_+ R_+ e^{\lambda^{\theta,\mu} x} + c_- R_- e^{-\lambda^{\theta,\mu} x} \qquad H \leq x \tag{4.31}$$

where $R_+ \in \mathbb{C}^3$ and $R_- \in \mathbb{C}^3$ are constant vectors and $(c_+, c_-) \in \mathbb{C}^2$ are arbitrary complex numbers. Regarding the structure of the matrix and using the second equation of the system (4.6), one gets that $R_+ = (r_+^1, r_+^2, r_+^3)$ with

$$r_+^1 = \frac{i\theta r_+^3 - i\gamma(H) r_+^2}{\alpha(H) + i\mu},\ r_+^2 = 1 - \frac{\theta^2}{\alpha(H) + i\mu},\ r_+^3 = \sqrt{-\det(A_\infty^{\theta,\mu})} - \frac{\theta\gamma(H)}{\alpha(H) + i\mu}.$$

The other vector $R_- = (r_-^1, r_-^2, r_-^3)$ is characterized by

$$r_-^1 = \frac{i\theta r_-^3 - i\gamma(H) r_-^2}{\alpha(H) + i\mu},\ r_-^2 = 1 - \frac{\theta^2}{\alpha(H) + i\mu},\ r_-^3 = -\sqrt{-\det(A_\infty^{\theta,\mu})} - \frac{\theta\gamma(H)}{\alpha(H) + i\mu}.$$

One notices that R_+ and R_- are well defined for all $\mu \in \mathbb{R}$, in particular even for $\mu = 0$.

4.4.2 The first basis function

The first basis function

$$\mathbf{U}_1^{\theta,\mu} = \left(U_1^{\theta,\mu}, V_1^{\theta,\mu}, W_1^{\theta,\mu}\right) \in \mathbb{X}^{\theta,\mu} \tag{4.32}$$

is the natural one that vanishes at the origin : $U_1^{\theta,\mu}(0) = 0$. For that reason G is chosen to be the origin in this subsection, so that the corresponding integral equation has a bounded right-hand side and a bounded kernel. It is naturally characterized by

$$V_1^{\theta,\mu}(0) = i\theta, \qquad \text{and } W_1^{\theta,\mu}(0) = i\gamma(0) \qquad (\neq 0). \tag{4.33}$$

Proposition 4.6. *The basis function (4.32) is uniformly bounded with respect to μ : for any $\theta \in [\theta_-, \theta_+]$ any interval, and any $H \in]L, \infty[$, there exists a constant C independent of μ such that*

$$\left\|U_1^{\theta,\mu}\right\|_{L^\infty(-L,H)} + \left\|V_1^{\theta,\mu}\right\|_{L^\infty(-L,H)} + \left\|W_1^{\theta,\mu}\right\|_{L^\infty(-L,H)} \leq C. \tag{4.34}$$

Proof. The right hand side in the integral equation (4.15) is

$$g^\mu(x) = \frac{h^\mu(x)}{\alpha(x) + i\mu} \text{ with } h^\mu(x) = i\theta \mathcal{D}_x^\theta A_\mu(x) + i\gamma(0)\mathcal{D}_x^\theta B_\mu(x).$$

With the choice (4.32) one has

$$h^\mu(0) = i\theta(-i\gamma(0)) + i\gamma(0)(i\theta) = 0 \qquad \forall \mu.$$

Therefore the right hand side of the integral equation

$$g^\mu(x) = \frac{h^\mu(x) - h^\mu(0)}{\alpha(x) + i\mu}$$

is bounded in $L^\infty(-L, H))$ uniformly with respect to μ. The solution $U_1^{\theta,\mu}$ (4.17) is also bounded, since from Subsection 4.2 the kernel $\mathcal{K}^{\theta,\mu}(x, z, 0)$ is also uniformly bounded. These bounds are uniform with respect to μ. The integral representation (4.18) of the $V_1^{\theta,\mu}$ yields that $V_1^{\theta,\mu}$ is also bounded. The boundedness of $W_1^{\theta,\mu}$ stems from similar arguments and from the integral representation (4.19). □

Proposition 4.7. *The first basis function (4.32) is exponentially growing at large scale ($\mu \neq 0$).*

Proof. For the sake of simplicity, denote $\mathbf{U}_1^{\theta,\mu} = (U_1, V_1, W_1)$, dropping the θs and μs. Then from system (4.6) one gets

$$\left\{ \begin{array}{cccc} W_1 & +i\theta U_1 & -V_1' & = 0, \\ i\theta W_1 & -(\alpha + i\mu)U_1 & -i\gamma V_1 & = 0, \\ -W_1' & +i\gamma U_1 & -(\alpha + i\mu)V_1 & = 0. \end{array} \right.$$

Multiplying the second equation by $\overline{U_1}$ and the third one by $\overline{V_1}$, the sum reads

$$i\theta W_1\overline{U_1} - W_1'\overline{V_1} - \left(\alpha|U_1|^2 + \alpha|V_1|^2 + i\gamma V_1\overline{U_1} - i\gamma U_1\overline{V_1}\right) - i\mu\left(|U_1|^2 + |V_1|^2\right) = 0.$$

On the other hand an integration over the interval $]M, N[$ yields

$$\begin{aligned} \int_M^N \left(i\theta W_1\overline{U_1} - W_1'\overline{V_1}\right) dx \\ &= \int_M^N \left(i\theta W_1\overline{U_1} + W_1\overline{V_1}'\right) dx - W_1(N)\overline{V_1}(N) + W_1(M)\overline{V_1}(M) \\ &= \int_M^N |W_1|^2 dx - W_1(N)\overline{V_1}(N) + W_1(M)\overline{V_1}(M), \end{aligned}$$

thanks to the first equation of the system. As a result

$$\int_M^N \left(|W_1|^2 - \alpha|U_1|^2 - \alpha|V_1|^2 - i\gamma V_1\overline{U_1} + i\gamma U_1\overline{V_1}\right) dx - i\mu \int_M^N \left(|U_1|^2 + |V_1|^2\right) dx \tag{4.35}$$

$$= W_1(N)\overline{V_1}(N) - W_1(M)\overline{V_1}(M).$$

Splitting between the real and imaginary parts, one gets the important relation

$$\mu \int_M^N \left(|U_1|^2 + |V_1|^2\right) dx = \mathrm{Im}\left(W_1(M)\overline{V_1}(M)\right) - \mathrm{Im}\left(W_1(N)\overline{V_1}(N)\right) \tag{4.36}$$

which holds in fact for any element (U, V, W) in $\mathbb{X}^{\theta,\mu}$ and for any $M < N$.

For $M = 0$ it reads $V_1(0) = \frac{\theta}{\gamma(0)} W_1(0)$ and $\mathrm{Im}\left(W_1(M)\overline{V_1}(M)\right) = 0$. Therefore

$$\mu \int_M^N \left(|U_1|^2 + |V_1|^2\right) dx = -\mathrm{Im}\left(W_1(N)\overline{V_1}(N)\right).$$

It shows that $W_1(N)\overline{V_1}(N) \not\to 0$ for $N \to \infty$. In other words the first basis function does not decrease exponentially at infinity. Considering (4.31) it means that this function is exponentially increasing at infinity. □

4.4.3 The second basis function

The eigenvalue $-\lambda^{\theta,\mu}$ corresponds to the physical behavior of a wave absorbed in the coercive zone $x \geq H$. The second basis function is meant to combine the associated non propagative behavior at infinity with a normalization at the origin that will ensure the linear independence with the first basis function, uniformly with respect to μ.

As a consequence, the second basis function

$$\mathbf{U}_2^{\theta,\mu} = (U_2^{\theta,\mu}, V_2^{\theta,\mu}, W_2^{\theta,\mu}) \in \mathbb{X}^{\theta,\mu}$$

is built with two requirements.

- It is exponentially decreasing at infinity, that is

$$\mathbf{U}_2^{\theta,\mu}(x) = c_- R_- e^{-\lambda^{\theta,\mu} x}, \qquad H \leq x, \tag{4.37}$$

  for some $c_- \in \mathbb{C}$.
- Its value at the origin is normalized with the requirement

$$i\mu U_2^{\theta,\mu}(0) = 1. \tag{4.38}$$

To ensure that these conditions are compatible, consider the third function

$$\mathbf{U}_3^{\theta,\mu} = (U_3^{\theta,\mu}, V_3^{\theta,\mu}, W_3^{\theta,\mu})(x) = R_- e^{-\lambda^{\theta,\mu} x} \qquad H \leq x, \tag{4.39}$$

where R_- and λ_- are defined in Section 4.4.1, smoothly extended so that $\mathbf{U}_3^{\theta,\mu} \in \mathbb{X}^{\theta,\mu}$. The identity

$$\mu \int_M^N \left(|U_3^{\theta,\mu}|^2 + |V_3^{\theta,\mu}|^2\right) dx$$

$$= \mathrm{Im}\left(W_3^{\theta,\mu}(M)\overline{V_3^{\theta,\mu}}(M)\right) - \mathrm{Im}\left(W_3^{\theta,\mu}(N)\overline{V_3^{\theta,\mu}}(N)\right)$$

with $N \to \infty$ and $M = 0$ shows that

$$\mu \int_0^\infty \left(|U_3^{\theta,\mu}|^2 + |V_3^{\theta,\mu}|^2\right) dx = \mathrm{Im}\left(W_3^{\theta,\mu}(0)\overline{V_3^{\theta,\mu}}(0)\right).$$

However, from (4.6), $V_3^{\theta,\mu}(0) = \frac{\theta}{\gamma(0)} W_3(0) - \frac{\mu}{\gamma(0)} U_3^{\theta,\mu}(0)$, so one gets

$$\mu \int_0^\infty \left(|U_3^{\theta,\mu}|^2 + |V_3^{\theta,\mu}|^2 \right) dx = -\frac{\mu}{\gamma(0)} \mathrm{Im} \left(W_3^{\theta,\mu}(0) \overline{U_3^{\theta,\mu}}(0) \right).$$

Since $\gamma(0) \neq 0$, it shows that $U_3^{\theta,\mu}(0) \neq 0$. This is why it is always possible to consider a renormalized function

$$\mathbf{U}_2^{\theta,\mu} = c_- \mathbf{U}_3^{\theta,\mu}, \qquad c_- = \frac{1}{i\mu U_3^{\theta,\mu}(0)} \tag{4.40}$$

so as to enforce (4.38).

Proposition 4.8. *With the normalizations (4.33) and (4.37-4.38), the Wronskian relation reads*

$$V_1^{\theta,\mu}(x) W_2^{\theta,\mu}(x) - W_1^{\theta,\mu}(x) V_2^{\theta,\mu}(x) = 1 \qquad \forall x \geq -L. \tag{4.41}$$

Proof. It is sufficient to compute the Wronskian at the origin

$$\begin{aligned} V_1^{\theta,\mu}(0) W_2^{\theta,\mu}(0) - W_1^{\theta,\mu}(0) V_2^{\theta,\mu}(0) &= i\theta W_2^{\theta,\mu}(0) - i\gamma(0) V_2^{\theta,\mu}(0) \\ &= (\alpha(0) + i\mu) U_2^{\theta,\mu}(0) = i\mu U_2^{\theta,\mu}(0) = 1 \end{aligned}$$

using (4.6) and thanks to (4.38). □

Remark 4. The value of the Wronskian (4.41) is independent of μ. It will be of major interest in the limit regime $\mu \to 0$.

The non zero Wronskian (4.41) shows that the two basis functions are indeed linearly independent. So they span the whole space

$$\mathbb{X}^{\theta,\mu} = \mathrm{Span} \left\{ \mathbf{U}_1^{\theta,\mu}, \mathbf{U}_2^{\theta,\mu} \right\}, \qquad \mu > 0.$$

4.5 Singularity continuity estimates

The integral equation (4.15) is singular at the limit. By comparison with the standard literature [Tri85, Mus92, Vek67b, DL85, Pic11, BW73, Shu97] no immediate convenient mathematical tool to analyze its properties can be found. That is why the following new continuity estimates with respect to the parameters of the problem are developed.

The careful analysis of the singularity that follows will be used in the next section to show that one basis function - more precisely its U component - is the sum of a singular part $\frac{1}{\alpha(x)+i\mu}$ plus a term which is bounded in L^p $(1 \leq p < \infty)$ uniformly with respect to μ.

Consider a general solution $\mathbf{U} = (U, V, W) \in \mathbb{X}^{\theta,\mu}$ of the integral equation (4.15) with prescribed data in H under the form

$$V(H) = a_H \text{ and } W(H) = b_H.$$

Introduce the compact notation

$$\|H\| = |a_H| + |b_H|.$$

The underlying idea is to obtain some sharp continuity estimates on the solution $\mathbf{U}$ with respect to $\|H\|$. The main point is to bound the constants uniformly with respect to $0 < \mu \leq 1$. The reference point can be different from H as well, but not equal to zero. Once these continuity estimates are proved, they will provide enough information to define the limit $\mu \to 0$ of the second basis function.

Proposition 4.9. *There exists a constant C_θ continuously depending on θ such that*

$$|U(x)| \le \frac{C_\theta}{\sqrt{r^2x^2+\mu^2}}\|H\|, \qquad 0 < x \le H. \tag{4.42}$$

Proof. Consider

$$\gamma_\theta = \left(\sup_{0\le\mu\le1}\|A_\mu\|_{W^{1,\infty}(0,H)} + \sup_{0\le\mu\le1}\|B_\mu\|_{W^{1,\infty}(0,H)}\right)(\|\gamma\|_\infty + |\theta|).$$

The integral equation (4.15) with $G = H$ implies that

$$|U(x)| \le \frac{\gamma_\theta\|H\|}{\sqrt{r^2x^2+\mu^2}} + \int_x^H \frac{|\mathcal{D}_x^\theta\mathcal{D}_z^\theta k(x,z)|}{\sqrt{r^2x^2+\mu^2}}|U(z)|dz,$$

where (H2) is used. Since $\mathcal{D}_x^\theta\mathcal{D}_z^\theta k(x,x) = 0$ for all x, there exists a constant β_θ such that

$$\left\|\mathcal{D}_x^\theta\mathcal{D}_z^\theta k(x,z)\right\|_{L^\infty]0,H[} \le \beta_\theta|x-z| \le \beta_\theta z \quad \text{ for } 0 \le x \le z.$$

So

$$\sqrt{r^2x^2+\mu^2}|U(x)| \le \gamma_\theta\|H\| + \beta_\theta\int_x^H z|U(z)|dz$$

and

$$rx|U(x)| \le \gamma_\theta\|H\| + \beta_\theta\int_x^H z|U(z)|dz, \qquad 0 \le x \le H.$$

The Gronwall lemma is useful to study this inequality. Indeed set $g(x) = \int_x^H |zU(z)|dz$, so that the previous inequality reads

$$-rg'(x) \le \gamma_\theta\|H\| + \beta_\theta g(x).$$

Therefore $0 \le \gamma_\theta\|H\| + rg'(x) + \beta_\theta g(x)$, that is

$$0 \le \gamma_\theta\|H\|e^{\frac{\beta_\theta}{r}x} + r\left(e^{\frac{\beta_\theta}{r}x}g(x)\right)'.$$

Next an integration on the interval $[x, H]$ together with the fact that $g(H) = 0$ by definition yields

$$0 \le \gamma_\theta\|H\|\frac{e^{\frac{\beta_\theta}{r}H} - e^{\frac{\beta_\theta}{r}x}}{\frac{\beta_\theta}{r}} - re^{\frac{\beta_\theta}{r}x}g(x),$$

that is

$$g(x) \le \frac{e^{\frac{\beta_\theta}{r}(H-x)} - 1}{\beta_\theta}\,\gamma_\theta\|H\|. \tag{4.43}$$

Finally one checks that

$$\sqrt{r^2x^2+\mu^2}|U(x)| \le \gamma_\theta\|H\| + \beta_\theta g(x) \le e^{\frac{\beta_\theta}{r}(H-x)}\gamma_\theta\|H\|,$$

which proves (4.42). □

4.5.1 A preliminary comment

Next define $\|0\| = |V(0)| + |W(0)|$.

Proposition 4.10. *There exists a constant C_θ with continuous dependence with respect to θ such that*

$$\|0\| \leq C_\theta(1 + |\ln \mu|)\|H\|. \tag{4.44}$$

Proof. The same notation as above are adopted. The integral expression of V (4.18) with $G = H$ yields the inequality

$$|V(0)| \leq \gamma_\theta \|H\| + \left| \int_0^H \mathcal{D}_z^\theta k(0,z).U(0) \right| dz$$

Notice that $\mathcal{D}_z^\theta k(0,z) = \Big(\mathcal{D}_z^\theta k(0,z) - \mathcal{D}_z^\theta k(0,0)\Big) + \mathcal{D}_z^\theta k(0,0)$. Since

$$\mathcal{D}_z^\theta k(0,0) = i\theta \partial_z k(0,0) - i\gamma k(0,0) = i\theta$$

one gets

$$\Big|\mathcal{D}_z^\theta k(0,z) - i\theta\Big| \leq \eta_\theta |z|$$

for some constant $\eta_\theta > 0$. It gives

$$|V(0)| \leq \underbrace{\gamma_\theta \|H\| + \eta_\theta \int_0^H z|U(z)|dz}_{Q} + |\theta| \underbrace{\left| \int_0^H U(z)dz \right|}_{R}.$$

By (4.42) $Q \leq C_\theta \|H\|$, and moreover

$$R := \left| \int_0^H U(z)dz \right| \leq C_\theta \|H\| \left| \int_0^H \frac{1}{r|x| + \mu} dz \right| \leq C_\theta \|H\| |\ln \mu|.$$

This completes the proof for $|V(0)|$. The term $|W(0)|$ is bounded with the same method starting from the integral (4.19) and using the identity $\partial_x \mathcal{D}_z^\theta k_\mu(x,x) = i\gamma(z)A_\mu(z)$. □

An interesting question is the following. Consider the integral equation (4.15) with $G = 0$. That is the starting point of the integral is the singularity. One may wonder if a direct use of the Gronwall lemma may yield valuable estimates, or not. It appears that a pollution with $\log \mu$ terms render the result of little interest.

Firstly consider for simplicity $0 \leq x$. Then (4.15) with $G = 0$ turns into

$$|U(x)| \leq C_\theta \frac{\|0\|}{\sqrt{r^2x^2 + \mu^2}} + C \int_0^x |U(z)|\, dz, \tag{4.45}$$

equation 4.26) providing a bound for the kernel. The constant $C_\theta > 0$ is chosen large enough. Set $h(x) = \int_0^x |U(z)|\, dz$ so that

$$h'(x) \leq C_\theta \frac{\|0\|}{\sqrt{r^2x^2 + \mu^2}} + C_\theta h(x).$$

Since $h(0) = 0$ the Gronwall lemma yields the inequality

$$h(x) \leq C'_\theta \int_0^x \frac{\|0\|}{|z| + |\mu|} dz$$

that is after integration ($0 \leq x \leq H$) $|h(x)| \leq C''_\theta \|0\| (1 + |\ln \mu|)$, for some constant $C''_\theta > 0$ with continuous dependence with respect to θ. Considering the bound (4.44) and the symmetry between $0 < x$ and $x < 0$ in the integral (4.15) (with $G = 0$) one obtains the estimate

$$\left| \int_0^x U(z)dz \right| \leq C'''_\theta \|H\| (1 + |\ln \mu|)^2, \quad -L \leq x \leq H. \tag{4.46}$$

Going back to (4.45) which is easily generalized to $x < 0$, one gets

$$|U(x)| \leq C_\theta \left(\frac{1}{\sqrt{r^2x^2 + \mu^2}} + 1 + |\ln \mu| \right) (1 + |\ln \mu|) \|H\|, \quad -L \leq x \leq H. \tag{4.47}$$

By comparison of (4.42) and (4.47), it is clear that this technique generates spurious terms of order $\log \mu$ for positive x values. It spoils the possibility of having sharp estimates also for negative x values. With this respect, the rest of this section is devoted to the obtaining of various sharp inequalities which are free of such spurious terms.

4.5.2 Identifying the singularity

Define

$$Q(\mathbf{U}) = V_1^{\theta,\mu}(H)W(H) - W_1^{\theta,\mu}(H)V(H). \tag{4.48}$$

This quantity is the Wronskian of the current solution $\mathbf{U}$ against the first basis function. Thanks to Lemma 4.4 it is then independent of the position H of the evaluation point.

Proposition 4.11. *There exists a constant C_θ continuously depending on θ and a continuous function $\mu \mapsto \epsilon(\mu)$ with $\epsilon(0) = 0$ such that*

$$\left| |\mu| \|U\|^2_{L^2(-L,H)} - \left| \frac{\pi Q(\mathbf{U})^2}{\alpha'(0)} \right| \right| \leq C_\theta \epsilon(\mu) \|H\|^2. \tag{4.49}$$

Proof. Remark that the proof is easily adapted for negative μ.
Consider the integral equation (4.15) with $G = 0$. One gets

$$U(x) = \frac{a_0 \mathcal{D}^\theta_x A(x) + b_0 \mathcal{D}^\theta_x B(x)}{\alpha(x) + i\mu} + \int_0^x \frac{\mathcal{D}^\theta_x \mathcal{D}^\theta_z k(x,z)}{\alpha(x) + i\mu} U(z)dz.$$

Here (a_0, b_0) are a priori different from (a_H, b_H). Due to Lemma 4.4 one has

$$Q(\mathbf{U}) = V_1^{\theta,\mu}(0)W(0) - W_1^{\theta,\mu}(0)V(0)$$

and thanks to the normalization of $\mathbf{U}_1$, the second equation of system (4.6) together with the integral representation of U with $G = 0$ then

$$Q(\mathbf{U}) = i\theta W(0) - i\gamma(0)V(0) = i\mu U(0) = a_0 \mathcal{D}^\theta_x A(0) + b_0 \mathcal{D}^\theta_x B(0)$$

So the integral equation reads

$$U(x) = \underbrace{\frac{Q(\mathbf{U})}{\alpha(x) + i\mu}}_{S_1} \tag{4.50}$$

$$+ \underbrace{a_0 \frac{\mathcal{D}^\theta_x A(x) - \mathcal{D}^\theta_x A(0)}{\alpha(x) + i\mu} + b_0 \frac{\mathcal{D}^\theta_x B(x) - \mathcal{D}^\theta_x B(0)}{\alpha(x) + i\mu}}_{S_2} + \underbrace{\int_0^x \frac{\mathcal{D}^\theta_x \mathcal{D}^\theta_z k(x,z)}{\alpha(x) + i\mu} U(z)dz}_{S_3}.$$

- The L^2 norm of the first term S_1 depends upon the value of

$$D_\mu = \int_{-L}^{H} \frac{\mu}{\alpha(x)^2 + \mu^2} dx.$$

The change of variable $x = \mu w$ shows that $D_\mu = \int_{-\frac{L}{\mu}}^{\frac{H}{\mu}} \frac{1}{b_\mu(w)^2+1} dw$ and $b_\mu(w) = \frac{\alpha(\mu w)}{\mu}$. Using the hypothesis (H2) one has that $|b_\mu(w)| \geq rw$, $r > 0$. Since

$$\int_{\mathbb{R}} \frac{dw}{r^2 w^2 + 1} = \frac{\pi}{r} < \infty$$

and the point-wise limit of $b_\mu(w)$ is $\alpha'(0)w$, the Lebesgue dominated convergence theorem states that $\lim_{0^+} D_\mu = \frac{\pi}{|\alpha'(0)|}$. Considering that

$$|Q(\mathbf{U})| \leq C_\theta^1 \|H\| \tag{4.51}$$

using (4.48), there exists a continuous function $\mu \mapsto \epsilon^1(\mu)$ with $\epsilon^1(0) = 0$ such that

$$\left| \mu \|S_1\|^2_{L^2(-L-,H)} - \left| \frac{\pi Q(\mathbf{U})^2}{\alpha'(0)} \right| \right| \leq C_\theta^1 \epsilon^1(\mu) \|H\|^2. \tag{4.52}$$

- The functions $\frac{\mathcal{D}_x^\theta A_\mu(x) - \mathcal{D}_x^\theta A_\mu(0)}{\alpha(x)+i\mu}$ and $\frac{\mathcal{D}_x^\theta B_\mu(x) - \mathcal{D}_x^\theta B_\mu(0)}{\alpha(x)+i\mu}$ can bounded in L^∞ uniformly with respect to μ. So

$$\int_{-L}^{H} |S_2(z)|^2 dz \leq c_\theta^2 \|0\|^2.$$

Estimate (4.44) yields

$$\mu \|S_2\|^2_{L^2(-L-,H)} \leq C_\theta^2 \mu (1 + |\ln \mu|)^2 \|H\|^2$$

for some constant $C_\theta^2 > 0$.

- The last term S_3 is

$$|S_3(x)| = \left| \int_0^x \frac{\mathcal{D}_x^\theta \mathcal{D}_z^\theta k^\mu(x,z)}{\alpha(x) + i\mu} U(z) dz \right| \leq c_3^\theta \left| \int_0^x |U(z)| dz \right|$$

since the kernel is bounded (4.26) with respect to θ and uniformly for $\mu \in [0, 1]$. Inequality (4.46) implies that $|S_3(x)| \leq c_\theta^3 (1 + |\ln \mu|)^2 \|H\|$. Therefore this term is bounded like

$$\mu \|S_3\|^2_{L^2(-L-,H)} \leq c_\theta^4 \mu (1 + |\ln \mu|)^4 \|H\|^2$$

for some constant c_θ^4.

Adding these three inequalities completes the proof. □

This result shows that the singularity will be linked to the quantity $Q(\mathbf{U})$.

4.5.3 Estimate on $(0, H)$

The next step starts by writing the general form of the integral equation (4.15), showing that the various singularities of the equation can be recombined under a more convenient form. This intermediate result is essential to obtain all following results. Indeed the integral equation for U (4.15) choosing $G = 0$ reads

$$(\alpha(x) + i\mu) U(x) = a_0 \mathcal{D}_x^\theta A_\mu(x) + b_0 \mathcal{D}_x^\theta B_\mu(x) + \int_0^x \mathcal{D}_x^\theta \mathcal{D}_z^\theta k^\mu(x,z) U(z) dz.$$

Definition 6. Define the functions $m^{\theta,\mu}$ and $n^{\theta,\mu}$ such as

$$m^{\theta,\mu} = \frac{\mathcal{D}_x^\theta A^\mu(x) - \mathcal{D}_x^\theta A^\mu(0)}{x} \text{ and } n^{\theta,\mu} = \frac{\mathcal{D}_x^\theta B^\mu(x) - \mathcal{D}_x^\theta B^\mu(0)}{x}.$$

Lemma 4.12. *The integral representation of U (4.15) reads*

$$(\alpha(x) + i\mu)U(x) = \tilde{a}m^{\theta,\mu}(x)x + \tilde{b}n^{\theta,\mu}(x)x + Q(\mathbf{U}) - \int_x^H \mathcal{D}_x^\theta \mathcal{D}_z^\theta k^\mu(x,0)U(z)dz \quad (4.53)$$

$$+ \int_0^x \left(\mathcal{D}_x^\theta \mathcal{D}_z^\theta k^\mu(x,z) - \mathcal{D}_x^\theta \mathcal{D}_z^\theta k^\mu(x,0)\right) U(z)dz, \quad \forall x \in [-L,\infty[.$$

Proof. Since by construction $a_0\mathcal{D}_x^\theta A_\mu(0) + b_0\mathcal{D}_x^\theta B_\mu(0) = Q(\mathbf{U})$ one has

$$(\alpha(x) + i\mu)U(x) = a_0\left(\mathcal{D}_x^\theta A_\mu(x) - \mathcal{D}_x^\theta A_\mu(0)\right) + b_0\left(\mathcal{D}_x^\theta B_\mu(x) - \mathcal{D}_x^\theta B_\mu(0)\right)$$

$$+Q(\mathbf{U}) + \int_0^x \mathcal{D}_x^\theta \mathcal{D}_z^\theta k^\mu(x,z)U(z)dz,$$

which has already been written in the previous proof. But one also has due to the integral equation for V (4.18) choosing $G = H$

$$V(0) = a_0 = a_H A_\mu(0) + b_H B_\mu(0) - \int_0^H \mathcal{D}_z^\theta k^\mu(0,z)U(z)dz.$$

Basic manipulations yield

$$a_0 = a_H - \int_0^H \left(\mathcal{D}_z^\theta k^\mu(0,z) - \mathcal{D}_z^\theta k^\mu(0,0)\right) U(z)dz - i\theta\int_0^H U(z)dz$$

because $\mathcal{D}_z^\theta k^\mu(0,0) = i\theta$. Since the function $\mathcal{D}_z^\theta k^\mu$ is continuous, there exists a constant C_4^θ independent of μ such that

$$\left|\mathcal{D}_z^\theta k^\mu(x,z) - \mathcal{D}_z^\theta k^\mu(x,x)\right| \leq C_4^\theta(z-x) \leq C_4^\theta z \quad \text{for } 0 \leq x \leq z \leq H.$$

Therefore the integral

$$\int_0^H \left|\mathcal{D}_z^\theta k^\mu(0,z) - \mathcal{D}_z^\theta k^\mu(0,0)\right| |U(z)|dz \leq C_4^\theta \int_0^H z|U(z)|dz$$

is bounded uniformly with respect to μ thanks to the bound provided (4.42). It is summarized as

$$a_0 = \tilde{a} - i\theta\int_0^H U(z)dz \quad (4.54)$$

where $|\tilde{a}| \leq C_5^\theta\|H\|$ is bounded uniformly with respect to μ. Similarly

$$b_0 = b_H - \int_0^H \partial_x\mathcal{D}_z^\theta k^\mu(0,z)U(z)dz \quad (4.55)$$

and since the function $\partial_x\mathcal{D}_z^\theta k^\mu$ is continuous and $\partial_x\mathcal{D}_z^\theta(0,0) = i\gamma(0)$

$$b_0 = \tilde{b} - i\int_0^H \gamma(0)U^{\theta,\mu}(z)dz \quad (4.56)$$

where $\tilde{b}$ is also bounded uniformly with respect to μ : $|\tilde{b}| \leq C_6^\theta \|H\|$. The integral equation then gives

$$(\alpha(x)+i\mu)U(x) = \tilde{a}\left(\mathcal{D}_x^\theta A_\mu(x) - \mathcal{D}_x^\theta A_\mu(0)\right) + \tilde{b}\left(\mathcal{D}_x^\theta B_\mu(x) - \mathcal{D}_x^\theta B_\mu(0)\right)$$

$$+Q(\mathbf{U}) - \int_0^H Q(x,z)U(z)dz + \int_0^x \mathcal{D}_x^\theta \mathcal{D}_z^\theta k^\mu(x,z)U(z)dz$$

where the new kernel is

$$Q(x,z) = \left(\mathcal{D}_x^\theta A_\mu(x) - \mathcal{D}_x^\theta A_\mu(0)\right) i\theta + \left(\mathcal{D}_x^\theta B_\mu(x) - \mathcal{D}_x^\theta B_\mu(0)\right) i\gamma(0)$$

$$= \mathcal{D}_x^\theta A^\mu(x)\mathcal{D}_z^\theta B^\mu(0) - \mathcal{D}_x^\theta B^\mu(x)\mathcal{D}_z^\theta A^\mu(0) = \mathcal{D}_x^\theta \mathcal{D}_z^\theta k^\mu(x,0)$$

after evident simplifications. It ends the proof. □

A first property which shows that (4.53) is less singular that its initial form (4.15) is the following lemma which uses the point wise estimate (4.42) on U (so an important restriction is nevertheless that $x > 0$).

Lemma 4.13. *The first component U of any element $\mathbf{U} \in \mathbb{X}^{\theta,\mu}$ satisfies $\forall x > 0$*

$$(\alpha(x)+i\mu)U(x) = p^{\theta,\mu}(x)x + Q(\mathbf{U}) - \int_x^H \mathcal{D}_x^\theta \mathcal{D}_z^\theta k^\mu(x,0)U(z)dz \tag{4.57}$$

where

$$\|p^{\theta,\mu}\|_{L^\infty(0,H)} \leq C^\theta \|H\|, \qquad \forall \mu \in [0,1]. \tag{4.58}$$

Indeed, the limit $\mu \to 0$ of the term $p^{\theta,\mu}(x)x/(\alpha(x)+i\mu)$ is regular, so that the only singularity remaining in the expression (4.57) is the one coming from $Q(\mathbf{U})$ which actually is the real singularity as suggested by Proposition 4.11.

Proof. Focus on the second integral in (4.53). Continuity properties with respect to the second variable z imply that there exists a constant C_7^θ independent of μ such that

$$\left|\mathcal{D}_x^\theta \mathcal{D}_z^\theta k^\mu(x,z) - \mathcal{D}_x^\theta \mathcal{D}_z^\theta k^\mu(x,0)\right| \leq C_7^\theta z. \tag{4.59}$$

So, for $x \geq 0$,

$$\left|\int_0^x \left(\mathcal{D}_x^\theta \mathcal{D}_z^\theta k^\mu(x,z) - \mathcal{D}_x^\theta \mathcal{D}_z^\theta k^\mu(x,0)\right) U(z)dz\right| \leq C_7^\theta \int_0^x z|U(z)|dz \leq C_7^\theta C_\theta \|H\| x$$

using estimate (4.42). Set

$$p^{\theta,\mu}(x) = \tilde{a}m^{\theta,\mu}(x) + \tilde{b}n^{\theta,\mu}(x) + \frac{1}{x}\int_0^x \left(\mathcal{D}_x^\theta \mathcal{D}_z^\theta k^\mu(x,z) - \mathcal{D}_x^\theta \mathcal{D}_z^\theta k^\mu(x,0)\right) U(z)dz \tag{4.60}$$

which satisfies by construction (4.58). □

As a consequence one has the following estimate on $(0,H)$.

Proposition 4.14. *For all $1 \leq p < \infty$, there exists a constant C_p^θ independent of μ and which depends continuously on θ such that*

$$\left\|U - \frac{Q(\mathbf{U})}{\alpha(\cdot)+i\mu}\right\|_{L^p(0,H)} \leq C_p^\theta \|H\|. \tag{4.61}$$

Proof. From Lemma 4.13 one has that

$$U(x) - \frac{Q(\mathbf{U})}{\alpha(x)+i\mu} = \frac{x}{\alpha(x)+i\mu} p^{\theta,\mu}(x) - \frac{\mathcal{D}_x^\theta \mathcal{D}_z^\theta k^\mu(x,0)}{\alpha(x)+i\mu} \int_x^H U(z)dz,$$

which turns into

$$\left(U(x) - \frac{Q(\mathbf{U})}{\alpha(x)+i\mu}\right) + \frac{\mathcal{D}_x^\theta \mathcal{D}_z^\theta k^\mu(x,0)}{\alpha(x)+i\mu} \int_x^H \left(U(z) - \frac{Q(\mathbf{U})}{\alpha(z)+i\mu}\right) dz$$

$$= \frac{x}{\alpha(x)+i\mu} p^{\theta,\mu}(x) - Q(\mathbf{U}) \frac{\mathcal{D}_x^\theta \mathcal{D}_z^\theta k^\mu(x,0)}{\alpha(x)+i\mu} \int_x^H \frac{1}{\alpha(z)+i\mu} dz \tag{4.62}$$

By virtue of (H2) notice that

$$\left| \int_x^H \frac{1}{\alpha(z)+i\mu} dz \right| \leq \int_x^H \frac{1}{|\alpha(z)|} dz \leq \frac{1}{r} \log(H/x).$$

Since all powers of the function $x \mapsto \ln|x|$ are integrable, the right-hand side (4.62) is naturally bounded in any L^p, $1 \leq p < \infty$. Therefore the function $Z(x) = U(x) - \frac{Q(\mathbf{U})}{\alpha(x)+i\mu}$ is solution of an integral equation with a bounded kernel and a right hand side in L^p. The form of this integral equation is

$$Z(x) + \widetilde{K}^{\theta,\mu}(x) \int_x^H Z(z)dz = b^{\theta,\mu}(x)$$

with $\left\|\widetilde{K}^{\theta,\mu}(x)\right\|_{L^\infty(0,H)} \leq C_8^\theta$ independently of μ. One also uses $\left\|b^{\theta,\mu}\right\|_{L^p(0,H)} \leq c_p^\theta \|H\|$ for $0 \leq \mu \leq 1$: the key estimate is (4.58) which explains why the result is restricted to $x > 0$. See Remark 3. Since this is a standard non-singular integral equation, see [Tri85], the claim is proved. □

4.5.4 Estimate on $(-L, H)$

The last result (4.61) shows that some singularities of the integral equation can be blended in a less singular formulation, so that the leading part of U is $\frac{1}{\alpha(\cdot)+i\mu}$. An important restriction of this technique, for the moment, is the need for a priori estimate (4.42) on U. This explains why inequality (4.61) is restricted to $x > 0$. The goal of this section is to extend the range of the estimates to the entire interval $(-L, H)$.

By inspection of the structure of the algebra, it appears that one has the same kind of inequalities on the entire interval by replacing U directly by the function $\frac{1}{\alpha(\cdot)+i\mu}$ in the integrals. A preliminary and fundamental result in this direction concerns the function

$$D^{\theta,\mu}(x) = -\frac{\mathcal{D}_x^\theta \mathcal{D}_z^\theta k^\mu(x,0)}{\alpha(x)+i\mu} \int_0^H \frac{1}{\alpha(z)+i\mu} dz + \int_0^x \frac{\mathcal{D}_x^\theta \mathcal{D}_z^\theta k^\mu(x,z)}{\alpha(x)+i\mu} \frac{1}{\alpha(z)+i\mu} dz$$

which is nothing more than the integral part of (4.53), U being replaced by the function $\frac{1}{\alpha(\cdot)+i\mu}$.

Proposition 4.15. *Let $1 \leq p < \infty$. One has $\left\|D^{\theta,\mu}\right\|_{L^p(-L,H)} \leq C_p^\theta$ where the constant depends continuously on θ and does not depend on μ.*

Proof. Two cases occur.

- **Assume** $0 \leq x \leq H$. The analysis is similar to the one of Proposition 4.14. One has the same kind of rearrangement (4.53), that is

$$D^{\theta,\mu}(x) = -\frac{\mathcal{D}_x^\theta \mathcal{D}_z^\theta k^\mu(x,0)}{\alpha(x)+i\mu}\int_x^H \frac{1}{\alpha(z)+i\mu}dz$$

$$+\int_0^x \frac{\mathcal{D}_x^\theta \mathcal{D}_z^\theta k^\mu(x,z) - \mathcal{D}_x^\theta \mathcal{D}_z^\theta k^\mu(x,0)}{\alpha(x)+i\mu}\frac{1}{\alpha(z)+i\mu}dz.$$

  The first term is bounded like $C^\theta \frac{|\log x|}{r}$ which is in all L^p, $p<\infty$. The second term is immediately bounded using (4.59) : indeed

$$\left|\int_0^x \frac{\mathcal{D}_x^\theta \mathcal{D}_z^\theta k^\mu(x,z) - \mathcal{D}_x^\theta \mathcal{D}_z^\theta k^\mu(x,0)}{\alpha(x)+i\mu}\frac{1}{\alpha(z)+i\mu}dz\right|$$

$$\leq C_7^\theta \frac{1}{\sqrt{\alpha(x)^2+\mu^2}}\int_0^x \frac{z}{\sqrt{\alpha(z)^2+\mu^2}}dz \leq C_7^\theta \frac{1}{r^2}.$$

- **Assume** $-L \leq x \leq 0$. The decomposition is slightly different and uses some cancellations permitted by the symmetry properties of the kernels. One has

$$D^{\theta,\mu}(x) = -\frac{\mathcal{D}_x^\theta \mathcal{D}_z^\theta k^\mu(x,0)}{\alpha(x)+i\mu}\int_{-x}^H \frac{1}{\alpha(z)+i\mu}dz$$

$$+\int_0^x \frac{\mathcal{D}_x^\theta \mathcal{D}_z^\theta k^\mu(x,z)}{\alpha(x)+i\mu}\frac{1}{\alpha(z)+i\mu}dz - \frac{\mathcal{D}_x^\theta \mathcal{D}_z^\theta k^\mu(x,0)}{\alpha(x)+i\mu}\int_0^{-x}\frac{1}{\alpha(z)+i\mu}dz,$$

  which emphasizes the importance of some symmetry properties of the kernels. Indeed

$$\int_0^{-x}\frac{1}{\alpha(z)+i\mu}dz = -\int_0^x \frac{1}{\alpha(-w)+i\mu}dw$$

$$= \int_0^x \frac{1}{\alpha(w)+i\mu}dw + \int_0^x \left(\frac{1}{-\alpha(-w)-i\mu} - \frac{1}{\alpha(w)+i\mu}\right)dw.$$

  Notice that

$$\frac{1}{-\alpha(-w)-i\mu} - \frac{1}{\alpha(w)+i\mu} = \frac{\alpha(w)+\alpha(-w)+2i\mu}{(\alpha(w)+i\mu)(-\alpha(-w)-i\mu)}.$$

  So, since $\alpha(0)=0$,

$$\left|\frac{1}{-\alpha(-w)-i\mu} - \frac{1}{\alpha(w)+i\mu}\right| \leq \frac{2\left\|\alpha\right\|_{W^{2,\infty}(-L,H)} w^2 + 2\mu}{r^2w^2+\mu^2},$$

  because $\alpha \in W^{2,\infty}(-L,H)$. One can bound

$$\left|\int_0^x \frac{1}{-\alpha(-w)-i\mu}dw - \int_0^x \frac{1}{\alpha(w)+i\mu}dw\right|$$

$$\leq \frac{\left\|\alpha\right\|_{W^{2,\infty}(-L,H)}}{r^2}|x| + \int_0^x \frac{2\mu}{r^2z^2+\mu^2}dz$$

$$\leq \frac{\left\|\alpha\right\|_{W^{2,\infty}(-L,H)}}{r^2}\max(L,H) + \int_0^\infty \frac{2\mu}{r^2z^2+\mu^2}dz \leq \frac{\left\|\alpha\right\|_{W^{2,\infty}(-L,H)}}{r^2}\max(L,H) + \frac{\pi}{r}.$$

As a consequence $D^{\theta,\mu}$ can be expressed as

$$\begin{aligned} D^{\theta,\mu}(x) &= -\frac{\mathcal{D}_x^\theta \mathcal{D}_z^\theta k^\mu(x,0)}{\alpha(x)+i\mu}\int_{-x}^{H}\frac{1}{\alpha(z)+i\mu}dz \\ &+\int_0^x \frac{\mathcal{D}_x^\theta \mathcal{D}_z^\theta k^\mu(x,z)-\mathcal{D}_x^\theta \mathcal{D}_z^\theta k^\mu(x,0)}{\alpha(x)+i\mu}\frac{1}{\alpha(z)+i\mu}dz + R(x) \end{aligned}$$

with $\|R\|_\infty(-L,H) \le C_{10}^\theta$. The two integrals have the same structure as for the first case. So the same result holds.

□

Proposition 4.16. *For all $1 \le p < \infty$, there exists a constant C_p^θ independent of μ such that*

$$\left\| U - \frac{Q(\mathbf{U})}{\alpha(\cdot)+i\mu}\right\|_{L^p(-L,H)} \le C_p^\theta \|H\|. \tag{4.63}$$

Proof. Equation (4.53) reads

$$U(x) = \frac{Q(\mathbf{U})}{\alpha(x)+i\mu} + \frac{x}{\alpha(x)+i\mu}\widetilde{p}^{\theta,\mu}(x)$$

$$-\int_0^H \frac{\mathcal{D}_x^\theta \mathcal{D}_z^\theta k^\mu(x,0)}{\alpha(x)+i\mu}U(z)dz + \int_0^x \frac{\mathcal{D}_x^\theta \mathcal{D}_z^\theta k^\mu(x,z)}{\alpha(x)+i\mu}U(z)dz.$$

Here $\widetilde{p}^{\theta,\mu}(x) = \tilde{a}m^{\theta,\mu}(x) + \tilde{b}n^{\theta,\mu}(x)$, so that $\|\widetilde{p}^{\theta,\mu}\|_{L^\infty(-L,H)} \le C^\theta\|H\|$ over the whole interval $(-L,H)$. Notice that $\widetilde{p}^{\theta,\mu}$ is the first part of $p^{\theta,\mu}$ defined in (4.60). Setting $u(x) = U(x) - \frac{Q(\mathbf{U})}{\alpha(x)+i\mu}$ one gets

$$u(x) - \int_0^x \frac{\mathcal{D}_x^\theta \mathcal{D}_z^\theta k^\mu(x,z)}{\alpha(x)+i\mu}u(z)dz$$

$$= \frac{x}{\alpha(x)+i\mu}\widetilde{p}^{\theta,\mu}(x) - Q(\mathbf{U})D^{\theta,\mu}(x) - \int_0^H \frac{\mathcal{D}_x^\theta \mathcal{D}_z^\theta k^\mu(x,0)}{\alpha(x)+i\mu}u(z)dz.$$

The left-hand side is a non singular integral operator of the second kind with and with a bounded kernel thanks to the fundamental property (4.26). The right-hand side is bounded in L^p with a continuous dependence with respect to $\|H\|$, see Lemma 4.13, estimation (4.51) and estimation (4.61).

□

4.6 Passing to the limit $\mu \to 0$

An important result is that the first basis function admits a limit which is defined as a continuous function in $\mathcal{C}^0[-L,\infty[$ and is independent of the sign of μ. On the other hand the second basis function admits a limit which is singular at $x = 0$. Moreover the limit is different for $\mu \to 0^+$ and for $\mu \to 0^-$. The linear independence of these limits will be establish with a transversality condition.

4.6.1 The first basis function

There is no difficulty for this case which is easily treated passing to the limit in the integral equation (4.17), choosing $G = 0$. The limit basis function is referred to as

$$\mathbf{U}_1^\theta = (U_1^\theta, V_1^\theta, W_1^\theta)$$

$\mathbf{U}_1^\theta$ is and will be called the regular solution by analogy with the terminology in scattering on the half-line. It is defined as the solution of a limit version of (4.15), the V and W component being defined by limit versions of (4.18) and (4.19) :

$$\begin{cases} U_1^\theta(x) - \displaystyle\int_0^x \bar{K}^\theta(x,z)U_1^\theta(z)dz = \bar{F}^\theta(x), \\ V_1^\theta(x) = i\theta A(x) + i\gamma(0)B(x) + \displaystyle\int_0^x \mathcal{D}_z^\theta k(x,z)U_1^\theta(z)dz, \\ W_1^\theta(x) = i\theta A'(x) + i\gamma(0)B'(x) + \displaystyle\int_0^x \partial_x\mathcal{D}_z^\theta k(x,z)U_1^\theta(z)dz, \end{cases}$$

where

$$\bar{K}^\theta(x,z) = \begin{cases} \dfrac{\mathcal{D}_x^\theta\mathcal{D}_z^\theta k(x,z)}{\alpha(x)} & \forall x \neq 0 \text{ and } 0 \leq z \leq x \text{ or } x \leq z \leq 0, \\ 0 & \text{in all other cases,} \end{cases}$$

is the limit kernel described in Proposition 4.3 and

$$\bar{F}^\theta(x) = \begin{cases} \dfrac{i\theta\mathcal{D}_x^\theta A(x) + i\gamma(0)\mathcal{D}_x^\theta B(x)}{\alpha(x)} & \forall x \neq 0, \\ \dfrac{\left(i\theta\mathcal{D}_x^\theta A + i\gamma(0)\mathcal{D}_x^\theta B\right)'(0)}{\alpha'(0)} & \text{otherwise.} \end{cases}$$

The right hand side $\bar{F}^\theta$ together with the kernel $\bar{K}^\theta$ considered in the integration domain are continuous, because $\mathcal{D}_x^\theta A(0) = -i\gamma(0)$, $\mathcal{D}_x^\theta B(0) = i\theta$ and see Proposition 4.3.

A preliminary pointwise convergence will be used to obtain an L^p convergence result.

Lemma 4.17. *There is pointwise convergence of the first component*

$$\left\| \left(U_1^{\theta,\mu}(x) - \frac{F^{\theta,\mu}(x)}{\alpha(x)+i\mu} \right) - \left(U_1^\theta - \bar{F}^\theta \right)(x) \right\|_{L^\infty(]-L,H[)} \to 0$$

which yields $\left\| U_1^{\theta,\mu} - U_1^\theta \right\|_{L^\infty_{\mathrm{loc}}(]-L,0[\cup]0,H[)} \to 0$.

As a result the other components satisfy

$$\left\| V_1^{\theta,\mu} - V_1^\theta \right\|_{L^\infty(]-L,H[)} \to 0, \text{ and } \left\| W_1^{\theta,\mu} - W_1^\theta \right\|_{L^\infty(]-L,H[)} \to 0.$$

Proof. **Convergence away from zero**

From the integral equations satisfied by $U_1^{\theta,\mu}$ and U_1^θ one has for all $x \in (-L,\infty)$ and all $\mu \neq 0$ the following integral equation on $U_1^{\theta,\mu} - U_1^\theta$:

$$\begin{aligned} &\left(U_1^{\theta,\mu} - U_1^\theta \right)(x) - \int_0^x \frac{\mathcal{D}_x^\theta\mathcal{D}_z^\theta k^\mu(x,z)}{\alpha(x)+i\nu} \left(U_1^{\theta,\mu} - U_1^\theta \right)(z)dz \\ &\qquad = \underbrace{\frac{F^{\theta,\mu}(x)}{\alpha(x)+i\mu} - \bar{F}^\theta(x)}_{T_1} + \int_0^x \left(\underbrace{\frac{\mathcal{D}_x^\theta\mathcal{D}_z^\theta k^\mu(x,z)}{\alpha(x)+i\mu} - \bar{K}(x,z)}_{T_2} \right) U_1^\theta(z)dz. \end{aligned} \tag{4.64}$$

Since the kernel of equation (4.64) is bounded, the resolvent kernel $\mathcal{K}^{\theta,\mu}$ is bounded, see Remark 2.

Denote $\mathcal{F}_\mu$ the right hand side of equation (4.64). Since $F^{\theta,\mu}(0) = 0$ and $\mathcal{D}_x^\theta\mathcal{D}_z^\theta k^\mu(0,0) = 0$, then $\mathcal{F}_\mu$ is bounded on $]-L,H[$.

The T_1 term converges pointwise to 0 at any $x \neq 0$ thanks to the definition of $\bar{F}^\theta$. Since T_2 pointwise converges to 0 and because it is bounded as indicated in Remark 2, the dominated convergence theorem shows that the integral term in $\mathcal{F}_\mu$ pointwise converges to 0 as long as $x \neq 0$ - note that it is obviously true for $x = 0$. Thus $\mathcal{F}_\mu$ pointwise converges to 0 as long as $x \neq 0$.

As a result, the dominated convergence theorem shows that

$$\left|U_1^{\theta,\mu}(x) - U_1^\theta(x)\right| \leq |\mathcal{F}_\mu(x)| + \left\|\mathcal{K}^{\theta,\mu}(x,z)\right\|_{L^\infty(\mathcal{D}_0 \cap \{x \in]-L,H[\})} \int_0^x |\mathcal{F}_\mu(z)|\, dz$$

pointwise converges to zero as long as $x \neq 0$ as well.

Note that at $x = 0$, (4.64) reads $U_1^{\theta,\mu}(0) - U_1^\theta(0) = \frac{F^{\theta,\mu}(0)}{i\mu} - \bar{F}^\theta(0) = -\bar{F}^\theta(0)$. Then, if $\bar{F}^\theta(0) = \gamma'(0)\theta + \gamma(0)^2 \neq 0$ the pointwise convergence of $U_1^{\theta,\mu} - U_1^\theta$ at $x = 0$ does not hold. Indeed, the term $\bar{F}^\theta(0)$ does not depend on μ. However, if $\gamma'(0)\theta + \gamma(0)^2 = 0$ we have pointwise convergence at $x = 0$ since in this case $U_1^{\theta,\mu}(0) - U_1^\theta(0) = 0$ for all μ.

Convergence on $]-L, H[$

Despite the last remark, a convergence in $L^\infty(]-L,H[)$ can be obtained subtracting the appropriate quantities to the first component and its limit. By (4.64)

$$\left|\left(\left(U_1^{\theta,\mu} - \frac{F^{\theta,\mu}(x)}{\alpha(x) + i\mu}\right) - \left(U_1^\theta\right)(x) - \bar{F}^\theta(x)\right)\right|$$

$$\leq \int_0^x \left|\frac{\mathcal{D}_x^\theta \mathcal{D}_z^\theta k^\mu(x,z)}{\alpha(x) + i\mu}\right| \left|U_1^{\theta,\mu} - U_1^\theta\right|(z) dz + \int_0^x \left|\frac{\mathcal{D}_x^\theta \mathcal{D}_z^\theta k^\mu(x,z)}{\alpha(x) + i\mu} - \bar{K}(x,z)\right| \left|U_1^\theta(z)\right| dz$$

Then, by the dominated convergence theorem, the function $\left(U_1^{\theta,\mu} - \frac{F^{\theta,\mu}}{\alpha + i\mu}\right) - \left(U_1^\theta - \bar{F}^\theta\right)$ converges to zero in $L^\infty(]-L,H[)$.

The convergence of $V_1^{\theta,\mu}$ and $W_1^{\theta,\mu}$ then stems from the dominated convergence theorem again. Indeed, since

$$\begin{cases} V_1^{\theta,\mu}(x) - V_1^\theta(x) & = i\theta(A_\mu - A)(x) + i\gamma(0)(B_\mu - B)(x) \\ & \quad + \int_0^x \left(\mathcal{D}_z^\theta k^\mu(x,z) U_1^{\theta,\mu}(z) - \mathcal{D}_z^\theta k(x,z) U_1^\theta(z)\right) dz, \\ W_1^{\theta,\mu}(x) - W_1^\theta(x) & = i\theta(A_\mu - A)'(x) + i\gamma(0)(B_\mu - B)'(x) \\ & \quad + \int_0^x \left(\partial_x \mathcal{D}_z^\theta k^\mu(x,z) U_1^{\theta,\mu} - \partial_x \mathcal{D}_z^\theta k(x,z) U_1^\theta(z)\right) dz, \end{cases}$$

the L^∞ convergence of both terms $\mathcal{D}_z^\theta k^\mu(x,z) U_1^{\theta,\mu}(z) - \mathcal{D}_z^\theta k(x,z) U_1^\theta(z)$ and $\partial_x \mathcal{D}_z^\theta k^\mu(x,z) U_1^{\theta,\mu} - \partial_x \mathcal{D}_z^\theta k(x,z) U_1^\theta(z)$ on $]-L,0[$ and $]0,H[$ ensures that the hypothesis of the dominated convergence theorem are satisfied. The convergence then holds on $]-L,H[$ since at $x = 0$ it is guaranteed by the convergence of A_μ and B_μ. □

Proposition 4.18. *The first basis functions satisfies for all* $p \in \mathbb{N}^*$

$$\left\|\mathbf{U}_1^{\theta,\mu} - \mathbf{U}_1^\theta\right\|_{L^p(-L,H)} \to 0.$$

Proof. The L^1 convergence is a consequence of the pointwise convergence obtained in Lemma 4.17 thanks to the dominated convergence theorem. Moreover Proposition 4.6 yields an L^∞ bound for $\mathbf{U}_1^{\theta,\mu} - \mathbf{U}_1^\theta$. The result is thus straightforward. □

The next result establishes that $\mathbf{U}_1^\theta$ is still - as its regularized approximation - exponentially increasing at infinity with a technical condition.

Proposition 4.19. *Assume hypothesis* (H5). *Then* $\mathbf{U}_1^{\theta=0}$ *increases exponentially at infinity.*

Remark 5. The constant 4 in the condition (H5) is probably non optimal.

Proof. Drop the super-index $\cdot^{\theta=0}$ to simplify : that is (U_1, V_1, W_1) stands for (U_1^0, V_1^0, W_1^0). Consider the identity (4.35) which holds true at the limit $\mu = 0$

$$\int_0^N \left(|W_1|^2 - \alpha|U_1|^2 - \alpha|V_1|^2 - i\gamma V_1\overline{U_1} + i\gamma U_1\overline{V_1}\right) dx$$

$$= W_1(N)\overline{V_1}(N) - W_1(0)\overline{V_1}(0), \quad 0 < N < \infty.$$

Since $\theta = 0$, then $V_1(0) = 0$. Notice also that $W_1 = V_1'$, so the relation reads

$$\int_0^N \left(|V_1'|^2 - \alpha|U_1|^2 - \alpha|V_1|^2 - i\gamma V_1\overline{U_1} + i\gamma U_1\overline{V_1}\right) dx = W_1(N)\overline{V_1}(N).$$

Proceed by contradiction : assume that the function is exponentially decreasing at infinity. It yields

$$\int_0^\infty \left(|V_1'|^2 - \alpha|U_1|^2 - \alpha|V_1|^2 - i\gamma V_1\overline{U_1} + i\gamma U_1\overline{V_1}\right) dx = 0.$$

Notice that $-\alpha|U_1|^2 - \alpha|V_1|^2 - i\gamma V_1\overline{U_1} + i\gamma U_1\overline{V_1} \geq 0$ for $x \geq H$ due to the coercivity property (H4). Therefore it implies that

$$\int_0^H \left(|V_1'|^2 - \alpha|U_1|^2 - \alpha|V_1|^2 - i\gamma V_1\overline{U_1} + i\gamma U_1\overline{V_1}\right) dx \leq 0.$$

Next observe that $U_1 = -i\frac{\gamma}{\alpha}V_1$, so that

$$\int_0^H \left(|V_1'|^2 + \frac{\gamma^2}{\alpha}|V_1|^2 - \alpha|V_1|^2\right) dx \leq 0.$$

Since $V_1(0) = 0$ and $\alpha(x) \approx \alpha'(0)x$ with $\alpha'(0) < 0$ (see hypothesis H1), it is convenient to notice the proximity with the famous Hardy inequality that we recall,

$$\int_0^H \frac{u(x)^2}{x^2} < 4\int_0^H u'(x)^2, \qquad u \in H^1(0,H), \quad u(0) = 0, \quad u \neq 0.$$

Since, thanks to hypothesis (H2),

$$\int_0^H \frac{\gamma^2}{|\alpha|}|V_1|^2 = \int_0^H \gamma^2 x \frac{x}{|\alpha|}\frac{|V_1|^2}{x^2} \leq \frac{\|\gamma\|_\infty^2 H}{r}\int_0^H \frac{|V_1|^2}{x^2},$$

it yields the inequality

$$0 \leq \left(1 - 4\frac{\|\gamma\|_\infty^2 H}{r}\right)\int_0^H |V_1'|^2 dx \leq \int_0^H \left(|V_1'|^2 + \frac{\gamma}{\alpha}|V_1|^2 - \alpha|V_1|^2\right) dx \leq 0,$$

where we used (H5). Therefore V_1 vanishes on the interval $[0, H]$. So at that point U_1 vanishes and W_1 also vanishes on the interval which is not compatible with $W_1(0) = i\gamma(0) \neq 0$. □

Proposition 4.20. *Assume hypothesis* (H5). *Then* $\mathbf{U}_1^\theta$ *increases exponentially at infinity.*

Denote by $(U_3^\theta(H), V_3^\theta(H), W_3^\theta(H))$ the solution to (4.6) for $x > 0$ that satisfies (4.39) with $\mu = 0$. *Proof.* Consider the function

$$\eta(\theta) = V_1^\theta(H) W_3^\theta(H) - W_1^\theta(H) V_3^\theta(H) \tag{4.65}$$

By definition

$$(V_3(H), W_3(H)) = \left(1 - \frac{\theta^2}{\alpha_\infty}, -\frac{\theta\gamma_\infty}{\alpha_\infty} - \sqrt{-\alpha_\infty + \theta^2 + \frac{\gamma_\infty^2}{\alpha_\infty}}\right) e^{-\sqrt{-\det A_\infty^{\theta,\mu=0}} H}.$$

This vector is real and always non zero. Therefore the function $\theta \mapsto f(\theta)$ is well defined. This function naturally satisfies two properties :

- $\eta(0) \neq 0$ since (V_1^0, W_1^0) is exponentially increasing by virtue of the previous property. Indeed $\eta(0) = 0$ if and only if the functions $x \mapsto (V_1^0(x), W_1^0(x))$ and $x \mapsto (V_3^0(x), W_3^0(x))$ are linearly dependent, which is not true.
- the function η is continuous since the first basis function is continuous wit respect to θ.

Therefore there exists an interval around 0 in which $\eta(\theta)$ is non zero, which in turn yields the fact that $\mathbf{U}_1^\theta$ is linearly independent of $\mathbf{U}_3^\theta$. Therefore $\mathbf{U}_1^\theta$ is exponentially increasing. □

4.6.2 The transversality condition

The transversality condition appears to be a sufficient condition of linear independence for the limits of the two basis functions. In Section 4.9 some cases where this linear independence is not true are studied.

Passing to the limit in the second basis function near the origin is more complicated. Indeed the limit U_2^θ is expected to satisfy $U_2^\theta \approx \frac{C}{x}$ for some local constant C. Therefore the limit is singular and special care has to be provided to avoid any artifacts in the analysis.

Define the special Wronskian between the first and third basis functions

$$\eta(\theta, \mu) = V_1^{\theta,\mu}(H) W_3^{\theta,\mu}(H) - W_1^{\theta,\mu}(H) V_3^{\theta,\mu}(H).$$

It is the natural continuous extension with respect to μ of the function $\theta \mapsto \eta(\theta)$. Then (4.40) reads

$$\mathbf{U}_2^{\theta,\mu} = \xi^{\theta,\mu} \mathbf{U}_3^{\theta,\mu}.$$

Plugging this relation in the Wronskian (4.41) one gets that

$$1 = \xi^{\theta,\mu} \eta(\theta, \mu)$$

This function is continuous with respect to μ. Moreover the function defined in (4.65) satisfies $\eta(\theta) = \eta(\theta, 0)$. The transversality condition is defined as the condition

$$\eta(\theta) \neq 0. \tag{4.66}$$

If the transversality condition is not satisfied, that is $\eta(\theta) = 0$, then by continuity $|\xi^{\theta,\mu}| \to \infty$ for $\mu \to 0$. If $\eta(\theta) = 0$, then the first basis function and the third function are linearly dependent at the limit $\mu = 0$. It is of course possible to develop the theory in this direction, but it is not the concern of this work. Therefore the transversality condition will always be assumed from now on. As mentioned earlier some aspects of the case where is not satisfied are postponed to Section 4.9.

Proposition 4.21. *Assume the transversality condition (4.66). Then for all $\epsilon > 0$ one has the limit*

$$\left\| \mathbf{U}_2^{\theta,\mu} - \frac{1}{\sigma(\theta)} \mathbf{U}_3^{\theta} \right\|_{(L^\infty[\epsilon,\infty[)^3} \to 0.$$

Proof. Evident. □

In order to show that the second basis function admits a continuous limit for $x < 0$, the strategy is to solve the integral equation (4.15) from $G = H$ backward, and to show that sharp estimates on the solution lead to information concerning the limit even for $x < 0$.

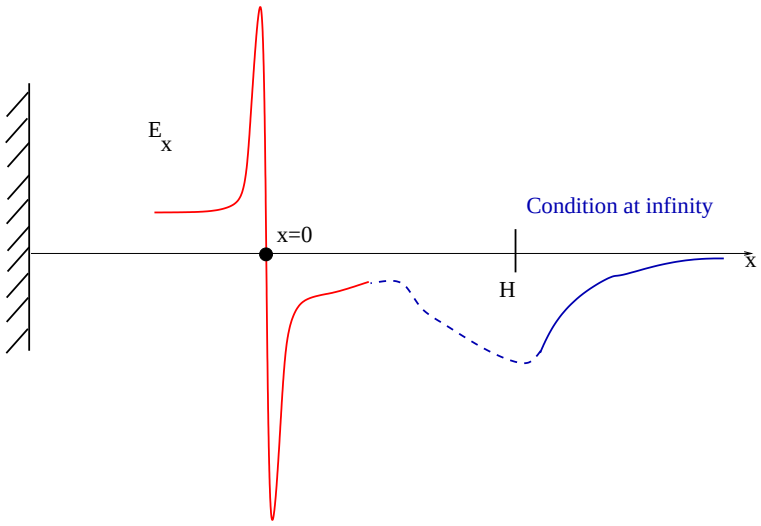


FIGURE 4.7 – Schematic representation of the real part of the limit electric field of the second basis function $U_2^{\theta,\mu}$, $\mu > 0$. Here the transversality condition $\sigma(\theta) \neq 0$ is satisfied, which turns into a singular behavior at the limit $\mu \to 0$.

4.6.3 The second basis function

The study of the second basis function for which $Q(\mathbf{U}_2^{\theta,\mu}) = 1$ will be based on continuity estimates from Section 4.5. The inequality (4.63) reads

$$\left\| U_2^{\theta,\mu} - \frac{1}{\alpha(\cdot) + i\mu} \right\|_{L^p(-L,H)} \leq C_p^{\theta} \left(\left| V_2^{\theta,\mu}(H) \right| + \left| W_2^{\theta,\mu}(H) \right| \right), \tag{4.67}$$

for $1 \leq p < \infty$.

Proposition 4.22. *Assume the transversality condition (4.66). There exists a constant C^θ independent of μ and continuous with respect to θ such that*

$$\left| V_2^{\theta,\mu}(H) \right| + \left| W_2^{\theta,\mu}(H) \right| \leq C^{\theta}. \tag{4.68}$$

Proof. Indeed, regarding relation (4.40), (4.41) the pair $(v, w) = (V_2^{\theta,\mu}(H), W_2^{\theta,\mu}(H))$ is solution of the linear system

$$\begin{cases} -vW_1^{\theta,\mu}(H) + wV_1^{\theta,\mu}(H) = 1, \\ vW_3^{\theta,\mu}(H) - wV_3^{\theta,\mu}(H) = 0. \end{cases}$$

The determinant of this linear system is equal to the value of the function $-\eta(\theta,\mu)$. So the transversality condition establishes that

$$\det \begin{pmatrix} -W_1^{\theta,\mu}(H) & V_1^{\theta,\mu}(H) \\ W_3^{\theta,\mu}(H) & -V_3^{\theta,\mu}(H) \end{pmatrix} = -\eta(\theta,\mu) \neq 0.$$

Therefore the solution of the linear system

$$v = -\frac{V_3^{\theta,\mu}(H)}{\eta(\theta,\mu)}, \quad w = -\frac{W_3^{\theta,\mu}(H)}{\eta(\theta,\mu)}$$

is bounded uniformly with respect to μ. □

Theorem 4.6.1. Assume the same transversality condition. For any $\theta \in [\theta_-, \theta_+]$ any interval, the second basis function satisfies the following estimates for some C_p^θ and C^θ which are continuous with respect to θ

$$\left\| U_2^{\theta,\mu} - \frac{1}{\alpha(\cdot) + i\mu} \right\|_{L^p(-L,H)} \leq C_p^\theta, \quad 1 \leq p < \infty, \tag{4.69}$$

$$\left\| \mathbf{U}_2^{\theta,\mu} \right\|_{H^1_{\text{loc}}[-L,0)\cup(0,H]} \leq C^\theta. \tag{4.70}$$

Proof. The first estimate is a straightforward consequence of (4.67) and (4.68). The integral representations (4.18)-(4.19) yield

$$\left\| V_2^{\theta,\mu} \right\|_{L^\infty_{\text{loc}}[-L,0)\cup(0,H]} + \left\| W_2^{\theta,\mu} \right\|_{L^\infty_{\text{loc}}[-L,0)\cup(0,H]} \leq C^\theta \tag{4.71}$$

for some C^θ. Then the second equation of (4.6) shows that one has the same bound for $U_2^{\theta,\mu}$

$$\left\| U_2^{\theta,\mu} \right\|_{L^\infty_{\text{loc}}[-L,0)\cup(0,H]} \leq C^\theta. \tag{4.72}$$

The bound on the derivatives follows from (4.6) □

Remark 6. Set $H' = -L$. From (4.71) one gets that $\|H'\|$ is bounded uniformly as well, therefore (4.42) can be generalized for $x < 0$ (resp. H') instead of $x > 0$ (resp. H). As a summary one has for a constant K^θ that can be further specified

$$\left| U_2^{\theta,\mu}(x) \right| \leq \frac{K^\theta}{r^2x^2 + \mu^2}, \qquad x \in (-L, H).$$

Now one passes to the limit $\mu \to 0^+$.

Proposition 4.23. *Assume the transversality condition. The second basis function admits a limit in the sense of in the sense of distributions for $\mu = 0^\pm$ as follows :*

$$\mathbf{U}_2^{\theta,\mu} \to \mathbf{U}_2^{\theta,\pm} = \left(P.V.\frac{1}{\alpha(x)} \pm \frac{i\pi}{\alpha'(0)}\delta_D + u_2^{\theta,\pm},\ v_2^{\theta,\pm},\ w_2^{\theta,\pm} \right)$$

where $u_2^{\theta,\pm}, v_2^{\theta,\pm}, w_2^{\theta,\pm} \in L^2(-L,\infty)$ and δ_D is the Dirac function at the origin.

Remark 7. The limits $\mathbf{U}_2^{\theta,\pm}$ are solutions of (4.6) in the sense of distributions. They will be called the singular solutions.

Proof. Consider the case $\mu \downarrow 0$. Some parts of the proof are already evident, essentially for quantities which are regular enough, namely $V_2^{\theta,\mu}$ and $W_2^{\theta,\mu}$, or for regions where all functions are regular, typically $x > 0$. Therefore the whole point is to pass to the limit in the singular part of the solution $U_2^{\theta,\mu}$. The equivalence between the integral formulation of Proposition 4.1 and the differential formulation (4.6) will be widely used.

• **Passing to the weak limit :** By continuity of the first basis function with respect to μ, one can pass to the limit concerning $(V_2^{\theta,\mu}(H), W_2^{\theta,\mu}(H))$. One gets that $(v, w) = \left(V_2^{\theta,0^+}(H), W_2^{\theta,0^+}(H)\right)$ is the unique solution of the linear system

$$\begin{cases} -vW_1^\theta(H) + wV_1^\theta(H) = 1, \\ vW_3^\theta(H) - wV_3^\theta(H) = 0, \end{cases} \tag{4.73}$$

where the coefficients are defined in terms of the first basis function for $\mu = 0$. By continuity away from the singularity at $x = 0$, one has that $\mathbf{U}_2^{\theta,\mu} \to \mathbf{U}_2^\theta$ in $L^\infty(\epsilon, H)$ for all $\epsilon > 0$. Using (4.69) it is clear that $U^{\theta,\mu} - \frac{1}{\alpha(\cdot)+i\mu}$ is bounded in $L^2(-L, H)$ uniformly with respect to μ. Therefore there exists a limit function denoted as u_2^{θ,O^+} such that for a subsequence

$$U_2^{\theta,\mu} - \frac{1}{\alpha(\cdot) + i\mu} \rightharpoonup_{weak} u_2^{\theta,O^+} \text{ in } L^2(-L, H).$$

Moreover the first derivative of $U_2^{\theta,\mu}$ is bounded in $L^2(-L, -\epsilon)$ by virtue of (4.70). Therefore

$$U_2^{\theta,\mu} \to_{strong} \frac{1}{\alpha(\cdot)} + u_2^{\theta,O^+} \text{ in } L^2(-L, -\epsilon)$$

at least for a subsequence. Considering the integral relations (4.18)-(4.19), these subsequences are such that

$$V_2^{\theta,\mu}(x) \to v_2^{\theta,O^+}(x), \tag{4.74}$$

and

$$W_2^{\theta,\mu}(x) \to w_2^{\theta,O^+}(x), \tag{4.75}$$

with uniform convergence on compact sets of $(-L, H) \setminus \{0\}$. Denote by $\epsilon > 0$ a small parameter such that α is invertible on $]-\epsilon, \epsilon[$ and denote by β the function defined as $\beta(z) = 1/\alpha'(\alpha^{-1}(z))$. Then

$$\begin{aligned} v_2^{\theta,0^+} &:= a_H\, A_0(x) + b_H\, B_0(x) + \int_H^x \mathcal{D}_x^\theta k^0(x,0) u_2^{\theta,+}(z) + \tilde{v}(x) \\ &\quad + \int_H^x \mathcal{D}_x^\theta (k^0(x,z) - k^0(x,0)) \left(\frac{1}{\alpha(z)} + u_2^{\theta,+}(z)\right), \end{aligned}$$

with

$$\tilde{v}(x) := \int_H^x \mathcal{D}_x^\theta k^0(x,0)\, \frac{1}{\alpha(z)}\, dz, \quad \text{for } x > 0,$$

$$\tilde{v}(x) := \mathcal{D}_x^\theta k^0(x,0) \left[\int_x^{-\epsilon} \frac{1}{\alpha(z)}\, dz + \ln \alpha(\epsilon)\, \beta(\epsilon) - \right.$$

$$\left. \ln \alpha(-\epsilon)\, \beta(-\epsilon) + \int_{\alpha(\epsilon)}^{\alpha(-\epsilon)} \ln(z)\, \beta'(z)\, dz \right] + \int_\epsilon^H \mathcal{D}_x^\theta k^0(x,0)\, \frac{1}{\alpha(z)}\, dz \quad \text{for}, x < 0,$$

$$w_2^{\theta,0^+} := a_H\, A_0'(x) + b_H\, B_0'(x) + \int_H^x \partial_x \mathcal{D}_x^\theta (k^0(x,z) - k^0(x,0)) \left(\frac{1}{\alpha(z)} + u_2^{\theta,+}(z)\right)$$

$$+ \int_H^x \partial_x \mathcal{D}_x^\theta k^0(x,0) u_2^{\theta,+}(z) + \tilde{w}(x),$$

with

$$\tilde{w}(x) := \int_H^x \partial_x \mathcal{D}_x^\theta k^0(x,0)\, \frac{1}{\alpha(z)}\, dz, \quad \text{for } x > 0,$$

$$\tilde{w}(x) := \partial_x \mathcal{D}_x^\theta k^0(x,0) \left[\int_x^{-\epsilon} \frac{1}{\alpha(z)}\, dz + \ln \alpha(\epsilon)\, \beta(\epsilon) - \right.$$

$$\left. \ln \alpha(-\epsilon)\, \beta(-\epsilon) + \int_{\alpha(\epsilon)}^{\alpha(-\epsilon)} \ln(z)\, \beta'(z)\, dz \right] + \int_\epsilon^H \partial_x \mathcal{D}_x^\theta k^0(x,0)\, \frac{1}{\alpha(z)}\, dz \quad \text{for}, x < 0.$$

The limits in (4.74), (4.75) also hold in the strong topology of $L^2(-L,H)$.

These weak or strong limits are naturally weak solution of the initial system (4.6) : denoting for simplicity

$$(u_2, v_2, w_2) = (u_2^{\theta,0+}, v_2^{\theta,0+}, w_2^{\theta,0+}),$$

these functions are solutions of

$$\begin{cases} \int w_2 \varphi_1 dx + i\theta \; P.V. \int \left(\frac{1}{\alpha} + u_2 \right) \varphi_1 dx - \frac{\theta\pi}{\alpha'(0)} \varphi_1(0) + \int v_2 \varphi_1' dx = 0, \\ i\theta \int w_2 \varphi_2 dx - \int (\alpha u_2 + 1) \varphi_2 dx - i \int \gamma v_2 \varphi_2 dx = 0, \\ \int w_2 \varphi_3' dx + i \; P.V. \int \gamma \left(\frac{1}{\alpha} + u_2 \right) \varphi_3 dx - \frac{\gamma(0)\pi}{\alpha'(0)} \varphi_3(0) \\ \qquad\qquad - \int \alpha v_2 \varphi_3 dx = 0, \end{cases} \tag{4.76}$$

for any sufficiently smooth test functions with compact support, for example $(\varphi_1, \varphi_2, \varphi_3) \in \mathcal{C}_0^1(-L,H)$. Passing to the limit makes use of the fact that in the sense of distributions, $\lim_{\mu \to 0+} \frac{1}{\alpha(x)+i\mu} = P.V \frac{1}{\alpha(x)} + i\pi \frac{1}{\alpha'(0)} \gamma_D$. The signs of $-\frac{\theta\pi}{\alpha'(0)}\varphi_1(0)$ and $-\frac{\gamma(0)\pi}{\alpha'(0)}\varphi_3(0)$ are compatible [1] with the fact the limit is for positive μ. The principal value is defined as :

$$P.V. \int \frac{1}{\alpha(x)}\, \varphi(x)\, dx := \lim_{\epsilon \downarrow 0} \left(\int_{-L}^{\rho(-\epsilon)} \frac{1}{\alpha(x)}\, \varphi(x) + \int_{\rho(\epsilon)}^H \frac{1}{\alpha(x)}\, \varphi(x) \right) dx,$$

where $\alpha(\rho(\mp\epsilon)) = \pm\epsilon$.

• **Uniqueness of the weak limit :** If there is another triplet $(\widetilde{u_2}, \widetilde{v_2}, \widetilde{w_2})$ solution of the same weak formulation (4.76), then the difference

$$(\widehat{u_2}, \widehat{v_2}, \widehat{w_2}) = (\widetilde{u_2} - u_2, \widetilde{v_2} - v_2, \widetilde{w_2} - w_2)$$

satisfies

$$\begin{cases} \int \widehat{w_2} \varphi_1 dx + i\theta \int \widehat{u_2} \varphi_1 dx + \int \widehat{v_2} \varphi_1' dx = 0, \\ i\theta \int \widehat{w_2} \varphi_2 dx - \int \alpha \widehat{u_2} \varphi_2 dx - i \int \gamma \widehat{v_2} \varphi_2 dx = 0, \\ \int \widehat{w_2} \varphi_3' dx + i \int \gamma \widehat{u_2} \varphi_3 dx - \int \alpha \widehat{v_2} \varphi_3\, dx = 0, \end{cases} \tag{4.77}$$

By construction $(\widehat{u_2}, \widehat{v_2}, \widehat{w_2}) = (0,0,0)$ for $x > 0$. For $x < 0$, the fact that $(\widehat{u_2}, \widehat{v_2}, \widehat{w_2})$ is a solution of the X-mode equations stems from (4.77). Therefore these functions can be

1. If one takes the limit $\mu \uparrow 0$, the signs of these terms are changed.

expressed as a linear combination of the first and second basis functions for $x < 0$. Since $\widehat{u_2} \in L^2(-L, 0)$ is non singular, only the first basis function is involved that is

$$(\widehat{u_2}, \widehat{v_2}, \widehat{w_2}) = \lambda \left(U_1^\theta, V_1^\theta, W_1^\theta\right) \qquad x < 0.$$

The system (4.77) yields for example

$$\int_{-L}^{0} \widehat{w_2}\varphi_3' dx + i \int_{-L}^{0} \gamma \widehat{u_2}\varphi_3 dx - \int_{-L}^{0} \alpha(x)\widehat{v_2}\varphi_3 \, dx = 0$$

where $\varphi_3(-L) = 0$ and $\varphi_3(0)$ is arbitrary. An integration by parts gives

$$\int_{-L}^{0} \left(-\widehat{w_2}' + i\gamma\widehat{u_2} - \alpha\widehat{v_2}\right) \varphi_3 dx + \widehat{\omega_2}(0)\varphi_3(0) = 0.$$

Since $(\widehat{u_2}, \widehat{v_2}, \widehat{w_2})$ is a non singular solution of the X-mode equations, one has that $-\widehat{w_2}' + i\gamma\widehat{u_2} - \alpha\widehat{v_2} = 0$. Finally

$$\widehat{\omega_2}(0)\varphi_3(0) = 0.$$

Choosing φ such that $\varphi_3(0) \neq 0$, it follows that $0 = \widehat{\omega_2}(0) = \lambda W_1^\theta(0)$. Considering the normalization (4.33) one gets that $\lambda = 0$. Therefore $(\widehat{u_2}, \widehat{v_2}, \widehat{w_2}) = (0, 0, 0)$. It means that the weak limit is unique : the whole sequence tends to the same weak limit.

• **Regularity :** By Theorem 4.6.1 the limit belongs to $H^1\left([-L, -\epsilon] \cup [\epsilon, \infty)\right)^3$.

• **Limit $\mu \uparrow 0$:** The sign of the Dirac function is changed in the final result of the proposition since $\lim_{\mu \to 0^-} \frac{1}{\alpha(x)+i\mu} = P.V \frac{1}{\alpha(x)} - i\pi \frac{1}{\alpha'(0)}\delta_D$. □

4.6.4 The limit spaces $\mathbb{X}^{\theta,\pm}$

Here both cases $\mu \leq 0$ and $\mu \geq 0$ will be considered.

The space $\mathbb{X}^{\theta,+}$

Passing to the limit $\mu \to 0^+$, the limit space $\mathbb{X}^{\theta,+}$ is

$$\mathbb{X}_\varepsilon^{\theta,+} = \text{Span}\left\{\mathbf{U}_1^\theta, \mathbf{U}_2^{\theta,+}\right\} \subset H^1_{loc}\left((-L, \infty) \setminus \{0\}\right). \tag{4.78}$$

The space $\mathbb{X}^{\theta,-}$

It is of course possible do all the analysis with negative $\mu < 0$ and to study the limit $\mu \to 0^-$. The first basis function is exactly the same. The second basis function is chosen exponentially decreasing at infinity and such that

$$i\mu U_2^{\theta,\mu} = 1 \qquad \mu < 0.$$

The generalization of the preliminary result (4.69) is straightforward

$$\left\| U_2^{\theta,\mu} - \frac{1}{\alpha(\cdot) + i\mu} \right\|_{L^p(-L,H)} \leq C_p^\theta, \qquad -1 \leq \mu < 0. \tag{4.79}$$

Passing to the limit $\mu \to 0^-$, it defines the limit space $\mathbb{X}^{\theta,-}$

$$\mathbb{X}_\varepsilon^{\theta,-} = \text{Span}\left\{\mathbf{U}_1^\theta, \mathbf{U}_2^{\theta,-}\right\} \subset H^1_{loc}\left((-L, \infty) \setminus \{0\}\right). \tag{4.80}$$

Observe of course that the first basis function belongs to $\mathbb{X}_\varepsilon^{\theta,+} \cap \mathbb{X}_\varepsilon^{\theta,-}$. Since the limit equation is the same, and the normalization at $x = H$ is also the same we readily observe that the second basis functions are identical for $0 < x$

$$U_2^{\theta,+}(x) = U_2^{\theta,-}(x) \qquad 0 < x. \tag{4.81}$$

The main point is to determine the difference between the limit of the two singular functions for $x < 0$.

Proposition 4.24. *One has*

$$U_2^{\theta,+}(x) - U_2^{\theta,-}(x) = \frac{-2i\pi}{\alpha'(0)} U_1^\theta(x) \qquad x < 0. \tag{4.82}$$

Proof. Notice that the Wronskian relations (4.41) are the same at the limit $\mu = 0^\pm$. By subtraction

$$V_1^\theta(x)\left(W_2^{\theta,+}(x) - W_2^{\theta,-}(x)\right) - W_1^\theta(x)\left(V_2^{\theta,\mu}(x) - V_2^{\theta,-}(x)\right) = 0.$$

It shows that the difference is proportional to the first basis function

$$U_2^{\theta,+}(x) - U_2^{\theta,-}(x) = \zeta U_1^\theta(x) \qquad x < 0. \tag{4.83}$$

Determining ζ will complete the proof. It is already known that the limit $\mu \to 0^+$ can be characterized by (4.76). The third equation writes

$$\int w_2^+ \varphi_3' dx + i\ P.V. \int \gamma \left(\frac{1}{\alpha} + u_2^+\right) \varphi_3 dx - \frac{\gamma(0)\pi}{\alpha'(0)} \varphi_3(0) - \int \alpha v_2^+ \varphi_3 dx = 0$$

where (u_2^+, v_2^+, w_2^+) refers to the non singular part of the limit $\mu \to 0^+$. The equivalent equation for the non singular part (u_2^-, v_2^-, w_2^-) of the limit $\mu \to 0^-$ is

$$\int w_2^- \varphi_3' dx + i\ P.V. \int \gamma \left(\frac{1}{\alpha} + u_2^-\right) \varphi_3 dx + \frac{\gamma(0)\pi}{\alpha'(0)} \varphi_3(0) - \int \alpha v_2^- \varphi_3 dx = 0.$$

By subtraction, one gets

$$\int (w_2^+ - w_2^-)\varphi_3' dx + i \int \gamma (u_2^+ - u_2^-)\varphi_3 dx$$
$$-\frac{2\gamma(0)\pi}{\alpha'(0)} \varphi_3(0) - \int \alpha (v_2^+ - v_2^-)\varphi_3 dx = 0.$$

Due to (4.81) these differences vanishes for $x > 0$. One gets

$$\int_{-L}^0 (w_2^+ - w_2^-)\varphi_3' dx + i \int_{-L}^0 \gamma (u_2^+ - u_2^-)\varphi_3 dx$$
$$-\frac{2\gamma(0)\pi}{\alpha'(0)} \varphi_3(0) - \int_{-L}^0 \alpha (v_2^+ - v_2^-)\varphi_3 dx = 0$$

where φ_3 is a smooth test function that vanishes at $-L$. Integration by part yields

$$\int_{-L}^0 \left(-(w_2^+ - w_2^-)' + i\gamma(u_2^+ - u_2^-) - \alpha(v_2^+ - v_2^-)\right) \varphi_3 dx$$
$$-\frac{2\gamma(0)\pi}{\alpha'(0)} \varphi_3(0) + (w_2^+ - w_2^-)(0)\varphi_3(0) = 0.$$

Due to (4.83) one has that

$$-(w_2^+ - w_2^-)' + i\gamma(u_2^+ - u_2^-) - \alpha(v_2^+ - v_2^-) = 0 \qquad x < 0.$$

Since $\varphi_3(0)$ is arbitrary, it means that $w_2^+(0) - w_2^-(0) = \frac{2\gamma(0)\pi}{\alpha'(0)}$. One obtains $\zeta W_1^\theta(0) = \frac{2\gamma(0)\pi}{\alpha'(0)}$, that is $i\gamma(0)\zeta = \frac{2\gamma(0)\pi}{\alpha'(0)} = 0$. Therefore $\zeta = \frac{-2i\pi}{\alpha'(0)}$. The claim is proved. □

4.7 Numerical validation

This section focuses on the discretization of the first order system on $(V^{\theta,\mu}, W^{\theta,\mu})$ given in (4.29), with $\mu \neq 0$. Note that because of the resonance at the origin, one has to use a stiff solver since a classical Euler solver would not be accurate. The method used is based on a modified Rosenbrock formula of order 2 proposed by Matlab as *ode*23*s*. The $U^{\theta,\mu}$ component is then recovered thanks to the relation

$$i\theta W^{\theta,\mu} - (\alpha + i\mu)U^{\theta,\mu} - i\gamma V^{\theta,\mu} = 0.$$

The following numerical results were obtained with the parameters

- $H = 2$ and $D = 5$,
- $\theta = 1$ and $\mu = 10^{-2}$,
- $\alpha(x) = -x$ for $x < H$ and $\alpha(x) = -2$ elsewhere,
- $\gamma(x) = 0.25$.

4.7.1 The first basis function

The first basis function is computed starting from the origin, since

$$(V_1^{\theta,\mu}, W_1^{\theta,\mu}) = (i\theta, i\gamma(0)). \tag{4.84}$$

One can observe on Figures 4.8 and 4.9 the propagative behavior of the solution for $x < 0$, and the exponential blow up for $x > 0$ described in Proposition 4.7 for the first basis function. Each of them is obtained for a different value of θ. No $1/\mu$ singularity appears around the resonance point as the regularization parameter decreases toward zero.

Since the first basis function satisfies (4.84), $(V_1^{\theta,\mu}, W_1^{\theta,\mu})$ belongs to $i\mathbb{R}^2$ and does not depend on μ. Moreover the imaginary part of system (4.29)'s matrix, namely (4.28), goes to zero with μ. As a result, for all $\theta \in \mathbb{R}$, the real parts of $(V_1^{\theta,\mu}, W_1^{\theta,\mu})$ go to zero as μ goes to zero as well. For this reason both real and imaginary parts of the solutions are represented on Figures 4.8 and 4.9.

4.7.2 The second basis function

The second basis function is computed as suggested by the theory : it is actually the third basis functions scaled to ensure the normalization at the origin $i\mu U_2^{\theta,\mu} = 1$. The third basis function is computed starting from $x = D >> H$, with the exact condition $(V_3^{\theta,\mu}, W_3^{\theta,\mu})$.

One can observe on Figure 4.10 the singularity at the origin together with the exponential decrease for $x > 0$. The function $x \mapsto \frac{1}{\alpha+i\mu}$ is plotted for comparison : the functions $U_2^{\theta,\mu}$ obviously fit perfectly the singularity $1/(\alpha + i\mu)$ at the origin as suggested in Proposition 4.16. As the regularization parameter μ goes to zero the solution computed converges to a singular solution.

Figure 4.11 evidences that the weak limit from Poposition 4.23 is actually a strong limit everywhere but at the resonance point. It displays the same numerical results as Figure 4.10, but with a fixed scaling.

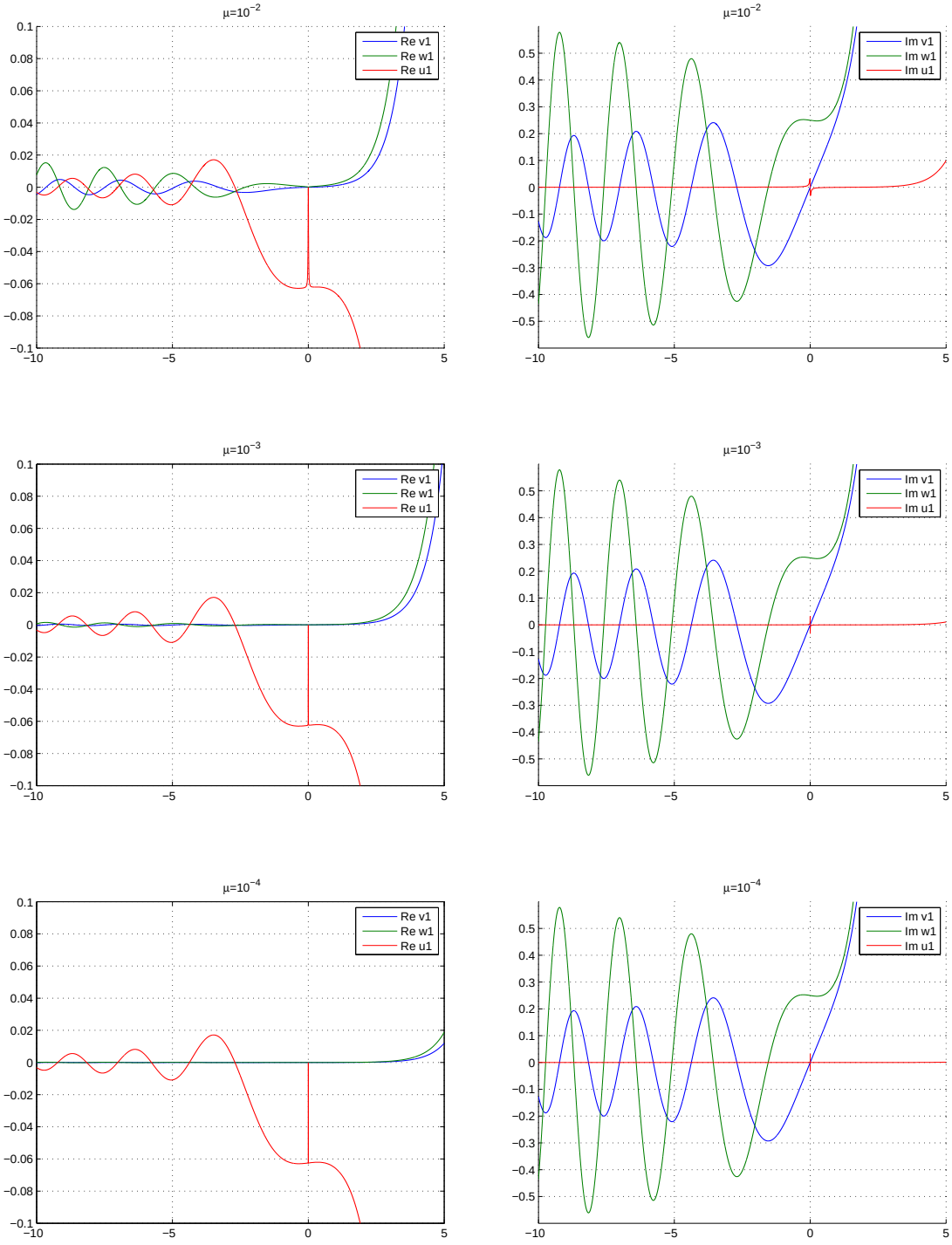


FIGURE 4.8 – First basis function of $\mathbb{X}^{\theta,\mu}$ computed for $\theta = 0$ and $\mu = 10^{-2}$, $\mu = 10^{-3}$, $\mu = 10^{-4}$. The real and imaginary parts of the three components $(U_1^{\theta,\mu}, V_1^{\theta,\mu}, W_1^{\theta,\mu})$ are represented. No significant change of behavior is observed as μ goes to zero.

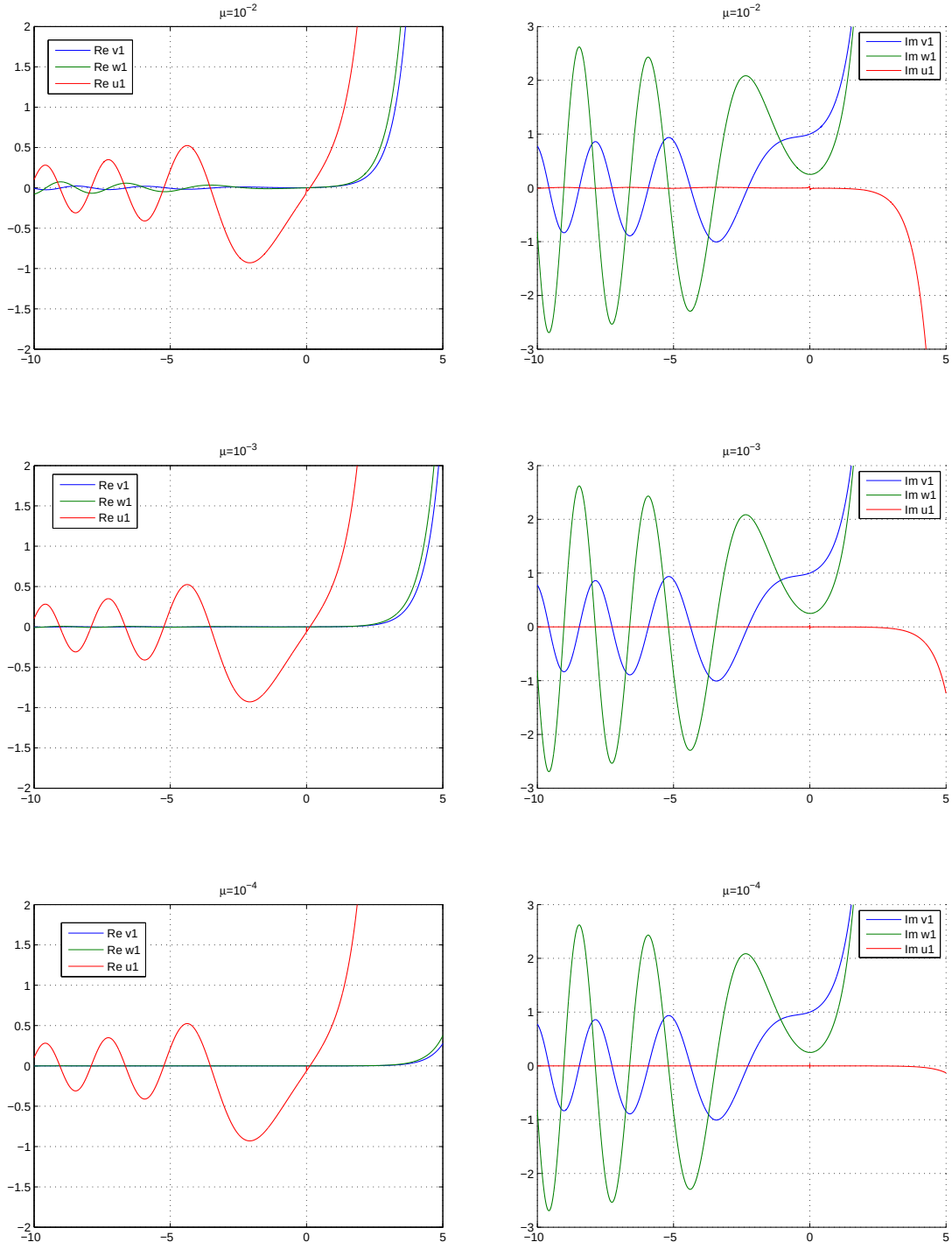


FIGURE 4.9 – First basis function of $\mathbb{X}^{\theta,\mu}$ computed for $\theta = 1$ and $\mu = 10^{-2}$, $\mu = 10^{-3}$, $\mu = 10^{-4}$. The real and imaginary parts of the three components $(U_1^{\theta,\mu}, V_1^{\theta,\mu}, W_1^{\theta,\mu})$ are represented. Again no significant change of behavior is observed as μ goes to zero.

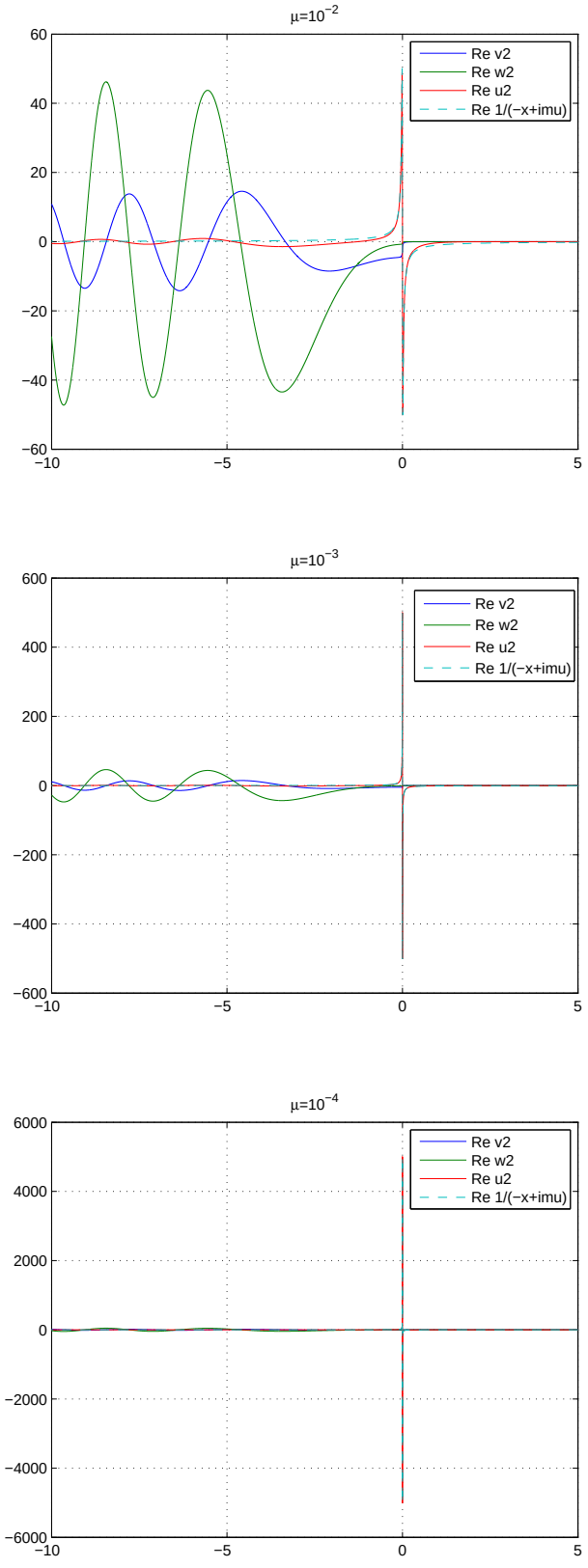


Figure 4.10 – Second basis function of $\mathbb{X}^{\theta,\mu}$ computed for $\theta = 1.5$ and $\mu = 10^{-2}$, 10^{-3}, 10^{-4}. The real parts of the three component $(U_2^{\theta,\mu}, V_2^{\theta,\mu}, W_2^{\theta,\mu})$ are represented. The real part of $(\alpha + i\mu)^{-1}$ is also represented, and evidences the blow up of the solution at the resonance as μ goes to zero.

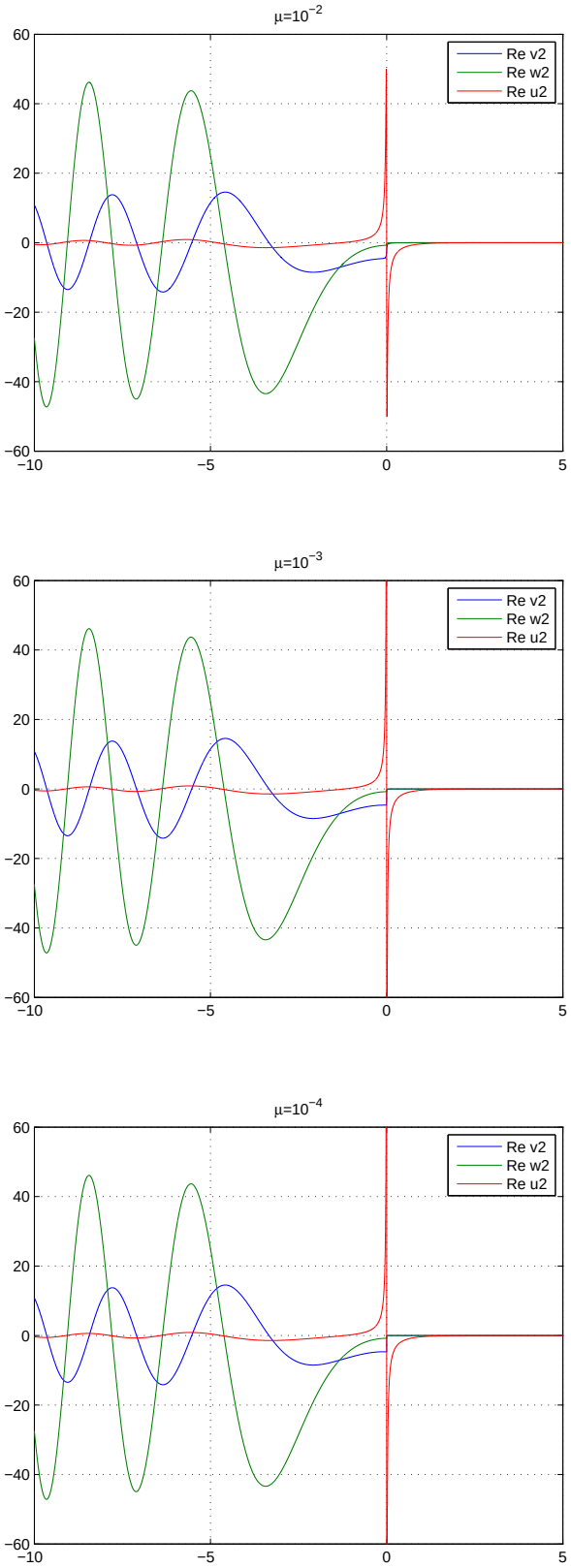


FIGURE 4.11 – Second basis function of $\mathbb{X}^{\theta,\mu}$ computed for $\theta = 1.5$ and $\mu = 10^{-2}$, 10^{-3}, 10^{-4}. The real parts of the three component $(U_2^{\theta,\mu}, V_2^{\theta,\mu}, W_2^{\theta,\mu})$ are represented. Here the scale is fixed to evidence the strong convergence observed on $]-L, D[\backslash\{0\}$: the convergence observed is strong everywhere but at the resonance point $x = 0$.

4.7.3 Difference between positive and negative value of μ

The second basis function can be computed for both positive and negative values of μ in order to verify the result given in Proposition 4.24 for $x < 0$ and the equality of the two functions for $x > 0$ given in (4.81).

Figures 4.12 and 4.13 illustrate these theoretical results for two different values of θ, and show satisfying fit, for $x > 0$ and $x < 0$. Again the convergence with respect to the regularization parameter μ is observed.

4.8 Proof of the main theorem

All the information about the first and second basis functions is now used to construct the solution of the system (4.6) with the boundary condition (4.2). The function g depends only of the vertical variable y. Under convenient condition g admits the Fourier representation

$$g(y) = \frac{1}{2\pi} \int_{\mathbb{R}} \widehat{g}(\theta) e^{i\theta y} d\theta. \tag{4.85}$$

First consider a small but non zero regularization parameter $\mu > 0$. For the sake of simplicity assume that the transversality condition is satisfied for all θ in the support of $\widehat{g}$

$$|\eta(\theta)| \geq c > 0 \quad \forall \theta \in \operatorname{supp}(\widehat{g}). \tag{H6}$$

It is just a convenient uniform version of the point-wise transversality condition (4.66). Additional comments are to be found in Section 4.8.3.

4.8.1 One Fourier mode

For one Fourier mode, one needs to consider the solution of (4.6) with boundary condition

$$\widehat{W}^{\mu}(-L) + i\sigma \widehat{V}^{\mu}(-L) = \widehat{g}.$$

Since the solution must decrease (exponentially) at $x \approx \infty$ to guarantee that no energy comes from infinity, the solution is proportional to the second basis function. That is there is a coefficient $\gamma^{\theta,\mu}$ such that $\widehat{\mathbf{U}}^{\mu} = \gamma^{\theta,\mu} \mathbf{U}_2^{\theta,\mu}$. The coefficient satisfies the equation

$$\gamma^{\theta,\mu} \left(W_2^{\theta,\mu}(-L) + i\sigma V_2^{\theta,\mu}(-L) \right) = \widehat{g}(\theta)$$

that is

$$\gamma^{\theta,\mu} = \frac{\widehat{g}(\theta)}{\tau^{\theta,\mu}}$$

from which it is clear that we must study the coefficient

$$\tau^{\theta,\mu} = W_2^{\theta,\mu}(-L) + i\mathrm{sgn}(\mu)\sigma \mathrm{V}_2^{\theta,\mu}(-\mathrm{L}). \tag{4.86}$$

A last technical result concerns this coefficient $\tau^{\theta,\mu}$.

Proposition 4.25. *Assume (H6). For every compact set $S \subset \mathbb{R}$, there exists $\epsilon > 0$, τ^+ and $\tau_- > 0$ such that $\tau^- \leq \left|\tau^{\theta,\mu}\right| \leq \tau^+$ for $0 < \mu \leq \epsilon$ and $\theta \in S$.*

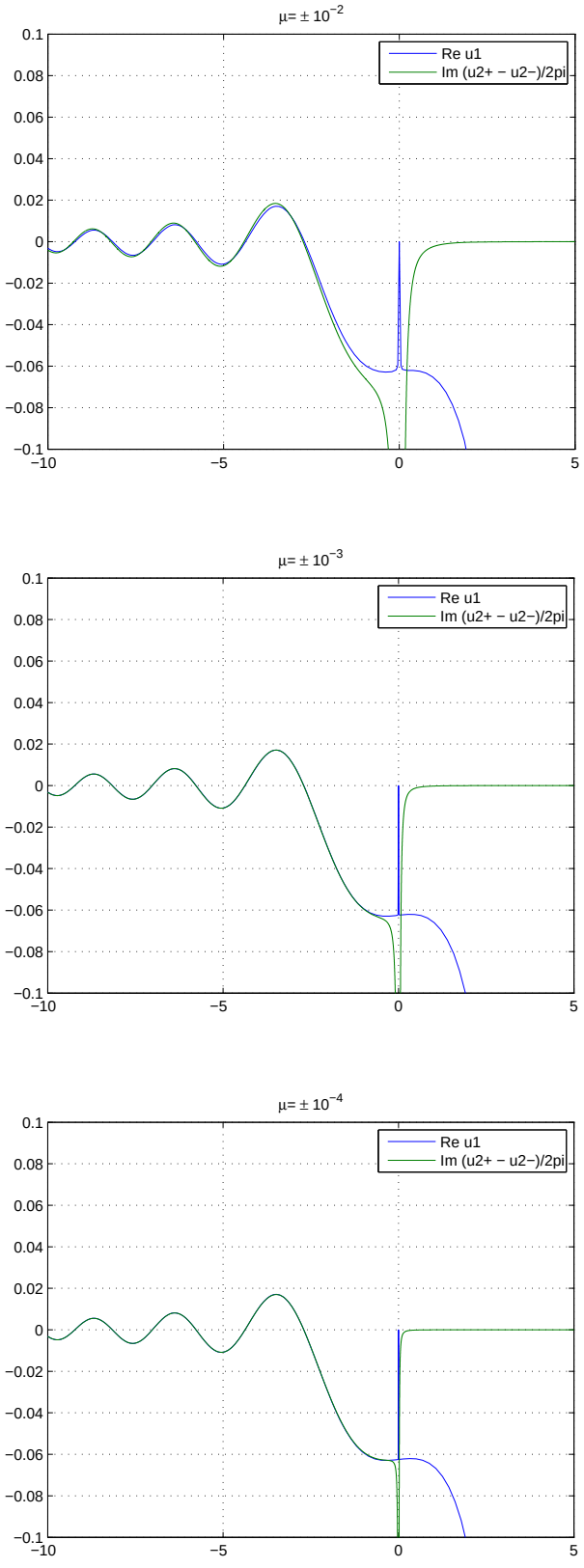


FIGURE 4.12 – Difference between the imaginary parts of the second basis function computed for $\theta = 0$ and $\mu = -10^{-2}$, $\mu = -10^{-3}$, $\mu = -10^{-4}$, compared to the real part of $U_1^{\theta,\mu}$.

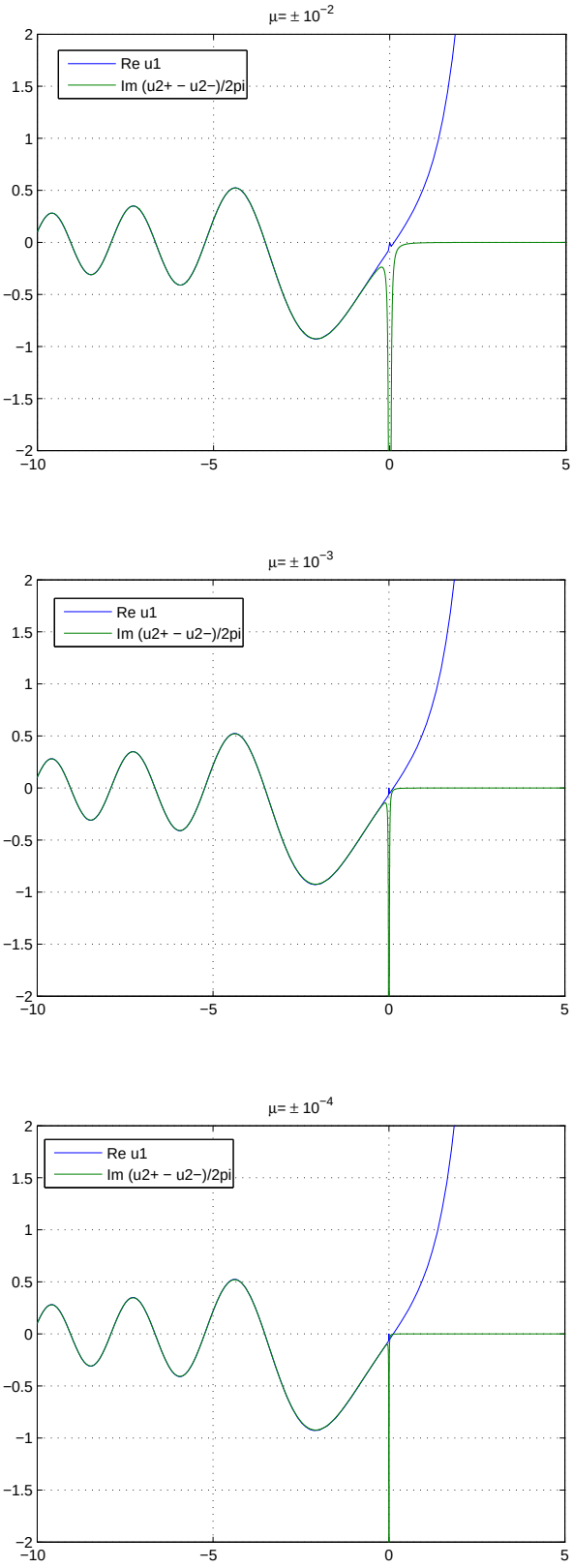


Figure 4.13 – Difference between the imaginary parts of the second basis function computed for $\theta = 1$ and $\mu = -10^{-2}$, $\mu = -10^{-3}$, $\mu = -10^{-4}$, compared to the real part of $U_1^{\theta,\mu}$.

Proof. The upper bound is a direct consequence of (4.71). To prove the lower bound, a useful result is the formula which comes from (4.36)

$$\mathrm{Im}\left(W_2^{\theta,\mu}(-L)\overline{V_2^{\theta,\nu}(-L)}\right) \geq \mu \int_{-L}^{\infty} \left|U_2^{\theta,\mu}(x)\right|^2 dx$$

Combining with (4.49) and $Q\left(\mathbf{U}^{\theta,\mu}\right) = 1$ (by construction), it yields

$$\mathrm{Im}\left(W_2^{\theta,\mu}(-L)\overline{V_2^{\theta,\nu}(-L)}\right) \geq \tau_- > 0.$$

Plugging the definition of $\tau^{\theta,\mu}$ inside this inequality, one gets

$$\mathrm{Im}\left(\tau^{\theta,\mu}\overline{V_2^{\theta,\nu}(-L)}\right) \geq \tau_- + \mathrm{sgn}(\mu)\sigma \left|\mathrm{V}^{\theta,\mu}(-\mathrm{L})\right|^2 \geq \tau_- > 0.$$

Therefore $\left|V_2^{\theta,\mu}(-L)\right| \times |\tau(\theta,\mu)| \geq \tau_-$. The L^∞ bounds (4.71) shows that there exists $C > 0$ such that $C\,|\tau(\theta,\mu)| \geq \tau_-$. □

By (4.74), (4.75),

$$\tau^{\theta,+} := W_2^{\theta,0^+}(-L) + i\mathrm{sgn}(\mu)\sigma \mathrm{V}_2^{\theta,0^+}(-\mathrm{L}) = \lim_{\mu\to 0^+} \tau^{\theta,\mu}. \tag{4.87}$$

Proposition 4.26. *For every compact set $S \subset \mathbb{R}$, there exists τ^+ and $\tau_- > 0$ such that $\tau^- \leq \left|\tau^{\theta,+}\right| \leq \tau^+$ for $\theta \in S$.*

Proof. This is immediate from Proposition 4.25 and (4.87). □

4.8.2 Fourier representation of the solution

The solution of (4.5) with the boundary condition (4.2) is given by the inverse Fourier formula

$$\begin{pmatrix} E_x^\mu \\ E_y^\mu \\ W^\mu \end{pmatrix}(x,y) = \frac{1}{2\pi}\int_{\mathbb{R}} \frac{\widehat{g}(\theta)}{\tau^{\theta,\mu}} \mathbf{U}_2^{\theta,\mu}(x) e^{i\theta y} d\theta \tag{4.88}$$

where it is assumed that $g \in L^2(\mathbb{R})$ and that $\widehat{g}$ has compact support. Passing to the limit in (4.88) one gets

$$\begin{pmatrix} E_x^+ \\ E_y^+ \\ W^+ \end{pmatrix}(x,y) = \frac{1}{2\pi}\int_{\mathbb{R}} \frac{\widehat{g}(\theta)}{\tau^{\theta,+}} \begin{pmatrix} P.V.\frac{1}{\alpha(x)} + \frac{i\pi}{\alpha'(0)}\delta_D + u_2^{\theta,+} \\ v_2^{\theta,+} \\ w_2^{\theta,+} \end{pmatrix} e^{i\theta y} d\theta \tag{4.89}$$

This formula has to be understood in the sense of distributions. Since by Theorem 4.6.1 $\left\|u^{\theta,+}\right\| \leq C_2^\theta$, $\left\|v^{\theta,+}\right\| \leq C_2^\theta$, $\left\|w^{\theta,+}\right\| \leq C_2^\theta$ with C_2^θ a continuous function of θ, and considering that $\tau^{\theta,\mu}$ converges to $\tau^{\theta,+}$ there is sufficient regularity to pass to the limit. The value of the heating is

$$\mathcal{Q} = \omega\varepsilon_0 \lim_{\mu\to 0^+} \mu \int |E_x^\mu(x,y)|^2 dxdy = \frac{\omega\varepsilon_0}{2}\int_{\mathbb{R}} \frac{|\widehat{g}(\theta)|^2}{|\alpha'(0)|^2 \left|\tau^{\theta,+}\right|^2} d\theta. \tag{4.90}$$

The result follows from $\omega = \varepsilon_0 = 1$.

To our knowledge this is the first time that such a formula is written where all terms are explicitly given. A similar but much less precise formula can be found in [CW74] derived by means of analogies, see also [PT05]. All these integrals are convergent provided $\widehat{g}$ decays sufficiently fast at infinity. Provided the transversality condition is satisfied, and it is always the case in a neighborhood of $\theta = 0$ under hypothesis (H5), the heating is generically positive. This is of course related to the presence of the strong Dirac singularity in the solution (4.89).

Remark 8. An essential consequence of this analysis is the physical heating $\mathcal{Q}$ which is related to the singularity $P.V.\frac{1}{\alpha(x)} \pm \frac{i\pi}{\alpha'(0)}\delta_D$ of the mathematical solution. The singularity is not an artifact of the model. It is on the contrary a direct way to measure the amount of heating provided to the plasma by the electromagnetic wave.

Remark 9. Observe that the singular solutions $\mathbf{U}_2^{\theta,\pm}$ are the unique solutions of the following initial value problem.

Find a triplet $(u_2^{\theta,\pm}, v_2^{\theta,\pm}, w_2^{\theta,\pm}) \in L^2(-L,\infty)^3$ which satisfies the constraints $v_2^{\theta,\pm}(H) = V_3^{\theta,0}(H)$, $w_2^{\theta,\pm}(H) = W_3^{\theta,0}(H)$, and

$$w_2^{\theta,\pm} - \frac{d}{dx}v_2^{\theta,\pm} + i\theta u_2^{\theta,\pm} = -i\theta P.V.\frac{1}{\alpha} \pm \frac{\theta\pi}{\alpha'(0)}\delta_D,$$

$$i\theta w_2^{\theta,\pm} - \alpha u_2^{\theta,\pm} - i\gamma v_2^{\theta,\pm} = 1,$$

$$-\frac{d}{dx}w_2^{\theta,\pm} + i\gamma u_2^{\theta,\pm} - \alpha(x) v_2^{\theta,\pm} = -iP.V.\frac{\gamma}{\alpha} \pm \frac{\gamma(0)\pi}{\alpha'(0)}\delta_D.$$

This problem has an unique solution, it is proved by the argument given to prove the uniqueness of the weak limits. For this purpose observe that

$$\left(\frac{1}{\alpha(x)} + u_2^{\theta,\pm}, v_2^{\theta,\pm}, w_2^{\theta,\pm}\right)(x) = \left(U_3^{\theta,0}, V_3^{\theta,0}, W_3^{\theta,0}\right)(x) \text{ for } x > 0.$$

Observe then the similarity with the standard limit absorption principle in scattering theory. In scattering theory the solutions obtained by the limit absorption principle are characterized as the unique solutions that satisfy the radiation condition, i.e., they are uniquely determined by the behavior at infinity. Here, the singular solutions are uniquely determined by their behavior at $+\infty$ and by their singular part $P.V.\frac{1}{\alpha(x)} \pm \frac{i\pi}{\alpha'(0)}\delta_D$. Note that it is natural that to specify the singularity at $x = 0$ because the equations are degenerate at $x = 0$. This principle could be used for practical computations. It is however a little more subtle since a boundary condition at finite distance $x = -L$ must be prescribed. That is the singular part is itself dependent on the boundary condition where the energy comes in the system. Mathematically it corresponds to the coefficient $\tau^{\theta,+}$ in the representation formula (4.89).

4.8.3 What happens if the transversality condition is not satisfied

An interesting question is to determine what happens if the transversality condition is not satisfied. Some simple remarks follow. Firstly the point-wise transversality condition (4.66) or the uniform one (H6) greatly simplify the analysis. They are satisfied at least for θ close to zero provided the transition zone is small enough (H5). Secondly some technical intermediate results may be wrong if these conditions are not satisfied : for example it is not clear whether σ is still regular at points θ such that $\sigma(\theta) = 0$.

The purpose of this paragraph is not to answer to the questions raised by this possibilities, but only to give understanding of the physical situation hidden behind and to explain what is the limit value of the heating (4.90). The analysis is as follows.

Physical picture : If $\sigma(\theta) = 0$ then the first basis function $\mathbf{U}_1^\theta$ is proportional to $\mathbf{U}_3^\theta$ (at least for $x \geq H$). It means that it is also exponentially decaying at infinity. That is the excitation provided by the boundary condition catches this non singular first basis function which is also the physical one. We remark that no heating is provided by $\mathbf{U}_1$ because $Q(\mathbf{U}_1) = 0$ by definition. Therefore a shortcut is : if $\sigma(\theta) = 0$, then physical heating vanishes.

Mathematical picture : On the other hand it is also clear that the function $\tau^{\theta,+}$ is upper bounded by virtue of the analysis provided in section 4.8.1. A consequence of the transversality condition is the fact that $V_2^{\theta,\mu}$ is uniformly bounded, see (4.71). So if $\sigma(\theta) = 0$, it is possible that $\lim_{\mu\to 0} \left|V_2^{\theta,\mu}\right| = \infty$. In this case $|\tau(\theta)| = \infty$ which yields in turn once again that the associated heating vanishes in (4.90).

In summary it is possible to conjecture that (4.89) is still valid even if the transversality condition is wrong : in this case the heating associated to the Fourier mode vanishes. The limit of (4.88) may a priori be more singular. More research is nevertheless needed to provide a rigorous basis to this analysis.

4.9 An eigenvalue problem

In the case $\theta = 0$, it is possible to get an interpretation of the hypothesis (H5) as an eigenvalue problem. Consider the simple case of the X mode equation for γ constant and

$$\alpha(x, y) = \begin{cases} -x \ \forall x \leq H, \\ -H \ \forall x > H, \end{cases}$$

so that a solution E of the corresponding X mode equation satisfies

$$\begin{cases} -E_y'' + \eta E_y &= 0. \\ i\gamma E_y &= \alpha E_x \end{cases}$$

with $\eta = \frac{\gamma^2}{\alpha} - \alpha$.

Definition 7. A solution of the X mode equation will be called *smooth at zero* if there is a constant C independent of μ such that $\|E_x\|_\infty \leq C$.

A solution of the X mode equation will be called *smooth at infinity* if there is a constant C independent of μ such that $E_y = Ce^{-\sqrt{\eta}x}$ for all $x > H$.

Because of this definition, a solution smooth at zero satisfies $E_y(0) = 0$ whereas a solution smooth at infinity satisfies $E_y'(H) + \sqrt{\eta}E_y(H) = 0$.

It has been shown that the so-called first solution $\mathbf{U}_1$ is smooth at zero while the second solution $\mathbf{U}_2$ is smooth at infinity, assuming that (H5) that reads here $4\gamma H < 1$. The question at stake is whether it is possible to find a solution that would be at the same time smooth at zero and at infinity. That would necessarily imply violating hypothesis (H5).

Consider a solution $u = E_y$ of the following problem on $\Omega =]0, H[$

$$\begin{cases} -u''(x) - \alpha(x)u(x) = \dfrac{\gamma^2}{x}u(x) \\ u(0) = 0 \\ u'(H) + \sqrt{\eta(H)}u(H) = 0 \end{cases} \tag{4.91}$$

Definition 8. Let $\mathcal{X}$ be the function space $\{v \in H^1(\Omega), v(0) = 0\}$ equipped with the norm $\|v\|^2_{\mathcal{X}} = \int_0^H |v'(x)|^2 dx$ and $\mathcal{H}$ be the Hilbert space

$$\mathcal{H} = \left\{ v \in L^2(\Omega), \int_0^H \frac{|v|^2}{x} dx < \infty \right\}.$$

The classical variational formulation for (4.91) reads

$$\int_0^H (u'\overline{v}' - \alpha u\overline{v})dx + \sqrt{\eta(H)}u(H)\overline{v}(H) = \gamma^2 \int_0^H \frac{u\overline{v}}{x} dx \ \forall v \in \mathcal{X},$$

or

$$a(u, v) = \gamma^2 (u, v)_{\mathcal{H}} \ \forall v \in \mathcal{X}.$$

where a is the sesquilinear form defined as the right hand side of the variational formulation and $(,)_H$ is the natural weighted scalar product on $\mathcal{H}$. This will now be considered as an eigenvalue problem, γ^2 playing the part of the eigenvalue.

Definition 9. The operator T is defined on $\mathcal{H}$ such that $a(T(u), v) = (u, v)_{\mathcal{H}}$. Its image is included in $\mathcal{X}$.

Lemma 4.27. *The operator T is a self adjoint and compact operator from $\mathcal{H}$ to $\mathcal{H}$.*

Proof. Since there is a constant C such that for all $v \in \mathcal{H}$

$$\int_0^H \frac{|v|^2}{x} dx \leq C \int_0^H \frac{|v|^2}{x^2} dx,$$

Hardy inequality implies that $\mathcal{X} \subset \mathcal{H}$, so that T is indeed defined from $\mathcal{H}$ to $\mathcal{H}$.

Since a is sesquilinear, one has for all $(u, v) \in \mathcal{H}^2$

$$(u, T(v))_{\mathcal{H}} = a(T(u), T(v)) = \overline{a(T(v), T(u))} = \overline{(v, T(u))_{\mathcal{H}}},$$

so that T is self adjoint.

Then consider a sequence v_n in $\mathcal{X}$ such that there is a bound C such that for all $n \in \mathbb{N}$: $\|v_n\| \leq C$. Because H^1 is compactly embedded in L^2 there is $w \in H^1$ such that

$$\begin{cases} v_n \to w \text{ weakly in } H^1, \\ v_n \to w \text{ strongly in } L^2. \end{cases}$$

Moreover for all $\epsilon > 0$

$$\begin{cases} \|v_n - w\|_{\mathcal{H}} & \leq \epsilon \displaystyle\int_0^\epsilon \frac{|v_n - w|^2}{x^2} + \frac{1}{\epsilon^2}\|v_n - w\|^2_{L^2}, \\ & \leq C\epsilon + \frac{1}{\epsilon^2}\|v_n - w\|^2_{L^2}. \end{cases}$$

For ϵ small enough the first term is as smaller than half any positive number, and thanks to the strong convergence in L^2 for n high enough the second is as smaller than half any positive number as well. In other words v_n converges strongly to w in $\mathcal{H}$. □

Proposition 4.28. *There are a sequence ζ_n in $\mathbb{R}^*_+$ and a basis u_n of $\mathcal{H}$ such that $\zeta_n \to \infty$ and for all $n \in \mathbb{N}$ u_n is solution of the eigenvalue problem (4.91) for $\gamma^2 = \zeta_n$.*

Proof. This is the direct consequence of Lemma 4.27 that ensures the hypothesis of a classical spectral decomposition theorem for self-adjoint operators, see [Bré83]. □

As a result, the hypothesis (H5) implies that $1/(4H) \leq \zeta_1$.

4.9.1 Numerical approximation of the eigenvalues

It is interesting to confirm Proposition 4.28 with a numerical approximation of the eigenvalues ζ_n. The procedure implemented here is the following. We solve the problem (4.28)-(4.6) presented in Section 4.3.1 with the initial condition corresponding to the first basis function (V_1, W_1). The coefficients of the equation are set as $\alpha(x) = -x$ and γ constant. Then we vary γ in the interval $[0, 8]$, and compute $|V_1(H)|$ as a function of γ :

$$\phi(\gamma) = |V_1(H)|.$$

We expect ϕ to be small for $\gamma = \sqrt{\zeta_n}$ since the exact solution of the eigenvalue problem is exponentially decreasing for $x \geq H$. This establishes an indicator of the position of the eigenvalues $\sqrt{\zeta_n}$. The results for three different values of the regularization parameter μ are presented in Figure 4.14. The positions of the eigenvalues are observed at the minima of ϕ, except $\gamma = 0$. Indeed in the case $\gamma = 0$ the method computes the trivial solution $V = 0$ since the initial condition is $(0, 0)$, see (4.84).

4.10 Comments

The technique developed in this chapter is limited to the cases of coefficients that do not depend on the y variable since it starts with considering the Fourier transform of the initial system with respect to the y variable.

Embracing another point of view, it is possible to analyze this problem as a linear system with an isolated singularity, see [CL55]. The study of corresponding turning points for systems described in [CW74] is a work in progress with Olivier Lafitte and Bruno Després.

It is also important to emphasize that the singularity here is in the equation, unlike in some cases where the singularity appears from a corner of the domain.

It is different as well from the T coercivity which is studied in [BCC12, Cia12, Che12], where the singularity appears inside the domain.

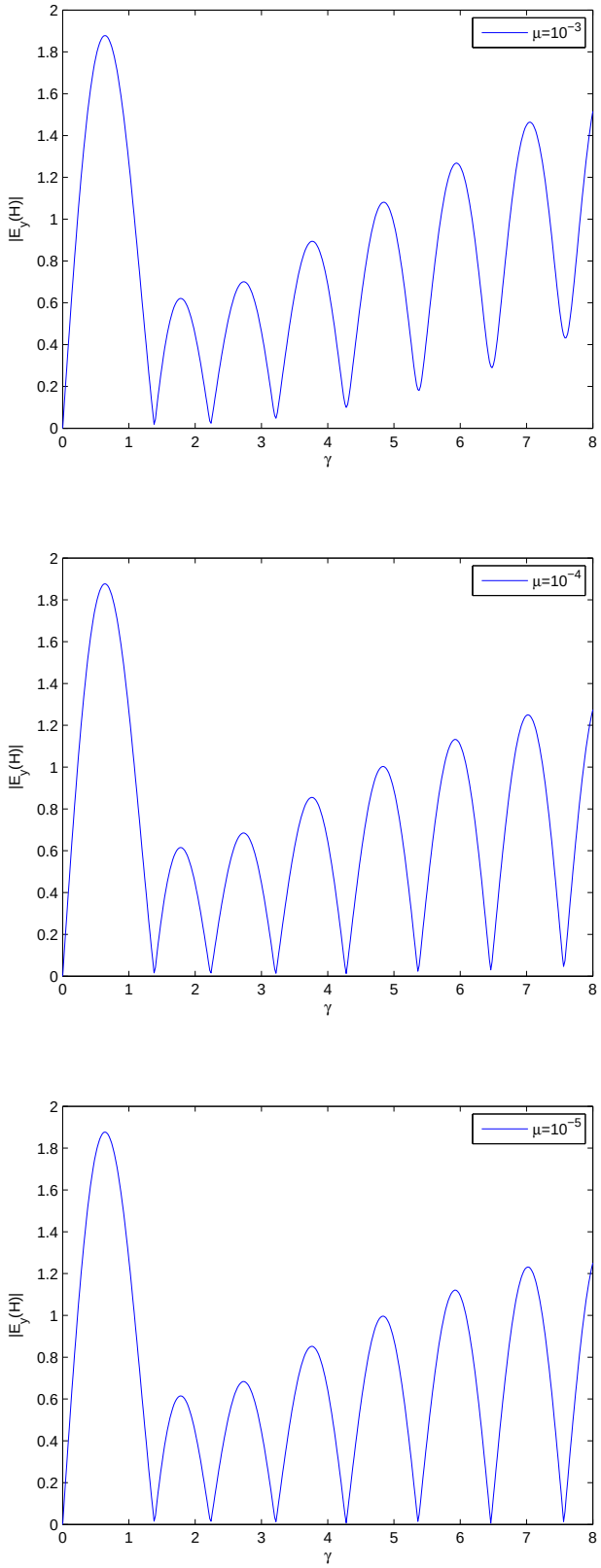


Figure 4.14 – For decreasing values of the regularization parameter μ, from top to bottom : 10^{-3}, 10^{-4} and 10^{-5}.

Part II

Numerical approximation

Chapter 5

Numerical approximation with generalized plane waves

Contents

5.1 Introduction

Wave problems include a large range of topics, such as elastic, acoustic and electromagnetic waves, bounded and unbounded domains, boundary conditions, isotropic, anisotropic, homogeneous and heterogeneous media, nonlinear waves, time and frequency domain. The numerical simulation of wave propagation has long interested the Applied Mathematics community. See for instance the book entitled *Finite element methods for*

Maxwell's equations by P. Monk, [Mon03]. Concerning reflectometry and heating applications, different numerical methods are already implemented, among which finite differences are used by Stéphane Heuraux, see [KGH09, HdSG+11], and finite elements are used by Simon Labrunie, see [Lab] and by Rémi Dumont, see [Dum09].

A pollution effect may appear when using the finite element method for the Helmholtz equation, as evidenced by Babuska and Sauter in [BS97]. To overcome this numerical pollution and diminish the numerical burden, one possibility is to incorporate information about the problem in the basis functions, as in the so called Trefftz methods, see [PHVD07] and [GHP09, HMP11]. The idea motivating these methods is to use the information given by the equation to design basis functions that are solutions of the homogeneous equation. In the case of Plane Wave methods, it aims at obtaining a more accurate approximation since the basis functions of the trial space contain information about the oscillatory behavior of the solution, information encoded in the wavenumber. A short bibliography of such methods follows, in an attempt of chronology.

The Ultra Weak Variational Formulation was introduced and then developed by B. Després and O. Cessenat starting from 1994 [Des94, CD98], for general linear problem. More details are to be found in the introduction of Section 5.2. The Partition of Unity Method was introduced in 1996 by I. Babuska and J.M. Melenk [MB96]. It was presented as a Finite Element Method based on conforming ansatz spaces, and offers a great flexibility in the choice of the local approximation. It was later developed to address scattering problems in [OS01, LBA02, PDLBT04]. The Discontinuous Enrichment Method uses a set of basis functions of polynomials enriched with Plane Waves. It was presented by C. Farhat, I. Harari and L. Franca in 2001 [FHF01] and some applications to the Helmholtz problem can be found in [FTWG04, TF06, ACR09]. The Plane Wave Discontinuous Galerkin method is a more general type of methods, that includes the UWVF. The analysis of its p and h versions, together with error estimates can be found in [BM08, GHP09, HMP11]. Overviews of some of these methods are proposed in [GGH11, WTTF12].

5.1.1 Notation and hypothesis

This chapter focuses on a new numerical method for the numerical approximation of the O-mode equation. The model problem is

$$\begin{cases} -\Delta u + \beta u &= f, & (\Omega), \\ (\partial_\nu + i\sigma)\, u &= Q\,(-\partial_\nu + i\sigma)\, u + g, & (\Gamma). \end{cases} \tag{5.1}$$

where β is a smooth function, vanishing on a hypersurface. Smooth is to be understood here as $C^r(\overline{\Omega})$, but it has to be noted that the more general case of a piecewise C^r coefficient - which would for instance model a lens in a propagative medium - would present no mathematical complication. The parameter r has to satisfy $r \geq 1$ for the theoretical results. The σ function can be a variable physical parameter satisfying $0 < \sigma_m \leq \sigma \leq \sigma_M$, but for the sake of simplicity we will consider it constant - and positive. The data g and f are L^2 functions respectively on the boundary and in the domain. Q is a piecewise constant function allowing to fit the condition : if $Q = -1$ it gives a Dirichlet condition, if $Q = 1$ a Neumann condition or if $Q = 0$ a Robin condition. However, the fact that $|Q| < 1$ on a non empty part of the boundary is mandatory to ensure the well-posedness of the problem.

The new method relies on a local approximation of the coefficient β, described by a parameter q. The procedure to design the basis functions requires that $r \geq q - 1$.

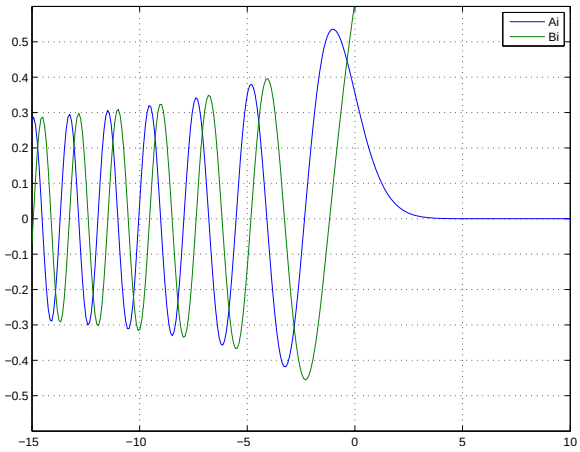


Figure 5.1 – Physical illustration of the change of sign in the coefficient of equation (5.2) : the first and second Airy functions.

5.1.2 A specificity

A classical wave propagation problem would correspond to $\beta > 0$. Yet the Ordinary propagation mode presents a cut-off : when β is either negative or positive the nature of the equation (5.1) is either elliptic coercive or elliptic propagative. Since the coefficient β is a continuous function, the transition between the two regimes is a hyper surface : the coefficient vanishes continuously in the domain at a point in dimension one, and along a line in dimension two. As a consequence the local wave number goes to zero toward the cut-off, and is equal to zero along the cut-off.

Such a continuous transition between elliptic coercive or elliptic propagative zones is less common in the framework of numerical simulation than a constant coefficient. In order to illustrate the specificity of the problem at stake, consider the further simplified 1D model

$$-\frac{d^2}{dx^2}u + xu = 0. \tag{5.2}$$

The fundamental solutions are the two Airy functions Ai and Bi. The first Airy function Ai represents the physical solution : this solution oscillates in the propagative zone $\beta(x) = x < 0$, and is absorbed in the non propagative zone $\beta(x) = x > 0$. One can observe on Figure 5.1 the cut-off point at $x = 0$ and the corresponding change of behavior.

Generally speaking, the most common model considered to approximate smooth coefficients is a piecewise constant approximation. Problems corresponding to wave transmission at an interface in time-harmonic regime with piecewise constant coefficients are for instance one main concern of A.-S. Bonnet-Bendhia, L. Chesnel, P. Ciarlet and X. Claeys, from the French ANR project Metamath. See for instance the recent thesis of Lucas [Che12].

The originality of the present work is to propose a high order local approximation of the coefficient. This approach gives an approximation tool not only in the vicinity of the cut-off, but also in the non propagative zone, which is not the case of many classical numerical methods adapted to wave propagation.

5.1.3 Plan

This chapter compiles two papers. The first one [IGD11] was a joint work with Bruno Després, and contains the 1D analysis of the method. It has been accepted for publication

in the IMA Journal of Numerical Analysis. The second one is in preparation and concerns the 2D properties of the generalized plane waves. It follows a question of Peter Monk about the interpolation properties of these new basis functions. The current work is an attempt to put into perspective the different features of the new numerical method.

5.2 A numerical method adapted to the O-mode equation

The UWVF alluded to in the introduction is the initial numerical method studied herein. A more detailed bibliography follows.

The UWVF was first introduced in a short note in 1994 by B. Després [Des94], as a method for generic linear problems with basis functions solutions of the homogeneous adjoint equation. The first application appears in the thesis of O. Cessenat [Ces96b], including Helmholtz problem in 2D and Maxwell's problem in 3D with plane wave basis functions. They also co-authored together a couple of papers on Helmholtz and acoustic problems, addressing numerical and theoretical convergence [CD98, Ces96b]. P. Monk and T. Huttunen starting from 2002 also co-authored, with different other collaborators, a series of papers more oriented toward computational aspects of the UWVF with plane wave basis functions, on Helmholtz's problems [HMK02], elastic waves [HMCK04], perfectly matching layers [HKM04], Maxwell's problems [HMM07], fluid-solid interaction [HKM08]. More recently they worked with T. Luostari on the use of Bessel basis functions [LHM12] and on linear elasticity [LHM13]. P. Monk also published with E. Darrigrand on the coupling of the UWVF with Fast Multipole methods [DM07, DM12] for integral representations. A more general point of view, deriving the UWVF as a Discontinuous Galerkin method was introduced by P. Monk and A. Buffa [BM08]. This perspective was followed by R. Hiptmair and his co-authors [GHP09, HMP11], developing explicit estimates thanks to Vekua theory.

Since explicit solutions of the homogeneous adjoint equation are not available for a general varying coefficient, new basis functions will be designed to adapt the UWVF for such a varying coefficient. Generalized plane wave basis functions are non-classical alternative functions, tailored for the model problem (5.1) including the cutoff regions.

5.2.1 The classical Ultra Weak Variational Formulation

Consider the initial problem (5.1), defined by the Helmholtz equation with a smooth coefficient β defined on a domain $\Omega \subset \mathbb{R}^d$, $d = 1$ or 2, together with a mixed boundary condition on the $\Gamma = \partial\Omega$.

The concern is not focused on the approximation of a smooth boundary by a refined mesh. As a consequence, for $d = 2$ the domain Ω considered here is a rectilinear polygon. This means first that it will be perfectly meshed. It also means that this domain has a Lipschitz boundary, so that the classical results for Lipschitz domains will hold.

The numerical approximation of a curved domain have been studied for instance in [CS12] to solve Dirichlet boundary-value problems for second-order elliptic equations, and in [CSS12] to solve second-order elliptic equations in exterior domains subject to a Dirichlet boundary condition on the interface of a scattering object.

Unlike a classical variational formulation, the UWVF requires the meshing of the domain as a preliminary step. The mesh domain, denoted $\mathcal{T}_h = \{\Omega_k\}_{k\in[\![1,N_h]\!]}$ where N_h is

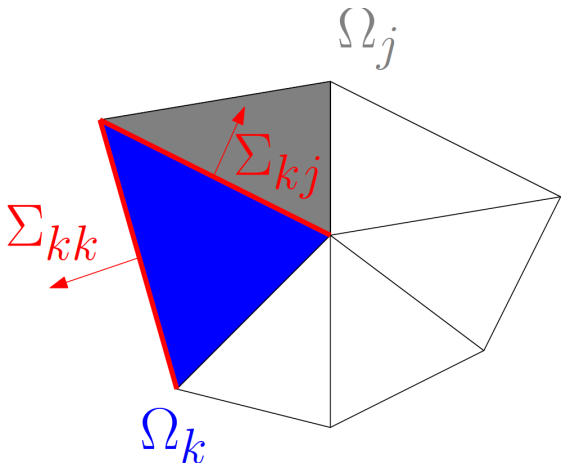


FIGURE 5.2 – An example of a simple domain meshed with $N_h = 6$ triangles. The outward normals of an element Ω_k are represented at the boundary of the domain Γ_k and at the interface Σ_{kj} with a neighboring element Ω_j.

the number of elements, is such that :

$$\begin{array}{rl} \overline{\Omega} & = \cup\overline{\Omega}_k, \Omega_k \cap \Omega_j = \emptyset \ \forall k \neq j, \\ \Gamma_k & = \overline{\Omega}_k \cap \Gamma, \text{ oriented from } \Omega_k \text{ toward the exterior of the domain,} \\ \Sigma_{kj} & = \overline{\Omega_k} \cap \overline{\Omega_j}, \text{ oriented from } \Omega_k \text{ to } \Omega_j, \\ \partial\Omega_k & = (\cup_j \Sigma_{kj}) \cup \Gamma_k. \end{array}$$

See Figure 5.2 as an illustration.

The function space for the UWVF is defined on the boundary of the elements of the mesh : in dimension one it is defined on the vertices of the elements, in dimension two it is defined on the edges of the elements.

Definition 10. The function space is denoted by V and defined as

$$V = \prod_{k \in [\![1, N_h]\!]} L^2(\partial\Omega_k).$$

It is equipped with the an Hermitian product

$$(X, Y) = \sum_k \int_{\partial\Omega_k} \frac{1}{\sigma} X_k \overline{Y_k} \quad \forall (X, Y) \in V,$$

and the corresponding norm :

$$\|X\| = \sqrt{(X, X)} \quad \forall X \in V.$$

The norm of an operator is

$$\|A\| = \sup_{X \neq 0} \frac{\|AX\|}{\|X\|} \quad \forall A \in \mathcal{L}(V).$$

Elements of V will be denoted by capital letters, whereas the corresponding functions defined on $\cup\Omega_k$ will be denoted by lowercase letters.

If $\Omega \subset \mathbb{R}$ the dimension of V is finite. It will simplify radically the one dimensional case. On the contrary, if $\Omega \subset \mathbb{R}^d$ with $d \geq 2$, the dimension of V is infinite.

Elementary properties of the trace operators

Since the function space is defined on the boundary of the mesh elements, it is useful to have in mind the definition of the traces in dimension one and two.

In dimension one, H^1 functions are actually continuous, so that the trace operator is trivially well-defined.

Theorem 5.2.1. Suppose Ω is a bounded interval in $\mathbb{R}$. Then H^1 is compactly embedded in $C(\overline{\Omega})$.

In dimension $d > 1$ the smoothness of the boundary of the domain is crucial. The case Ω bounded and $\partial\Omega$ C^1 is the most common case. One can cite two main results in this particular case. The Rellich Kondrachov compactness theorem only states the compact embedding of $W^{1,p}(\Omega)$ in $C(\overline{\Omega})$ for $p > d$, see [Bré83] (p.169). The trace theorem of [Eva10] (p.272) defines the trace as a bounded linear operator from $W^{1,p}(\Omega)$ to $L^p(\partial\Omega)$ for $1 \leq p < \infty$.

Some results with less restrictive hypothesis on the domain are given in [Gri85] in dimension $d = 2$. A precise result is stated therein in Theorem 1.5.2.3 (p.43) supposing Ω is a bounded open set of $\mathbb{R}^2$ whose boundary is a curvilinear polygon of class C^1. But in the scope of this work no hypothesis more restrictive than a Lipschitz boundary is necessary. The following result states Theorem 1.5.1.3 (p.38) with $d = 2$ and $p = 2$.

Theorem 5.2.2. Let Ω be a bounded open set of $\mathbb{R}^2$ with a Lipschitz boundary Γ. The the trace mapping which is defined for $u \in C^{0,1}(\overline{\Omega})$, has a unique continuous extension as an operator from $H^1(\Omega)$ onto $H^{1/2}(\Gamma)$.

Since all the functions u defined on a subset of $\mathbb{R}^d$ considered in this chapter are either in $H^1(\Omega)$ or in $\prod H^1(\Omega_k)$, the trace operator is well-defined. However there is a restriction to consider the quantities $(\pm\partial_\nu + i\sigma)u_{|\partial\Omega}$, because the derivatives might be less regular. A specific hypothesis is then needed to state the UWVF, see Theorem 5.2.4.

A preliminary weak result

It is useful to give a meaning to a more general problem (5.3) in a weak sense on a Lipschitz domain $\mathcal{O}$,

$$\begin{cases} -\Delta u + \beta u & = f, \quad (\mathcal{O}), \\ (\pm\partial_\nu + i\zeta)\, u & = \tilde{g}, \quad (\partial\mathcal{O}). \end{cases} \tag{5.3}$$

It will be used for two different purposes :

- Applied on the whole domain Ω with the boundary condition defined by

$$(\partial_\nu + i\zeta)\, u = \tilde{g}, \text{ with } \zeta = \sigma\frac{1-Q}{1+Q} \text{ and } \tilde{g} = \frac{g}{1+Q},$$

  it proves the existence of a unique solution to the initial problem (5.1). This solution is the one that is referred to in Theorem 5.2.4.

- Secondly, applied on each element on the mesh Ω_k with the boundary condition defined by

$$(-\partial_\nu + i\zeta)\, u = \tilde{g}, \text{ with } \zeta = \sigma,$$

  it provides a technical tool to define the UWVF. See Definition 12.

A classical result is :

Theorem 5.2.3. Denote by $\mathcal{O} \subset \Omega$ a bounded polygonal domain in $\mathbb{R}$ or $\mathbb{R}^2$. To define the boundary condition denote by f and g elements of $L^2(\mathcal{O})$ and $L^2(\partial\mathcal{O})$. Suppose that Q is a constant real number and $|Q| < 1$. Then there exists a unique solution $u \in H^1$ to the variational formulation corresponding to (5.3), i.e. such that

$$\int_{\mathcal{O}} \nabla u \cdot \overline{\nabla v} + \int_{\mathcal{O}} \beta u\overline{v} \pm i\zeta \int_{\partial\mathcal{O}} u\overline{v} = \int_{\mathcal{O}} f\overline{v} \pm \int_{\partial\mathcal{O}} \tilde{g}\overline{v}, \forall v \in H^1. \tag{5.4}$$

Moreover, there exists a constant C such that :

$$\|u\|_{L^2(\mathcal{O})} \leq C \left(\|f\|_{L^2(\mathcal{O})} + \|\tilde{g}\|_{L^2(\partial\mathcal{O})} \right). \tag{5.5}$$

The functional analysis details concerning the regularity of the boundary in dimension two are to be found in [Gri85], a book that focuses on non smooth domains.

Proof. This proof relies on classical methods for variational formulations. Let us introduce an intermediate problem

$$\begin{cases} -\Delta w + w & = \hat{f}, (\mathcal{O}), \\ (\pm\partial_\nu + i\zeta)\, w & = \hat{g}, (\partial\mathcal{O}). \end{cases} \tag{5.6}$$

Let a and l be the corresponding sesquilinear and anti linear forms, so that for any u and v in $H^1(\mathcal{O})$

$$a(u,v) = \int_{\mathcal{O}} \nabla u \cdot \overline{\nabla v} + \int_{\mathcal{O}} u\overline{v} \pm i\zeta \int_{\partial\mathcal{O}} u\overline{v}, \tag{5.7}$$

$$b(v) = \int_{\mathcal{O}} \hat{f}\overline{v} \pm \int_{\partial\mathcal{O}} \hat{g}\overline{v}. \tag{5.8}$$

As a is sesquilinear and continuous, b is antilinear continuous and $Re(a(v,v))$ is coercive, there exists a unique $u \in H^1$ such that

$$a(u,v) = l(v), \forall v \in H^1, \tag{5.9}$$

for any couple $(\hat{g}, \hat{f}) \in L^2(\mathcal{O}) \times L^2(\partial\mathcal{O})$. See [DL84] for this version of Lax-Milgram theorem. Then let us define the linear operator $\mathcal{A}$ by

$$\mathcal{A} : (\hat{f}, \hat{g}) \in L^2(\mathcal{O}) \times L^2(\partial\mathcal{O}) \mapsto u \in L^2(\mathcal{O}), \tag{5.10}$$

where u is the solution given by (5.9). Moreover, note that from a classical a priori estimate one has

$$\|u\|_{H^1} \leq \|\hat{f}\|_{L^2} + \|\hat{g}\|_{L^2(\partial\mathcal{O})}. \tag{5.11}$$

The operator A is compact since the injection of H^1 in L^2 is compact since Ω is bounded. This compact injection holds regardless of the dimension d, see [Nec67], and for even weaker hypothesis on the boundary, see [Ami78] (p.83). Remark that

$$u \text{ is solution of } (5.1) \Leftrightarrow u = \mathcal{A}\left((id - \beta)u + f, \tilde{g}\right), \tag{5.12}$$

$$\Leftrightarrow \left[I - A((id - \beta)\cdot, 0)\right] u = \mathcal{A}\left(f, \tilde{g}\right). \tag{5.13}$$

Since β is bounded, the operator $K := A((id - \beta)\cdot, 0))$ is also compact, and the Fredholm alternative holds, see [Bré83]. So uniqueness is equivalent to existence of a solution for the problem (5.4). Then suppose $u \in L^2$, actually also in H^1, is such that $(I - K)u = 0$, which means

$$\int_{\mathcal{O}} \nabla u \cdot \overline{\nabla v} + \int_{\mathcal{O}} \beta u\overline{v} \pm i\zeta \int_{\partial\mathcal{O}} u\overline{v} = 0, \forall v \in H^1. \tag{5.14}$$

Choosing $v = u$ as test function, and considering the imaginary part of (5.14) one gets that $u = 0$ on $\partial\mathcal{O}$. As $\mathcal{O}$ in bounded and as β is $C(\overline{\mathcal{O}})$, the method of translations presented in [Bré83] (p.182) and credited to L. Nirenberg can be adapted to prove that $u \in H^2$. Then an integration by part of the first term shows that $\partial_\nu u = 0$ on $\partial\mathcal{O}$ since u is solution of the homogeneous variational formulation (5.4). Here in dimension one the Cauchy-Lipschitz theorem gives that $u = 0$ in $\mathcal{O}$, but in dimension two a unique continuation theorem is required. An appropriate version can be found in [Hor76]. As a result, in both cases the solution is unique, and so, thanks to Fredholm alternative : there exists a unique solution to (5.1).

Besides, as a consequence of the open mapping theorem, if $(I - K)u = f$ there exists a constant C such that $\|u\|_{L^2} \leq C\|f\|_{L^2}$. So if u is solution of (5.1), then

$$\|u\|_{L^2} \leq C \left\| \mathcal{A}\left(f, \frac{1}{1+Q}g\right) \right\|_{L^2}, \tag{5.15}$$

so that thanks to (5.11) it yields :

$$\|u\|_{L^2(\mathcal{O})} \leq C \left(\|f\|_{L^2(\mathcal{O})} + \left\| \frac{1}{1+Q}g \right\|_{L^2(\partial\mathcal{O})} \right). \tag{5.16}$$

□

A recent work presents a completely different method based on a coercive formulation of the Helmholtz equation, see [MS]. Since it is restricted to constant coefficients, this approach does not apply to the current problem.

The standard formulation

The UWVF is a reformulation of the initial problem that considers functions defined on the boundary of the mesh elements, and couples the different contributions from the different mesh elements through fluxes at the interfaces.

Note that the definition of the formulation itself does not require to explicit the shape functions. Instead, as in the case of a classical variational formulation, the trial space has a infinite dimension. However the trial space introduced to derive the UWVF is more specific than the classical Sobolev spaces used to derive classical variational formulations. It is indeed the function space of local solution of the homogeneous equation.

Definition 11. For all $k \in \mathbb{N}^*$ such that $1 \leq k \leq N_h$ define the local trial space

$$H_k(\beta) = \left\{ v_k \in H^1(\Omega_k), \left| \begin{array}{l} (-\Delta + \beta)v_k = 0 \ (\Omega_k), \\ ((-\partial_\nu + i\sigma)v_k)_{|\partial\Omega_k} \in L^2(\partial\Omega_k) \end{array} \right. \right\} \tag{5.17}$$

and the corresponding global trial space $H = \prod_{k=1}^{N_h} H_k(\beta)$.

As a consequence, the derivation of the UWVF for the initial problem (5.1) is stated as follows.

Theorem 5.2.4. Let $u \in H^1(\Omega)$ be the unique solution of problem (5.1)such that $\partial_{\nu_k} u \in L^2(\partial\Omega_k)$ for any k. Let $\sigma, Q > 0$ be given real numbers. Then $X \in V$ defined by $X_{|\partial\Omega_k} = X_k$

with $X_k = ((-\partial_\nu + i\sigma)u_{|\Omega_k})_{|\partial\Omega_k}$ satisfies

$$\begin{aligned}&\sum_k \left(\int_{\partial\Omega_k} \frac{1}{\sigma} X_k \overline{(-\partial_\nu + i\sigma)e_k} - \sum_{j,j\neq k} \int_{\Sigma_{kj}} \frac{1}{\sigma} X_j \overline{(\partial_\nu + i\sigma)e_k} \right) \\ &- \sum_{k,\Gamma_k\neq\emptyset} \int_{\Gamma_k} \frac{Q}{\sigma} X_k \overline{(\partial_\nu + i\sigma)e_k} = -2i \sum_k \int_{\Omega_k} f\overline{e} + \sum_k \int_{\Gamma_k} \frac{1}{\sigma} g \overline{(\partial_\nu + i\sigma)e_k},\end{aligned} \tag{5.18}$$

for any $e = (e_k)_{k\in[\![1,N_h]\!]} \in H$. Conversely, if $X \in V$ is solution of (5.18) then the function, u defined locally by

$$\begin{cases} u_{|\Omega_k} = u_k \in H^1(\Omega_k), \\ (-\Delta + \beta)u_k = f_{|\Omega_k}, \\ (-\partial_{\nu_k} + i\sigma)u_k = X_k, \end{cases} \tag{5.19}$$

is the unique solution of the problem (5.1).

This result is classical in the context of UWVF. For the reason quoted earlier, adapting the proof to the case of a general smooth coefficient β regardless its sign is straightforward.

Proof. Consider $e \in H$ and $u \in H^1(\Omega)$ solution (5.1). A simple computation shows that for a given $k \in [\![1, N_h]\!]$

$$\begin{aligned}&\int_{\partial\Omega_k} \frac{1}{\gamma}(-\partial_\nu + i\sigma)u \cdot \overline{(-\partial_\nu + i\sigma)e_k} - \int_{\partial\Omega_k} \frac{1}{\gamma}(\partial_\nu + i\sigma)u \cdot \overline{(\partial_\nu + i\sigma)e_k} \\ &\qquad = -2i \int_{\partial\Omega_k} (u\overline{\partial_\nu e_k} - \partial_\nu u \overline{e}_k).\end{aligned} \tag{5.20}$$

Moreover the definition of H, together with the initial problem (5.1), yields

$$\begin{cases} (-\Delta + \beta)u = f, & (\Omega_k), \\ (-\Delta + \beta)e_k = 0, & (\Omega_k). \end{cases} \tag{5.21}$$

Performing two integrations by part - justified on a Lipschitz bounded domain in [Gri85] (p.52) - the following holds for all $k \in [\![1, N_h]\!]$

$$\begin{cases} \int_{\Omega_k} \nabla u \cdot \nabla\overline{e_k} + \int_{\Omega_k} \beta u \cdot \overline{e_k} - \int_{\partial\Omega_k} \partial_\nu u \cdot \overline{e_k} &= \int_{\Omega_k} f \cdot \overline{e_k}, \\ \int_{\Omega_k} \nabla u \cdot \nabla\overline{e_k} + \int_{\Omega_k} \beta u \cdot \overline{e_k} - \int_{\partial\Omega_k} u \cdot \overline{\partial_\nu e_k} &= 0. \end{cases}$$

So using the boundary conditions together with the smoothness of the solution u, namely for all $k \in [\![1, N_h]\!]$

$$\begin{cases} (\partial_\nu + i\sigma)u_{|\Sigma_{kj}} &= (-\partial_\nu + i\sigma)u_{|\Sigma_{jk}}, \\ (\partial_\nu + i\sigma)u_{|\Gamma_k} &= Q(-\partial_\nu + i\sigma)u_{|\Gamma_k} + g, \end{cases} \tag{5.22}$$

from the identity (5.20) stems for all $k \in [\![1, N_h]\!]$

$$\begin{aligned}&\left(\int_{\partial\Omega_k} \frac{1}{\sigma} X_k \overline{(-\partial_\nu + i\sigma)e_k} - \sum_{j,j\neq k} \int_{\Sigma_{kj}} \frac{1}{\sigma} X_j \overline{(\partial_\nu + i\sigma)e_k} \right) \\ &-\mathbf{1}_{\Gamma_k\neq\emptyset} \int_{\Gamma_k} \frac{Q}{\sigma} X_k \overline{(\partial_\nu + i\sigma)e_k} \\ &= -2i \int_{\partial\Omega_k} f\overline{e} + \int_{\Gamma_k} \frac{1}{\sigma} g \overline{(\partial_\nu + i\sigma)e_k}.\end{aligned}$$

Summing over k then gives the UWVF (5.18).

Conversely, let X be a solution of (5.18) and let u satisfy (5.19) on every Ω_k. The hypothesis on u and e, gives (5.21) and then for all $k \in [\![1, N_h]\!]$

$$\int_{\partial\Omega_k} \frac{1}{\sigma}(-\partial_\nu + i\sigma)u \cdot \overline{(-\partial_\nu + i\sigma)e_k} - \int_{\partial\Omega_k} \frac{1}{\sigma}(\partial_\nu + i\sigma)u \cdot \overline{(\partial_\nu + i\sigma)e_k} = -2i \int_{\Omega_k} f\overline{e_k}.$$

Summing over k and combining the result with (5.18) satisfied by X one gets for all $e = (e_k) \in H$

$$\begin{aligned}&\sum_{k,j\neq k} \int_{\Sigma_{kj}} \frac{1}{\sigma} X_k \cdot \overline{(\partial_\nu + i\sigma)e_k} + \sum_{k,\Gamma_k\neq\emptyset} \int_{\Gamma_k} \frac{1}{\sigma} X_k \cdot \overline{(\partial_\nu + i\sigma)e_k} \\ &= \sum_{k,j\neq k} \int_{\Sigma_{kj}} \frac{1}{\sigma} X_j \cdot \overline{(\partial_\nu + i\sigma)e_k} + \sum_{k,\Gamma_k\neq\emptyset} \int_{\Gamma_k} \frac{1}{\sigma}(QX_k + g) \cdot \overline{(\partial_\nu + i\sigma)e_k}.\end{aligned}$$

Therefore u satisfies (5.22). It shows that u is the unique smooth solution of (5.1) given by Theorem 5.2.3. □

In order to give a more compact formulation useful for further developments, some definitions are required.

Definition 12. For any function $f \in L^2(\Omega)$, let E_f be the extension mapping defined by :

$$E_f : \left\{ \begin{array}{ccc} V & \to & H, \\ Z & \mapsto & e = (e_k)_{k\in[\![1,N_h]\!]}, \end{array} \right.$$

where e is defined $\forall k \in [\![1, N_h]\!]$ by the unique solution of the following problem :

$$\left\{ \begin{array}{rl} (-\Delta + \beta)e_k & = f \quad (\Omega_k), \\ (-\partial_{\nu_k} + i\sigma)e_k & = Z_k \quad (\partial\Omega_k). \end{array} \right.$$

Also define E which is the homogeneous extension operator with vanishing right hand side, namely $E = E_0$.

Notice that E_f is well defined, see 5.2.1.

Remark 10. The operator E is the inverse of the trace operator defined on $H \subset \prod H^1(\Omega_k)$ as $u \mapsto ((-\partial_{\nu_k} + i\sigma)u_k)$, into the subset of V defined by $V_H = \{(-\partial_\nu + i\sigma)(v_k)_{|\partial\Omega_k}, v_k \in H_k\}$. This will provide a convenient tool for forthcoming computations : every basis function can be considered either as a set of functions defined in the volumes Ω_k, i.e. an element of V_H, or a set of functions defined on the boundaries $\partial\Omega_k$, i.e. an element on H. See figure 5.3, and the proof of Proposition 5.4 for an illustration.

Definition 13. Let F be the mapping defined by

$$F : \left\{ \begin{array}{ccc} V & \to & V, \\ Z & \mapsto & \left((\partial_\nu + i\sigma)E(z)_{|\partial\Omega_k}\right)_{k\in[\![1,N_h]\!]}. \end{array} \right.$$

This operator F relates the outgoing and incoming traces on the boundaries $\partial\Omega_k$.

Definition 14. Let Π be the mapping defined by

$$\Pi : \left\{ \begin{array}{ccc} V & \to & V, \\ Z_{|\Sigma_{kj}} & \mapsto & Z_{|\Sigma_{jk}}, \\ Z_{|\Gamma_k} & \mapsto & QZ_{|\Gamma_k}. \end{array} \right.$$

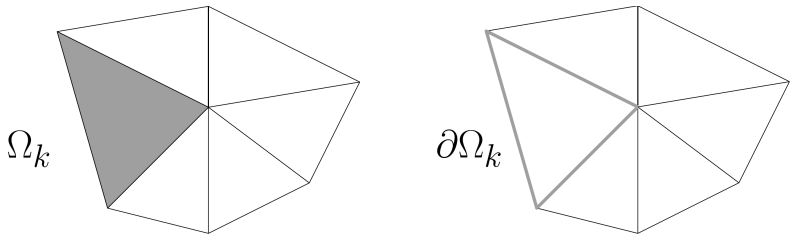


FIGURE 5.3 – A correspondence between H and V_H.

This operator Π changes the incoming trace at an interface Σ_{kj} to the outgoing trace toward the corresponding neighbor Ω_j.

Definition 15. If F^* denotes the adjoint operator of the operator F, let A be the operator $F^*\Pi$.

With this notation the problem (5.18) is equivalent [CD98] to

$$\begin{cases} \text{Find } X \in V \text{ such that } \forall Y \in V \\ (X,Y) - (\Pi X, FY) = (B,Y), \end{cases} \tag{5.23}$$

where the right hand side $B \in V$ is given by the Riesz theorem

$$(B,Y) = -2i \int_\Omega f\overline{E(Y)} + \int_\sigma \frac{1}{\sigma} g\overline{F(Y)} \qquad \forall Y \in V.$$

More precisely

- If u is solution of the initial problem (5.1) such that

$$\left((-\partial_\nu + i\sigma)u_{|\partial\Omega_k}\right)_{k\in[\![1,N_h]\!]} \in V,$$

  then $X = \left((-\partial_\nu + i\sigma)u_{|\partial\Omega_k}\right)_{k\in[\![1,N_h]\!]}$ is solution in V of (5.23).
- Conversely if X is solution of (5.23) then $u = E_f(X)$ is the unique solution of (5.18). The problem (5.23) is equivalent to

$$\begin{cases} \text{For } B \in V, \text{ find } X \in V \\ (I - A)X = B. \end{cases} \tag{5.24}$$

Some basic properties of these operators follow. Again this is classic in the UWVF literature.

Lemma 5.1. *Assume the boundary coefficient is such that* $|Q| \leq 1$. *The operator* Π *satisfies* $\|\Pi\| \leq 1$.

Proof. This is obvious from the definition of Π. □

Lemma 5.2. *The operator* F *is an isometry.*

Proof. For any $Y \in V$, let $e \in H$ be $E(Y)$. Then

$$\begin{aligned}\|FY\|^2 &= \sum_{k\in[\![1,N_h]\!]} \int_{\partial\Omega_k} \frac{1}{\sigma}|(\partial_\nu + i\sigma)e_k|^2, \\ &= \sum_{k\in[\![1,N_h]\!]} \int_{\partial\Omega_k} \frac{1}{\sigma}|\partial_\nu e_k|^2 + \sigma|e_k|^2 + 2\Im(\partial_\nu e_k \cdot \overline{e}_k),\end{aligned}$$

$$\begin{aligned}\|Y\|^2 &= \sum_{k\in[\![1,N_h]\!]} \int_{\partial\Omega_k} \frac{1}{\sigma}|(-\partial_\nu + i\sigma)e_k|^2, \\ &= \sum_{k\in[\![1,N_h]\!]} \int_{\partial\Omega_k} \frac{1}{\sigma}|\partial_\nu e_k|^2 + \sigma|e_k|^2 - 2\Im(\partial_\nu e_k \cdot \overline{e}_k).\end{aligned}$$

Integration by parts are justified because the domain has a Lipschitz boundary, see [Gri85] (p.58). Since $\int_{\partial\Omega_k} \partial_\nu e_k \cdot \overline{e}_k = \int_{\Omega_k} |\nabla e_k|^2 + \beta|e_k|^2 \in \mathbb{R}$, one gets that $\|FY\|^2 = \|Y\|^2$. This implies the result. □

Proposition 5.3. *The operator $A = F^*\Pi$ satisfies $\|A\| \leq 1$.*

Proof. It is a direct consequence of Lemmas 5.1 and 5.2. □

This operator also satisfies the following property.

Proposition 5.4. *The operator $I - A$ is injective.*

Proof. Let $X \in V$ such that $(I - A)X = 0$, which means $X = F^*\Pi X$. Define $Z \in V$ such that $Z = \Pi X$, then $F^*Z = X$ so that $\Pi F^* Z = Z$. Then define $u = (u_k)_{1\leq k\leq N_h} \in H$ such that for all $k \in [\![1, N_h]\!]$

$$\left\{\begin{array}{rll} -\Delta u_k + \beta u_k &= 0, & (\Omega_k), \\ (\partial_\nu + i\sigma)u_k &= Z_{|\partial\Omega_k}, & (\partial\Omega_k). \end{array}\right. \tag{5.25}$$

In order to identify F^*Z, define $Y \in V$ such that

$$\forall k \in [\![1, N_h]\!], Y_k = (-\partial_\nu + i\sigma)u_k.$$

It is known that for all $W \in V$, there exists $w = (w_k)_{1\leq k\leq N_h} \in H$ such that $w = E(W)$, which means w satisfies for all $k \in \mathbb{N}$ such that $1 \leq k \leq N_h$

$$\left\{\begin{array}{rll} -\Delta w_k + \beta w_k &= 0, & (\Omega_k), \\ (-\partial_\nu + i\sigma)w_k &= W_k, & (\partial\Omega_k). \end{array}\right. \tag{5.26}$$

Then

$$\begin{aligned}(Y, W) &= \sum_{k\in[\![1,N_h]\!]} \int_{\partial\Omega_k} \frac{1}{\sigma}(-\partial_\nu + i\sigma)u_k \cdot \overline{(-\partial_\nu + i\sigma)w_k}, \\ &= \sum_{k\in[\![1,N_h]\!]} \int_{\partial\Omega_k} \frac{1}{\sigma}\partial_\nu u_k \cdot \partial_\nu \overline{w_k} + \sigma u_k \cdot \overline{w_k} + i\partial_\nu u_k \cdot \overline{w_k} - iu \cdot \partial_\nu \overline{w_k},\end{aligned}$$

$$\begin{aligned}(Z, FV) &= \sum_{k\in[\![1,N_h]\!]} \int_{\partial\Omega_k} \frac{1}{\sigma}(\partial_\nu + i\sigma)u_k \cdot \overline{(\partial_\nu + i\sigma)w_k}, \\ &= \sum_{k\in[\![1,N_h]\!]} \int_{\partial\Omega_k} \frac{1}{\sigma}\partial_\nu u_k \cdot \partial_\nu \overline{w_k} + \sigma u_k \cdot \overline{w_k} - i\partial_\nu u_k \cdot \overline{w_k} + iu_k \cdot \partial_\nu \overline{w_k}.\end{aligned}$$

On the other hand, from (5.25) and (5.26) for all $k \in [\![1, N_h]\!]$

$$\left\{ \begin{array}{ll} \int_{\partial\Omega_k} \partial_\nu u_k \cdot \overline{w_k} & = \int_{\Omega_k} \nabla u_k \cdot \nabla \overline{w_k} + \int_{\Omega_k} \beta u_k \cdot \overline{w_k}, \\ \int_{\partial\Omega_k} u_k \cdot \partial_\nu \overline{w_k} & = \int_{\Omega_k} \nabla u_k \cdot \nabla \overline{w_k} + \int_{\Omega_k} \beta u_k \cdot \overline{w_k}, \end{array} \right.$$

so that $\int_{\partial\Omega_k} -\partial_\nu u_k \cdot \overline{w_k} + u_k \cdot \partial_\nu \overline{w_k} = 0$. As a consequence $\forall W \in V, (Y, W) = (Z, FW)$, which exactly means that $Y = F^* Z$. Since $\Pi F^* Z = Z$, it leads to $\Pi Y = Z$.

To conclude let's read this last equation in terms of the functions u_k defined in (5.25).

$$\forall (k,j) \in [\![1, N_h]\!]^2, \left\{ \begin{array}{ll} (-\partial_\nu + i\sigma)\,(u_k)_{|\Sigma_{jk}} & = (\partial_\nu + i\sigma)\,(u_k)_{|\Sigma_{kj}}\,, \\ Q(-\partial_\nu + i\sigma)\,(u_k)_{|\Gamma_k} & = (\partial_\nu + i\sigma)\,(u_k)_{|\Gamma_k}\,, \end{array} \right.$$

so that both u and $\partial_\nu u$ are continuous along every interface Σ_{kj}, and now

$$\left\{ \begin{array}{lll} -\Delta u + \beta u & = 0, & (\Omega), \\ (\partial_\nu + i\sigma) u & = Q(-\partial_\nu + i\sigma) u, & (\sigma). \end{array} \right.$$

Then u is the unique solution of the corresponding problem (5.1) provided by Theorem 5.2.3 : it is the 0 solution. Then $Z = 0$, and so $X = 0$. This ends the proof. □

An abstract discretization procedure

The next step consists in the discretization of Equation (5.23). This could be treated thanks to a standard Galerkin method : consider a subspace $V_h \subset V$ with finite dimension, and seek the discrete solution $X_h \in V_h$ such that

$$\forall Y_h \in V_h, (X_h, Y_h) - (\Pi X_h, F Y_h) = (b, Y_h). \tag{5.27}$$

Before describing in the next section a strategy to make such a Galerkin method effective, this section shows that the Galerkin approach (5.27) yields a well posed discrete problem. The following analysis of this well known fact is slightly different from what can be found in the literature [Des94, CD98, GHP09, BM08, HMP11, HMP13].

Definition 16. Consider the operator A introduced in Definition 15. Define the norm $|||U||| = \|(I - A)U\|$ for all $U \in V$, and the bilinear form of the formulation (5.23) : $a(X, Y) = (X, Y) - (\Pi X, FY)$.

Since $I - A$ is injective, $||| \cdot |||$ is indeed a norm. Two fundamental properties are coercivity and bicontinuity.

Lemma 5.5. *The bilinear form is coercive with respect to the norm* $||| \cdot |||$

$$|||X|||^2 \leq 2\Re\,(a(X, X)) \quad \forall X \in V,$$

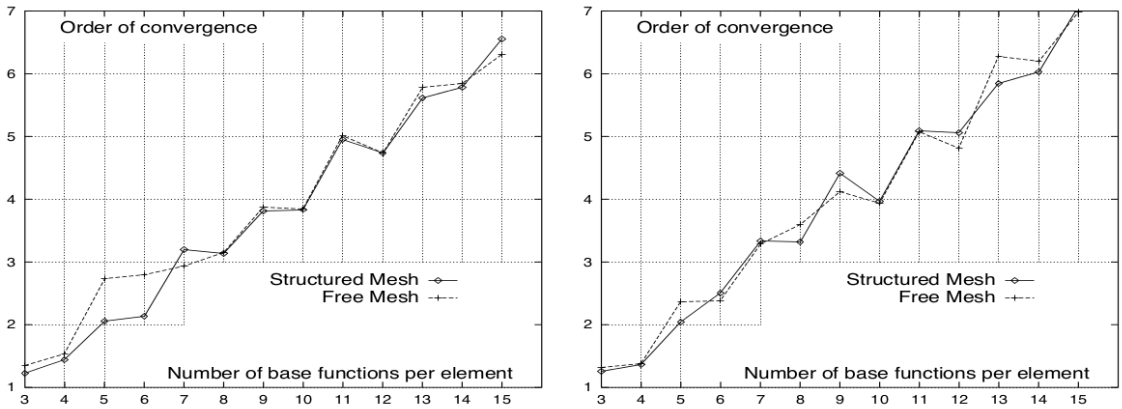
and is bicontinuous in the sense that

$$|a(X, Y)| \leq |||X||| \times \|Y\| \quad \forall X, Y \in V.$$

*Proof.*Consider X and Y in V. One has by definition $|||X|||^2 = \|X\|^2 + \|AX\|^2 - 2\Re(X, AX)$. Since $\|A\| \leq 1$ then

$$|||X|||^2 \leq 2\left(\|X\|^2 - \Re(X, AX)_V\right) = 2\Re\left((I - A)\,X, X\right)_V = 2\Re\left(a(X, X)\right).$$

The coercivity is proved. The skewed bicontinuity is evident from Cauchy-Schwarz inequality applied to $a(X, Y) = ((I - A)\,X, Y)$. □



Norme sur V Norme $L^2(\Gamma)$

FIGURE 5.4 – Example of numerical orders of convergence with respect to p the local number of basis functions per element, computed for two different norms. From Cessenat's thesis [Ces96b] paragraph I.3.2.3.

Proposition 5.6. *Assume there exists X solution of the problem (5.24). Then any solution X_h of the discrete problem* (5.27) *satisfies the inequality*

$$|||X - X_h||| \leq 2 \inf_{Z_h \in V_h} \|X - Z_h\|. \tag{5.28}$$

Proof. By construction $a(X - X_h, Y_h) = 0 \ \forall Y_h \in V_h$. So

$$a(X - X_h, X - X_h) = a(X - X_h, X - Z_h) \quad \text{with } Z_h = Y_h - Z_h.$$

It ends the proof with the coercivity and skewed bicontinuity of lemma 5.5. □

Lemma 5.7. *For all $B \in V$, the discrete solution X_h exists and is unique.*

Proof. If X_h exists, it is solution of a linear system, the dimension of the system being the dimension of the discrete subspace V_h. Therefore it is sufficient to check that if $a(X_h, Y_h) = 0$ for all $Y_h \in V_h$, then $X_h = 0$. Apply the inequality (5.28) with the choice $X = B = 0$. It yields $\|X_h\| \leq 2 \inf_{Z_h \in V_h} \|Z_h\| = 0$. □

Note that this property means that the discrete problem (5.27) is well-posed with no restriction on the size of the mesh. As a direct application, one can think of the p convergence of the method : instead of refining the mesh, convergence is achieved increasing the number of basis functions on each element. This has been studied for instance in [Ces96b] from the numerical point of view, see an example in Figure 5.4, and in [HMP11] with a Discontinuous Galerkin approach.

5.2.2 A method adapted to vanishing coefficients

Until that point no significant difference has to be noted between the classical UWVF for constant coefficients and the case of a general smooth coefficient. The discretization procedure requires the explicit definition of a finite dimension trial space of homogeneous solutions of the initial equation. For constant coefficients, such solutions have mainly been chosen as classical plane waves in [CD98, HMK02], but generalized harmonic polynomials are studied with the Vekua transform proposes in [HMP11] and more recently Bessel basis functions have been considered in [LHM12]. However it turns out that none of the corresponding trial spaces could be used in the case of an analytic vanishing coefficient. As a result the definition of an adapted trial space is mandatory to obtain a new formulation.

The adapted basis functions

The technical reason for the need of adapted test functions is the following. The definition of the operator F is linked to the functional space H through the operator E. Then in order to compute the $(\Pi\cdot, F\cdot)$ term in (5.27), some local solutions of the homogeneous equation would be needed. But such exact solutions are not available. A constructive procedure to design the basis functions to generate the discrete trial space follows.

Shape functions φ are sought as solutions of the homogeneous equation

$$(-\Delta + \beta)\varphi = 0, \tag{5.29}$$

locally on Ω_k. For a constant and negative coefficient β in the cell, it is sufficient to use plane waves, that is

$$\begin{cases} \varphi(x) &= e^{\sqrt{\beta}x} \text{ in dimension one,} \\ \varphi(x,y) &= e^{\sqrt{\beta}(\kappa_1 x + \kappa_2 y)} \text{ in dimension two,} \end{cases} \tag{5.30}$$

with $\kappa_1^2 + \kappa_2^2 = 1$. In dimension two, if the vector (κ_1, κ_2) is real, it simply specifies the direction of the wave. This is the basic idea of all plane wave methods. If $\beta = x$ is linear with respect to the first variable, it is possible to construct φ from the Airy functions Ai and Bi. But these are highly transcendental. However if β is non constant in the cell, then no simple and general analytic formula holds for φ.

By comparison with the plane wave formula (5.30) it is natural to consider a generalized plane wave such as $\varphi = e^P$, P being a polynomial. One gets that φ is solution of (5.29) if and only if

$$\begin{cases} \partial_x^2 P + (\partial_x P)^2 &= \beta(x) \text{ in dimension one,} \\ \partial_x^2 P + (\partial_x P)^2 + \partial_y^2 P + (\partial_y P)^2 &= \beta(x,y) \text{ in dimension two.} \end{cases} \tag{5.31}$$

However many attempts showed that such a representation is not sufficient. The explanation is simple, and here is an illustration dimension two : as P can be expanded as a finite series of monomial $x^i y^j$, the result is a finite series of term

$$\begin{aligned} &\partial_x^2(x^i y^j) + \left(\partial_x(x^i y^j)\right)^2 + \partial_y^2(x^i y^j) + \left(\partial_y(x^i y^j)\right)^2 \\ &= \left(i(i-1)x^{i-2} + i x^{2i-2}\right) y^j + x^i \left(j(j-1)y^{j-2} + j^2 y^{2j-2}\right). \end{aligned} \tag{5.32}$$

For example consider the case $\beta(x,y) = x$ and look for a polynomial

$$P = \sum_{i \le K} \sum_{j \le L} a_{ij} x^i y^j, \qquad a_{KL} \neq 0,$$

solution of (5.31). If either $K \geq 2$ or $L \geq 2$, the maximal degree of (5.32) cannot decrease which is contradictory with the fact that $\beta = x$ is a polynomial of degree one. So $K \leq 1$ and $L \leq 1$. In this case the degrees of (5.32) with respect to x and y are 0. In summary there is no solution of the functional equation (5.31) in the general case, P being a polynomial.

Therefore a modification is needed. Since a polynomials cannot be an exact solution of (5.31) in the general case, it could be a solution of a modified version of (5.31). Test functions are sought as exact solutions of an approximated equation in which the coefficent β is replaced by a local approximation. The parameter l will number the $p_k \in \mathbb{N}$ basis functions designed in the cell Ω_k, so that $1 \leq l \leq p_k$. On Ω_k the test function φ_k^l satisfies

$$\left(-\Delta + \beta_k^l\right) \varphi_k^l = 0 \text{ in } \Omega_k$$

where β_k^l is an approximation of β in Ω_k such that (5.31) with right-hand side β_k^l has polynomial solutions. As a consequence β_k^l itself is necessarily a polynomial. A priori $\beta_k^l \neq \beta_k^{l'}$ for $l \neq l'$.

Definition 17. Generalized plane waves A generalized plane wave will be understood as any function of the form $\varphi = e^P$ where P is a polynomial solution on Ω_k of

$$\left\{ \begin{array}{rl} \partial_x^2 P + (\partial_x P)^2 & = \tilde{\beta}(x) \text{ in dimension one,} \\ \partial_x^2 P + (\partial_x P)^2 + \partial_y^2 P + (\partial_y P)^2 & = \tilde{\beta}(x,y) \text{ in dimension two.} \end{array} \right. \tag{5.33}$$

and $\tilde{\beta}$ is a suitable approximation of β.

The meaning of *suitable* will be specified in the design procedure section.

Remark 11. Consider the vectorial space spanned by a set of generalized plane waves. Denote by φ an element of this space. A priori this element will not be the solution of an approximated equation since $\varphi = \sum_{k,l} c_{k,l}\varphi_k^l$ only implies that

$$-\Delta\left(\sum_{k,l} c_{k,l}\varphi_k^l\right) + \sum_{k,l} c_{k,l}\beta_k^l\varphi_k^l = 0.$$

The adapted formulation

Foreseeing the numerical application of the method, the basis functions of the finite dimension test space are designed with support reduced to one element of the mesh : the matrix resulting from the system (5.27) will have a sparse profile.

Definition 18. For all $(k,l) \in \mathbb{N}^2$ such that $1 \leq k \leq N_h$ and $1 \leq l \leq p_k$, consider that the generalized plane waves φ_k^l have been constructed. The local discrete space is

$$W_k = Span\left\{(-\partial_\nu + i\sigma)\varphi_k^l\right\}_{1\leq l\leq p_k} \subset L^2(\partial\Omega_k).$$

The global discrete space $V^q \subset V$ is defined by :

$$V^q = \prod_{1\leq k\leq N_h} W_k.$$

From the local support of φ_k^l, the corresponding shape function in V^q defined from φ_k^l has support in $\partial\Omega_k$ and vanishes in $\partial\Omega_{k'}$ for all k' such that $k' \neq k$. The trace $Z_k^l \in V$ is

$$\left\{ \begin{array}{l} Z_k^l = (-\partial_\nu + i\sigma)\varphi_k^l \text{ on } L^2(\partial\Omega_k), \\ Z_k^l = 0 \text{ on } L^2(\partial\Omega_{k'}) \quad k' \neq k. \end{array} \right.$$

Since the generalized plane waves are C^∞ in Ω_k, the corresponding traces are actually piecewise C^∞_{loc} on $\partial\Omega_k$.

An equivalent way to define W_k and V^q could be

$$W_k = Span\left(Z_k^l\right)_{1\leq l\leq p_k} \text{ and } V^q = Span\left(Z_k^l\right)_{1\leq l\leq p_k,\ 1\leq k\leq N_h}.$$

Definition 19. Let $E^q \in \mathcal{L}\left(V^q, \prod_{k=1}^{N_h} H^1(\Omega_k)\right)$ be the discrete mapping defined for all $k \in [\![1, N_h]\!]$ and for all $l \in [\![1, p_k]\!]$ by

$$\begin{cases} E^q(Z_k^l) = \varphi_k^l \text{ on } L^2(\partial\Omega_k), \\ E^q(Z_k^l) = 0 \text{ on } L^2(\partial\Omega_{k'}) \text{ if } \ k' \neq k. \end{cases} \tag{5.34}$$

Similarly define $F^q \in \mathcal{L}(V^q, V)$ for all $k \in [\![1, N_h]\!]$ and for all $l \in [\![1, p_k]\!]$, by

$$\begin{cases} F^q\left(Z_k^l\right) = (\partial_\nu + i\sigma)(\varphi_k^l) \text{ on } L^2(\partial\Omega_k), \\ F^q(Z_k^l) = 0 \text{ on } L^2(\partial\Omega_{k'}) \text{ if } \ k' \neq k. \end{cases}$$

The extension mapping E^q is defined on $V^q \subset V$.

- In dimension one $\dim V^q = \dim V = 2N_h$ so $V^q = V$. As a result E^q is defined on the whole space V.
- In dimension two $\dim V = \infty$ so V^q is included in, but different from V. As a consequence, Definition 19 does not define E^q on the whole space V. However, since V^q is a linear subspace of V, there is a projector P_h from V to V^q. So the operator E^q can be extended as an operator on V in the following way : for all $Y \in V$

$$E^q(Y) = E^q(P_h Y) + E((I - P_h)Y).$$

With this notation and definitions, the abstract UWVF with generalized plane waves is defined as follows.

Definition 20. (UWVF method with generalized plane waves) Find $X_h \in V^q$ such that

$$\forall Y_h \in V^q, (X_h, Y_h)_V - (\Pi X_h, F^q Y_h)_V = (B^q, Y_h)_V \tag{5.35}$$

with the right hand side given by

$$(B^q, Y_h)_V = -2i \int_\Omega f\overline{E^q(Y_h)} + \int_\Gamma \frac{1}{\sigma} g\overline{F^q(Y_h)} \qquad \forall Y_h \in V^q. \tag{5.36}$$

5.2.3 Explicit design procedure of a Generalized Plane Wave

The different axis of analysis that will follow this section strongly rely on the shape function design. It is then necessary to precede any further analysis by an explicit procedure to design the generalized plane waves.

In dimension one the design of the local set of basis functions is straightforward whereas in dimension two the design of a shape function deserves closer attention.

To ensure the local approximation of the coefficient β by a coefficient $\tilde{\beta}$ mentioned in the definition of generalized plane waves, namely Definition 17, one solution is to fit the coefficients of the polynomial P to approximate the Taylor expansion of β at $G \in \mathbb{R}^d$ with respect to the parameter h, that will later represent the size of the mesh. Then the polynomial P will ensure that

$$\beta = \tilde{\beta} + O(h^q), \tag{5.37}$$

which defines the desired meaning of the "suitable approximation" $\tilde{\beta}$. Here q is a new parameter, and is expected to affect the convergence rate of the new method. The parameter q will denote the order of approximation of β, and is consider such as $q \geq 1$.

The constant coefficient of P is not involved in equation (5.33), it is fixed to zero : $P(0) = 0$, with no further comment. Since the procedure is based on Taylor expansions, the function φ from now on will represent $e^{P(\cdot - G)}$. This implies that the amplitude of the corresponding shape function is normalized at G : $\varphi(G) = e^0 = 1$.

Remark 12. Focus on the non linear system provided by (5.33), which right hand side is the Taylor expansion of β up to the order q. The unknowns of this system are the coefficients of P. For a fixed value of q, increasing the global degree of P, noted $\mathrm{d}P$, provides more degrees of freedom while the number of equations is given. There exists a threshold beyond which the system is under determined. Then it is likely that a high enough value of $\mathrm{d}P$ can provide an invertible system with additional unknowns to be fixed. At the same time it is reasonable to choose $\mathrm{d}P$ as small as possible to minimize the amount of computations. The main question is therefore to determine the optimal value of the $\mathrm{d}P$ with respect to the approximation parameter q.

The paragraph concerning the case $d = 1$ is presented as a gradual tutorial about what stems from the definition of generalized plane waves, and the next paragraph concerns the case $d = 2$. Some properties of the basis functions are also provided in this section, foreseeing Sections 5.3 and 5.4.

In dimension one

Only two basis functions per cell are needed. It is a common property of plane wave methods in dimension one, since the number of elementary solutions of a second order differential equation is two. The local basis functions will be denoted $\varphi_\pm = e^{P_\pm(x-G)}$.

Consider first a simple case. If β is locally constant, that is

$$\beta(x) = \beta(G) \in \mathbb{R}, x \in [G - h/2, G + h/2],$$

then $P^\pm(x) = \pm\sqrt{\beta(G)}x$ are two natural solutions which correspond to the two local plane waves $\varphi_\pm(x) = e^{P_\pm(x-G)}$ in the case $\beta(G) < 0$.

Then if β is not constant, its local Taylor expansion reads

$$\beta = \sum_{i=0}^{q-1} \frac{d^i\beta}{dx^i}(G)\,(x-G)^i + O(h^q), \qquad x \in [G - h/2, G + h/2].$$

Using the finite expansion $P_\pm = \sum_{i \le \mathrm{d}P} \lambda_i^\pm (x-G)^i$, one obtains

$$\beta_\pm = P_\pm'' + (P_\pm')^2 = \left(\sum_{i=0}^{\mathrm{d}P} \lambda_i^\pm (x-G)^i\right)'' + \left(\left(\sum_{i=0}^{\mathrm{d}P} \lambda_i^\pm (x-G)^i\right)'\right)^2.$$

In order to satisfy (5.37) both the degree $\mathrm{d}P \in \mathbb{N}$ and the coefficients $(\lambda_i^\pm)_{0\le i\le \mathrm{d}P}$ of P have to be defined such that

$$\left(\sum_{i=0}^{\mathrm{d}P} \lambda_i^\pm (x-G)^i\right)'' + \left(\left(\sum_{i=0}^{\mathrm{d}P} \lambda_i^\pm (x-G)^i\right)'\right)^2 = \sum_{i=0}^{q-1} \frac{d^i\beta}{dx^i}(G)\ (x-G)^i + O(h^q). \quad (5.38)$$

Identifying the coefficients in the polynomial part of the previous equation leads to a non linear system of q equations with $\mathrm{d}P$ unknowns.

Some remarks and examples follow.

- **Trivial case : $q = \mathrm{d}P = 1$.** In this case equation (5.38) reads $\left(\lambda_1^\pm\right)^2 = \beta\,(G)$, so that

$$\begin{cases} \lambda_1^\pm = \pm\sqrt{\beta\,(G)} \\ P_\pm(x) = \pm\sqrt{\beta\,(G)}\,(x-G) \end{cases}$$

As already remarked, if $\beta\,(G) < 0$, it yields two plane waves with opposite directions.

- **Counter-example :** $q = \mathrm{d}P = 2$**.** The non linear system obtained from the first two terms in (5.38) is

$$\begin{cases} 2\lambda_2 + (\lambda_1)^2 = \beta(G) \equiv a, \\ 4\lambda_1\lambda_2 = \beta'(G) \equiv b. \end{cases} \tag{5.39}$$

Elimination of λ_2 yields

$$-2(\lambda_1)^3 + 2a\lambda_1 = b. \tag{5.40}$$

It is of course possible to compute λ_1 as any root of this polynomial, λ_2 will then be computed as a ratio, i.e. $\lambda_2 = \frac{b}{4\lambda_1}$. So in principle this method can generate at least two different polynomials P. But in the case $b = 0$, $\lambda_1 = 0$ is one root. In such a case λ_2 would be singular : $\lambda_2 = +\infty$. As a consequence the property 2 of Lemma 5.8 would not satisfied.

It must be noticed that the use of such a method in numerical tests revealed a singularity near $\beta(x) \approx 0$.

Another problem is the generalization to high order. This procedure requires the exact computation of the roots of a high order polynomial which generalizes (5.40). This is not possible for orders ≥ 5. This is why this method is not used.

- **Example :** $q = 2$ **and** $\mathrm{d}P = 3$**.** This is the direct application of Remark 12. Taking into account λ_3, the system becomes

$$\begin{cases} 2\lambda_2 + (\lambda_1)^2 = a, \\ 6\lambda_3 + 4\lambda_1\lambda_2 = b. \end{cases} \tag{5.41}$$

This system now has 3 unknowns and 2 equations. So it has a priori an infinite number of solutions. A natural normalization condition arises by considering that the two basis functions should be linearly independent. Since $\varphi_+(G) = \varphi_-(G) = 1$ it is sufficient to ensure that $\varphi'_+(G) \neq \varphi'_-(G)$, for instance imposing that

$$\begin{cases} \dfrac{d}{dx} e^{P_-(G)} = 0 \Longleftrightarrow P'_-(G) = 0 \\ \dfrac{d}{dx} e^{P_+(G)} = 1 \Longleftrightarrow P'_+(G) = 1. \end{cases}$$

The first case corresponds to $\lambda_1^- = 0$, the second one to $\lambda_1^+ = 1$. With this normalization it is evident that λ_2 and λ_3 can be computed explicitly from (5.41) and that the resulting formulas are just polynomial expressions with respect to all coefficients. One obtains two sets of coefficients which are

$$\begin{cases} \lambda_1^+ = 1, \quad \lambda_2^+ = \dfrac{1}{2}(a-1), \quad \lambda_3^+ = \dfrac{1}{3}(b - 2a + 2) \\ \lambda_1^- = 0, \qquad \lambda_2^- = \dfrac{a}{2}, \qquad \lambda_3^- = \dfrac{1}{3}(b - 2a). \end{cases}$$

These formulas hold for all $(a, b) \in \mathbb{C}^2$. Notice that since $\beta_\pm = P''_\pm + (P'_\pm)^2$, then $\lambda_1^+ \neq \lambda_1^-$ implies that $\beta_+ \neq \beta_-$.

This method can be generalized to any order $q > 1$. The first thing is to chose a convenient degree $\mathrm{d}P$. For any order q, consider the system obtained from (5.38)

$$\begin{cases} (i+2)(i+1)\lambda_{i+2} = \dfrac{d^i}{dx^i}\beta(G) - \displaystyle\sum_{0 \leq l, l' \leq \mathrm{d}P-1}^{l+l'=i} (l+1)(l'+1)\lambda_{l+1}\lambda_{l'+1}, \\ \displaystyle\sum_{0 \leq l, l' \leq \mathrm{d}P-1}^{l+l'=i} (l+1)(l'+1)\lambda_{l+1}\lambda_{l'+1} = \dfrac{d^i}{dx^i}\beta(G), \end{cases}$$

where the first equation holds for $i \leq q-1$ and the second one for $i \geq q$ and $i \leq r$, as long as $\mathrm{d}P > q+1$. Choosing $\mathrm{d}P \leq q+1$ one gets rid of this second type of equations and obtains an invertible system

$$\begin{cases} \forall i \in \mathbb{N} \text{ such that } i \leq q-1 \\ (i+2)(i+1)\lambda_{i+2} = \dfrac{d^i}{dx^i}\beta(G) - \displaystyle\sum_{0\leq l,l'\leq \mathrm{d}P-1}^{l+l'=i} (l+1)(l'+1)\lambda_{l+1}\lambda_{l'+1}. \end{cases} \tag{5.42}$$

In fact, the corresponding system of q equations with $q+1$ unknowns is obtained identifying the first q coefficients in both parts of the expansion (5.38) with the normalization $\lambda_1^+ = 0$ which corresponds to $P'_+(G) = 0$ and $\lambda_1^- = 1$ which corresponds to $P'_-(G) = 1$.

The resulting functions $\varphi_+ = e^{P_+}$ and $\varphi_- = e^{P_-}$ are linearly independent functions, and

$$\beta_+ = P''_+ + (P'_+)^2 \text{ and } \beta_- = P''_- + (P'_-)^2. \tag{5.43}$$

By construction the first q coefficients of these polynomials $\beta_\pm$ coincide. But of course all other coefficients have no reason to be equal, so $\beta_+ \neq \beta_-$ in the general case .

The definition of the basis functions space requires some notation :

$$\begin{cases} \Omega =]a,b[\subset \mathbb{R} \\ \overline{\Omega} = \displaystyle\bigcup_{k\in[\![1,N_h]\!]} [x_k, x_{k+1}] \text{ with } x_k < x_{k+1} \\ x_{k+1/2} = \dfrac{x_k + x_{k+1}}{2}, \\ h = \max_{k\in[\![1,N_h]\!]}(x_{k+1} - x_k). \end{cases}$$

As mentioned previously $p_k = 2$ for all $k \in [\![1, N_h]\!]$. On each cell of the mesh, the two basis functions $\varphi_{k,\pm}$ are defined with $G = x_{k+1/2}$.

One can summarize as follows.

Lemma 5.8. *The corresponding functions $\beta_{k,\pm}$ defined in (5.43) satisfy the following statements.*

1. *These functions, together with all the coefficients of the polynomials $P_\pm$, are bounded independently from the cell number k, as well as all their derivatives.*
2. *By construction there exists a constant C_q such that $\|\beta_{k,\pm} - \beta\|_{L^\infty(\Omega_k)} \leq C_q h^q$ where q is the approximation parameter.*
3. *This construction is valid even if the sign of β changes or if β vanishes.*

The last point is essential to be able to address the numerical approximation of the Airy equation.

The normalization $\lambda_\pm \in \{0, 1\}$ is arbitrary. It is also possible to choose another normalization such as $\lambda_1^\pm = \pm\sqrt{x_{k+1/2}}$. This choice gives two basis functions as long as $x_{k+1/2} \neq 0$. It will be illustrated as a numerical example in the dedicated chapter.

Item 2 of Lemma 5.8 establishes a property of approximation with respect to h. One of the numerical tests shows that a similar property of convergence holds with respect to the order parameter q. In practice the q-convergence is a highly desirable property since it allows to use large cells.

In dimension two

Consider a general point $G = (x_G, y_G) \in \Omega \subset \mathbb{R}^2$. Again the design of the polynomial P starts with the choice of its degree, and next focuses on giving an explicit expression to compute its coefficients, using a Taylor expansion around the given point G.

The shape function will be designed to ensure that

$$\beta - P_\Delta = O(h^q), \tag{5.44}$$

where P_Δ stands for $\partial_x^2 P + (\partial_x P)^2 + \partial_y^2 P + (\partial_y P)^2$. A precise analysis of equation (5.44) in terms of coefficients of P leads to chose the degree of P such that the computation of the coefficients appears to be straightforward. For the same reasons as in the one dimensional case the degree $\mathrm{d}P$ will be $q+1$, but the normalization will necessarily involve more coefficients.

The Taylor expansion of P_Δ is necessarily more involved than the corresponding term in dimension one. Since

$$\left\{\begin{array}{rcl} \partial_x^2 P(x,y) & = & \displaystyle\sum_{0\le i+j\le \mathrm{d}P-2} (i+2)(i+1)\lambda_{i+2,j} x^i y^j, \\ \partial_y^2 P(x,y) & = & \displaystyle\sum_{0\le i+j\le \mathrm{d}P-2} (j+2)(j+1)\lambda_{i,j+2} x^i y^j, \end{array}\right.$$

for any $(i_0, j_0) \in [\![1, \mathrm{d}P-2]\!]^2$ such that $i_0 + j_0 \le \mathrm{d}P - 2$, the term

$$(i_0+2)(i_0+1)\lambda_{i_0+2,j_0} + (j_0+2)(j_0+1)\lambda_{i_0,j_0+2}$$

will appear in the coefficient of $x^{i_0} y^{j_0}$. Since

$$\begin{array}{rcl} (\partial_x P(x,y))^2 & = & \displaystyle\sum_{1\le i+j\le \mathrm{d}P} \sum_{1\le i'+j'\le \mathrm{d}P} (i)(i')\lambda_{i,j}\lambda_{i',j'} x^{i+i'-2} y^{j+j'}, \\ & = & \displaystyle\sum_{0\le i+j\le \mathrm{d}P-1} \sum_{0\le i'+j'\le \mathrm{d}P-1} (i+1)(i'+1)\lambda_{i+1,j}\lambda_{i'+1,j'} x^{i+i'} y^{j+j'}, \end{array}$$

then for any $(i_0, j_0) \in [\![1, \mathrm{d}P-2]\!]^2$ such that $i_0+j_0 \le \mathrm{d}P-2$ the corresponding contribution from $\lambda_{i+1,j}\lambda_{i_0-i+1,j_0-j}$ to the term $x^{i_0} y^{j_0}$ could only be such that

$$\left\{\begin{array}{rl} j & \le j_0, \\ j_0 - j & \le j_0, \\ i+1 & \le i_0+1, \\ i_0 - i + 1 & \le i_0+1. \end{array}\right.$$

That means that the λs contributing to P_Δ from $(\partial_x P)^2$ can only be of index (i,j) such that $i + j \le i_0 + j_0 + 1$. For similar reasons the λs contributing to P_Δ from $(\partial_y P)^2$ can only be of index (i,j) such that $i + j \le i_0 + j_0 + 1$. So since overall all the terms except $\partial_x^2 P$ and $\partial_y^2 P$ will only involve $\lambda_{i,j}$ satisfying $i + j \le i_0 + j_0 + 1$: suppose λ_{i_0,j_0+2} is given, then λ_{i_0+2,j_0} is easily determined. For this reason the degree is again set as $\mathrm{d}\mathbf{P} = \mathbf{q} + \mathbf{1}$.

Since

$$\left\{\begin{array}{rcl} (\partial_x P)^2 & = & \displaystyle\sum_{0\le i+j\le q-1} \left(\sum_{k=0}^{i} \sum_{l=0}^{j} (i-k+1)(k+1)\lambda_{i-k+1,j-l}\lambda_{k+1,l} \right) x^i y^j + O(h^q), \\ (\partial_y P)^2 & = & \displaystyle\sum_{0\le i+j\le q-1} \left(\sum_{k=0}^{j} \sum_{l=0}^{i} (j-k+1)(k+1)\lambda_{i-l,j-k+1}\lambda_{l,k+1} \right) x^i y^j + O(h^q), \\ \partial_x^2 P & = & \displaystyle\sum_{0\le i+j\le q-1} (i+2)(i+1)\lambda_{i+2,j} x^i y^j, \\ \partial_y^2 P & = & \displaystyle\sum_{0\le i+j\le q-1} (j+2)(j+1)\lambda_{i,j+2} x^i y^j, \end{array}\right.$$

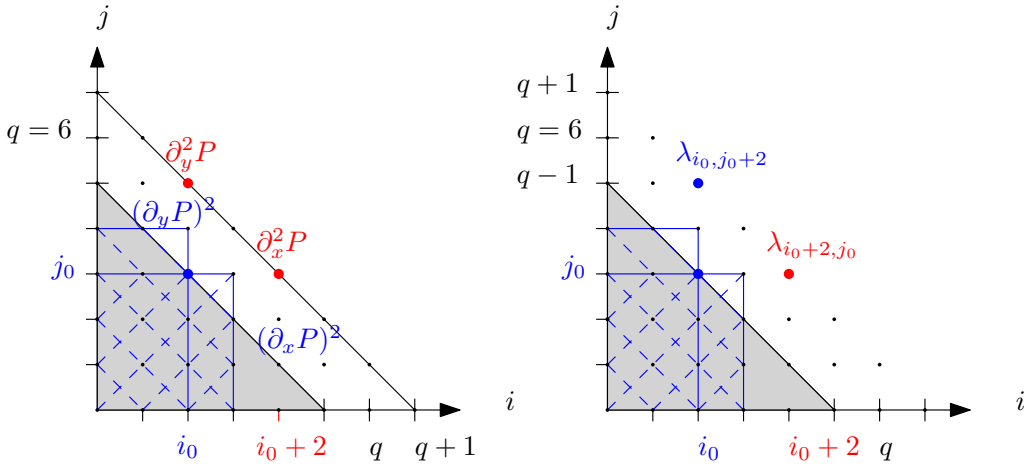


FIGURE 5.5 – For a given (i_0, j_0), indices (i, j) such that $\lambda_{i,j}$ contributes to the $x^{i_0}y^{j_0}$ term in $\beta - P_\Delta$. Left : Contributions from P_Δ to the $x^{i_0}y^{j_0}$ term in $\beta - P_\Delta$. The second order derivatives are represented in red. Right : Evidence of the invertibility of the system (5.45) : λ_{i_0+2,j_0} (in red) can be explicitly expressed as long as $\lambda_{k,l}$ (in blue) are known for all $k \leq i_0 + 1$ and $l \leq \mathrm{d}P - 2 - k$.

and since the Taylor expansion of β in two dimensions reads

$$\beta(x,y) = \sum_{(i,j)/0\leq i+j\leq q-1} \frac{\partial_x^i \partial_y^j \beta(G)}{i!j!}(x-x_G)^i(y-y_G)^j + O\left(\|(x,y)-(x_G,y_G)\|^q\right),$$

then to satisfy (5.44) the coefficients of P necessarily satisfy the following system

$$\begin{aligned} &\forall (i,j) \text{ s.t. } 0 \leq i+j \leq q-1, \\ &\frac{\partial_x^i \partial_y^j \beta(G)}{i!j!} = (i+2)(i+1)\lambda_{i+2,j} + (j+2)(j+1)\lambda_{i,j+2} \\ &\qquad + \sum_{k=0}^{i}\sum_{l=0}^{j}(i-k+1)(k+1)\lambda_{i-k+1,j-l}\lambda_{k+1,l} \\ &\qquad + \sum_{k=0}^{j}\sum_{l=0}^{i}(j-k+1)(k+1)\lambda_{i-l,j-k+1}\lambda_{l,k+1}. \end{aligned} \tag{5.45}$$

This is an explicit formula which can be used to compute the whole set of coefficients of P by induction. It is equivalent to the one dimensional formula (5.42) from the previous paragraph.

Inspecting the formula (5.45), the choice to fix the set of coefficients

$$\{\lambda_{i,j}, i \in \{0,1\}, j \in [\![0, q+1-i]\!]\}$$

defines explicitly all the coefficients of P. Others choices to obtain an invertible system would give exactly the same theoretical results. For instance choosing to fix the set of coefficients $\{\lambda_{i,j}, j \in \{0,1\}, i \in [\![0, q+1-i]\!]\}$ is a possible choice as well. But numerically, as will be seen later on, there is no evidence of the lack of symmetry of the first proposed choice with respect to the two space variables. This is due to the fact that in any case the procedure ensures the approximation of β up to the order q, no matter what this choice can be. In practice a very accurate approximation of β is obtained.

However, even if explicit, the formula to compute all the coefficients is complicated, and the algebra associated is non trivial. A general type of shape functions corresponds to the following definition.

Definition 21. Consider $N \in \mathbb{C}$ such that $N \neq 0$. For a given $\theta \in \mathbb{R}$, the N-normalization is defined by

1. $(\lambda_{1,0}, \lambda_{0,1}) = N(\cos\theta, \sin\theta)$.
2. $\{\lambda_{i,j}, i \in \{0,1\}, 1 < i + j \leq q + 1\}$ are set to zero.

For a given value of θ the formula (5.45) provides an explicit function $\varphi = e^{P(\cdot - G)}$ depending on N and θ. Varying θ then provides different functions φ, only as long as $N \neq 0$. This condition $N \neq 0$ is mandatory to define a set of linearly independent shape functions.

The first point comes from the idea of considering a classical plane wave plus higher degree terms since each angle $\theta \in \mathbb{R}$ gives a shape function. It justifies the name given to the corresponding shape function : **generalized plane wave**.

The second point greatly simplifies both the analysis of the method - since everything will only depend on the two quantities $(\lambda_{1,0}, \lambda_{0,1})$ - and the numerical computations by a substantial decrease of basic operations necessary to evaluate a shape function.

Section 5.4 is dedicated to the analysis in dimension two. The dependence of the coefficients $\lambda_{i,j}$ with respect to $(\lambda_{1,0}, \lambda_{0,1})$ will be specified.

Definition 22 (Finite dimension approximation space)**.** Suppose $p \in \mathbb{N}$ is such that $p \geq 3$, consider equispaced directions $\theta_l = 2\pi(l-1)/p$ for all $l \in [\![1, p]\!]$ and the corresponding $(\lambda^l_{1,0}, \lambda^l_{0,1})$ and φ_l defined by the N-normalization. The local set of shape functions denoted $\mathcal{E}(G, p)$ is defined by $\{\varphi_l\}_{l \in [\![1,p]\!]}$.

From the explicit expression of $(\lambda_{1,0}, \lambda_{0,1})$ fixed in the N-normalization, the fact that the sum $(\lambda_{1,0})^2 + (\lambda_{0,1})^2$ does not depend on θ is important in the analysis proposed in Section 5.4, but not the value N of the sum. Indeed, as a result some crucial quantities do not depend on the $l \in [\![1, p]\!]$.

Remark 13. The dimension of the approximation space $Span\mathcal{E}(G, p)$ is p, this is a consequence of Lemma 5.20. Indeed, consider a linear combination $\sum_{l=1}^{p} x_l \varphi_l$ and suppose it is equal to zero. Then the vector $\mathcal{X} = (x_l)_{l \in [\![1,p]\!]}$ satisfies $M_n \mathcal{X} = 0$, where M_n is introduced in Definition 26. But since $rk(M_n) = p$ and $0 \in Im(M_n)$ then necessarily $\mathcal{X} = 0$, so that $\{\varphi_l\}_{l \in [\![1,p]\!]}$ is a linearly independent family.

This property is obviously uniform with respect to h.

The link with the approximation space V^q described in Section 5.2 is the following. Consider a mesh of the domain, $\overline{\Omega} = \cup \overline{\Omega_k}$. On a mesh element Ω_k which center is denoted G_k, the local trace space W_k is defined via the shape functions of $\mathcal{E}(G_k, p(k))$, and again

$$V^q = \prod_{1 \leq k \leq N_h} W_k,$$

as in Definition 18.

Two examples of normalizations

Until that point the design of $\mathcal{E}(G,p)$ is explicit up to the choice of the normalization parameter N. Different choices are then possible. Two different ones will be considered in this work.

The first normalization is the natural direct generalization of the classical plane waves.

Definition 23. The normalization corresponding to $N = \sqrt{\beta(G)}$ will be called the β-normalization.

Such a normalization leads indeed to $P = \sqrt{\beta(G)}(x\cos\theta + y\sin\theta)$+higher order terms, where $\tilde{P} = \sqrt{\beta(G)}(x\cos\theta + y\sin\theta)$ corresponds to the associated classical plane wave. However, keeping in mind the original problem to be addressed, one can notice that at the cut-off, $\beta(G) = 0$, the β-normalization does only provide one shape function which is constant equal to one regardless of the value of $\beta(G)$.

A second normalization can theoretically overcome this drawback.

Definition 24. The normalization corresponding to $N = i(= \sqrt{-1})$ will be called the constant-normalization, since it does not depend on $\beta(G)$.

In this case Proposition 5.21 and Theorem 5.4.1 still holds in this case if $\beta(G) = 0$. However the induction formula used to define the coefficients of P does involve the derivatives of β at G : even if N does not depend on G, the other coefficients of P do.

The performances of these two normalizations will be compared in the chapter dedicated to the numerical results.

5.3 Numerical analysis in dimension one

This section rolls out the steps of a numerical analysis of the h-convergence of the new method in dimension one, to result in an explicit theoretical order expressed with respect to the approximation parameter q.

5.3.1 Convenient global notation

Two polynomials $P_{k,1}$ and $P_{k,2}$ on Ω_k for all $k \in [\![1, N_h]\!]$ correspond to the basis functions and approximated coefficients denoted by $\varphi_{k,1}$, $\beta_{k,1}$ and $\varphi_{k,2}$, $\beta_{k,2}$. For the sake of simplicity, the basis functions space will be globally denoted by $\{\varphi_j\}_{j\in[\![1,2N_h]\!]}$ and the corresponding coefficients $\{\beta_j\}_{j\in[\![1,2N_h]\!]}$; $\{Z_j\}_{j\in[\![1,2N_h]\!]}$ will denote the corresponding traces, i.e.

$$\forall j \in [\![1, 2N_h]\!], Z_j = \{(-\partial_\nu + i\sigma)\varphi_{j|\partial\Omega_k}\}_{k\in[\![1,N_h]\!]}.$$

The family $\{Z_j\}_{j\in[\![1,2N_h]\!]}$ is a basis of the functional space V^q. A fundamental property

$$V^q = V \text{ only in dimension one}$$

was already pointed after the definition of V^q. It will radically reduce the technicalities of the proof. In fact, Lemmas 5.10 and 5.12 rely on the fact that in dimension one $dim\ V^q = 2N_h$.

5.3.2 Preliminary results

For the sake of completeness, here are classical results. Both Theorems 5.3.1 and 5.3.2 provide a priori estimates explicit with respect to the mesh parameter h. It has to be noted that such explicit estimates are not available in dimension two.

Theorem 5.3.1. Let $\mathcal{O}$ be a one-dimensional open interval with length h. Let w be the unique solution of

$$\left\{ \begin{array}{rl} -\Delta w + \beta w &= 0, \quad (\mathcal{O}), \\ (-\partial_\nu + i\sigma)\, w &= g, \quad (\partial\mathcal{O}). \end{array} \right. \tag{5.46}$$

Then there exists a constant C which depends of $\|\beta\|_{L^\infty(\mathcal{O})}$ and σ such that for h small enough

$$\|w\ \|_{L^2(\mathcal{O})} \leq C\sqrt{h}\, \|g\|_{L^2(\partial\mathcal{O})}\,. \tag{5.47}$$

The existence and uniqueness of the solution is given by theorem 5.2.3. A very classical Poincaré inequality in one dimension will be needed in the proof.

Proposition 5.9. *There exists a constant C such that for all $h > 0$, for all open interval $\mathcal{O} \subset \mathbb{R}$ which length is h, for all $u \in L^2(\mathcal{O})$*

$$\|u\|_{L^2(\mathcal{O})} \leq C\left(\sqrt{h}\|u\|_{L^2(\partial\mathcal{O})} + h\|u'\|_{L^2(\mathcal{O})}\right). \tag{5.48}$$

Proof. There exists $a \in \mathbb{R}$ such that $\mathcal{O} =]a, a+h[$. From $u(x) = u(a) + \int_a^x u'(t)dt$ it yields

$$\int_a^{a+h} |u(x)|^2 dx \leq 2h|u(a)|^2 + 2\int_a^{a+h} \left(\int_a^x |u'(t)|dt\right)^2 dx,$$

so that

$$\|u\|_{L^2(\mathcal{O})} \leq \sqrt{2h}\|u\|_{L^2(\partial\mathcal{O})} + \sqrt{2}h\|u'\|_{L^2(\mathcal{O})}.$$

It gives the result for $C = \sqrt{2}$. □

*Proof.*Of Theorem 5.3.1. A more general inequality than (5.47) can actually be proved, on the non homogeneous problem :

$$\left\{ \begin{array}{ll} -u'' + \beta u = f, & (\mathcal{O}) \\ (-\partial_\nu + i\sigma)u = g, & (\partial\mathcal{O}). \end{array} \right. \tag{5.49}$$

Using u as test function, one gets

$$\int_{\mathcal{O}} |u'|^2 + i\sigma \int_{\partial\mathcal{O}} |u|^2 = \int_{\mathcal{O}} f\overline{u} - \int_{\mathcal{O}} \beta|u|^2 + \int_{\partial\mathcal{O}} g\overline{u}.$$

Then

$$\left\{ \begin{array}{l} \|u\|^2_{L^2(\partial\mathcal{O})} \leq \dfrac{1}{\sigma}\|f\|_{L^2(\mathcal{O})}\|u\|_{L^2(\mathcal{O})} + \dfrac{1}{\sigma}\|g\|_{L^2(\partial\mathcal{O})}\|u\|_{L^2(\partial\mathcal{O})}, \\ \|u'\|^2_{L^2(\mathcal{O})} \leq \|g\|_{L^2(\partial\mathcal{O})}\|u\|_{L^2(\partial\mathcal{O})} + \|\beta\|_{L^\infty(\mathcal{O})}\|u\|^2_{L^2(\mathcal{O})} + \|f\|_{L^2(\mathcal{O})}\|u\|_{L^2(\mathcal{O})}. \end{array} \right.$$

Using the fact that $2ab \leq a^2/\sigma + \sigma b^2$ for all $a, b \in \mathbb{R}$ the first inequality yields

$$\|u\|^2_{L^2(\partial\mathcal{O})} \leq \frac{2}{\sigma}\|f\|_{L^2(\mathcal{O})}\|u\|_{L^2(\mathcal{O})} + \frac{1}{\sigma^2}\|g\|^2_{L^2(\partial\mathcal{O})},$$

so that

$$\|g\|_{L^2(\partial\mathcal{O})}\|u\|_{L^2(\partial\mathcal{O})} \leq \frac{1}{2\sigma}\|g\|^2_{L^2(\partial\mathcal{O})} + \frac{\sigma}{2}\|u\|^2_{L^2(\partial\mathcal{O})}$$

$$\leq \frac{1}{2\sigma}\|g\|^2_{L^2(\partial\mathcal{O})} + \|f\|_{L^2(\mathcal{O})}\|u\|_{L^2(\mathcal{O})} + \frac{1}{2\sigma}\|g\|^2_{L^2(\partial\mathcal{O})}.$$

Then

$$\|u'\|^2_{L^2(\mathcal{O})} \leq \frac{1}{\sigma}\|g\|^2_{L^2(\partial\mathcal{O})} + 2\|f\|_{L^2(\mathcal{O})}\|u\|_{L^2(\mathcal{O})} + \|\beta\|_{L^\infty(\mathcal{O})}\|u\|^2_{L^2(\mathcal{O})}$$

stems from the second inequality. So from (5.48)

$$\|u\|^2_{L^2(\mathcal{O})} \leq C\left(h\left(\frac{2}{\sigma}\|f\|_{L^2(\mathcal{O})}\|u\|_{L^2(\mathcal{O})} + \frac{1}{\sigma^2}\|g\|^2_{L^2(\partial\mathcal{O})}\right)\right.$$

$$\left.+h^2\left(\frac{1}{2\sigma}\|g\|^2_{L^2(\partial\mathcal{O})} + 2\|f\|_{L^2(\mathcal{O})}\|u\|_{L^2(\mathcal{O})} + \|\beta\|_{L^\infty(\mathcal{O})}\|u\|^2_{L^2(\mathcal{O})}\right)\right).$$

For h small enough it proves that

$$\|u\|^2_{L^2(\mathcal{O})} \leq C\left(\frac{h}{\sigma^2}\|g\|^2_{L^2(\partial\mathcal{O})} + \frac{h^2}{\sigma^2}\|f\|^2_{L^2(\mathcal{O})}\right). \tag{5.50}$$

One can notice that the scaling of this estimate is optimal. Indeed considering that σ has the inverse dimension of a length - which is evident from the boundary condition - all quantities have the same dimension at inspection of (5.49). Inequality (5.47) is obtained by taking $f = 0$ in the previous inequality. □

The next result concerns the approximation error between the problem

$$\left\{\begin{array}{rcll} -\Delta w + \beta w & = & f, & (\mathcal{O}), \\ (-\partial_\nu + i\sigma)\, w & = & g, & (\partial\mathcal{O}), \end{array}\right. \tag{5.51}$$

and the modified problem

$$\left\{\begin{array}{rcll} -\Delta w + \beta_h w & = & f, & (\mathcal{O}), \\ (-\partial_\nu + i\sigma)\, w & = & g, & (\partial\mathcal{O}), \end{array}\right. \tag{5.52}$$

where $\mathcal{O}$ represents any open set which length is h, included in Ω.

Theorem 5.3.2. Let $\mathcal{O}$ be a one-dimensional open interval which length is h, such that $\mathcal{O} \subset \Omega$. If u is solution of the problem (5.51) and u_h is solution of the problem (5.52), then for small h there exists a constant C such that

$$\|u - u_h\|_{L^2(\mathcal{O})} \leq C\left(h^{\frac{3}{2}}\|g\|_{L^2(\partial\mathcal{O})} + h^2\|f\|_{L^2(\mathcal{O})}\right)\|\beta - \beta_h\|_{L^\infty(\mathcal{O})}. \tag{5.53}$$

Proof. Suppose that u and u_h are the solutions of the two following problems

$$\left\{\begin{array}{ll} -u'' + \beta u = f, & (\mathcal{O}) \\ (-\partial_\nu + i\sigma)u = g, & (\partial\mathcal{O}). \end{array}\right.$$

and

$$\left\{\begin{array}{ll} -u_h'' + \beta_h u_h = f, & (\mathcal{O}) \\ (-\partial_\nu + i\sigma)u_h = g, & (\partial\mathcal{O}). \end{array}\right.$$

Then $e_h := u - u_h$ satisfies

$$\left\{\begin{array}{ll} -e_h'' + \beta_h e_h = (\beta_h - \beta)u, & (\mathcal{O}) \\ (-\partial_\nu + i\sigma)e_h = 0, & (\partial\mathcal{O}). \end{array}\right.$$

Inequality (5.50) yields

$$\|e_h\|_{L^2(\mathcal{O})} \leq C\frac{h}{\sigma}\|(\beta_h-\beta)u\|_{L^2(\mathcal{O})} \leq C\frac{h}{\sigma}\|\beta_h-\beta\|_{L^\infty(\mathcal{O})}\|u\|_{L^2(\mathcal{O})}.$$

Using one more time (5.50) to estimate u and regarding σ which is a positive number, then

$$\|e_h\|_{L^2(\mathcal{O})} \leq C\left(h^{\frac{3}{2}}\|g\|_{L^2(\partial\mathcal{O})}+h^2\|f\|_{L^2(\mathcal{O})}\right)\|\beta_h-\beta\|_{L^\infty(\mathcal{O})}.$$

□

5.3.3 Approximation of the operator F

This paragraph is devoted to showing that the new operator F^q introduced in Section 5.2.2 is an approximation of the original operator F up to the order $q+1$ in h. Consider the following problem

$$\left\{\begin{array}{l}\text{Find } X_h \in V_q \text{ such that}\\ (I-A^q)X_h = B,\end{array}\right. \tag{5.54}$$

where $A^q = (F^q)^*\Pi$. Here h and q are given. This result relies on a preliminary lemma.

Lemma 5.10. *Let $q \geq 2$. Suppose h is small enough and basis functions are constructed as described in paragraph 5.2.3. There exists a constant C independent of k such that for all $k \in [\![1, N_h]\!]$, for all $(x_1, x_2) \in \mathbb{C}^2$,*

$$\sum_{j\in\{1,2\}}|x_j|\|Z_j\|_{L^2(\partial\Omega_k)} \leq C\left\|\sum_{j\in\{1,2\}}x_jZ_j\right\|_{L^2(\partial\Omega_k)},$$

where (Z_1, Z_2) is a local notation for (Z_k^1, Z_k^2).

Proof. Set $k \in [\![1, N_h]\!]$ and $Z = x_1Z_1 + x_2Z_2 \in V$ which support is $\partial\Omega_k$. First x_j can be written as a function of Z. This is a priori possible using $\{Y_j\}_{j\in\{1,2\}}$ which is the dual basis of $\{Z_j\}_{j\in\{1,2\}}$. For all $(j,l) \in \{1,2\}^2$, the dual function Y_j is defined by

$$(Y_j, Z_l)_V = \delta_{jl}, \tag{5.55}$$

where δ denotes the Kronecker symbol. The proof proceeds in several steps.

First step. One has that $x_j = (Z, Y_j)_V$, therefore

$$\sum_{j\in\{1,2\}}|x_j|\|Z_j\| \leq \left(\sum_{j\in\{1,2\}}\|Z_j\|\|Y_j\|\right)\|Z\|.$$

So the claim is proved if the right hand side term can be estimated.

Second step : estimate of $\sum_{j\in\{1,2\}}\|Z_j\|\|Y_j\|$. From (5.55) it turns out that

$$\left\{\begin{array}{rl}Y_1 &= \dfrac{-\|Z_2\|^2}{|(Z_1,Z_2)|^2-\|Z_1\|^2\|Z_2\|^2}Z_1 + \dfrac{(Z_1,Z_2)}{|(Z_1,Z_2)|^2-\|Z_1\|^2\|Z_2\|^2}Z_2,\\ Y_2 &= \dfrac{(Z_1,Z_2)}{|(Z_1,Z_2)|^2-\|Z_1\|^2\|Z_2\|^2}Z_1 - \dfrac{\|Z_1\|^2}{|(Z_1,Z_2)|^2-\|Z_1\|^2\|Z_2\|^2}Z_2,\end{array}\right.$$

which yields

$$\sum_{j\in\{1,2\}} \|Z_j\|\|Y_j\| \leq 2\frac{\|Z_1\|^2\|Z_2\|^2}{\|Z_1\|^2\|Z_2\|^2 - |(Z_1,Z_2)|^2}.$$

Set for convenience $D = \dfrac{|(Z_1,Z_2)|}{\|Z_1\|\|Z_2\|}$ so that obviously $D<1$ and

$$\sum_{j\in\{1,2\}} \|Z_j\|\|Y_j\| \leq 2\frac{1}{1-D^2}.$$

It means that the whole proof relies on a more precise upper bound for D when h goes to zero.

Third step : end of the proof. By definition $(Z_j)_{|\partial\Omega_k} = \left((-\partial_\nu + i\sigma)e^{P_j(\cdot - x_{k+1/2})}\right)_{|\partial\Omega_k}$. By construction

$$\begin{cases} P_j(0) = 0 \ \forall j = 1,2, \\ P_1'(0) = 0, \\ P_2'(0) = 1, \end{cases}$$

even if it means exchanging the numbering of Z_js. Since by construction all derivatives of P_1 and P_2 are uniformly bounded - as mentioned in point 1 of Lemma 5.8 - one has

$$\begin{cases} P_j(x - x_{k+1/2}) = O(h) \ \forall j = 1,2, \\ P_1'(x - x_{k+1/2}) = O(h), \\ P_2'(x - x_{k+1/2}) = 1 + O(h) \end{cases}$$

when h goes to 0 and for all $x \in [x_k, x_{k+1}]$. So one can estimate

$$\begin{aligned} ||Z_1||^2 &= \frac{1}{\sigma}\left|-P_1'(x_{k+1} - x_{k+1/2}) + i\sigma\right|^2 \left|e^{P_1(x_{k+1} - x_{k+1/2})}\right|^2 \\ &\quad + \frac{1}{\sigma}\left|P_1'(x_k - x_{k+1/2}) + i\sigma\right|^2 \left|e^{P_1(x_k - x_{k+1/2})}\right|^2 \\ &= \frac{1}{\sigma}\left|-P_1'(x_{k+1} - x_{k+\frac{1}{2}}) + i\sigma\right|^2 + \frac{1}{\sigma}\left|P_1'(x_k - x_{k+\frac{1}{2}}) + i\sigma\right|^2 + O(h), \end{aligned}$$

that is $||Z_1||^2 = 2\sigma + O(h)$. With the same method one obtains

$$||Z_2||^2 = 2\frac{1+\sigma^2}{\sigma} + O(h) = \frac{1+\sigma^2}{\sigma^2}2\sigma + O(h),$$

and

$$\begin{aligned} (Z_1,Z_2) &= \frac{1}{\sigma}(-P_1'(x_{k+1} - x_{k+1/2}) + i\sigma)\overline{(-P_2'(x_{k+1} - x_{k+1/2}) + i\sigma)} \\ &\quad + \frac{1}{\sigma}(P_1'(x_k - x_{k+1/2}) + i\sigma)\overline{(P_2'(x_k - x_{k+1/2}) + i\sigma)} + O(h) \end{aligned}$$

that is $(Z_1,Z_2) = 2\sigma + O(h)$. Therefore $D^2 = \frac{\sigma^2}{1+\sigma^2} + O(h)$. It proves the claim for h sufficiently small.

□

By construction the polynomials designed in dimension one in Section 5.2.3 by the approximation of the Taylor expansion (5.38) are such that all their coefficients are uniformly bounded up to order q for all cells in the domain. This is why the error $O(h)$ in the above analysis is uniform with respect to the cell index k, which is therefore not indicated. Note that this is not true if one constructs the polynomials with the method used in the counter example (5.39).

Lemma 5.11. *For small h and considering the basis functions constructed as described in Paragraph 5.2.3, there exists a constant C such that*

$$\|F^q - F\| \leq Ch^{q+1}. \tag{5.56}$$

Proof. For all $j \in [\![1, 2N_h]\!]$, the function φ_j is by construction $\varphi_j = E^q(Z_j)$ such that

$$\varphi_j \in \{\varphi_l\}_{l\in[\![1,2N_h]\!]} \text{ satisfies } \forall k \in [\![1, N_h]\!] \begin{cases} Z_j = (-\partial_\nu + i\sigma)\varphi_j, & (\partial\Omega_k), \\ (-\mathrm{dx}^2 + \beta_\mathrm{j})\,\varphi_j = 0, & (\Omega_k). \end{cases}$$

Also define $\psi_j = E(Z_j)$ such that and the equation with the exact coefficient β

$$\psi_j \in H \text{ satisfies } \forall k \in [\![1, N_h]\!] \begin{cases} Z_j = (-\partial_\nu + i\sigma)\psi_j, & (\partial\Omega_k), \\ (-\mathrm{dx}^2 + \beta)\,\psi_j = 0, & (\Omega_k). \end{cases}$$

So

$$\begin{aligned} |(F^q - F)Z_j|^2 &= |(\partial_\nu + i\sigma)(\varphi_j - \psi_j)|^2, \\ &= |(-\partial_\nu + i\sigma)(\varphi_j - \psi_j)|^2 + 2\Re\left(i\sigma(\varphi_j - \psi_j)\partial_\nu\overline{(\varphi_j - \psi_j)}\right), \\ &= -2\sigma\Im\left((\varphi_j - \psi_j)\partial_\nu\overline{(\varphi_j - \psi_j)}\right), \end{aligned}$$

since φ_j and ψ_j satisfy the same boundary condition : $(-\partial_\nu + i\sigma)(\varphi_j - \psi_j) = 0$. Then on the only element where Z_j is non zero, numbered $k = k(j)$, the following holds

$$\begin{aligned} \int_{\partial\Omega_k} \frac{1}{\sigma}|(F^q - F)Z_j|^2 &= -2\Im\left(\int_{\partial\Omega_k} (\varphi_j - \psi_j)\partial_\nu\overline{(\varphi_j - \psi_j)}\right), \\ &= -2\Im\left(\int_{\Omega_k} (\varphi_j - \psi_j)\partial_x^2\overline{(\varphi_j - \psi_j)} - 2\Im\int_{\Omega_k}\left|\frac{d}{dx}(\varphi_j - \psi_j)\right|^2\right), \\ &\leq -2\Im\left(\int_{\Omega_k} (\varphi_j - \psi_j)(\beta_j\overline{\varphi_j} - \beta\overline{\psi_j})\right), \end{aligned}$$

since both φ_j and ψ_j satisfy homogeneous equations. So

$$\begin{aligned} \int_{\partial\Omega_k} \frac{1}{\sigma}|(F^q - F)Z_j|^2 &\leq -\Im\left(\int_{\Omega_k} (\beta_j + \beta)|\varphi_j - \psi_j|^2 + \int_{\Omega_k} (\beta_j - \beta)(\varphi_j - \psi_j)\overline{(\varphi_j + \psi_j)}\right), \\ &\leq \|\beta_j + \beta\|_{L^\infty(\Omega_k)} \left\|\overline{(\varphi_j - \psi_j)}\right\|^2_{L^2(\Omega_k)} \\ &\quad + \|\beta_j - \beta\|_{L^\infty(\Omega_k)} \left\|\overline{(\varphi_j - \psi_j)}\right\|_{L^2(\Omega_k)} \left(\|\varphi_j\|_{L^2(\Omega_k)} + \|\psi_j\|_{L^2(\Omega_k)}\right), \end{aligned}$$

thanks to Cauchy-Schwarz inequality. On the other hand, from general estimate (5.5) on the initial problem and specific one dimensional estimate (5.53), for small h

$$\begin{aligned} \|\varphi_j - \psi_j\|_{L^2(\Omega_k)} &\leq Ch^{\frac{3}{2}}\|Z_j\|_{L^2(\partial\Omega_k)}\|\beta - \beta_j\|_{L^\infty(\Omega_k)}, \\ \|\varphi_j\|_{L^2(\Omega_k)} &\leq C\sqrt{h}\|Z_j\|_{L^2(\partial\Omega_k)}, \\ \|\psi_j\|_{L^2(\Omega_k)} &\leq C\sqrt{h}\|Z_j\|_{L^2(\partial\Omega_k)}, \end{aligned}$$

and $\|\beta_j + \beta\|_{L^\infty(\Omega_k)}$ is bounded as noticed in Remark 5.8. So for small h

$$\|(F^q - F)Z_j\|^2_{L^2(\partial\Omega_k)} \leq C'h^2\|\beta_j - \beta\|^2_{L^\infty(\Omega_k)}\|Z_j\|^2_{L^2(\partial\Omega_k)},$$

where still k denotes $k(j)$. Now for all $k \in [\![1, N_h]\!]$ let $L(k)$ be the set of indexes $l \in [\![1, 2N_h]\!]$ such that Ω_k is the support of Z_l. Hence, for all $Z \in V^q$ then $Z_{|\partial\Omega_k} = \sum_{l\in L(k)} x_l Z_l$ where both Z_ls vanish on $\partial\Omega_j$ for all $j \neq k$, it yields

$$\begin{aligned}\|(F^q - F)Z\|_{L^2(\partial\Omega_k)} &\leq \sum_{l\in L(k)} |x_l| \|(F^q - F)Z_l\|_{L^2(\partial\Omega_k)} \\ &\leq Ch \max_{l\in L(k)} \|\beta_k^l - \beta\|_{L^\infty(\Omega_k)} \left(\sum_{l\in\{1,2\}} |x_l| \|Z_l\|_{L^2(\partial\Omega_k)} \right).\end{aligned}$$

Thanks to Lemma 5.10 it means that

$$\|(F^q - F)Z\|_{L^2(\partial\Omega_k)} \leq \sqrt{C'}h \max_{l\in L(k)} \|\beta_k^l - \beta\|_{L^\infty(\Omega_k)} \|Z\|_{L^2(\partial\Omega_k)}.$$

Going back to the definition of the V norm for all $Z \in V$

$$\|(F^q - F)Z\| \leq Ch \max_{j\in[\![1,2N_h]\!]} \|\beta_j - \beta\|_{L^\infty(\Omega_{k(j)})} \|Z\|,$$

which exactly means $\|F^q - F\| \leq Ch \max_{j\in[\![1,2N_h]\!]} \|\beta - \beta_j\|_{L^\infty(\Omega_{k(j)})}$. The index $k(j)$ denotes the number of the element which is the support of Z_j. The result then comes from equation (5.37) ensured by the construction of approximated coefficients β_js. □

5.3.4 A convergence result

This paragraph starts with the definition of a useful norm to adapt the second Strang lemma.

Lemma 5.12. *There exists a constant C such that for all $X \in V$:*

$$Ch^{3/2}\|X\| \leq \|(I - A)X\|.$$

As a consequence, one can see that $|||\cdot||| = \|(I - A)\cdot\|$ separates points, so that it defines a norm. A second norm, adapted to the new method, will later be defined replacing A by A^q.

Remark 14. Since in dimension one the dimension of the space V is finite, all the norms are equivalent on V. But the constants in the continuity inequalities depend on h, and this lemma specifies the dependence in this mesh parameter between $\|\cdot\|$ and $|||\cdot|||$.

The first step of the following proof is onerous, but it somehow provides an explicit inverse of the operator $(I - A)$.

Proof.

First step. Take $X \in V - \{0\}$, and define $B = (I - A)X$. In order to read into this equation in V, define $u = E(X)$ and $w = E(B)$, so that $(u, w) \in H \times H$ and

$$\forall k \in [\![1, N_h]\!] \begin{cases} \left(-\dfrac{d}{dx} + \beta\right) u_k &= 0, \quad (\Omega_k), \\ (-\partial_\nu + i\sigma)u_k &= X_k, \quad (\partial\Omega_k), \end{cases}$$

$$\forall k \in [\![1, N_h]\!] \begin{cases} \left(-\dfrac{d}{dx} + \beta\right) w_k &= 0, \quad (\Omega_k), \\ (-\partial_\nu + i\sigma)w_k &= B_k, \quad (\partial\Omega_k). \end{cases}$$

Since F is an isometry one has $FX - \Pi X = FB$. It means that on every interface a condition is satisfied : for all $k \in [\![1, N_h]\!]$

$$\left\{ \begin{array}{ll} (-\partial_\nu + i\sigma)u_k(x_k) - \mathbf{1}_{k\neq 1}(-\partial_\nu + i\sigma)u_{k-1}(x_k) & = (-\partial_\nu + i\sigma)w_k(x_k), \\ (\partial_\nu + i\sigma)u_k(x_{k+1}) - \mathbf{1}_{k\neq N_h}(\partial_\nu + i\sigma)u_{k+1}(x_{k+1}) & = (\partial_\nu + i\sigma)w_k(x_{k+1}). \end{array} \right.$$

This leads to a system of jump conditions on the interfaces

$$\left\{ \begin{array}{l} (-\partial_\nu + i\sigma)u_1(x_1) = (-\partial_\nu + i\sigma)w_1(x_1), \\ \forall k \in [\![2, N_h]\!], \left| \begin{array}{l} \left(\dfrac{d}{dx}u_{k-1} - \dfrac{d}{dx}u_k\right)(x_k) = \dfrac{1}{2}\left((-\partial_\nu + i\sigma)w_k - (\partial_\nu + i\sigma)w_{k-1}\right)(x_k), \\ (u_k - u_{k-1})(x_k) = \dfrac{1}{2i\sigma}\left((-\partial_\nu + i\sigma)w_k - (\partial_\nu + i\sigma)w_{k-1}\right)(x_k), \end{array} \right. \\ (\partial_\nu + i\sigma)u_{N_h}(x_{N_h+1}) = (\partial_\nu + i\sigma)w_{N_h}(x_{N_h+1}). \end{array} \right. \tag{5.57}$$

But in dimension one any second order differential equation has exactly two solutions. Considering U_0 and U_1 these two fundamental solutions of the homogeneous equation, the solution u of $\left(-\dfrac{d}{dx} + \beta\right)u = 0$ on Ω satisfies

$$\forall k \in [\![1, N_h]\!], u_k = \delta_0^k U_0 + \delta_1^k U_1, \tag{5.58}$$

where $\left(\delta_0^k, \delta_1^k\right)_{k\in[\![1,N_h]\!]}$ completely determine $u \in H$. Plugging (5.58) in (5.57), and defining

$$\left\{ \begin{array}{l} \lambda_0 = (-\partial_\nu + i\sigma)w_1(x_1), \\ \forall k \in [\![2, N_h]\!], \left| \begin{array}{l} \lambda_{k-1} = \dfrac{1}{2}\left((-\partial_\nu + i\sigma)w_k - (\partial_\nu + i\sigma)w_{k-1}\right)(x_k), \\ \mu_{k-1} = \dfrac{1}{2i\sigma}\left((-\partial_\nu + i\sigma)w_k - (\partial_\nu + i\sigma)w_{k-1}\right)(x_k), \end{array} \right. \\ \mu_{N_h} = (\partial_\nu + i\sigma)w_{N_h}(x_{N_h+1}), \end{array} \right.$$

it yields

$$\left\{ \begin{array}{l} (-\partial_\nu + i\sigma)U_0(x_1)\delta_0^1 + (-\partial_\nu + i\sigma)U_1(x_1)\delta_1^1 = \lambda_0, \\ \forall k \in [\![2, N_h]\!], \left| \begin{array}{l} \partial_x U_0(x_k)(\delta_0^{k-1} - \delta_0^k) + \partial_x U_1(x_k)(\delta_1^{k-1} - \delta_1^k) = \lambda_{k-1}, \\ U_0(x_k)(\delta_0^{k-1} - \delta_0^k) + \quad U_1(x_k)(\delta_1^{k-1} - \delta_1^k) = \mu_{k-1}, \end{array} \right. \\ (\partial_\nu + i\sigma)U_0(x_{N_h+1})\delta_0^{N_h} + (\partial_\nu + i\sigma)U_1(x_{N_h+1})\delta_1^{N_h} = \mu_{N_h}. \end{array} \right. \tag{5.59}$$

Given the change of variable

$$\forall k \in [\![1, N_h - 1]\!] \left\{ \begin{array}{l} D_0^k = \delta_0^k - \delta_0^{k+1}, \\ D_1^k = \delta_1^k - \delta_1^{k+1}, \end{array} \right. \tag{5.60}$$

system (5.59) gives a linear system which unknowns are $(D_0^k, D_1^k)_{k\in[\![1,N_h-1]\!]}$. Defining the Wronskian $Wr_0 = U_1\dfrac{d}{dx}U_0 - U_0\dfrac{d}{dx}U_1$ - which is non zero - the solution reads

$$\forall k \in [\![1, N_h - 1]\!] \left\{ \begin{array}{l} D_0^k = \dfrac{1}{Wr_0}\left(\lambda_k U_1(x_{k+1}) - \mu_k \dfrac{d}{dx}U_1(x_{k+1})\right), \\ D_1^k = \dfrac{1}{Wr_0}\left(\mu_k \dfrac{d}{dx}U_0(x_{k+1}) - \lambda_k U_0(x_{k+1})\right). \end{array} \right.$$

Then the structure of system (5.59) is

$$\left\{\begin{array}{l}\beta\delta_0^1+\beta\delta_1^1=\lambda_0,\\ \delta_0^k-\delta_0^{k+1}=D_0^k\ \forall k\in[\![1,N_h-1]\!],\\ \delta_1^k-\delta_1^{k+1}=D_1^k\ \forall k\in[\![1,N_h-1]\!],\\ \sigma\delta_0^{N_h}+\eta\delta_1^{N_h}=\mu_{N_h}.\end{array}\right.$$

Eliminating $(\delta_0^k,\delta_1^k)_{k\in[\![1,N_h-1]\!]}$ it yields $\delta_0^1=\mathcal{D}_z^\theta+\delta_0^{N_h}$ and $\delta_1^1=\sum_{k=1}^{N_h-1}D_1^k+\delta_1^{N_h}$, and

$$\left\{\begin{array}{l}\beta\delta_0^{N_h}+\beta\delta_1^{N_h}=L,\\ \sigma\delta_0^{N_h}+\eta\delta_1^{N_h}=\mu_{N_h},\end{array}\right. \tag{5.61}$$

with

$$L=\lambda_0-(-\partial_\nu+i\sigma)U_0(a)\mathcal{D}_z^\theta-(-\partial_\nu+i\sigma)U_1(a)\sum_{k=1}^{N_h-1}D_1^k. \tag{5.62}$$

$Wr_1=(-\partial_\nu+i\sigma)U_0(a)(\partial_\nu+i\sigma)U_1(b)-(\partial_\nu+i\sigma)U_0(b)(-\partial_\nu+i\sigma)U_1(a)$ is the determinant of system (5.61). If it were zero, then its columns would be linearly dependent, say $a_0C_1+a_1C_2=0$; this would mean $(\partial_\nu+i\sigma)(a_0U_0+a_1U_1)(x_1)=0$ and $(\partial_\nu+i\sigma)(a_0U_0+a_1U_1)(x_{N_h})=0$ so that $u=a_0U_0+a_1U_0$ would satisfy

$$\left\{\begin{array}{l}-\dfrac{d}{dx}u+\beta u=0,\\ (\partial_\nu+i\sigma)u=0.\end{array}\right.$$

So u would be the unique solution (zero) of this last system, which is not possible since U_0 and U_1 are independent. Then Wr_1 is non zero. One finally obtains that

$$\left\{\begin{array}{l}\delta_0^{N_h}=\dfrac{1}{Wr_1}\Big(L(\partial_\nu+i\sigma)U_1(b)-\mu_{N_h}(-\partial_\nu+i\sigma)U_1(a)\Big),\\ \delta_1^{N_h}=\dfrac{1}{Wr_1}\Big(\mu_{N_h}(-\partial_\nu+i\sigma)U_0(a)-L(-\partial_\nu+i\sigma)U_1(a)\Big),\\ \left.\begin{array}{l}\\ \forall k\in[\![1,N_h-1]\!]\end{array}\right|\begin{array}{l}\delta_0^k=\delta_0^{N_h}+\displaystyle\sum_{j=k}^{N_h-1}D_0^j,\\ \\ \delta_1^k=\delta_1^{N_h}+\displaystyle\sum_{j=k}^{N_h-1}D_1^j.\end{array}\end{array}\right. \tag{5.63}$$

Now u is completely known.

Second step. The next step is the estimate of the coefficients $(\delta_0^k,\delta_1^k)_{k\in[\![1,N_h]\!]}$ using (5.63). Since F is an isometry, and λ_k and μ_k are linear combinations of the components of FB

$$\left\{\begin{array}{l}\forall k\in[\![0,N_h-1]\!],|\lambda_k|\le\sqrt{\sigma}\|B\|,\\ \forall k\in[\![1,N_h]\!],|\mu_k|\le\dfrac{1}{\sqrt{\sigma}}\|B\|.\end{array}\right. \tag{5.64}$$

Thus from (5.60) and (5.64), with C depending on U_0, U_1, σ and Wr_0,

$$\left\{\begin{array}{l}\left|\mathcal{D}_z^\theta\right|\le CN_h\|B\|,\\ \left|\displaystyle\sum_{k=1}^{N_h-1}D_1^k\right|\le CN_h\|B\|.\end{array}\right.$$

From (5.62), $|L| \leq CN_h\|B\|$, and since $|\mu_{N_h}| \leq C\|B\|$ one has from (5.63)

$$\left|\delta_i^{N_h}\right| \leq CN_h\|B\| \ \forall i \in \{0,1\},$$

and next for $k \in [\![1, N_h - 1]\!]$:

$$\begin{aligned}\left|\delta_i^k\right| &\leq \left|\delta_i^{N_h}\right| + \sum_{k=1}^{N_h-1} \left|D_i^k\right| \\ &\leq CN_h\|B\|.\end{aligned}$$

As a result all δ terms satisfy $\left|\delta_i^k\right| \leq CN_h\|B\|$ for $i \in \{0,1\}$ and $k \in [\![1, N_h]\!]$.

End of the proof. A last calculus leads to the following inequalities

$$\begin{aligned}\|x\|^2 &= \sum_{k\in[\![1,N_h]\!]} \left\|\delta_0^k(-\partial_\nu + i\sigma)U_0 + \delta_1^k(-\partial_\nu + i\sigma)U_1\right\|^2_{L^2(\partial\Omega_k)} \\ &\leq \sum_{k\in[\![1,N_h]\!]} \left(2C(|\delta_0^k| + |\delta_1^k|)\right)^2 \leq C \sum_{k\in[\![1,N_h]\!]} N_h^2\|b\|^2 \leq C\|b\|^2 N_h^3,\end{aligned}$$

so that $\|X\| \leq Ch^{-3/2}\|B\|$.

□

Now in order to analyze the new method, we turn to the adapted operator A^q.

Definition 25. Define $|||X|||_q = \|(I - A^q)X\|$ for all $X \in V$. This is a norm under the condition of the next proposition. The adapted bilinear form is defined for all $X, Y \in V$ by

$$a_q(X, Y) = ((I - A^q)X, Y).$$

Proposition 5.13. *Let $q \geq 2$ be given and let h be small enough. There exists a constant $C > 0$ such that*

$$Ch^{3/2}\|X\| \leq \|(I - A^q)X\| \ \forall X \in V. \tag{5.65}$$

Proof. One has for all $X \in V$

$$\begin{aligned}\|(I - A)X\| &\leq \|(I - A^q)X\| + \|(A^q - A)X\| \\ &\leq \|(I - A^q)X\| + Ch^{q+1}\|X\|,\end{aligned}$$

so that for all $X \in V$

$$\|(I - A)X\|_V - Ch^{q+1}\|X\| \leq \|(I - A^q)X\|.$$

Lemma 5.12 concludes the proof since $h^{q+1} < h^{3/2}$ for h small enough. □

Proposition 5.14. *For h small enough the bilinear form a_q is uniformly coercive, i.e. for all $X \in V$*

$$|||X|||_q^2 \leq 3\Re\left(a_q(X, X)\right).$$

Proof. One has $|||X|||_q^2 \leq \|X\|^2 - 2\Re\left(A_qX, X\right) + \|A_qX\|^2$. Since

$$\|A_qX\| \leq \|AX\| + \|(A_q - A)X\| \leq (1 + Ch^{q+1})\|X\| \tag{5.66}$$

there exists another constant denoted $C' > 0$ such that

$$\|A_qX\|^2 \leq (1 + C'h^{q+1})\|X\|^2.$$

Therefore

$$|||X|||_q^2 \leq 2\|X\|^2 + C'h^{q+1}\|X\|^2 - 2\Re\left(A_qX, X\right),$$

that is $|||X|||_q^2 - C'h^{q+1}\|X\|^2 \leq 2\Re\left(a_q(X,X)\right)$. For small h since $q > 3/2$ and due to Proposition 5.13 one has

$$C'h^q\|X\|^2 \leq C'h^{1/3}(h^{3/2} - h^q)\|X\|^2 \leq h^{1/3}|||X|||_q^2,$$

then

$$\frac{2}{3}|||X|||_q^2 \leq |||X|||_q^2 - C'h^q\|X\|^2.$$

Combined with the previous inequality it proves the claim. □

Remark that the following result was proved in (5.66).

Lemma 5.15. *The operator A^q satisfies $\|A^q\| \leq 1 + Ch^{q+1}$.*

The main convergence result is an adapted version of second Strang lemma with the $|||\cdot|||_q$ norm. This result stated in Theorem 5.3.3, combined with the additional estimate from Lemma 5.16, is the corner stone of the estimate (5.69) that concludes this numerical analysis.

Theorem 5.3.3. Suppose that $q \geq 2$ and h is small enough to satisfy hypothesis of Porpositions 5.13 and 5.14. Denote $X \in V$ the solution of the exact problem (5.24) in dimension one and $X_h \in V$ the solution of the discrete problem (5.54). Then there exists a constant $C > 0$ such that

$$|||X - X_h|||_q \leq Ch^{-3/2}\left(\inf_{Y_h \in V}|||X - Y_h|||_q + \sup_{W_h \in V\setminus\{0\}}\frac{|a_q(X, W_h) - f_q(W_h)|}{\|W_h\|}\right), \quad (5.67)$$

where $f_q(Y) = (B^q, Y)_V$.

Proof.

- The first ingredient is the uniform coercivity with respect to $|||.|||_q$ needed in the second Strang lemma. It is proved in Proposition 5.14.
- The second step consists in characterizing the uniform continuity of a_q. For all $(X, Y) \in V^2$

  $$|a_q(X,Y)| = |((I - A^q)X, Y)| \leq |||X|||_q\,\|Y\|.$$

  Using (5.65) one has $\|W_h\| \leq Ch^{-3/2}|||W_h|||_q$ for some constant C, so that for small h

  $$\forall (X,Y) \in V^2,\ |a_q(X,Y)| \leq Ch^{-3/2}|||X|||_q|||Y|||_q.$$

- The last step is the inequality itself. The triangle inequality yields

  $$|||X - X_h|||_q \leq |||X - Y_h|||_q + |||X_h - Y_h|||_q\ \forall Y_h \in V.$$

  On the other hand Proposition 5.6 shows that

  $$\begin{aligned}\frac{1}{3}|||X_h - Y_h|||_q^2 &\leq |a_q(X_h - Y_h, X_h - Y_h)|\,,\\ &\leq |a_q(X - Y_h, X_h - Y_h)| + |a_q(X - X_h, X_h - Y_h)|\,,\\ &\leq Ch^{-3/2}|||X - Y_h|||_q|||X_h - Y_h|||_q\\ &\qquad + |a_q(X, X_h - Y_h) - f_q(X_h - Y_h)|\,.\end{aligned}$$

As $W_h = X_h - Y_h \in V$, then

$$\frac{1}{3}|||X_h - Y_h|||_q \leq Ch^{-3/2}|||X - Y_h|||_q + \frac{|a_q(X, W_h) - f_q(W_h)|}{\|W_h\|}\frac{\|W_h\|}{|||W_h|||_q}.$$

Using one more time $\|W_h\| \leq Ch^{-3/2}|||W_h|||_q$, it yields the desired result. □

The residual defined by

$$D_h(X, W_h) = |a_q(X, W_h) - f_q(W_h)| \ \forall W_h \in V.$$

can be estimate with respect to the right hand sides of (5.36).

Lemma 5.16. *There exists a constant $C > 0$ such that*

$$\forall W_h \in V\backslash\{0\}, \ \frac{D_h(X, W_h)}{\|W_h\|} \leq Ch^{q+1}\left(\|X\| + \|g\|_{L^2(\Gamma)} + h\|f\|_{L^2(\Omega)}\right). \tag{5.68}$$

Proof. For all $W_h \in V\backslash\{0\}$,

$$\begin{aligned} D_h(X, W_h) &= |((I - A^q)X, W_h)_V - (B^q, W_h)_V|, \\ &\leq |((A - A^q)X, W_h)_V| + |((I - A)X, W_h)_V - (B, W_h)_V| \\ &\qquad + |(B - B_q, W_h)|, \\ &\leq Ch^{q+1}\|X\| \ \|W_h\| + Ch^{q+1}(\|g\|_{L^2(\Gamma)} + h\|f\|_{L^2(\Omega)})\|W_h\|, \end{aligned}$$

since the second term vanishes because $(I - A)X = B$. The third term is bounded using (5.36) and (5.5) like

$$\begin{aligned} |(B - B_q, W_h)| &\leq C(\|F - F_q\| \ \|g\|_{L^2(\Gamma)} + \|E - E_q\| \ \|f\|_{L^2(\Omega)})\|W_h\| \\ &\leq Ch^{q+1}(\|g\|_{L^2(\Gamma)} + h\|f\|_{L^2(\Omega)})\|W_h\|. \end{aligned}$$

This gives exactly (5.68). □

It is now easy to prove the theoretical convergence of the method in dimension one.

Theorem 5.3.4. One has the estimate

$$|||X - X_h|||_q = O(h^{q-1/2}). \tag{5.69}$$

Proof. In dimension one the discrete space of approximation is equal to V whatever the method of construction of basis functions is. This is why one can choose $Y_h = X$ in (5.67). So $\inf_{Y_h \in V} |||X - Y_h|||_q = 0$. The remaining term is bounded with (5.68). □

It is useful to rewrite this inequality using a norm with the usual scaling

$$\overline{\|Z\|} = \sqrt{\sum_{k \in [\![1, N_h]\!]} h|Z_k|^2}.$$

By construction $\overline{\|Z\|} = h^{\frac{1}{2}}\|Z\|$. Using (5.65) one gets $\overline{\|Z\|} \leq Ch^{-1}|||Z|||_q$. Therefore a corollary of the theorem is the estimate of convergence

$$\overline{\|X - X_h\|} = O(h^{q-3/2}). \tag{5.70}$$

5.4 Dimension two : interpolation property of the GPW

Since the comprehensive numerical analysis presented in dimension one does not hold in dimension two, a less involved path will be followed in this case. The main result, Theorem 5.4.1, states that the h-convergence of the projection of the solution X on the finite dimension space V^q - which is explicitly designed in paragraph 5.2.3 - converges toward X. The operator P_h still denotes the projection from V on V^q.

Theorem 5.4.1. Let u be a solution of a homogeneous Helmholtz problem (5.1). Assume that u is of class C^{n+1} with $n \geq 1$. Let $X \in V$ satisfy $X_{|\partial\Omega_k} = (-\partial_{\nu_k} + i\sigma)u_{|\partial\Omega_k}$. The number p of basis functions $Z_{k,l} = (-\partial_\nu + i\sigma)\varphi_k^l$ per element Ω_k is fixed for every element, $p = 2n+1$. Then a constant $C > 0$ exists, depending on n and the problem's data σ such that

$$\|(I - P_h)X\|_V \leq Ch^{n-1/2}\|u\|_{C^{n+1}(\Omega)}.$$

The uniform independence of the shape functions with respect to h has already been commented in Remark 13.

Since in this section everything is local, k will be used here as a summation index, hopefully bringing no confusion with the index of the mesh elements.

5.4.1 Preliminary : Chain rule

This section is dedicated to describing the formula to derive a composition of two functions, in dimensions one and two. A wide bibliography about this formula is to be found in [Ma09]. It is linked to the notion of partition of an integer.

Faa Di Bruno formula gives the mth derivative of a composite function with a single variable. It is named after Francesco Faa Di Bruno, but was stated in earlier work of Louis Antoine François Arbogast around 1800, see [Cra05].

If f and g are functions with enough derivatives, then

$$\frac{d^m}{dx^m}f(g(x)) = m!\sum f^{(\sum_k b_k)}(g(x))\prod_{k=1}^{m}\frac{1}{b_k!}\left(\frac{g^{(k)}(x)}{k!}\right)^{b_k},$$

where the sum is over all nonnegative integers $(b_k)_{k\in[\![1,m]\!]}$ such that $\sum_k kb_k = m$. These solutions are actually the partitions of m.

The multivariate formula has been widely studied, the version described here is the one from [CS96] applied to dimension 2. A linear order on $\mathbb{N}^2$ is defined by : for all $(\rho,\xi) \in (\mathbb{N}^2)^2$, the relation $\rho \prec \xi$ holds provided that

- $\rho_1 + \rho_2 < \xi_1 + \xi_2$; or
- $\rho_1 + \rho_2 = \xi_1 + \xi_2$ and $\rho_1 < \xi_1$.

If f and g are functions with enough derivatives, then

$$\partial_x^i\partial_y^j f(g(x,y)) = i!j!\sum_{1\leq\mu\leq i+j} f^\mu(g(x,y))\sum_{s=1}^{i+j}\sum_{p_s((i,j),\mu)}\prod_{l=1}^{s}\frac{1}{k_l!}\left(\frac{1}{i_l!j_l!}\partial_x^{i_l}\partial_y^{j_l}(g(x,y))\right)^{k_l}, \tag{5.71}$$

where the partitions of (i,j) are defined by the following sets : for all $\mu \in [\![1, i+j]\!]$, for all $s \in [\![1, i+j]\!]$,

$$p_s((i,j),\mu) = \Big\{(k_1,\cdots,k_s;(i_1,j_1),\cdots,(i_s,j_s)) : k_i > 0, 0 \prec (i_1,j_1) \prec \cdots \prec (i_s,j_s),$$

$$\left.\sum_{l=1}^{s} k_l = \mu, \sum_{l=1}^{s} k_l i_l = i, \sum_{l=1}^{s} k_l j_l = j \right\}.$$

See [Har06] for a proof of the formula interpreted in terms of collapsing partitions.

5.4.2 A fundamental property of the shape functions

Since the design and the interpolation study are based on different Taylor expansions, the derivatives of the shape function φ are important quantities. Both

- the coefficients $\lambda_{i,j}$s defining a shape function φ
- the derivatives of φ

are here expressed as polynomials with two variables with respect to $(\lambda_{1,0}, \lambda_{0,1})$. The following Lemma 5.17 and Proposition 5.18 give a description of these quantities with respect to the only non zero coefficients fixed by the N-normalization, namely $(\lambda_{1,0}, \lambda_{0,1})$.

Lemma 5.17. *The coefficients*

$$\{\lambda_{i,j}, 0 \leq i \leq q+1, 0 \leq j \leq q+1-i\}$$

defined by the induction formula (5.45) *can be described as polynomials with two variables with respect to* $(\lambda_{1,0}, \lambda_{0,1})$ *:*

$$\begin{cases} \forall i \geq 2 \\ \lambda_{i,j} \text{ is of total degree at most } i-2. \end{cases} \tag{5.72}$$

The whole proof relies on an investigation of the induction formula (5.45).

Proof. The existence and uniqueness of a solution stems from the induction relation (5.45) defining the system. See Figure 5.5.

Because of the point 2 of the N-normalization, formula (5.45) for $i = 0$ and $i = 1$ reads

$$\begin{cases} \beta(G) &= 2\lambda_{2,0} + (\lambda_{1,0})^2 + (\lambda_{0,1})^2, \\ \dfrac{\partial_y^j \beta(G)}{j!} &= 2\lambda_{2,j} & \forall j > 0, \\ \partial_x \beta(G) &= 6\lambda_{3,0} + 4\lambda_{2,0}\lambda_{1,0}, \\ \dfrac{\partial_x \partial_y^j \beta(G)}{j!} &= 6\lambda_{3,j} + 4\lambda_{2,j}\lambda_{1,0} & \forall j > 0. \end{cases} \tag{5.73}$$

Then (5.72) for $i = 2$ stems from point 1 of the normalization. Indeed for $j = 0$ the sum $(\lambda_{1,0})^2 + (\lambda_{0,1})^2$ does not depend on $(\lambda_{1,0}, \lambda_{0,1})$ themselves but only on N. Afterwards (5.72) for $i = 3$ is clear from (5.73).

Now set $i \geq 2$ and suppose that the statement (5.72) holds true for all $\tilde{i} \in [\![3, i+1]\!]$. Then, isolating $\lambda_{i+2,j}$ in (5.45), the highest possible degree of each term is

- $i-2$ for the term in $\lambda_{i,j+2}$,
- $(i-1)+1$ for the term in $\lambda_{i+1,j}\lambda_{1,0}$,
- $(i-k-1)+(k-1)$ for the terms in $\lambda_{i-k+1,j-l}\lambda_{k+1,l}$ with $k \neq 0$ and $k \neq i$,
- $(i-2)+1$ for the term in $\lambda_{i,j+1}\lambda_{0,1}$,
- $(i-l-2)+(l-2)$ for the term in $\lambda_{i-l,j-k+1}\lambda_{l,k+1}$ with $l \neq 0$ and $l \neq i$, note that $\lambda_{i-l,j-k+1}\lambda_{l,k+1} = 0$ with $l \neq 1$ and $l \neq i-1$ because of the point 2 of the normalization.

As a consequence the terms with higher degree appearing in the expression of $\lambda_{i+2,j}$ have degree at most equal to i. It completes the proof of (5.72) for $i > 2$ by induction. □

Proposition 5.18. *Consider a shape function $\varphi = e^P$ designed with the N-normalization. Then for all $(i,j) \in \mathbb{N}^2$ such that $i+j \leq q+1$ there is a polynomial $R_{i,j} \in \mathbb{C}[X,Y]$ such that $\mathrm{d}R_{i,j} \leq i-2$ and such that*

$$\partial_x^i \partial_y^j \varphi(G) = (\lambda_{0,1})^j (\lambda_{1,0})^i + R_{i,j}(\lambda_{1,0}, \lambda_{0,1}). \tag{5.74}$$

The coefficients of $R_{i,j}$ only depend on N and on the derivatives of β.

Remark 15. Since $(\lambda_{1,0})^2 + (\lambda_{0,1})^2$ is given, none of the polynomial expressions that are at stake can be unique. For instance, considering the specificity of the β-normalization, any occurrence of $(\lambda_{1,0})^2$ could be replaced by $\beta(G) - (\lambda_{0,1})^2$ which would change the term of higher degree. This is the reason why $R_{i,j}$ is not unique : see Subsection 5.4.3 for a different point of view. However, the computations described in the proof of Lemma 5.17 give an explicit procedure for the computation of all $\lambda_{i,j}$s : this is the important point that will be used for practical implementation.

One could have expected the degree of $R_{i,j}$ to be smaller than $i+j-1$. The fact that it does not depend on j is clearly due to the choice of $\{\lambda_{i,j}, i \in \{0,1\}, i+j>1\}$ to be zero, the point 2 of the normalization. The fact that it is smaller than $i-2$ and not even $i-1$ is due to the fact that the degree of $\lambda_{2,j}$ is 0, since $(\lambda_{1,0})^2 + (\lambda_{0,1})^2 = N$ is constant.

Proof. Applying the chain rule introduced in equation (5.71) to $\varphi = e^P$ one gets for all $(i,j) \in \mathbb{N}^2$,

$$\partial_x^i \partial_y^j \varphi(G) = i!j! \sum_{\mu=1}^{i+j} \sum_{s=1}^{i+j} \sum_{p_s((i,j),\mu)} \prod_{l=1}^{s} \frac{(\lambda_{i_l,j_l})^{k_l}}{k_l!}, \tag{5.75}$$

where $p_s((i,j),\mu)$ is the set of partitions of (i,j) with length μ :

$$\left\{ (k_l,(i_l,j_l))_{l \in [\![1,s]\!]} : k_l \in \mathbb{N}^*, 0 \prec (i_1,j_1) \prec \cdots \prec (i_l,j_l), \sum_{l=1}^{s} k_l = \mu, \sum_{l=1}^{s} k_l (i_l,j_l) = (i,j) \right\}. \tag{5.76}$$

Now consider such a partition to be given and focus on the degree of the corresponding product term, namely $\prod_{l=1}^{s} (\lambda_{i_l,j_l})^{k_l}$. Thanks to Lemma 5.17 one can split this product into different terms regarding their degree as polynomials with respect to $(\lambda_{1,0}\lambda_{0,1})$. As a result, since $Deg \prod_{l=1}^{s} (\lambda_{i_l,j_l})^{k_l} = \sum_{l=1}^{s} k_l Deg\ \lambda_{i_l,j_l}$, this quantity is also at most equal to

$$\sum_{i_l=0, j_l=1} k_l j_l + \sum_{i_l=1, j_l=0} k_l i_l + \sum_{i_l=2} k_l \cdot 0 + \sum_{i_l \geq 3} k_l (i_l - 2), \tag{5.77}$$

where the two first sums contain at most one term each.

Obviously the leading term in $\partial_x^i \partial_y^j \varphi(G)$ is $(\lambda_{0,1})^j (\lambda_{1,0})^i$, it corresponds to the partition $(i,j) = j(0,1) + i(1,0)$. Indeed, as long as a partition contains at least one term such that $i_l \geq 2$, the resulting degree computed from (5.77) will contain at least one term $k_l \cdot 0$ or $k_l(i_l - 2)$, and any of them is at most $k_l(i_l + j_l) - 2$; as a consequence the degree computed in (5.77) is then strictly lower than $\sum_{l=1}^{s} k_l (i_l + j_l) - 2 = i + j - 2$.

Since the product term corresponding to the partition $j(0,1) + i(1,0)$ is $(\lambda_{0,1})^j (\lambda_{1,0})^i / (j!i!)$ it completes the proof. □

5.4.3 A more algebraic viewpoint

This paragraph presents a more algebraic point of view on the non uniqueness mentioned in Proposition 5.18 and commented in Remark 15.

The point 1 of the N-normalization gives that $P_N = (\lambda_{1,0})^2 + (\lambda_{0,1})^2 - N$ satisfies $P_N = 0$ for the p different functions of $\mathcal{E}(G,p)$. From then on, considering other quantities as polynomials with two variables in $(\lambda_{0,1}, \lambda_{1,0})$ is in fact computing in the quotient ring $\mathbb{C}[\lambda_{1,0}, \lambda_{0,1}]/(P_N)$ of $\mathbb{C}[\lambda_{1,0}, \lambda_{0,1}]$ modulo the ideal generated by P_N. For instance, the system (5.73) reads

$$\left\{\begin{array}{lll} \lambda_{2,0} &= \dfrac{\beta(G)-N}{2} & (P_N), \\ \lambda_{2,j} &= \dfrac{\partial_y^j \beta(G)}{2(j!)} & (P_N),\ \forall j>0, \\ \lambda_{3,0} &= \dfrac{\partial_x \beta(G) - 2\lambda_{1,0}(\beta(G)-N)}{6} & (P_N), \\ \lambda_{3,j} &= \dfrac{\partial_x \partial_y^j \beta(G)}{6(j!)} + 2\dfrac{\partial_y^j \beta(G)}{j!}\lambda_{1,0} & (P_N),\ \forall j>0. \end{array}\right. \tag{5.78}$$

Of course in this quotient ring, each equivalence class has an infinite number of elements, and all the computations of the previous subsection are performed on elements of these classes. Thus any equality applies to all the elements of the same class. Note that since the ring considered here is the ring of polynomials with two variables, there is no such thing as the Euclidean division. As a result there is nothing like a canonical element of a class used for computations. One can easily see that for $q \geq 4$

$$\begin{aligned} \partial_x^4 \partial_y \varphi(G) &= (\lambda_{1,0})^4(\lambda_{1,0}) + 2\partial_y\beta(G)\left((\lambda_{1,0})^2 - (\lambda_{0,1})^2\right) + 2\partial_x\beta(G)\lambda_{0,1}\lambda_{1,0} + 2\partial_x\partial_y\beta(G)\lambda_{1,0} \\ &\quad + (-3\partial_y^2\beta(G) + \partial_x\beta(G))\lambda_{0,1} - \partial_y^3\beta(G) + \partial_x^2\partial_y\beta(G), \\ &= (\lambda_{1,0})^4(\lambda_{1,0}) + 2\partial_y\beta(G)\left((\lambda_{1,0})^2 + (\lambda_{0,1})^2\right) + 2\partial_x\beta(G)\lambda_{0,1}\lambda_{1,0} + 2\partial_x\partial_y\beta(G)\lambda_{1,0} \\ &\quad + (-3\partial_y^2\beta(G) + \partial_x\beta(G))\lambda_{0,1} - \partial_y^3\beta(G) + \partial_x^2\partial_y\beta(G) - 2\beta(G)\partial_y\beta(G), \end{aligned}$$

which gives two possible $R_{4,1} \in \mathbb{C}[\lambda_{1,0}, \lambda_{0,1}]$ satisfying (5.74) in Proposition 5.18.

This work does definitely not aim at profiting from this algebraic point of view.

5.4.4 Interpolation

This subsection focuses on the interpolation property of the set of shape functions $\mathcal{E}(G,p)$. The sketch of the proof follows the one developed by Cessenat in [CD98], but it is adapted to the generalized plane wave shape functions. The proof of Theorem 5.4.1 finally represents the application of this result from the UWVF perspective.

Definition 26. For all $l \in \mathbb{N}$ let e_l be the classical plane wave defined as

$$e_l(x,y) = e^{i\kappa((x-x_G)\cos\theta_l + (y-y_G)\sin\theta_l)},$$

and φ_l the shape function defined with the N-normalization with $N = i\kappa \in \mathbb{C}^*$, θ_l being $2\pi(l-1)/(2n+1)$. Suppose $n \in \mathbb{N}^*$. The $(n+1)(n+2)/2 \times (2n+1)$ matrices M_n^C and M_n are defined as follows : for all $(k_1,k_2) \in \mathbb{N}^2$, such that $k_1 + k_2 \leq n$

$$\left\{\begin{array}{l} \left(M_n^C\right)_{\frac{(k_1+k_2)(k_1+k_2+1)}{2}+k_2+1,l} = \frac{\partial_x^{k_1}\partial_y^{k_2} e_l(G)}{k_1!k_2!}, \\ \left(M_n\right)_{\frac{(k_1+k_2)(k_1+k_2+1)}{2}+k_2+1,l} = \frac{\partial_x^{k_1}\partial_y^{k_2} \varphi_l(G)}{k_1!k_2!}. \end{array}\right.$$

Their lth columns contain respectively the Taylor expansion coefficients of the functions e_l and φ_l.

For instance, one has $M_1 = \begin{pmatrix} \varphi_1(G) & \varphi_2(G) & \varphi_3(G) \\ \partial_x\varphi_1(G) & \partial_x\varphi_2(G) & \partial_x\varphi_3(G) \\ \partial_y\varphi_1(G) & \partial_y\varphi_2(G) & \partial_y\varphi_3(G) \end{pmatrix}$, with the classical plane waves $M_1^C = \begin{pmatrix} 1 & 1 & 1 \\ i\kappa\cos\theta_1 & i\kappa\cos\theta_2 & i\kappa\cos\theta_3 \\ i\kappa\sin\theta_1 & i\kappa\sin\theta_2 & i\kappa\sin\theta_3 \end{pmatrix}$ and

$$M_2 = \begin{pmatrix} \varphi_1(G) & \varphi_2(G) & \varphi_3(G) & \varphi_4(G) & \varphi_5(G) \\ \partial_x\varphi_1(G) & \partial_x\varphi_2(G) & \partial_x\varphi_3(G) & \partial_x\varphi_4(G) & \partial_x\varphi_5(G) \\ \partial_y\varphi_1(G) & \partial_y\varphi_2(G) & \partial_y\varphi_3(G) & \partial_y\varphi_4(G) & \partial_y\varphi_5(G) \\ \partial_x^2\varphi_1(G)/2 & \partial_x^2\varphi_2(G)/2 & \partial_x^2\varphi_3(G)/2 & \partial_x^2\varphi_4(G)/2 & \partial_x^2\varphi_5(G)/2 \\ \partial_x\partial_y\varphi_1(G) & \partial_x\partial_y\varphi_2(G) & \partial_x\partial_y\varphi_3(G) & \partial_x\partial_y\varphi_4(G) & \partial_x\partial_y\varphi_5(G) \\ \partial_y^2\varphi_1(G)/2 & \partial_y^2\varphi_2(G)/2 & \partial_y^2\varphi_3(G)/2 & \partial_y^2\varphi_4(G)/2 & \partial_y^2\varphi_5(G)/2 \end{pmatrix}.$$

The rank of the matrix M_n^C is computed for a general set of classical plane waves in Lemma 5.19, which profits from the fact that the result proved by Cessenat and Després in [CD98] for $\kappa > 0$ is actually still valid for $\kappa \in \mathbb{C}^*$. The proof of Proposition 5.21 relies on Lemma 5.20 that explicits the link between the matrix M_n^C and the corresponding matrix M_n built with the generalized plane waves. The parameter κ is likely to become $-iN$ for in the study of the N-normalized case.

Lemma 5.19. *There are two matrices : a rectangle matrix P_n only depending on $\beta(G)$ and a square invertible matrix S_n only depending on the directions θ_l such that $S_n = P_n \cdot M_n^C$. Moreover $rk(M_n^C) = 2n+1$.*

As previously announced, the proof follows exactly the steps of [CD98]. *Proof.* Consider M_n^C be the matrix introduced in Definition 26 so that for all $(k_1, k_2) \in \mathbb{N}^2$, such that $k_1 + k_2 \leq n$

$$\left(M_n^C\right)_{\frac{(k_1+k_2)(k_1+k_2+1)}{2}+k_2+1,l} = \frac{\partial_x^{k_1}\partial_y^{k_2} e_l(G)}{k_1!k_2!} = \frac{(i\kappa)^{k_1+k_2}}{k_1!k_2!}\cos^{k_1}\theta_l \sin^{k_2}\theta_l.$$

Define for all $k \in [\![0, n]\!]$

$$(S_n)_{n\pm k+1,l} = \frac{1}{(i\kappa)^k}\left(\partial_x \pm i\partial_y\right)^k e_l(G) = \frac{k!}{(i\kappa)^k}\sum_{s=0}^{k}\frac{(-i)^s\partial_x^{(k-s)}\partial_y^s e_l(G)}{(k-s)!s!}.$$

Thanks to the definition of M_n^C one can check that

$$(S_n)_{n\pm k+1,l} = \frac{k!}{(i\kappa)^k}\sum_{s=0}^{k}(\pm i)^s (M_n^C)_{\frac{((k-s)+s)((k-s)+s+1)}{2}+s+1,l},$$

so that S_n is a $(2n+1)\times(2n+1)$ matrix that is a linear transform of M_n^C. More precisely, define P_n as an $(2n+1)\times\frac{(n+1)(n+2)}{2}$ matrix such that

$$(P_n)_{l,\frac{k(k+1)}{2}+s+1} = k!(\pm i)^s/(i\kappa)^k.$$

Then $S_n = P_n \cdot M_n^C$. As a consequence, $rk(M_n^C) \geq rk(S_n)$.

The rank of S_n is now to be evaluated thanks to the definition of the plane waves e_l. Since $e_l(x,y) = e^{(i\kappa)((x-x_G)\cos\theta_l+(y-y_G)\sin\theta_l)}$ then

$$\left(\partial_x \pm i\partial_y\right)^k e_l = (i\kappa)^k(\cos\theta_l \pm i\sin\theta_l)^k e_l.$$

Consider that $z_l = \cos\theta_l + i\sin\theta_l = (\cos\theta_l - i\sin\theta_l)^{-1}$ because $|z_l| = 1$, and since $e_l(G) = 1$ it yields

$$(\partial_x \pm i\partial_y)^k e_l(G) = (i\kappa)^k (z_l)^{\pm k} \Rightarrow (S_n)_{n\pm k+1,l} = (z_l)^{\pm k}.$$

Thus S_n's columns are proportional to the one of a Vandermonde matrix and

$$det\ S_n = \prod_{i=1}^{n} z_i^{-n} \prod_{i<j} (z_i - z_j).$$

From the choice of θ_ls, for all $i \neq j : z_i \neq z_j$ so that S_n is invertible and $rk(M_n^C) \geq rk(S_n) = 2n+1$. Since

$$rk(M_n^C) \leq \min\left(2n+1, \frac{(n+1)(n+2)}{2}\right) = 2n+1$$

the proof is then completed. □

Lemma 5.20. *Consider $\mathcal{E}(G,p)$ introduced in Definition 22 and M_n^C built with $\kappa = -iN$. Then there is a lower triangular matrix L_n, which diagonal coefficients are all equal to 1 and which other coefficients are linear combinations of the derivatives of β evaluated at G, such that*

$$M_n = L_n \cdot M_n^C. \tag{5.79}$$

As a consequence $rk(M_n) = rk(M_n^C)$ and both $\|L_n\|$ and $\|(L_n)^{-1}\|$ are bounded by a constant only depending on β.

The following proof is straightforward considering the feature of the derivatives of φ_l described in Proposition 5.18.

Proof. From (5.74) there exists a polynomial $R_{i,j} \in \mathbb{C}[X,Y]$ with $Deg\ R_{i,j} \leq i-2$ such that

$$\forall (i,j) \in \mathbb{N}^2, \partial_x^i \partial_y^j \varphi_l(G) = \partial_x^i \partial_y^j e_l(G) + R_{i,j}(\partial_x e_l(G), \partial_y e_l(G)). \tag{5.80}$$

The coefficients of $R_{i,j}$ do not depend on the shape function considered, but only depends on β and its derivatives evaluated at G. By construction of the classical plane wave e_l, one has

$$\begin{cases} \partial_x^k \partial_y^m e_l(G) &= (\partial_x e_l(G))^k (\partial_y e_l(G))^m, \\ &= (i\kappa)^{k+m} \cos(\theta)^k \sin(\theta)^m. \end{cases}$$

The numbering of the rows in matrices M_n^C and M_n is set up such that the derivatives of smaller order appear higher in the matrix, which proves (5.79). Indeed (5.80) shows that any coefficient of M_n is the sum of the corresponding coefficient in M_n^C plus a linear combination - which coefficients do not depend on the column that is considered but only on β and its derivatives evaluated at G - of terms that appear higher in the corresponding column of M_n.

The rank of M_n is then equal to the rank of M_n^C, and $\|L_n\|$ and $\|(L_n)^{-1}\|$ do only depend on the coefficients of $R_{i,j}$. As a result they do not depend on the shape functions but only on the coefficient β and its derivatives at G. □

For all $n \in \mathbb{N}$, denote by $\|\cdot\|_{\mathcal{C}^n}$ the norm defined by $\sum_{j=0}^{n} \|\hat{A}\ \cdot\|_\infty$.

Proposition 5.21. *Suppose u is a solution of Helmholtz equation (5.1) in a vicinity $\mathcal{V}_G$ of $G \in \mathbb{R}^2$, h denoting the size of $\mathcal{V}_G$. In addition, suppose that $n \in \mathbb{N}$, $q \geq n+1$, $p = 2n+1$ and that u satisfies $u \in \mathcal{C}^{n+1}$. Then there are a function $u_a \in Span\mathcal{E}(G,p)$ depending on β and n, and a constant C depending on β and n such that for all $\overrightarrow{x} \in \mathcal{V}_G$*

$$\begin{cases} |u(\overrightarrow{x}) - u_a(\overrightarrow{x})| \leq Ch^{n+1} \|u\|_{\mathcal{C}^{n+1}}, \\ \|\nabla u(\overrightarrow{x}) - \nabla u_a(\overrightarrow{x})\| \leq Ch^n \|u\|_{\mathcal{C}^{n+1}}. \end{cases} \tag{5.81}$$

Proof. The idea of the proof is to look for $u_a = \sum_{l=1}^{2n+1} x_l \varphi_l$ by fitting its Taylor expansion to the one of u. This will be done by solving a linear system concerning the unknowns $(x_l)_{l\in[\![1,2n+1]\!]}$.

Since u belongs to $\mathcal{C}^{n+1}$ and for all $l \in [\![1, 2n+1]\!]$ the shape function φ_l belongs to $\mathcal{C}^\infty$, their Taylor expansions read for all $(x, y) \in \mathcal{V}_G$

$$\left| u(x,y) - \sum_{m=0}^{n} \sum_{k_1+k_2=m} B_{k_1k_2} x^{k_1} y^{k_2} \right| \leq C h^{n+1} \|u\|_{\mathcal{C}^{n+1}},$$

$$\left| \varphi_l(x,y) - \sum_{m=0}^{n} \sum_{k_1+k_2=m} M^l_{k_1k_2} x^{k_1} y^{k_2} \right| \leq C h^{n+1} \|\varphi_l\|_{\mathcal{C}^{n+1}},$$

where for the sake of simplicity $M^l_{k_1k_2}$ stands for the coefficient of M_n that corresponds to $\partial_x^{k_1}\partial_y^{k_2}\varphi_l/(k_1!k_2!)$, namely the coefficient $(M_n)_{\frac{(k_2+k_1)(k_2+k_1+1)}{2}+k_2+1,l}$, and in the same way B_{k_1,k_2} stands for $(B_n)_{\frac{(k_2+k_1)(k_2+k_1+1)}{2}+k_2+1}$. The system to be solved is then

$$\begin{cases} \text{Find } (x_l)_{l\in[\![1,2n+1]\!]} \in \mathbb{C}^{2n+1} \text{ s. t.} \\ \sum_{l=1}^{2n+1} M^l_{k_1,k_2} x_l = B_{k_1,k_2}, \ \forall m \in [\![0,n]\!], \ \forall (k_1,k_2) \in [\![0,n]\!]^2 \text{ s. t. } k_1+k_2 = m. \end{cases}$$

In order to study the system's matrix, the equations depending on (k_1, k_2) have to be numbered : they will be considered with increasing $m = k_1 + k_2$, and with decreasing k_1 for a fixed value of m. Defining the corresponding vector $B_n \in \mathbb{C}^{\frac{(n+1)(n+2)}{2}}$, together with the unknown $X^n = (x_1, x_2, \cdots, x_{2n+1}) \in \mathbb{C}^{2n+1}$, the system now reads

$$\begin{cases} \text{Find } X^n \in \mathbb{C}^{2n+1} \text{ such that} \\ M_n \cdot X^n = B_n \end{cases}$$

where $M_n \in \mathbb{C}^{\frac{(n+1)(n+2)}{2}\times(2n+1)}$ is the matrix from Definition 26.

Since the system is not square, there is a solution if and only if $B_n \in Im(M_n)$.

i) The technical point is to prove that $rk(M_n) = 2n+1$. It is straightforward from Lemmas 5.20 and 5.19.

ii) There exists a subset $K \subset \mathbb{C}^{\frac{(n+1)(n+2)}{2}}$ such that $Im(M_n) \subset K$ and $B_n \in K$. This subspace K is built from the fact that the shape functions are designed to fit the Taylor expansion of the Helmholtz equation :

$$K := \Bigg\{ (C_{k_1,k_2}) \in \mathbb{C}^{\frac{(n+1)(n+2)}{2}}, \forall (k_1,k_2) \in \mathbb{N}^2, k_1+k_2 \leq n-2,$$

$$(k_1+1)(k_1+2)C_{k_1+2,k_2} + (k_2+1)(k_2+2)C_{k_1,k_2+2} = \sum_{i=0}^{k_1}\sum_{j=0}^{k_2} \frac{\partial_x^i \partial_y^j \beta(G)}{i!j!} C_{k_1-i,k_2-j} \Bigg\} \tag{5.82}$$

All shape function φ_l, $l \in [\![1, 2n+1]\!]$, satisfy $(-\Delta + \beta)\varphi_l = (-P_{\Delta,l} + \beta)\varphi_l$. From the equation (5.44) with $q \geq n+1$, it is then straightforward to see that $Im(M_n) \subset K$. The fact that $B_n \in K$ simply stems from plugging the Taylor expansions of u and β into Helmholtz equation.

iii) The dimension of K defined by (5.82) is $2n+1$. Indeed, one can check - using the same numbering as previously for the equations - that K is defined by $n(n+1)/2$ linearly independent relations on $\mathbb{C}^{\frac{(n+1)(n+2)}{2}}$, so that its dimension is $(n+1)(n+2)/2-n(n+1)/2$.

As a consequence, from the solution to the system $M_n \cdot X^n = B_n$ that now is known to exist, one can define $u_a = \sum_{l=1}^{2n+1} x_l \varphi_l$. Thanks to that definition and to the Taylor expansions of u and the φ_ls it yields

$$|u(\overrightarrow{x}) - u_a(\overrightarrow{x})| \leq Ch^{n+1} \left(\|u\|_{\mathcal{C}^{n+1}} + \|u_a\|_{\mathcal{C}^{n+1}} \right).$$

Moreover one has $X^n = (S_n^C)^{-1} P_n^C (L_n)^{-1} B_n$, where $(S_n^C)^{-1} P_n^C$ is bounded from above by $\sup_{l \in [\![1,2n+1]\!]} \|e_l\|_{\mathcal{C}^{n+1}}$, see Lemma 5.19, $(L_n)^{-1}$ is bounded from above by a constant depending only on β and its derivatives from Lemma 5.20, and B_n is bounded by $\|u\|_{\mathcal{C}^{n+1}}$. Since for all $l \in [\![1, 2n+1]\!]$ it yields $|x_l| \leq C\|u\|_{\mathcal{C}^{n+1}}$, it turns out to be the first part of (5.81) :

$$|u(\overrightarrow{x}) - u_a(\overrightarrow{x})| \leq C(2n+2)h^{n+1} \|u\|_{\mathcal{C}^{n+1}}.$$

At last, the second part of (5.81) stems from taking the Taylor Lagrange formula of the gradient of $u - u_a$, up to the order n, since

$$\sum_{m=0}^{n} \sum_{k_1+k_2=m} \left(B_{k_1 k_2} (x - x_G)^{k_1} (y - y_G)^{k_2} - \sum_{l=1}^{2n+1} \left(x_l M_{k_1 k_2}^l (x - x_G)^{k_1} (y - y_G)^{k_2} \right) \right) = 0.$$

That is : for all $\overrightarrow{x} = (x, y) \in V_G$ there is (ζ_1, ζ_2) on the segment line between $\overrightarrow{x}$ and G such that

$$\begin{cases} \partial_x (u - u_a)(\overrightarrow{x}) = \sum_{l=0}^{n} \dfrac{\partial_x^{l+1} \partial_y^{n-l} (u - u_a)(\zeta_1)}{l!(n-l)!} (x - x_G)^l (y - y_G)^{n-l}, \\ \partial_y (u - u_a)(\overrightarrow{x}) = \sum_{l=0}^{n} \dfrac{\partial_x^{l} \partial_y^{n-l+1} (u - u_a)(\zeta_2)}{l!(n-l)!} (x - x_G)^l (y - y_G)^{n-l} \end{cases}$$

which indeed leads to the desired inequality. □

Proof.Of Theorem 5.4.1. Applying Proposition 5.21 on every element of the mesh, one gets an element $u_a \in \mathcal{E}$, $u_a = \sum_{k,l} x_k^l \varphi_k^l$ such that for all $\overrightarrow{x} \in \Omega$:

$$\begin{cases} |u(\overrightarrow{x}) - u_a(\overrightarrow{x})| \leq Ch^{n+1} \|u\|_{\mathcal{C}^{n+1}(\Omega)}, \\ \|\nabla u(\overrightarrow{x}) - \nabla u_a(\overrightarrow{x})\| \leq Ch^{n} \|u\|_{\mathcal{C}^{n+1}(\Omega)}, \end{cases}$$

where C depends on Ω. If $X_a \in V$ is defined by $(X_a)_{|\partial\Omega_k} = (-\partial_{\nu_k} + i\omega)(u_a)_{|\partial\Omega_k}$, then

$$\begin{aligned} \|X - X_a\|^2_{L^2(\partial\Omega_k)} &\leq 2 \int_{\partial\Omega_k} \|\nabla u - \nabla u_a\|^2 + 2\gamma^2 \int_{\partial\Omega_k} |u - u_a|^2 \\ &\leq 2C^2 h^{2n} \int_{\partial\Omega_k} (1 + \gamma^2 h^2) \|u\|_{C^{n+1}(\Omega)}, \end{aligned}$$

so that for h small enough

$$\|X - X_a\| \leq Ch^{n+1/2} \sqrt{N_h} \|u\|_{C^{n+1}(\Omega)}.$$

The result then stems from the fact that, for a regular mesh, the total number of elements can be bounded by C/h^2. □

5.5 Comments

5.5.1 On the generalization of the explicit design procedure

The design of an approximated solution $u = e^P$ for a general - not necessarily linear - differential operator of order $m \in \mathbb{N}^*$ on $\mathbb{R}^d$ for $d \in \mathbb{N}^*$ is a justifiable question. Consider the operator

$$Au = \sum_{k=0}^{m} \sum_{|\eta|=k} a_\eta \prod_{\sum \rho^i = \eta} \partial_{\overrightarrow{x}}^{\rho^i} u, \tag{5.83}$$

where η, ρ_i are a multi indexes, and the coefficients a_η belong to C^r. Considering a linear order on $\mathbb{N}^d$ and $\eta_{\max}$ the coefficient of higher index, the condition $a_{\eta_{\max}} \neq 0$ is a sufficient condition to generalize the explicit design procedure of shape functions for the operator (5.83).

5.5.2 On the numerical analysis in dimension two

It is interesting to identify the crucial points of a potential analysis in dimension two. Two main points are highlighted here that would establish a basis to adapt the second Strang lemma in dimension two for the new method.

The estimate of $F - F^q$

This estimate proved in Subsection 5.3.3 relies on two main ingredients : the one dimensional estimates from Subsection 5.3.2 on the one hand, the estimate given in Lemma 5.10 on the other hand.

The first ingredient is based on the estimate of the L^2 norm of a function in the domain via its L^2 norm on the boundary on the domain and the L^2 norm of its Jacobian in the domain. The bounds are explicit with respect to the size of the domain. Obtaining such an estimate in two dimension requires more attention.

The second ingredient implies heavier computations in dimension two since p basis functions are then involved. Indeed, expressing the elements W_js of the dual basis in the Z_j basis means solving a system of p equations which matrix is hermitian since its coefficients are the scalar products (Z_i, Z_j). Considering the third step, one can easily check that $(Z_i, Z_j) = \sigma h + O(h)$, so that for instance for $p = 3$ one has

$$\sum_{j=1}^{p} \|Z_j\| \, \|W_j\| \leq 9$$

when h goes to zero.

The coercivity property

The coercivity of the operator $I - A$ is a key property in the study of the classical UWVF described in paragraph 5.2.1. However, this coercivity does not hold a priori in the space $(V, \|\cdot\|)$, but with the adapt norm $\|(I - A)\cdot\|$. This induces technical difficulties to study the coercivity property of the new method's operator $I - A^q$.

In dimension one, both proofs of Propositions 5.13 and 5.14 rely on the the estimate of $F - F^q$ provided in Lemma 5.11. As a result the bilinear form a_q the desired coercivity property is not directly obtained in dimension two.

It would then be interesting to pursue this work in the direction of continuity estimates in dimension two between the norms $\|\cdot\|$, $|||\cdot||| = \|(I - A)\cdot\|$ and $|||\cdot|||_q \|(I - A^q)\cdot\|$.

The adapted second Strang Lemma

If both the coercivity and the estimate on $F - F^q$ were known, one could then think of generalizing the second Strang Lemma to extend the dimension one result.

In the proof of Lemma 5.11 everything would be the same in dimension two as in dimension one except that the proof of the estimate (5.53) proposed for the dimension one does not hold in dimension two.

The proof of Lemma 5.12 in dimension two would require even more onerous computations.

Recent developments in the direction of the two dimensional case can be found in [MPS12].

5.5.3 Toward new horizons

Since the UWVF falls into the framework of Discontinuous Galerkin methods in [HMM07], it is plausible that considering the method that is presented in this chapter from a DG point of view could provide new tools for the theoretical study.

Another possibility would be to look at the Vekua theory [Vek67a, Hen57] that addresses general analytic coefficients in 2D. In [MHP11], an explicit study of the Helmholtz equation with constant coefficient is proposed. More generally Vekua's theory maps the solutions of second order elliptic partial differential equations to harmonic functions. During discussions with Peter Monk and Andrea Moiola it was pointed out that the kernel functions are not explicit in the case of varying coefficients, so that further investigations would be unavoidable.

Chapter 6

Numerical Results

Contents

6.1 Introduction

This chapter gathers the computational aspects of this work. The main part of the implementing work started during a visit to Peter Monk at the University of Delaware. The visit was funded by a grant of the Fondation Pierre Ledoux. The implementation started from a 2D code for the classical UWVF for elastic equations provided by Temmu Luostari, see [HMCK04], that has been adapted to the new method in collaboration with Peter Monk. I would like to thank both of them for this.

The resulting code is a Matlab 2D code with versions for both structured and unstructured meshes generated with Matlab *pdetoolbox*. All the linear systems are assembled and solved with Matlab. The goal of this code was not to get high performance but rather to demonstrate that the new method can reach high order convergence.

Section 6.2 concerns the h convergence of the new method in dimension one. It corresponds to Chapter 5. The results of this section were obtained thanks to a specific code in dimension one that I developed for my Master's thesis. Section 6.3 aims at highlighting the main difficulties that were met and overcome to produce a 2D code for the new method developed during this PhD work. Section 6.4 deals with the interpolation properties of the generalized plane waves in dimension two, and presents a range of cases for different types of wave propagation. Section 6.5 displays a comparison between numerical computations obtained for different choices, concerning either the quadrature formula or the other parameters of the method. It also presents some first test cases in a more physical test case, computed on a domain that contains around 50 wavelengths. A last section, Section 6.6, presents the first steps that were followed to adapt the code for the X mode together with a first test case.

6.2 O mode simulation in 1D

In one dimension there is no numerical integration required to compute the integrals : integrating on the skeleton of the mesh only requires the evaluation of the integrand at two points. As a result the error $u - u_h$ is not polluted by the numerical approximation of the integrals, like in two dimensions.

The test problem considered here is the following : the domain $\Omega =]a, b[\subset \mathbb{R}$ is given, and the system to be approximated is

$$\begin{cases} -u''(x) + x\ u(x) &= 0, & (]a,b[), \\ (\partial_\nu + i\sigma)u(x) &= (\partial_\nu + i\sigma)Ai(x), & (\{a,b\}), \end{cases}$$

where Ai is the first Airy function. It corresponds to $Q = 0$ and $g = (\partial_\nu + i\sigma)Ai$. The exact solution that is approximated is the first Airy function itself.

The flux parameter σ is set to be constant equal to 1 in the following examples. Simulations run with the classical choice $\sigma = \sqrt{\beta}$ - evaluated at the edge of the cells - give similar results, as long as the mesh is such that 0 is not part of any element boundary.

6.2.1 Validation of the theoretical convergence result

The domain is meshed uniformly by the set of points $\{x_k\}_{k\in[\![1,N_h+1]\!]}$, where N_h stands for the number of elements : $x_1 = a$ and $x_{N_h+1} = b$. The computed solution corresponds to an element $X_h \in V^q$, and since this is in dimension one then

$$V^q = Span\ \{Z_{k,1}, Z_{k,2}\}_{1\leq N_h} = V,$$

where $Z_{k,l}$ corresponds to the shape function $\varphi_{k,l}$ defined on Ω_k. A simple formula expresses the traces of u_h with respect to the solution of the discrete system $X_h = \sum_{k=1}^{N_h} x_{k,1}Z_{k,1} + x_{k,2}Z_{k,2} \in V$

$$\begin{cases} 2i\sigma u_h = (I + \Pi)X_h + g & (\{a,b\}), \\ 2i\sigma u_h = (I + \Pi)X_h & (\{x_k\}_{k\in[\![2,N_h]\!]}), \end{cases}$$

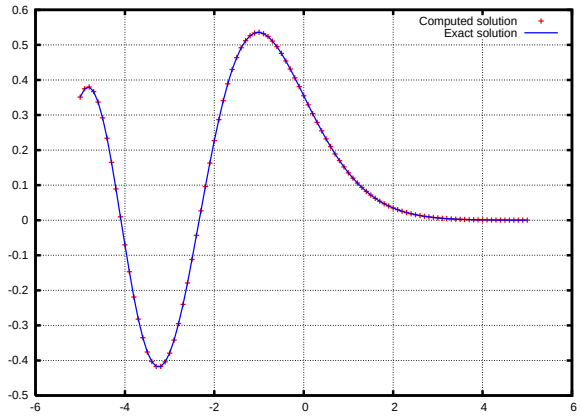


FIGURE 6.1 – Plot of the analytical Airy function on $\Omega =]-5,5[$, and comparison with the numerical solution computed with the generalized basis functions. Here we used 50 cells and two high order generalized plane waves per cell, precisely $q = 5$. One distinguishes between the propagative medium $x < 0$ and the non propagative medium $x > 0$.

so that u_h is explicitly reconstructed at the vertices of the mesh as

$$\begin{cases} u_h(a) = \dfrac{1}{2i\sigma}\left((1+Q)\displaystyle\sum_{l=1,2} x_{1,l}(\varphi'_{1,l} + i\sigma\varphi_{1,l}) + g\right)(a), \\ u_h = \dfrac{1}{2i\sigma}\left(\displaystyle\sum_{l=1,2} x_{k,l}(\varphi'_{k,l} + i\sigma\varphi_{k,l})(x_k) + \sum_{l=1,2} x_{k-1,l}(\varphi'_{k-1,l} + i\sigma\varphi_{k-1,l})(x_{k-1})\right), \\ \forall k \in [\![2, N_h]\!], \\ u_h(b) = \dfrac{1}{2i\sigma}\left((1+Q)\displaystyle\sum_{l=1,2} x_{N_h,l}(-\varphi'_{1,l} + i\sigma\varphi_{1,l}) + g\right)(b). \end{cases}$$

In the following, the accuracy is reported using a discrete l^2 norm

$$\sqrt{\frac{\sum_{k\in[\![1,N_h+1]\!]} |u_{ex}(x_k) - u_h(x_k)|^2}{\sum_{k\in[\![1,N_h+1]\!]} |u_{ex}(x_k)|^2}}.$$

A first computation is illustrated on Figure 6.1, representing the exact and approximated solutions of the Airy equation. It shows that the transition between the propagative and non propagative zones, at $x = 0$, is very well recovered by the method.

An investigation of the numerical convergence follows. The computational domain is $\Omega =]-5,5[$. The h-convergence is described in Table 6.1 and Figure 6.2. As expected, the rate of convergence increases with the parameter q. It highlights the fact that the numerical method reaches high order convergence on a case that includes a cut-off. Note that the numerical rates of convergence are better than the theoretical estimates.

Better convergence rates are observed for odd values of q compared to even values.

One can see that on the finest meshes the solution is accurate to machine precision for the highest values of parameter q.

Remark 16. The reason why Not Evaluated (*NE*) appears in Table 6.1 is the following. The computations require the evaluation of $e^{P(\pm h/2)}$. For $q = 6$, the degree of P is 7, and

TABLE 6.1 – h-convergence of the computed solution of the Airy equation on $\Omega =]-5, 5[$. Errors and orders of convergence for different orders of approximation q depending on the number of unknowns $N_u = 2N_h$, with $N_h = 10/h$ is the number of elements in the mesh. NE stands for Not Evaluated, see remark 16

	$q=2$		$q=3$		$q=4$		$q=5$		$q=6$	
N_u	Error	Rate	Error	Rate	Error	Rate	Error	Rate	Error	Rate
4	9.5e-01	-	9.9e-01	-	8.6e-01	-	8.6e-01	-	NE	-
8	9.2e-01	0.05	9.7e-01	0.03	9.7e-01	-0.18	9.9e-01	-0.20	9.9e-01	-
16	7.8e-01	0.23	9.5e-01	0.03	9.2e-01	0.09	9.6e-01	0.04	9.4e-01	0.04
32	6.0e-01	0.39	3.3e-01	1.51	2.5e-01	0.89	1.5e-01	2.65	1.1e-01	3.14
64	2.0e-01	1.59	3.2e-02	3.4	2.0e-02	3.61	3.2e-03	5.6	2.0e-03	5.75
128	5.4e-02	1.89	2.1e-03	3.91	1.3e-03	3.93	5.2e-05	5.94	3.2e-05	5.96
256	1.4e-02	1.97	1.3e-04	3.98	8.4e-05	3.98	8.2e-07	5.99	5.0e-07	5.99
512	3.4e-03	1.99	8.3e-06	4.00	5.3e-06	4.00	1.3e-08	6.00	7.9e-09	6.00
1024	8.6e-04	2.00	5.2e-07	4.00	3.3e-07	4.00	2.0e-10	6.00	1.2e-10	6.00
2048	2.2e-04	2.00	3.3e-08	4.00	2.1e-08	4.00	3.1e-12	5.99	1.9e-12	6.00
4096	5.4e-05	2.00	2.0e-09	4.00	1.3e-09	4.00	7.3e-14	5.43	7.5e-14	4.69
8192	1.3e-05	2.00	1.3e-10	4.00	8.1e-11	4.00	1.6e-14	2.21	5.8e-14	0.37
16384	3.4e-06	2.00	7.9e-12	4.01	5.0e-12	4.01	5.0e-14	-1.67	5.0e-14	0.20

for $N_u = 4$ unknowns the size of the mesh is $h = 5$. As a consequence the computations include the evaluation of a quantity close to $e^{(h^7)}$, which is considered as infinity up to the machine precision : it can not be evaluated.

6.2.2 Another normalization

Figure 6.3 and Table 6.2 display the same convergence results as in Section 6.2.1 but with basis functions designed with the normalization $\lambda_1^{\pm} = \pm\sqrt{\alpha(x_{k+1/2})}$. Comparing to Table 6.1 and Figure 6.2, one can see that the convergence rate is not modified by this new choice, however for a given number of mesh elements the error is smaller when the method is constructed with this new normalization than with the normalization $\lambda_1^{\pm} = 0$ or 1. In fact, for a given order q, the numerical results show that the constant underlying in estimate

$$\overline{\|X - X_h\|} = O(h^{q-3/2})$$

is much better : the related improvement observed in the numerical error is approximately $\approx 10^2$.

Once again the only difference between these two choices of basis functions relies on the fact that the leading coefficient in $P_{\pm}$ does depend or not on the equation's coefficient β. The theoretical tools previously developed could be adapted without difficulty to this new family of basis functions : nevertheless the vertical shift observed between the convergence plots in Figures 6.2 and 6.3 will require more research to be fully understood.

6.2.3 About q convergence

In Figure 6.2 as well as in Figure 6.3, for a fixed number of unknowns, the error decreases when the parameter $q \geq 2$ increases. To obtain better understanding of this phenomenon, a specific computation followed by a theoretical estimate illustrates the

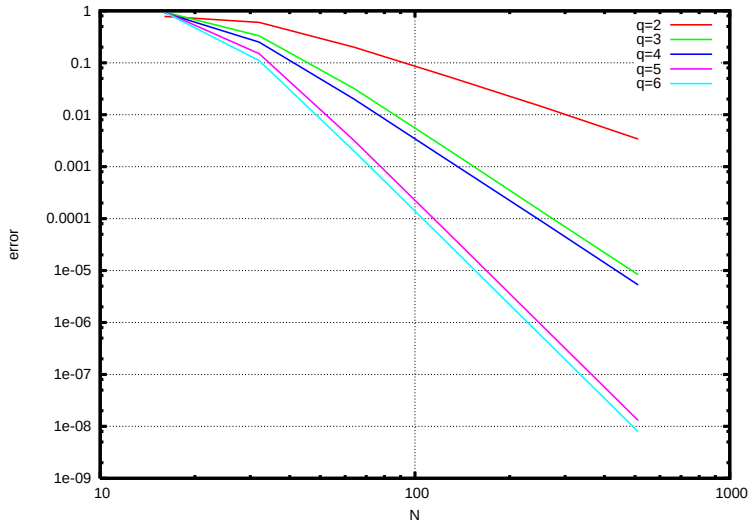


FIGURE 6.2 – h-convergence of the computed solution of the Airy equation on $\Omega =]-5,5[$. The relative discrete l^2 error is represented for different orders of approximation q depending on the number of unknowns $N = 2N_h$ with $N_h = 10/h$.

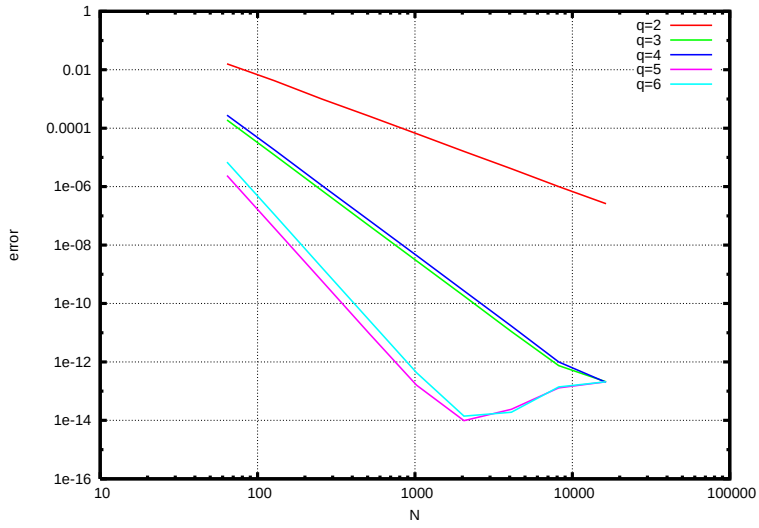


FIGURE 6.3 – h-convergence of the computed solution of the Airy equation on $\Omega =]-5,5[$ using the normalization $\lambda_1^{\pm} = \pm\sqrt{\beta(x_{k+1/2})}$. The relative discrete l^2 error is represented as a function of the number of elements defining the mesh for different orders of approximation q depending on the number of unknowns $N = 10/h$.

TABLE 6.2 – h-convergence of the computed solution of the Airy equation on $\Omega =]-5, 5[$ using the normalization described in subsection 6.2.2. Errors and orders of convergence for different orders of approximation q depending on the number of unknowns $N_u = 2N_h$ with $N_h = 10/h$.

	$q=2$		$q=3$		$q=4$		$q=5$		$q=6$	
N_u	Error	Rate	Error	Rate	Error	Rate	Error	Rate	Error	Rate
16	1.9e-01	1.92	3.9e-02	3.69	4.7e-02	5.65	5.4e-03	7.07	2.0e-02	5.19
32	6.2e-02	1.64	2.9e-03	3.75	4.2e-03	3.48	1.4e-04	5.28	4.2e-04	5.54
64	1.6e-02	1.93	1.9e-04	3.95	2.8e-04	3.92	2.4e-06	5.86	6.9e-06	5.93
128	4.2e-03	1.98	1.2e-05	3.99	1.8e-05	3.98	3.8e-08	5.97	1.1e-07	5.98
256	1.0e-03	1.99	7.4e-07	4.00	1.1e-06	3.99	6.0e-10	5.99	1.7e-09	6.00
512	2.6e-04	2.00	4.6e-08	4.00	7.0e-08	4.00	9.4e-12	6.00	2.7e-11	6.00
1024	6.5e-05	2.00	2.9e-09	4.00	4.4e-09	4.00	1.6e-13	5.92	4.3e-13	5.98
2048	1.6e-05	2.00	1.8e-10	4.00	2.7e-10	4.00	9.8e-15	3.99	1.4e-14	4.95
4096	4.1e-06	2.00	1.1e-11	4.00	1.7e-11	4.00	2.4e-14	-1.28	1.9e-14	-0.42
8192	1.0e-06	2.00	7.5e-13	3.90	1.0e-12	4.06	1.3e-13	-2.43	1.4e-13	-2.88
16384	2.6e-07	2.00	2.1e-13	1.84	2.0e-13	2.32	2.1e-13	-0.73	2.1e-13	-0.61

influence of q on the remainder of the Taylor expansion of $\beta - \beta_\pm$ that is naturally reflected in the approximation of Ai by the shape functions.

Figure 6.4 displays the Airy function and its approximations by the two basis functions φ constructed at a point x_0, for increasing values of q. Two points x_0 are chosen, the first one is the cut-off point whereas the second one is lying in the propagative zone. In order to analyze this apparent uniformity of the approximations observed on Figure 6.4 (uniformity with respect to the parameter q), one has to remind of the design process of the shape functions. Indeed, the process is based on a Taylor expansion and the fact that the theoretical order of convergence does depend on q, so that the analysis relies on an estimate of the rest of order q in $\beta - \beta_\pm$. For the sake of simplicity suppose here that $\beta \in C^\infty$. From the detailed design process one can see that this rest is actually

$$\beta - \beta_\pm = \sum_{i=q}^{\infty} \frac{d^i\beta(G)}{dx^i}\frac{(x-G)^i}{i!} - \sum_{i=q}^{2q}\left(\sum_{l=0}^{i}(l+1)(i-l+1)\lambda_{l+1}\lambda_{i-l+1}\right)(x-G)^i,$$

so that the coefficient of the leading order term is

$$C_q = \frac{1}{q!}\frac{d^q\beta(G)}{dx^q} - \sum_{l=0}^{q}(l+1)(q-l+1)\lambda_{l+1}\lambda_{q-l+1}. \tag{6.1}$$

The following result specifies how this term behaves with respect to q.

Proposition 6.1. *Assume that $q \geq 1$ is fixed and that $\beta(x) = x + a$, $a \in \mathbb{R}$. The coefficient of the leading order term, C_q can be estimated explicitly with respect to q : there is a constant C depending on λ_1 and on $\|\beta\|_{\infty,\Omega}$ but independent of q such that*

$$|C_q| \leq C^q. \tag{6.2}$$

Proof. Because of the definition of C_q, it is convenient to prove a first estimate of the terms $(l+2)\lambda_{l+2}$ such as : for all $l \in \mathbb{N}$ such that $0 \leq l \leq q-1$

$$|(l+2)\lambda_{l+2}| \leq \frac{C^{l+2}}{l+1}, \tag{6.3}$$

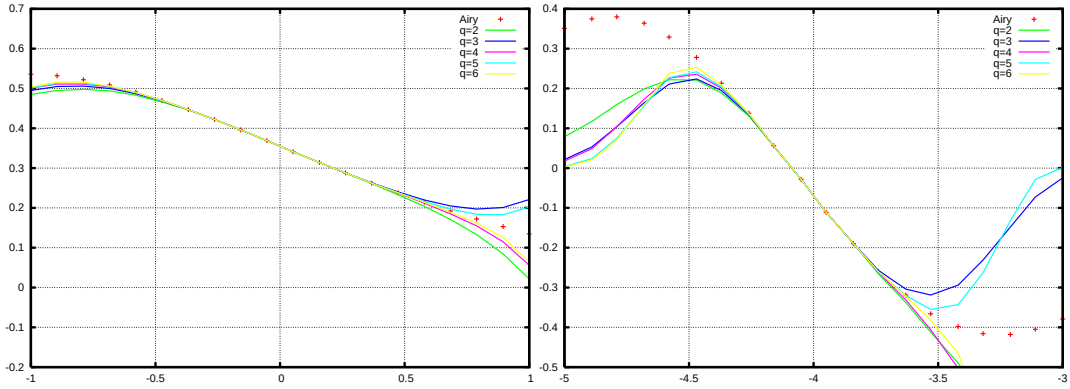


FIGURE 6.4 – Approximation of Airy function by corresponding basis functions for different values of q, around the cut-off on the left, $x_0 = 0$, and in the propagative zone in the right, $x_0 = -4$. See subsection 6.2.3.

where the constant C is a guess for the constant in the right hand side of (6.2). The induction formula defining the coefficients λs reads for all $i \in \mathbb{N}$ such that $i \leq q-1$:

$$\lambda_{i+2} = \frac{1}{(i+2)(i+1)} \left(\frac{d^i \beta(G)}{dx^i} - \sum_{l=0}^{i} (l+1)(i-l+1)\lambda_{l+1}\lambda_{i-l+1} \right).$$

One can see that

$$\begin{cases} \lambda_1 \text{depends on the normalization,} \\ 2\lambda_2 = \beta(G) - \lambda_1^2, \\ 3\lambda_3 = \dfrac{1}{2}\left(\dfrac{d\beta(G)}{dx} - 2\lambda_1\lambda_2 \right), \end{cases} \tag{6.4}$$

while for all $i \in \mathbb{N}$ such that $2 \leq i \leq q-1$ the hypothesis on β gives

$$(i+2)\lambda_{i+2,=} \frac{1}{i+1} \left(-\sum_{l=0}^{i} (l+1)(i-l+1)\lambda_{l+1}\lambda_{i-l+1} \right).$$

So the constant C has to ensure that (6.3) holds at least in the three cases described by (6.4) : this is why a first guess C_g is set as

$$C_g = \max \left\{ |\lambda_1|, \sqrt{|\beta(G)| + |\lambda_1|^2}, \sqrt[3]{1 + 2|\beta(G)||\lambda_1| + 2|\lambda_1|^3} \right\}.$$

Now proceed by induction. For $l \geq 2$ suppose that (6.3) holds for all $l' \in \mathbb{N}$ such that $0 \leq l' < l$. Then from the assumption on β one has

$$\begin{aligned} |(l+2)\lambda_{l+2}| &\leq \frac{1}{(l+1)} \left(\sum_{j=0}^{l} |(j+1)\lambda_{j+1}| \cdot |(l-j+1)\lambda_{l-j+1}| \right), \\ &\leq \frac{C_g^{q+2}}{l+1} \left(\frac{2}{l+1} + \sum_{j=1}^{l-1} \frac{1}{j(l-j)} \right), \end{aligned} \tag{6.5}$$

so that an upper bound for the sum term is now needed. The induction hypothesis shows that an estimate of

$$S_l = \sum_{j=1}^{l-1} \frac{1}{j(l-j)} = \frac{2}{l} \sum_{j=1}^{l-1} \frac{1}{j}$$

for all $l \in \mathbb{N}$ such that $2 \leq l < k$ would provide such a bound. On the other hand

$$S_{l+1} - S_l = \frac{2}{l(l+1)} \left(1 - \sum_{j=1}^{l-1} \frac{1}{j} \right) \leq 0$$

holds for all $l \geq 2$, so that $S_l \leq S_2 = 1$ for all $l \geq 2$. But

$$\frac{2}{l+1} + \sum_{j=1}^{l-1} \frac{1}{j(l-j)} \leq 1 \Leftrightarrow 4 \leq l, \tag{6.6}$$

so the inductive step only holds from $l \geq 4$. As a result the constant C actually has to ensure that the initial condition for the induction holds for all $l \in \mathbb{N}$ such that $l < 4$. The updated constant C can then be defined as

$$C = \max \left\{ C_g, \sqrt[4]{\lambda_1(1 + 2|\beta(G)||\lambda_1| + 2|\lambda_1|^3)}, \right.$$

$$\left. \sqrt[5]{|\beta(G)| + 2|\beta(G)|^2|\lambda_1| + \frac{5}{3}|\lambda_1|^2 + \frac{16|\beta(G)|}{3}|\lambda_1|^3 + \frac{10}{2}|\lambda_1|^5} \right\}.$$

With this updated constant, the base case is satisfied for $l \leq 3$ and the inductive step is proved combining equations (6.5) and (6.6). This completes the proof of (6.3) by induction. Then from the definition of C_q stated in (6.1) and applying again the same estimate process the conclusion is reached. □

This estimate holds for all $G \in \Omega$, however C does depend on the normalization, so that - for $\lambda_1 \in \{0, 1\}$ - the constant C can be bounded independently of G by considering that $|\beta(G)| \leq \|\beta\|$. For $\lambda_1 = \pm\sqrt{\beta(G)}$ as well, the same inequality provides an upper bound for C that does not depend on G.

The case of a more general coefficient β could be addressed with an appropriate condition on the derivatives of β. Indeed, if one can guarantee that the successive derivatives satisfy

$$\frac{1}{l!} \left| \frac{d^l \beta(G)}{dx^l} \right| + \frac{2}{l+1} + \sum_{j=1}^{l-1} \frac{1}{j(l-j)} \leq 1$$

instead of (6.6), then a similar proof would be valid in that more general case.

Interpretation

Proposition 6.1 states that

$$\beta - \beta_{\pm} = (Ch)^q + O(h^{q+1}),$$

where C does only depend on β, Ω and the normalization $\lambda_1^{\pm}$. The q-convergence observed on Figures 6.2 and 6.3 is then justified for the leading order term as soon as $hC < 1$: it also provides a guess of the length of the interval on which q convergence holds.

6.3 Code development and computational aspects in 2D

The main implementation work was to create a code for the UWVF supplemented with the new basis functions for the O mode problem. Since the initial code for the elastic

problem was based on classical plane waves, the integrals for the coefficients of the matrices were computed in close form. The introduction of the new basis functions required the implementation of quadrature formulas to approximate these integral terms. Practically, a computation starts from the definition of some parameters

$$\begin{cases} q \text{ the approximation order ,} \\ p \text{ the number of basis functions per element ,} \\ N \text{ the normalization ,} \\ N_h \text{ the mesh parameter .} \end{cases}$$

that are used to build locally the coefficients of the basis functions and so the assembly of the matrices. The formulas stemming from the Taylor expansion have resulted in a substantial burden.

As for the theoretical study, the numerical simulation for the X mode problem required specific attention. The first attempt in this direction was to use finite differences to compute the solution of a simple X mode problem, but reached no conclusive end at that time. This basic scheme did not include any kind of regularization because it was implemented prior to the theoretical understanding presented in Chapter 4. Nevertheless it evidenced the necessity of an appropriate regularization process to stabilize X mode computations. A 2D UWVF code was later adapted to the regularized X mode problem. Based on a potential formulation, this code is very close to the O mode one.

The following examples were computed on both personal and high performance computers. The high performance computer available at the Laboratoire Jacques Louis Lions has a Non Uniform Memory Architecture, with twenty 2.0 GHz Xeon octo-core processors, and 640 Go of shared memory. It was used to run in parallel the UWVF code for different sets of parameters (q, N_h) while (p, N) were fixed, each value of (q, N_h) being sent to each processor, all of them running the same UWVF code. Thanks to this process it was possible to generate Figures 6.12 and 6.14 to emphasize different features of the method. The case displayed on Figures 6.19 and 6.20 were computed on a classical bi-core processor machine within a few hours each. The corresponding matrices inverted in this case are complex 169792×169792 sparse matrices.

6.4 Validation of the interpolation result in 2D

Following the theoretical result, each of the numerical validation case is computed, for a given value of n, setting $q = n + 1$ and $p = 2n + 1$. The test case considered is $\beta(x, y) = x - 1$, with the exact solution $u_e(x, y) = Ai(x)e^{iy}$. As stated in the main theorem of Section 5.4.4 from Chapter 5, u_e can be approximated by a function u_a that belongs to the approximation space $Span\ \mathcal{E}(G, p)$, space that is built with either the β-normalization or the constant-normalization. The vector $\{c_l\}_{1 \leq l \leq p}$ such that $u_a = \sum_{1 \leq l \leq p} c_l \varphi_l$ is computed inverting a system, as $(M_n)^{-1}B$, where

- the matrix M_n contains the coefficients of the Taylor expansion of the shape functions,
- the vector B contains the coefficients of the Taylor expansion of u_e.

Both of them were built as theoretical tools to analyze the order of convergence, it is then natural to follow this path from the numerical point of view.

The procedure set up for this validation is the following : estimate the error $\max |u_e - u_a|$ on disks with decreasing radius h in order to observe the order of convergence with respect to h. This estimate is computed at the nodes of a grid that are situated along equi-angular

radius of the disk. The errors provided in the upcoming tables are then discrete errors results computed over 1300 points. Several different cases are proposed to validate the theoretical order of convergence

- in the propagative zone $x < 1$,
- in the absorbing zone $x > 1$,
- at the cut-off $x = 1$,

and an additional case concerns the behavior of the shape function designed with the β-normalization as the approximation point gets closer to the cut-off.

6.4.1 In the propagative zone

The point $G = [-3, 1]$ is in the propagative zone. Concentric disks are centered on G with radius $h = 1/2^i$. Following the interpolation theorem, the expected order of convergence is $n + 1$.

Figure 6.5 displays computed convergence results that fit perfectly the theory. A set of $p = 11$ classical plane waves is used as a control case, since $p = 11$ is the highest number of shape functions used in the different cases with the generalized plane waves. Note that machine precision is reached in some cases.

6.4.2 In the non propagative zone

The point $G = [2, 1]$ is in the non propagative zone. Again concentric disks are centered on G with radius $h = 1/2^i$, and the expected order of convergence is $n + 1$. There is no classical plane wave that can be computed here since $\beta(G) > 0$.

Figure 6.6 displays computed convergence results that fit perfectly the theoretical result as well. Again machine precision is reached in some cases.

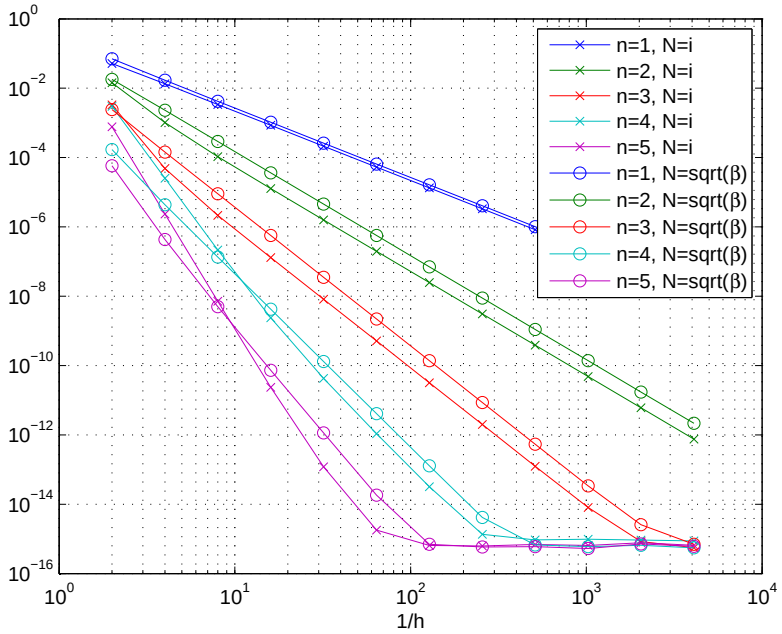
6.4.3 Toward the cut-off

Since at the cut-off $\beta(x = 1, y) = 0$, it is interesting to look at what happens with the normalization $N = \sqrt{\beta}$ long this line. Indeed, when applying the generalized plane waves with the UWVF and refining the mesh, the center of some mesh cells will get closer to the cut-off.

For this numerical test the point $G = [1 - h, 1]$ remains in the propagative zone. Then disks are here centered on a point G_h that stands at a distance h from the line $x = 1$, still with radius $h = 1/2^i$. Classical plane waves are compared to the normalization $N = \sqrt{\beta}$ with the same number of shape functions.

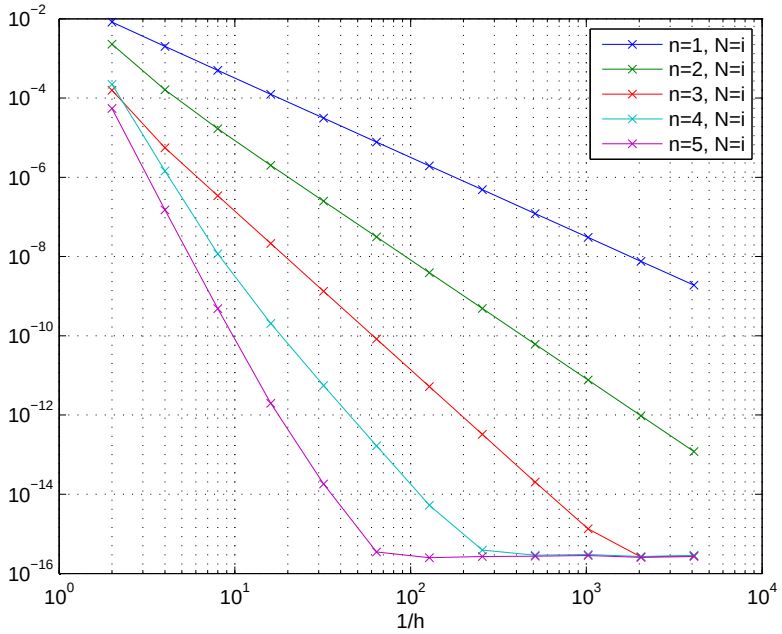
Figure 6.7 shows that the generalized plane waves normalized with $N = \sqrt{\beta}$ give a high order approximation of u even getting closer to the vanishing line $x = 1$, as long as h is not too small. Note that there does not seem to be a significant difference between the two type of functions. This is observed on numerical results even if there is no corresponding theoretical explanation.

Another possibility is to compare the influence of two parameters : the size of the disk h and the distance d between G and the line $x = 1$. The error e depends on both parameters, so one can write $e(h, d)$. Figure 6.8 displays the error computed for h and d convergence with the normalization $N = \sqrt{\beta}$. The h convergence is clearly damaged for decreasing values of d. This is linked to the low frequency limit when β goes to zero. However, looking at the h convergence with $d = h$, one can see that the error $e = e(h, h)$ converges as the error $e(h, 1/2)$ until $h = 1/2^5$.



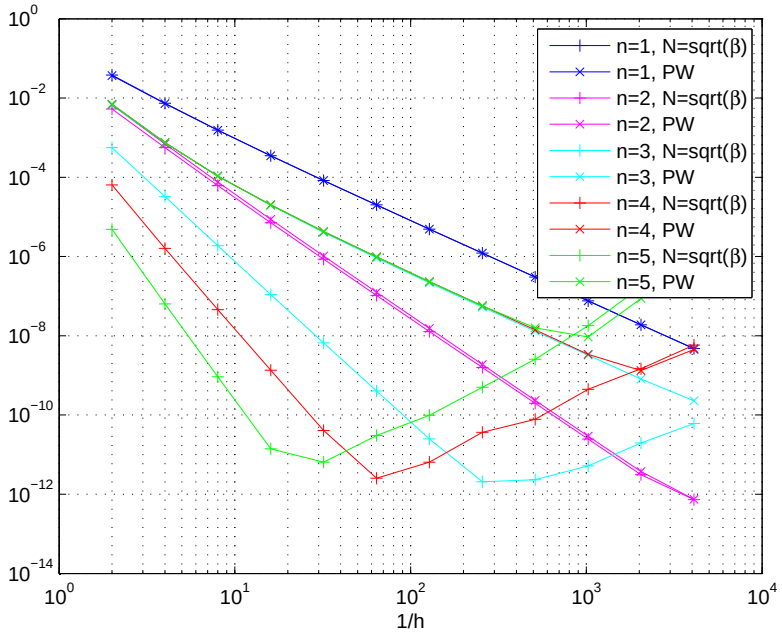
	$p=5$	$n=1$		$n=2$		$n=3$		$n=4$		$n=5$	
h	PW	β	CST	β	CST	β	CST	β	CST	β	CST
$1/2^2$	2.98	2.07	1.94	2.96	3.81	4.07	6.08	5.29	6.88	7.06	8.37
$1/2^3$	3.01	2.02	1.98	3.00	3.27	3.99	4.50	5.02	6.88	6.44	8.35
$1/2^4$	3.01	2.00	2.00	3.00	3.06	4.00	4.04	5.00	6.49	6.09	8.26
$1/2^5$	3.01	2.00	2.00	3.00	3.01	4.00	4.00	5.00	5.82	6.00	7.61
$1/2^6$	3.00	2.00	2.00	3.00	3.00	4.00	4.00	5.00	5.30	5.97	6.07

Figure 6.5 – Convergence results in the propagative zone, computed at $G = [-3, 1]$ with different shape functions. Comparison between classical plane waves and generalized plane waves for both constant and β normalizations. Some of the associated orders of convergence are also provided.



	$n=1$		$n=2$		$n=3$		$n=4$		$n=5$	
h	β	CST	β	CST	β	CST	β	CST	β	CST
$1/2^2$	2.16	2.03	3.05	3.82	4.14	4.82	5.09	7.26	6.24	8.50
$1/2^3$	2.07	2.01	3.03	3.27	4.05	4.03	5.04	6.96	6.07	8.28
$1/2^4$	2.03	2.00	3.02	3.07	4.02	4.00	5.02	5.83	6.02	7.93
$1/2^5$	2.02	2.00	3.01	3.01	4.01	4.00	5.01	5.21	6.00	6.76
$1/2^6$	2.01	2.00	3.00	3.00	4.00	4.00	5.00	5.05	5.87	5.70

FIGURE 6.6 – Convergence results in the non-propagative zone, computed at $G = [2, 1]$ with different shape functions. Comparison between generalized plane waves for both constant and β normalizations. Some of the associated orders of convergence are also provided.

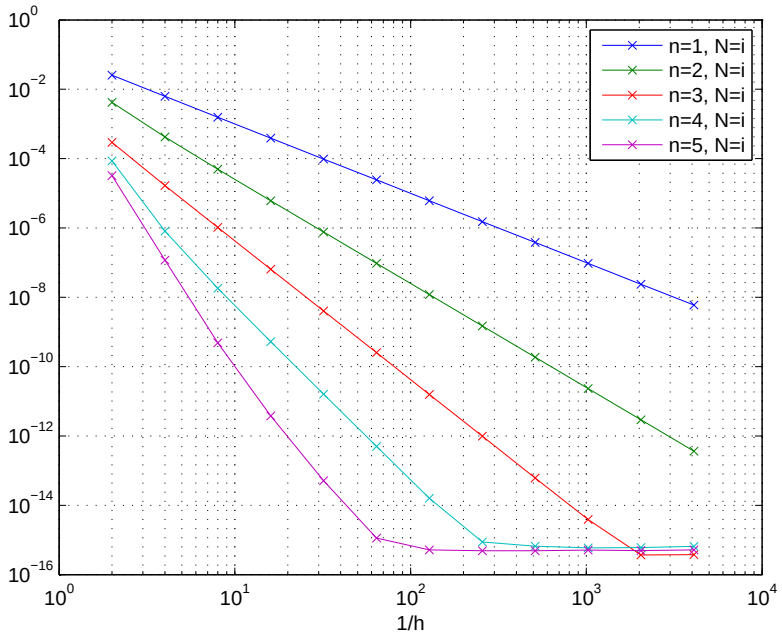


	$n=1$		$n=2$		$n=3$		$n=4$		$n=5$	
h	PW	β	PW	β	PW	β	PW	β	PW	β
$1/2^2$	2.35	2.37	3.27	3.24	3.17	4.09	3.21	5.32	3.21	6.25
$1/2^3$	2.23	2.24	3.19	3.19	2.77	4.13	2.82	5.14	2.82	6.11
$1/2^4$	2.14	2.15	3.12	3.11	2.47	4.09	2.40	5.08	2.40	6.03
$1/2^5$	2.08	2.09	3.08	3.07	2.27	4.06	2.24	5.05	2.24	1.13
$1/2^6$	2.05	2.04	3.05	3.04	2.15	4.04	2.13	4.00	2.13	-2.24

Figure 6.7 – Convergence results toward $\beta = 0$, computed at $G = [1-h, 1]$. Comparison between classical plane waves and generalized plane waves for both normalizations $N = \sqrt{\beta}$ and $N = i$. Some of the associated orders of convergence are also provided.

h\ d	$1/2^1$	$1/2^2$	$1/2^3$	$1/2^4$	$1/2^5$	$1/2^6$	$1/2^7$	$1/2^8$	$1/2^9$	$1/2^{10}$
1/2	4.8e-06	5.5e-06	5.5e-06	5.4e-06	5.4e-06	5.3e-06	5.2e-06	5.2e-06	5.2e-06	5.2e-06
$1/2^2$	5.7e-08	6.4e-08	6.4e-08	6.2e-08	6.1e-08	6.0e-08	5.9e-08	5.8e-08	5.8e-08	6.9e-08
$1/2^3$	8.3e-10	9.2e-10	9.2e-10	9.0e-10	8.8e-10	8.7e-10	9.2e-10	1.2e-09	3.5e-09	2.4e-08
$1/2^4$	1.3e-11	1.4e-11	1.4e-11	1.4e-11	1.8e-11	3.6e-11	1.0e-10	5.4e-10	3.2e-09	2.2e-08
$1/2^5$	2.0e-13	2.3e-13	3.5e-13	8.8e-13	6.4e-12	2.8e-11	1.2e-10	5.4e-10	3.8e-09	2.2e-08
$1/2^6$	4.3e-15	1.7e-14	1.6e-13	7.6e-13	6.2e-12	3.0e-11	1.0e-10	6.0e-10	3.1e-09	2.0e-08
$1/2^7$	2.4e-15	1.6e-14	1.6e-13	7.7e-13	6.2e-12	2.8e-11	9.8e-11	5.1e-10	2.9e-09	2.3e-08
$1/2^8$	2.3e-15	1.5e-14	1.6e-13	7.9e-13	6.1e-12	2.7e-11	1.0e-10	5.0e-10	2.5e-09	1.6e-08
$1/2^9$	2.0e-15	1.5e-14	1.6e-13	7.9e-13	5.4e-12	2.5e-11	9.7e-11	4.9e-10	2.5e-09	1.9e-08
$1/2^{10}$	1.9e-15	1.4e-14	1.5e-13	7.5e-13	6.0e-12	2.5e-11	8.7e-11	5.0e-10	2.5e-09	1.8e-08

Figure 6.8 – Error computed on a disk of radius h centered at $G = [1-d; 1]$. The approximation is computed with the β-normalization and with $n = 5$, $q = n+1$ and $p = 2n+1$.



h	$n=1$	$n=2$	$n=3$	$n=4$	$n=5$
$1/2^2$	2.01	3.33	4.15	6.72	8.10
$1/2^3$	2.00	3.09	4.01	5.49	7.93
$1/2^4$	2.00	3.02	4.00	5.11	7.01
$1/2^5$	2.00	3.00	4.00	5.03	6.20
$1/2^6$	2.00	3.00	4.00	5.01	5.50

FIGURE 6.9 – Convergence results computed at $G = [1, 1]$ using generalized plane waves with the constant-normalization. Some of the associated orders of convergence are also provided.

6.4.4 At the cut-off

The point $G = [1, 1]$ lies exactly on the vanishing line of β. Then again concentric disks are centered on G with radius $h = 1/2^i$. Both classical plane waves and generalized plane waves with the normalization $N = \sqrt{\beta}$ would provide only one function since $\beta(G) = 0$. However the normalization $N = i$ - as described previously - is well defined even at the cut-off.

As Figures 6.5 and 6.6, Figure 6.9 displays results that fit perfectly the theoretical result : the point G lays along the cut-off line and the computed orders of convergence for a given parameter n are exactly $n + 1$, as stated in the theorem.

6.4.5 Back to classical plane waves

This is a simple sanity check in the case of a coefficient β piecewise constant on each element of the mesh.

In dimension two as in dimension one that for $q = 1$, the new shape functions normalized with $N = \sqrt{\beta(G)}$ are exactly classical plane waves as long as $\beta < 0$, since in this

	$n=1$		$n=2$		$n=3$		$n=4$		$n=5$	
h	PW	β	PW	β	PW	β	PW	β	PW	β
$1/2^2$	1.93	1.93	2.94	2.94	3.94	3.94	4.96	4.96	5.97	5.97
$1/2^3$	1.98	1.98	2.99	2.99	3.99	3.99	4.99	4.99	5.99	5.99
$1/2^4$	2.00	2.00	3.00	3.00	4.00	4.00	5.00	5.00	6.00	6.00
$1/2^5$	2.00	2.00	3.00	3.00	4.00	4.00	5.00	5.00	6.00	6.00
$1/2^6$	2.00	2.00	3.00	3.00	4.00	4.00	5.00	5.00	5.99	6.00
$1/2^7$	2.00	2.00	3.00	3.00	4.00	4.00	5.00	5.00	5.02	5.46
$1/2^8$	2.00	2.00	3.00	3.00	4.00	4.00	4.97	4.98	0.78	1.43
$1/2^9$	2.00	2.00	3.00	3.00	4.00	4.00	4.22	4.54	0.14	-0.00
$1/2^{10}$	2.00	2.00	3.00	3.00	4.00	4.00	0.76	0.80	0.02	0.05
$1/2^{11}$	2.00	2.00	3.00	3.00	3.99	3.98	-0.10	-0.00	-0.15	-0.00
$1/2^{12}$	2.00	2.00	3.00	3.00	3.48	3.31	0.04	0.14	-0.02	-0.12

FIGURE 6.10 – Numerical validation of the fact that generalized and classical plane waves are the same when the coefficient β is constant. The computations are performed for $\beta = -4$ to approximate the exact solution $u_e(x,y) = e^{2iy}$. The differences between the convergence rates for $h < 1/2^8$ are due to the fact that machine precision is reached : then the behavior of the algorithm that inverts the interpolation matrix is not controlled anymore.

case

$$\lambda_{2,0} = \frac{1}{2}\left(-2\lambda_{0,2} - \lambda_{1,0}^2 - \lambda_{0,1}^2\right).$$

This corresponds to the classical fact of approximating a smooth coefficient by its piecewise constant value at the center of the cells. This is illustrated by Figure 6.10.

6.5 O mode simulation in 2D

Unlike in the one dimensional case, a quadrature formula is needed in this case, and since integrals are to be computed over the edge of triangles, only a one dimensional formula is required. Some formulas are detailed in this section. A study of the relative influence of the basis functions parameters follows. More physical material concludes the section.

The test problem considered here is the following : different domains will be considered, all of them meshed with triangles, and the system is

$$\begin{cases} -\Delta u + (x - \kappa^2)u &= 0, & (\Omega), \\ (\partial_\nu + i\sigma)u &= Q(-\partial_\nu + i\sigma)u + g, & (\Gamma), \end{cases}$$

with the parameter $\sigma = 1$.

Mainly h convergence results will be pointed out. The error is evaluated as a discrete l^2 relative norm evaluated at the center of the mesh cells : it means the unknown is not reconstructed at the edges of the mesh, but on the mesh elements.

6.5.1 Comparing performances of the quadrature formulas

The parameter is set as : $Q = 0$ on the boundary Γ, and g corresponds to the exact solution $u_e = Ai(x)e^{i\kappa y}$.

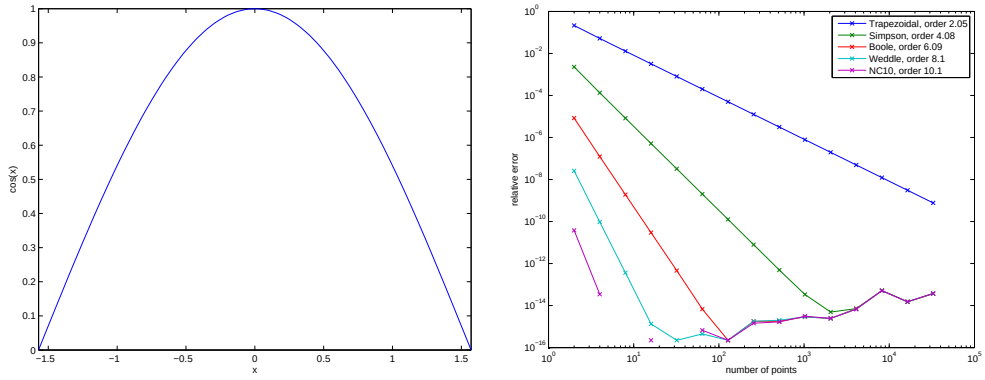


FIGURE 6.11 – Numerical test of different Newton-Cotes quadrature formulas, integrating the cos function on $\Omega =]\pi, \pi[$ and providing the numerical order of convergence. On the left is the function integrated, and on the right a convergence graph shows the different rates of h convergence.

Different quadrature formulas

The computation of the integrals is performed by the use of a one dimensional quadrature formula. Different Newton-Cotes formulas are implemented :

- with 5 points, called Boole formula, with corresponding weights

$$(7, 32, 12, 32, 7)/90,$$

- with 7 points, called Weddle formula, with corresponding weights

$$(41, 216, 27, 272, 27, 216, 41)/840$$

- with 10 points, with corresponding weights

$$\frac{(25713, 141669, 9720, 174096, 52002, 52002, 174096, 9720, 141669, 25713)}{806400}$$

The formula with 9 points is not used in order to avoid the use of negative weights. The trapezoidal formula is also considered as a benchmark case. The h-convergence for a test case is illustrated on Figure 6.11 with these different methods, refining the mesh of the domain $\Omega =]\pi, \pi[$.

A benchmark case

The domain is $\Omega =]-6, 3[\times]-1, 1[$, and as required by the theorem, the parameters of the method are set to satisfy the interpolation result :

$$\begin{cases} q &= n+1, \\ p &= 2n+1, \end{cases}$$

where two values of n are considered : $n = 3, 4$. Figure 6.12 describes the numerical results obtained in this case, using the normalization $N = \sqrt{\beta}$. In the case $n = 3$ the three different quadrature formulas compared in the figure give very close errors, and the convergence rates are slightly different between 4.5 and 5. However in the case $n = 4$ the difference is much more important. The orders of convergence for the Boole formula start from 4.5 and

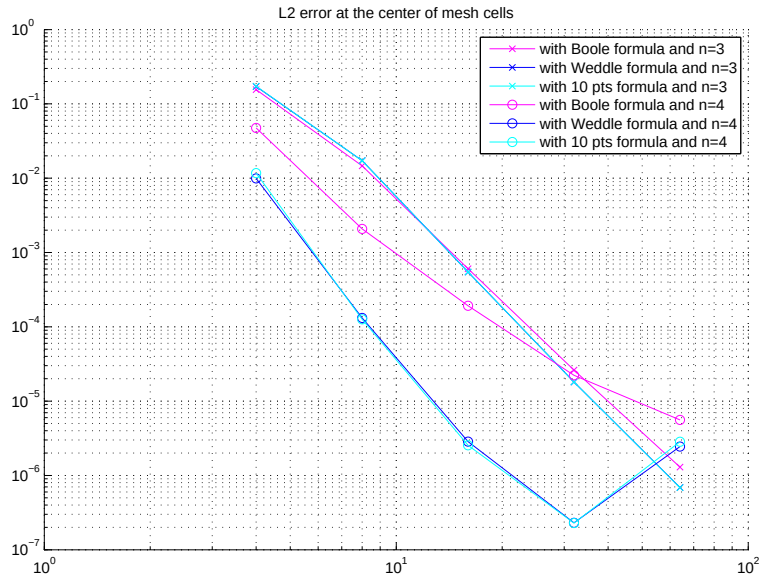


FIGURE 6.12 – Comparison of h convergence result for different quadrature formula : Boole, Weddle and the 10 points formula are displayed. The limitation of the Boole formula, that approximates an integral using 5 evaluating points, is obvious on the case $n = 4$ (round markers) while it is almost nonexistent in the case $n = 3$ ($\times$ markers).

decrease to 3.4, 3.1 and 1.9. For the Weddle and 10 points formulas the rates are much better : they decrease from 6.5 to 3.4.

These remarks evidence the fact that the choice of the quadrature formula has to match the parameters of interpolation : there is no reason for increasing the accuracy of the interpolation with n if the accuracy of the final result is limited by a low order quadrature formula. On the other hand the rate between the processing times for Boole and the 10 points formula is 5/7, so that the raise of computing time linked to the quadrature formula should not be neglected.

The fact that Weddle and the 10 points formula give similar results might be linked to the limitation to the square root of the machine precision $\sqrt{\epsilon}$ that will be described in the next section. Indeed, one can see in the test case presented in Figure 6.11 that these two formulas reach this precision $\sqrt{\epsilon}$ even for a small number of points discretizing the domain $[-5, 5]$.

6.5.2 A parameter study

Here the integration method is fixed, using the 10 points formula. The reference solution is $u_e(x, y) = Ai(x)e^{iy}$. It is displayed on Figure 6.13, the solution being reconstructed on each cell and represented on a refined mesh.

Figure 6.14 compares the results obtained with the normalization $N = \sqrt{\beta}$ and $N = i$, using parameters that satisfy the interpolation requirements. There is an obvious difference between the two normalization : the $N = \sqrt{\beta}$ gives much better results, by a factor 10^{-2}. It is interesting to see that for the normalization $N = \sqrt{\beta}$ the convergence rate seem to match the interpolation prediction until $n = 3$ but there is a deterioration of the convergence for

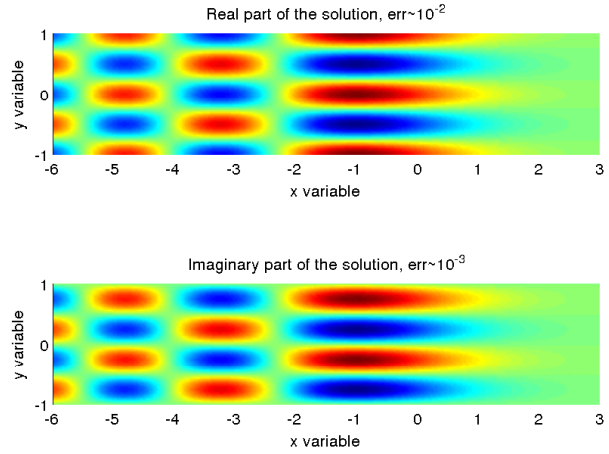


FIGURE 6.13 – The computed solution. Computed for $p = 7$, $q = 5$ and with the normalization $N = \sqrt{\beta}$, on a 576 triangles mesh. Top : the real part of the numerical solution. Bottom : the imaginary part of the numerical solution.

$n = 4$. The corresponding case for the constant normalization is not represented because it blows up.

Figure 6.15 displays the h convergence results for increasing values of the number of basis functions per element p, in order to see if the default of convergence observed on Figure 6.14 can be overcome. Starting from $p = 7$ the number of basis functions per elements is raised up to $p = 10$. In the case $p = 7$, the convergence does not pass 5 and is not as good for $q = 1, 2$ than for higher values of q, and the errors are the same for any value of q higher than 2. For $p = 8$ the rates in both cases $q = 1, 2$ is almost 2, both cases $q = 3, 4$ between 3.5 and 4.5 and both cases $q = 5, 6$ tat present very similar results are only slightly better. For $p = 9$ the results are even more separated for $q = 5, 6$ that reach order 6. The deterioration of the convergence results when the error reaches the threshold of the square root of the machine precision, namely $\sqrt{\epsilon} \approx 10^{-8}$, has already been observed with the classical UWVF and is due to poor conditioning of the matrix. A theoretical estimate on this conditioning number can be found in [Ces96b].

6.5.3 A first reflectometry test case

This case was proposed by Stéphane Heuraux as a first step toward real life problems. It models a wave sent in a plasma by an antenna from the wall of a reactor. The reactor is represented by a square domain, while the antenna is represented by a wave guide added outside the reactor on a wall plus a horn inside the reactor.

The geometry

The geometry is described in Figure 6.16. The domain Ω is a $L \times L$ square, the width and length of the waveguide are l_0 and $4l_0$. The boundaries, the subdomains are created with the Matlab toolbox *pdetool*, and they are described together with two examples of meshes in Figure 6.17. At that point the horn is not yet defined as a boundary since its edges belong to the interior of the domain : such an edge belongs to two different triangles

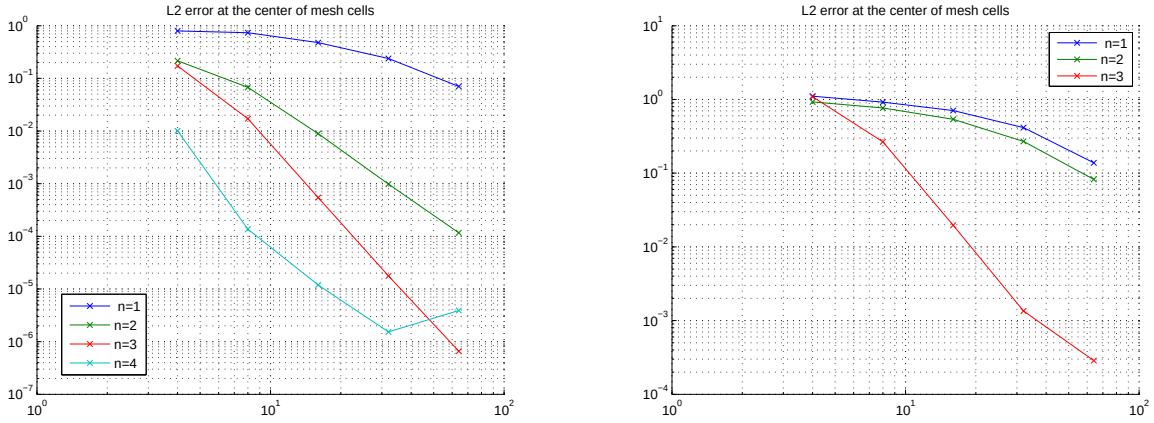


FIGURE 6.14 – Convergence results for different values of the interpolation parameter n, with $q = n + 1$ and $p = 2n + 1$. On the left : with basis functions normalized by $N = \sqrt{\beta}$. On the right : with basis functions normalized by $N = i$.

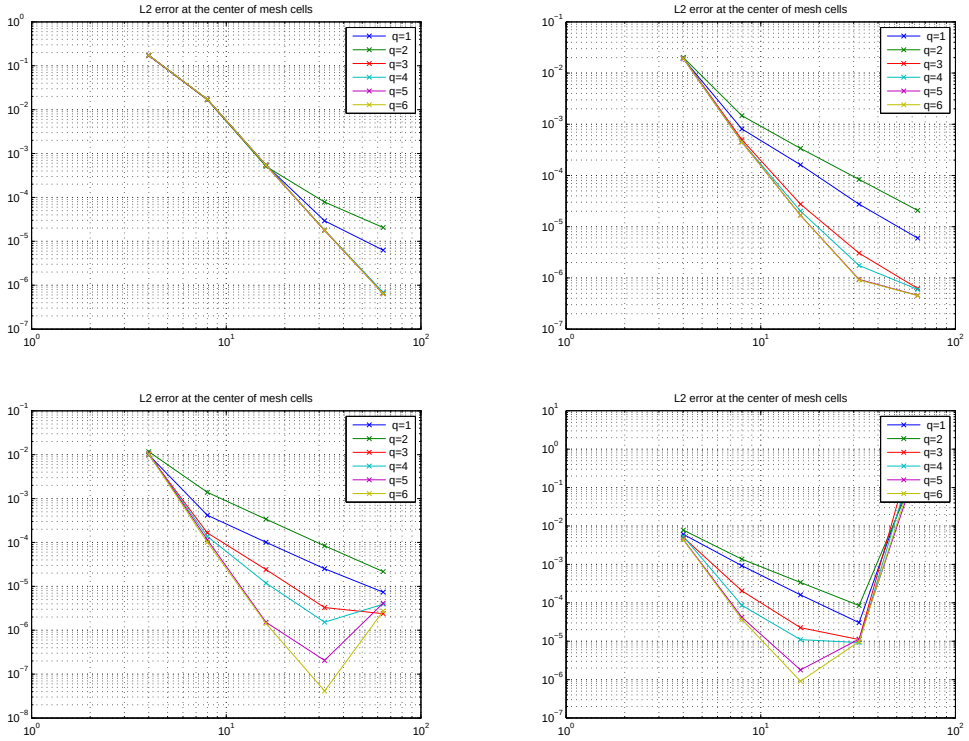


FIGURE 6.15 – h-convergence results for different values of parameters p and for q going from 1 to 6. Computed with the normalization $N = \sqrt{\beta}$ and the 10 points quadrature formula. From top left to bottom right : with $p = 7$, $p = 8$, $p = 9$ and $p = 10$ basis functions per element.

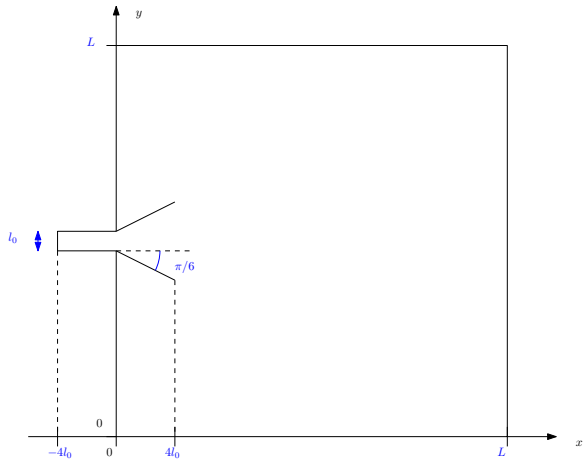


FIGURE 6.16 – Slice of tokamak, specifying the domain parameters : the wave guide width, the shape of the horn and the size of the main part of the domain.

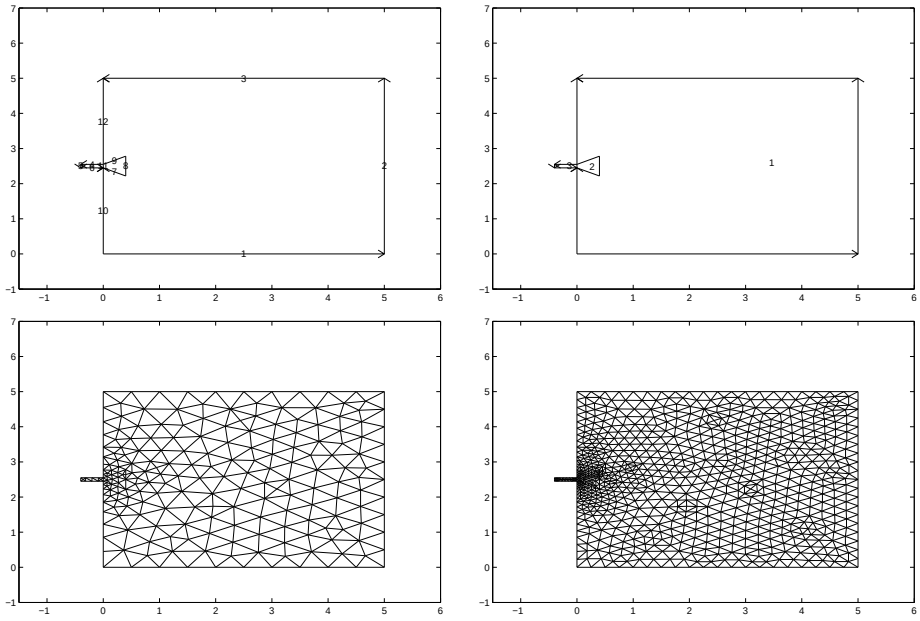


FIGURE 6.17 – Producing the domain and the mesh. The figures represent pdetool windows. Top left : the domain is created thanks to two polygons. Top right : the default edge labels. Bottom left : a coarse mesh of the domain. Bottom right : a refined mesh of the domain.

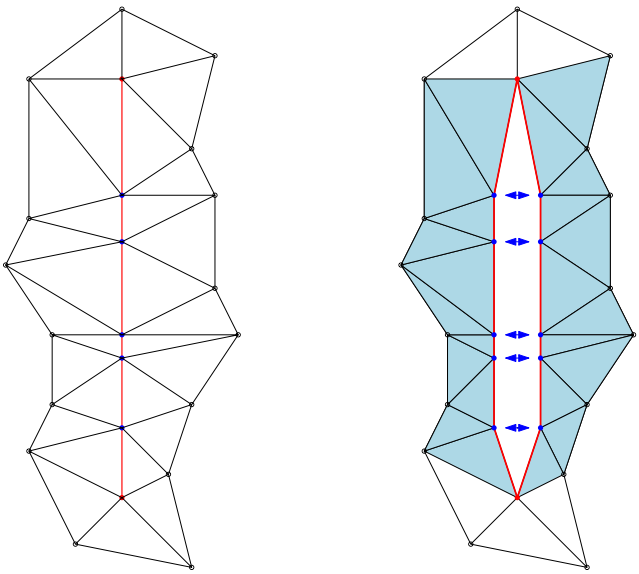


FIGURE 6.18 – Modifications of the initial mesh data : Points and triangles modified along a horn edge. Left : The red edges are lying on the horn boundary and the blue vertices are on the horn but not an end of the horn, they are to be modified. Right : Each blue point is doubled by a new point of the mesh with the same coordinates. Each red edge is doubled by a new edge which vertices are on one side of the horn. The vertices of each light blue triangle are updated with the new points numbers.

of the mesh, one lying inside the horn, the other one lying outside the horn but in the reactor.

Producing the crack along the edges of the horn has required to process the mesh with the following steps.

- Listing the points on the horn boundary.
- Creating new points with the same coordinates, except for the ends of the horn.
- For each of these new points, updating their number for each triangle that it belongs to, as long as the triangle is inside the horn.
- Updating the labels of the boundaries.

See figure 6.18 for a graphic illustration of the process.

First results

In both cases the size of the domain is set to be $L = 50l_0$, where l_0 is the wavelength of the incoming signal in the horn.

Wave propagation in a homogeneous medium

The size of the domain is $L = 50l_0$ and the coefficient is constant $\beta = -\kappa^2$. See Figure 6.19 that represents a wave propagating from an antenna in a propagative homogeneous domain. This result was computed with the normalization $N = \sqrt{\beta}$. As already mentioned, in such a case classical plane wave are exact solutions of the solution equation.

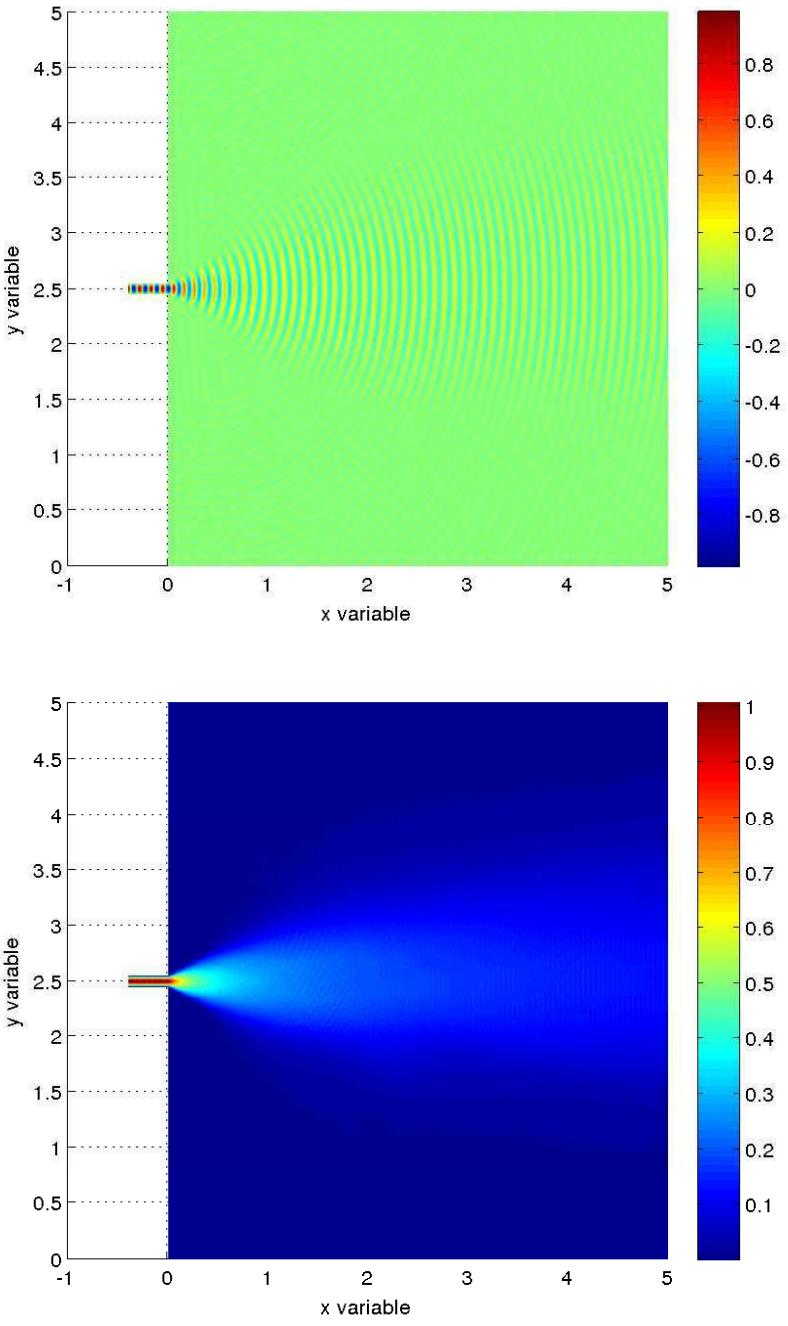


FIGURE 6.19 – Wave propagating in a homogeneous domain. Result computed using generalized plane waves designed with $N = \sqrt{\beta}$ and the UWVF, for $p = 7$ and $q = 4$. Top : real part of the computed solution. Bottom : modulus of the computed solution.

A wave reflection

The cut-off is set at $x = 40l_0$. The heterogeneous medium is modeled by the coefficient

$$\beta(x) = \begin{cases} -\kappa^2, & x < 2, \\ -\kappa^2(x-4)/(2), & x \geq 2. \end{cases} \tag{6.7}$$

One gets a wave propagating from the wave guide through the horn toward the right end of the domain, reflected by the cut off situated at $x = 4$. See Figure 6.20. This result was computed with the normalization $N = \sqrt{\beta}$. Comparing to Figure 6.19 it is clear that the wave does not propagate further than the cut off and bounces back toward the antenna.

Notice that simulations for this kind of problem can be found in [BLSS03, BLSS04, BS00] for caustics related problems and in [DNS08] for laser propagation topics.

6.5.4 Luneburg lens

Another application for this method adapted to smooth coefficients is the simulation of a Luneburg lens, examlpe suggested by Peter Monk. This is a gradient-index lens which refractive index is $\beta = \sqrt{2 - r/R}$, see [GJ99], where R stands for the radius of the lens and r stands for the distance to the center of the lens. It presents a focal point on the outer surface of the lens. Figure 6.21 displays the modulus of the electric field, highlighting a focal point. However, this focal point does not stand exactly on the surface of the lens. Further investigations would be required to get a solution with the correct focal point. For instance a quadrature formula adapted to the curvature of the lens surface could be implemented.

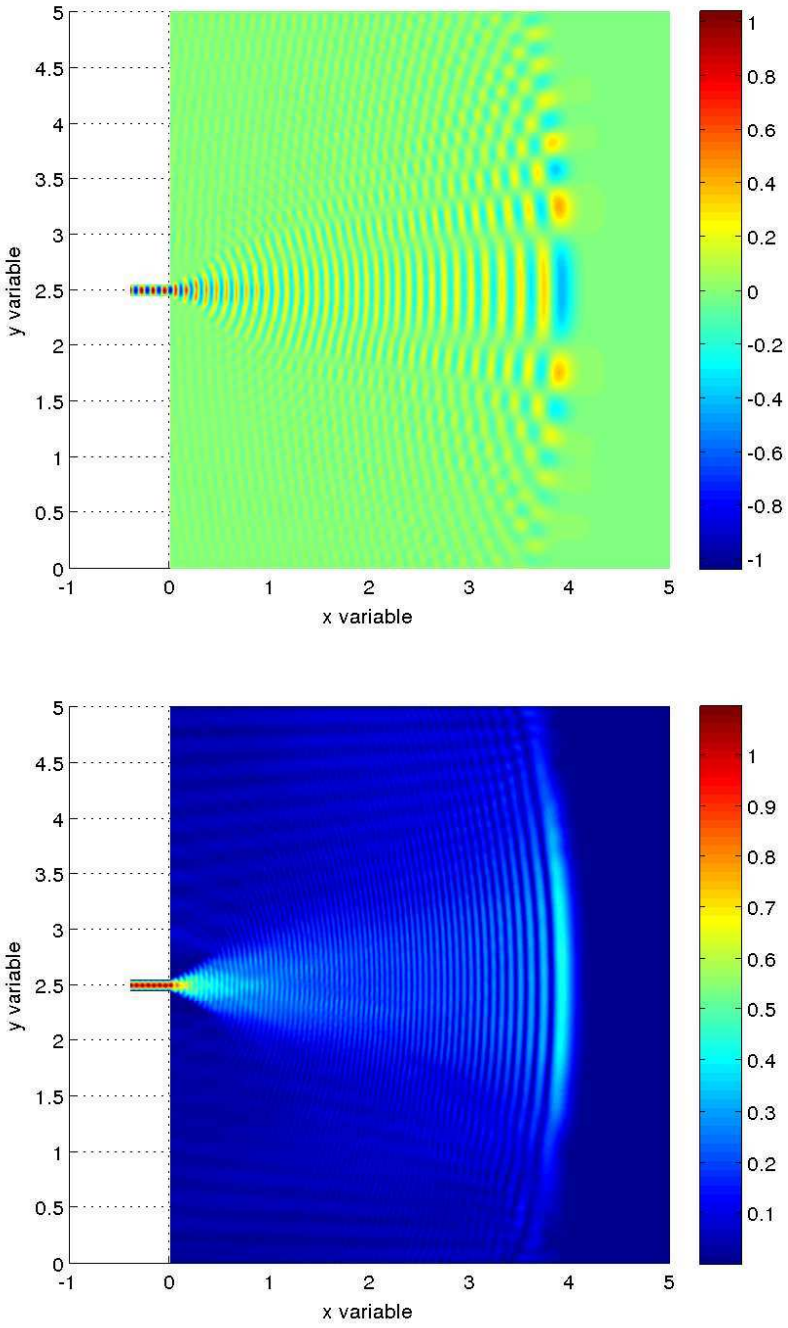


FIGURE 6.20 – Wave reflected by the cut off. Result computed using generalized plane waves designed with $N = \sqrt{\beta}$ and the UWVF, for $p = 7$ and $q = 4$. Top : real part of the computed solution. Bottom : modulus of the computed solution.

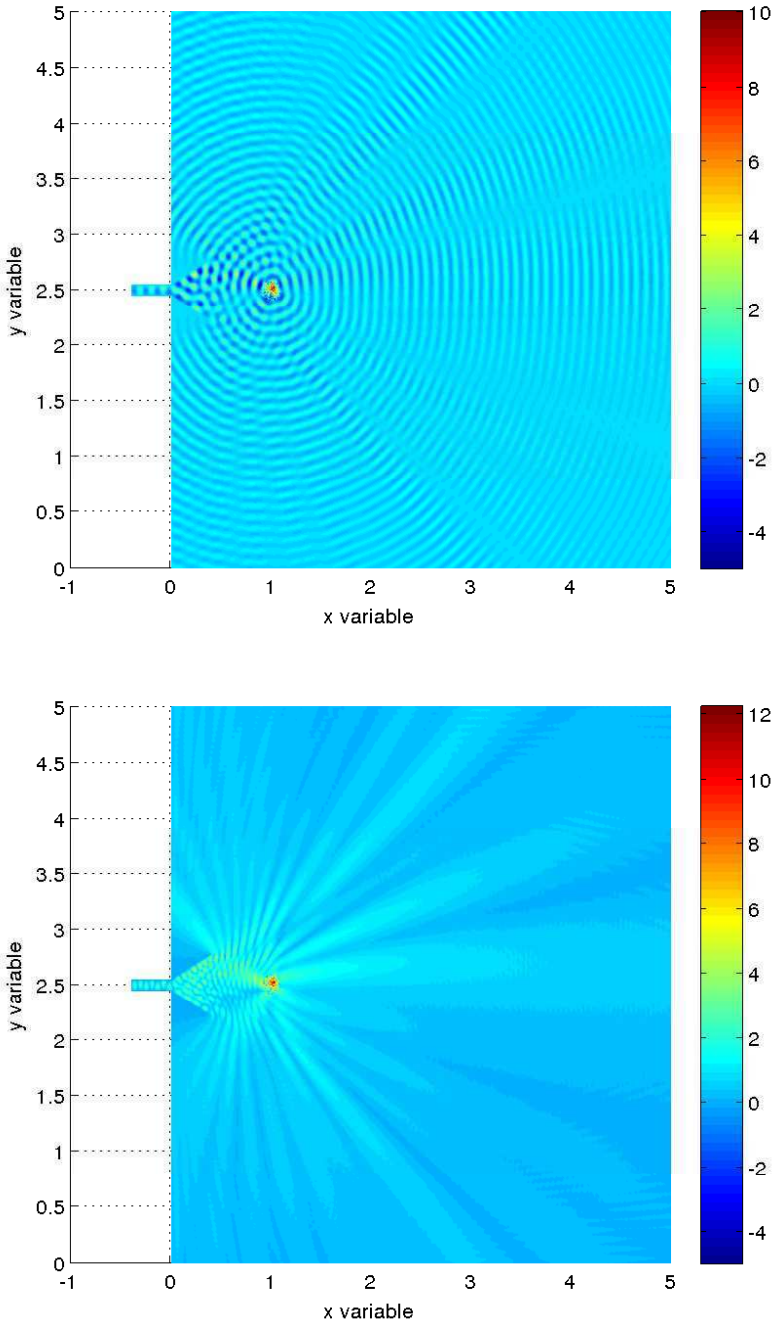


FIGURE 6.21 – Wave sent from an antenna toward a Luneburg lens which is centered at $(1, 2.5)$ and which radius is equal to the wave length of the incoming wave. Result computed using generalized plane waves designed with $N = \sqrt{\beta}$ and the UWVF, for $p = 7$ and $q = 4$. Top : the quantity represented is the real part of the signal. Bottom : the quantity represented is the modulus of the signal.

6.6 X mode simulation in 2D

Adapting the O mode code to the X mode case requires first to write the Ultra-Weak Variational Formulation in this case. In this section the electric field is denoted $E = (E_x, E_y)$.

$$\begin{cases} \nabla \wedge \nabla \wedge E - \varepsilon_\perp E = 0, \\ (\nabla \wedge E + i\sigma E \wedge \nu) = Q\,(\nabla \wedge E - i\sigma E \wedge \nu) + g, \end{cases} \tag{6.8}$$

where

$$\varepsilon_\perp(x) = \begin{pmatrix} \alpha & i\gamma \\ -i\gamma & \alpha \end{pmatrix}, \tag{6.9}$$

$\sigma \in \mathbb{R}$ and $Q \in \mathbb{R}$ satisfies $|Q| \leq 1$. Note that in dimension two, because of the definition of the *curl* operator, the equation of system (6.8) reads

$$\begin{cases} \partial_y\left(\partial_x E_y - \partial_y E_x\right) - \alpha E_x - i\gamma E_y &= 0, \\ -\partial_x\left(\partial_x E_y - \partial_y E_x\right) + i\gamma E_x - \alpha E_y &= 0. \end{cases}$$

6.6.1 The formulation

Let E be a solution of system (6.8), and F be a solution of the dual homogeneous equation, namely

$$\nabla \wedge \nabla \wedge F - \varepsilon_\perp^* F = 0, \tag{6.10}$$

such that for all $k \in [\![1, N_h]\!]$ $\nabla \wedge F$ and $F \wedge \nu$ belong to $L^2(\Omega_k)$. Then, since

$$\begin{aligned} &(\nabla \wedge E - i\sigma E \wedge \nu)\,\overline{(\nabla \wedge F - i\sigma F \wedge \nu)} - (\nabla \wedge E + i\sigma E \wedge \nu)\,\overline{(\nabla \wedge F + i\sigma F \wedge \nu)} \\ &\qquad = 2i\sigma\left(\nabla \wedge E \cdot \overline{F} \wedge \nu - E \wedge \nu \cdot \nabla \wedge \overline{F}\right), \end{aligned}$$

one has for all k

$$\begin{aligned} &\int_{\partial\Omega_k} \frac{1}{\sigma}\,(\nabla \wedge E - i\sigma E \wedge \nu)\,\overline{(\nabla \wedge F - i\sigma F \wedge \nu)} \\ &\qquad - \int_{\partial\Omega_k} \frac{1}{\sigma}\,(\nabla \wedge E + i\sigma E \wedge \nu)\,\overline{(\nabla \wedge F + i\sigma F \wedge \nu)} = 0. \end{aligned}$$

As a result, the UWVF of the X mode problem (6.8) reads

$$\begin{aligned} &\sum_{k\in[\![1,N_h]\!]} \int_{\partial\Omega_k} \frac{1}{\sigma}\,(\nabla \wedge E - i\sigma E \wedge \nu)_k\,\overline{(\nabla \wedge F - i\sigma F \wedge \nu)_k} \\ &- \sum_{k\in[\![1,N_h]\!]} \sum_j \int_{\Sigma_{kj}} \frac{1}{\sigma}\,(\nabla \wedge E - i\sigma E \wedge \nu)_j\,\overline{(\nabla \wedge F + i\sigma F \wedge \nu)_k} \\ &- \sum_{k\in[\![1,N_h]\!]} \int_{\Gamma_k} \frac{Q}{\sigma}\,(\nabla \wedge E - i\sigma E \wedge \nu)\,\overline{(\nabla \wedge F + i\sigma F \wedge \nu)} \\ &= \sum_{k\in[\![1,N_h]\!]} \int_{\Gamma_k} g\overline{(\nabla \wedge F + i\sigma F \wedge \nu)} \end{aligned} \tag{6.11}$$

6.6.2 Design of shape functions

The design of a shape function is more difficult for the X mode equation than for the O mode equation. In fact there is no direct generalization of classical plane waves for the solution of (6.8) because of the non constant coefficients α and γ. However, taking advantage of the fact that $div(\varepsilon_\perp^* F) = 0$, so that there exists a potential φ satisfying

$$\varepsilon_\perp^* F = \overrightarrow{curl}\varphi, \tag{6.12}$$

the initial equation on F can be turned into the following equation on φ :

$$curl(\varepsilon_\perp^{-*}\overrightarrow{curl}\varphi) - \varphi = 0. \tag{6.13}$$

Indeed, since

$$\overrightarrow{curl}(curl\ F - \varphi) = 0,$$

then $curl\ F - \varphi$ is a constant C. But then $\varphi = C + curl\ F$ shows that C can be chosen equal to zero because only the derivatives of φ take part in (6.12), and $F = \varepsilon_\perp^{-*}\overrightarrow{curl}\varphi$ plugged into $\varphi = curl\ F$ gives (6.13).

Taking into account the fact that

$$\varepsilon_\perp^{-*}(x) = \frac{1}{\overline{\alpha}^2 - \overline{\gamma}^2}\begin{pmatrix} \overline{\alpha} & -i\overline{\gamma} \\ i\overline{\gamma} & \overline{\alpha} \end{pmatrix},$$

then φ actually satisfies

$$\begin{aligned} \frac{\overline{\alpha}}{\overline{\alpha}^2 - \overline{\gamma}^2}(\partial_x^2\varphi + \partial_y^2\varphi) + &\left(\partial_x\frac{\overline{\alpha}}{\overline{\alpha}^2 - \overline{\gamma}^2} + \partial_y\frac{i\overline{\gamma}}{\overline{\alpha}^2 - \overline{\gamma}^2}\right)\partial_x\varphi \\ &+ \left(\partial_y\frac{\overline{\alpha}}{\overline{\alpha}^2 - \overline{\gamma}^2} - \partial_x\frac{i\overline{\gamma}}{\overline{\alpha}^2 - \overline{\gamma}^2}\right)\partial_y\varphi + \varphi = 0, \end{aligned} \tag{6.14}$$

so that $\varphi = e^P$ implies that

$$\begin{aligned} \Bigg[\frac{\overline{\alpha}}{\overline{\alpha}^2 - \overline{\gamma}^2}(\partial_x^2 P + \partial_y^2 P + (\partial_x P)^2 + (\partial_y P)^2) + &\left(\partial_x\frac{\overline{\alpha}}{\overline{\alpha}^2 - \overline{\gamma}^2} + \partial_y\frac{i\overline{\gamma}}{\overline{\alpha}^2 - \overline{\gamma}^2}\right)\partial_x P \\ &+ \left(\partial_y\frac{\overline{\alpha}}{\overline{\alpha}^2 - \overline{\gamma}^2} - \partial_x\frac{i\overline{\gamma}}{\overline{\alpha}^2 - \overline{\gamma}^2}\right)\partial_y P + 1\Bigg]e^P = 0. \end{aligned} \tag{6.15}$$

A shape function $\varphi = e^P$ is then sought as a solution of an approximated version of this last equation (6.15).

Following the idea developed in the design of shape functions for the O mode equation, the Taylor expansion of (6.15) at a fixed point $G \in \mathbb{R}^2$ is considered, and for the sake of clarity f_α and f_γ are defined by

$$f_\alpha = \frac{\overline{\alpha}}{\overline{\alpha}^2 - \overline{\gamma}^2} \text{ and } f_\gamma = \frac{i\overline{\gamma}}{\overline{\alpha}^2 - \overline{\gamma}^2}.$$

Their Taylor expansions then read

$$f_\alpha = \sum_{i,j=0}^{\infty} A_{i,j}(x - x_G)^i(y - y_G)^j \text{ and } f_\gamma = \sum_{i,j=0}^{\infty} B_{i,j}(x - x_G)^i(y - y_G)^j.$$

Looking for $\varphi = e^P$ with $P = \sum_{0\leq i+j\leq \mathrm{d}P} \lambda_{i,j}(x-x_G)^i(y-y_G)^j$ and $P(0,0)=0$, one gets a system that turns out to have the same feature as the system obtained for the O mode equation. As a result, the same choice $\mathrm{d}P = q+1$ provides the following system

$$\left\{\begin{array}{l} \forall (i,j)\in\mathbb{N}^2 \ s.t.\ 0\leq i+j\leq q-1 \\ \displaystyle\sum_{k=0}^{i}\sum_{l=0}^{j}\Big(A_{i-k,j-l}\big((k+2)(k+1)\lambda_{k+2,l}+(l+2)(l+1)\lambda_{k,l+2} \\ \qquad \displaystyle+\sum_{n_i=0}^{k}\sum_{n_j=0}^{l}(n_i+1)(k-n_i+1)\lambda_{n_i+1,n_j}\lambda_{k-n_i+1,l-n_j} \\ \qquad \displaystyle+\sum_{n_i=0}^{k}\sum_{n_j=0}^{l}(n_j+1)(l-n_j+1)\lambda_{n_i,n_j+1}\lambda_{k-n_i,l-n_j+1}\big) \\ \qquad +\big((i-k+1)A_{i-k+1,j-l}+(j-l+1)B_{i-k,j-l+1}\big)(k+1)\lambda_{k+1,l} \\ \qquad +\big((j-l+1)A_{i-k,j-l+1}-(i-k+1)B_{i-k+1,j-l}\big)(l+1)\lambda_{k,l+1}\Big) \\ \qquad +\mathbf{1}_{\{i=0,j=0\}}=0. \end{array}\right. \tag{6.16}$$

Here again, as for the O mode case, defining a set $\{\lambda_{i,j}, i\in\{0,1\}, j\in[\![0,q+1-i]\!]\}$ provides an explicit expression to compute all the coefficients of P as long as $A_{0,0}\neq 0$. Naturally the same choice stems from this remark to design a set of linearly independent shape functions :

Definition 27. Consider $N\in\mathbb{C}$ such that $N\neq 0$. The local set of shape functions $\mathcal{E}^X(G,p)=\{\varphi_l\}_{1\leq l\leq p}$ is defined by $\varphi = e^{P_l}$ with $P_l\in\mathbb{C}[X,Y]$ such that $P_l = \sum_{1\leq i+j\leq q+1}\lambda_{i,j}(x-x_G)^i(y-y_G)^j$. For all $l\in\mathbb{N}$ such that $1\leq l\leq p$, the coefficients of P_l are defined by

- $\theta_l = 2l\pi/p$

1. $(\lambda_{1,0},\lambda_{0,1}) = N(\cos\theta_l,\sin\theta_l)$.
2. $\{\lambda_{i,j}, i\in\{0,1\}, 1<i+j\leq q+1\}$ are set to zero,

- $\{\lambda_{i,j}, i\notin\{0,1\}, 1<i+j\leq q+1\}$ are solutions of the system (6.16).

Link with the theory

It is then manifest that the physics of the problem appears through the design process :

- the equation (6.12) has a meaning only if $\varepsilon_\perp$ is invertible, that is only if $\alpha^2(G)-\gamma^2(G)\neq 0$, i.e. not at the cut-off,
- the system (6.16) is invertible as long as $\alpha(G)\neq 0$, i.e. not at the resonance either.

So even in the simple case $\alpha=-x$ and γ real and constant, the regularization process proposed in the theoretical study is mandatory to give a meaning to the design process. Indeed, if $\alpha=-x+i\mu$, then at the cut-off one has

$$\Re(\alpha^2-\gamma^2) = x^2-\mu^2-\gamma^2\neq 0$$

and at the resonance one has $\Im(\alpha(G))=\mu\neq 0$, and as a result the shape functions are well-defined in all the domain.

6.6.3 A benchmark case

An analytic solution of the problem (6.8) can be obtained from the solution of the Budden problem. Looking for a solution that does not depend on the y variable, the coefficients are

$$\begin{cases} \alpha(x,y) = -x + i\mu, \\ \gamma(x,y) = \sqrt{\alpha^2(x,y) - \alpha(x,y)/4 + 1}, \end{cases} \tag{6.17}$$

so that the y component of the electric field satisfies

$$-\frac{d^2 E_y}{dx^2} + \left(\frac{1}{4} - \frac{1}{x - i\mu}\right) E_y = 0.$$

Considering such coefficients, a cut-off occurs at $x = 4$ and a resonance occurs at $x = 0$. Two corresponding solutions of the X mode problem (6.8) are available : a first solution is given by

$$E = e^{(-x+i\mu)/2} \begin{pmatrix} i\gamma \\ -\alpha \end{pmatrix}, \tag{6.18}$$

and is smooth as $\mu \to 0$, whereas a second solution given by

$$E = (-e^{(x-i\mu)/2} + (x - i\mu) e^{(-x+i\mu)/2} E i^{\mu}(x)) \begin{pmatrix} i\gamma \\ -\alpha \end{pmatrix}, \tag{6.19}$$

is singular as $\mu \to 0$. In this last expression the modified exponential integral function is defined by

$$E i^{\mu}(x) = \int_{-\infty}^{x} \frac{e^{t - i\mu}}{t - i\mu} dt.$$

In this case, since

$$\varepsilon_{\perp}^{-*} = \frac{1}{\overline{\alpha}^2 - \overline{\gamma}^2} \begin{pmatrix} \overline{\alpha} & i\overline{\gamma} \\ -i\overline{\gamma} & \overline{\alpha} \end{pmatrix},$$

the required Taylor expansions for the design process - see equations (6.13), (6.14) and (6.15) - are available :

$$\begin{cases} \dfrac{\overline{\alpha}}{\overline{\alpha}^2 - \overline{\gamma}^2} &= \dfrac{4(x_G + i\mu)}{4 - x_G - i\mu} + \dfrac{16}{4 - x_G - i\mu} \displaystyle\sum_{j=0}^{\infty} \left(\frac{x - x_G}{4 - x_G - i\mu}\right)^j, \\ \dfrac{\overline{\gamma}}{\overline{\alpha}^2 - \overline{\gamma}^2} &= \dfrac{4}{x_G + i\mu - 4} \displaystyle\sum_{j=0}^{\infty} \left(\frac{\partial_x^l \overline{\gamma}(x_G)}{l!(4 - x_G - i\mu)^{j-l}}\right) (x - x_G)^j. \end{cases}$$

6.6.4 Interpolation properties

The interpolation properties of the shape functions can be adapted to the X mode case. This paragraph is organized following two main steps : a first result gathers the properties of the potential shape functions φs while another result states the corresponding result for the interpolation of the electric field E.

The following definitions correspond to the ones given in the O mode case, where again e_l refers to the classical plane wave with wave length $\kappa = -iN$ and φ_l refers to the shape function of $\mathcal{E}(G, p)$ with $n \in \mathbb{C}^*$, both of them associated to the same θ_l.

Definition 28. Suppose $n \in \mathbb{N}^*$ and $N \in \mathbb{C}^*$. The $(n+1)(n+2)/2 \times (2n+1)$ matrices M_n^C and M_n^X are defined as follows : for all $(k_1, k_2) \in \mathbb{N}^2$, such that $k_1 + k_2 \leq n$ set

$$\begin{cases} \left(M_n^C\right)_{\frac{(k_1+k_2)(k_1+k_2+1)}{2}+k_2+1,l} = \frac{\partial_x^{k_1}\partial_y^{k_2} e_l(G)}{k_1!k_2!}, \\ \left(M_n^X\right)_{\frac{(k_1+k_2)(k_1+k_2+1)}{2}+k_2+1,l} = \frac{\partial_x^{k_1}\partial_y^{k_2} \varphi_l(G)}{k_1!k_2!}. \end{cases}$$

Their lth columns contain respectively the Taylor expansion coefficients of the functions e_l and φ_l.

Let φ_e be a solution of the homogeneous equation (6.14) in a vicinity $\mathcal{V}_G$ of $G \in \mathbb{R}^2$, h denoting the size of $\mathcal{V}_G$, to be approximated.

Proposition 6.2. *Any shape function $\varphi = e^P \in \mathcal{E}^X(G,p)$ satisfies :*

- *the coefficients of P $\{\lambda_{i,j}, 0 \leq i \leq q+1, 1 < i+j < q+1\}$ can be described as polynomials with two variables in $(\lambda_{1,0}, \lambda_{0,1})$ as follows.*

$$\forall i \geq 2 \ \lambda_{i,j} \text{ is of total degree at most } i-1. \tag{6.20}$$

- *for all $(i,j) \in \mathbb{N}^2$ such that $i+j \leq q+1$ there is a polynomial $S_{i,j} \in \mathbb{C}[X,Y]$ is such that $\mathrm{d}S_{i,j} \leq i-1$, and the coefficients of $S_{i,j}$ only depend on N and on the derivatives of α and γ that satisfies*

$$\partial_x^i \partial_y^j \varphi(G) = (\lambda_{0,1})^j (\lambda_{1,0})^i + S_{i,j}(\lambda_{1,0}, \lambda_{0,1}). \tag{6.21}$$

The rank of M_n^C is known : $rk(M_n^C) = 2n+1$. The matrix M_n^X satisfies :

- *there is a lower triangular matrix L_n, which diagonal coefficients are all equal to 1 and which other coefficients are linear combinations of the derivatives of α and γ evaluated at G, such that*

$$M_n^X = L_n^X \cdot M_n^C. \tag{6.22}$$

- *$rk(M_n^X) = rk(M_n^C)$ and both $\|L_n^X\|$ and $\|(L_n^X)^{-1}\|$ are bounded by a constant only depending on β.*

As a result, suppose that $\varphi_e \in C^{n+1}$ there are a function $\varphi_a \in Span\ \mathcal{E}^X(G,p)$ depending on α, γ and n, and a constant C depending on α, γ and n such that for all $\overrightarrow{x} \in \mathcal{V}_G$

$$\begin{cases} |\varphi_e(\overrightarrow{x}) - \varphi_a(\overrightarrow{x})| \leq C h^{n+1} \|\varphi_e\|_{\mathcal{C}^{n+1}}, \\ \|\nabla\varphi_e(\overrightarrow{x}) - \nabla\varphi_a(\overrightarrow{x})\| \leq C h^n \|\varphi_e\|_{\mathcal{C}^{n+1}}. \end{cases} \tag{6.23}$$

Only the elements that differ from the O mode case will be commented here. The rest of the proofs that do not need to be adapted will not be mentioned here.

*Proof.*Of the first property of the shape function, namely (6.20). From the system (6.16) one gets for $i = 0$

$$\begin{cases} A_{0,0}(2\lambda_{2,0} + \lambda_{1,0}^2 + \lambda_{0,1}^2) + (A_{1,0} + B_{0,1})\lambda_{1,0} + (A_{0,1} - B_{1,0})\lambda_{0,1} + 1 = 0, \\ \sum_{l=0}^{j} A_{0,j-l}(2\lambda_{2,l} + \mathbf{1}_{l=0}(\lambda_{1,0}^2 + \lambda_{0,1}^2)) \\ \qquad + (A_{1,j} + (j+1)B_{0,j+1})\lambda_{1,0} + ((j+1)A_{0,j+1} - B_{1,j})\lambda_{0,1} = 0, \end{cases}$$

so that $\lambda_{2,l}$ is a polynomial of total degree at most 1 with respect to $(\lambda_{1,0}, \lambda_{0,1})$, thanks to the normalization that imposes $\lambda_{1,0}^2 + \lambda_{0,1}^2$ to be constant. This proves the basis case of the induction. Then for all $i > 1$ suppose that the result holds for all $i' < i$. The system (6.16) gives

$$\begin{aligned}
\sum_{k=0}^{i}\sum_{l=0}^{j} A_{i-k,j-l}\Big((k+2)(k+1)\lambda_{k+2,l} + (l+2)(l+1)\lambda_{k,l+2} \\
+ \sum_{n_i=0}^{k}\sum_{n_j=0}^{l}(n_i+1)(k-n_i+1)\lambda_{n_i+1,n_j}\lambda_{k-n_i+1,l-n_j} \\
+ \sum_{n_i=0}^{k}\sum_{n_j=0}^{l}(n_j+1)(l-n_j+1)\lambda_{n_i,n_j+1}\lambda_{k-n_i,l-n_j+1}\Big) \\
+ ((i-k+1)A_{i-k+1,j-l} + (j-l+1)B_{i-k,j-l+1})(k+1)\lambda_{k+1,l} \\
+ ((j-l+1)A_{i-k,j-l+1} - (i-k+1)B_{i-k+1,j-l})(l+1)\lambda_{k,l+1} = 0,
\end{aligned}$$

Isolating the $\lambda_{i+2,j}$ term in the right hand side, the degree of the others terms are at most :

- $k+1 \leq i+1$ for the $\lambda_{k+2,l}$ terms,
- $k-1 \leq i-1$ for the $\lambda_{k,l+2}$ terms,
- $2 \leq i$ for the $\lambda_{1,0}\lambda_{1,l}$ terms corresponding to $k=0$,
- $k+1 \leq i+1$ for the $\lambda_{1,0}\lambda_{k+1,l}$ terms that arise only if $k > 0$,
- $k \leq i$ for the $\lambda_{n_i+1,n_j}\lambda_{k-n_i+1,l-n_j}$ terms that arise only if $k \geq 2$,
- $2 \leq i$ for the $\lambda_{0,1}\lambda_{0,l+1}$ terms corresponding to $k=0$,
- $k+1 \leq i+1$ for the $\lambda_{0,1}\lambda_{k,l+1}$ terms that arise only if $k > 0$,
- $k \leq i$ for the $\lambda_{n_i,n_j+1}\lambda_{k-n_i,l-n_j+1}$ terms that arise only if $k \geq 2$,
- $1 \leq i-1$ for the $\lambda_{1,l}$ terms corresponding to $k=0$,
- $k \leq i$ for the $luk+1l$ terms that arise only if $k > 0$,
- $1 \leq i-1$ for the $\lambda_{0,l+1}$ terms corresponding to $k=0$,
- $k-1 \leq i-1$ for the $\lambda_{k,l+1}$ terms that arise only if $k > 0$.

So altogether the total degree of $\lambda_{i+2,j}$ is at most $i+1$ and the induction step is proved. □

*Proof.*Of the second property of the shape function, namely (6.21). Consider a given partition of (i,j) with length μ, an element of $p_s((i,j),\mu)$ defined by

$$\left\{(k_l,(i_l,j_l))_{l\in[\![1,s]\!]} : k_l \in \mathbb{N}^*, 0 \prec (i_1,j_1) \prec \cdots \prec (i_l,j_l), \sum_{l=1}^{s} k_l = \mu, \sum_{l=1}^{s} k_l(i_l,j_l) = (i,j)\right\}.$$

The degree of the corresponding product term, namely $\prod_{l=1}^{s}(\lambda_{i_l,j_l})^{k_l}$ satisfies $Deg\ \prod_{l=1}^{s}(\lambda_{i_l,j_l})^{k_l} = \sum_{l=1}^{s} k_l Deg\ \lambda_{i_l,j_l}$, from (6.20) this quantity is also at most equal to

$$\sum_{i_l=0,j_l=1} k_l j_l + \sum_{i_l=1,j_l=0} k_l i_l + \sum_{i_l\geq 1} k_l(i_l-1)$$

where the two first sums contain at most one term each. And again this degree is maximal if and only if the partition contains no term such that $i_l \geq 1$. □ The claim concerning the matrix M_n^X represents no difficulty.

*Proof.*Of the final estimate (6.23). A linear system of unknowns $(x_l)_{1\leq l\leq 2n+1}$ and of matrix M_n^X stems from the idea of approximating the Taylor expansion of φ_e thanks to $\varphi_a = \sum_{l=1}^{2n+1} x_l\varphi_l$.

$$\left|\varphi_e(x,y) - \sum_{m=0}^{n}\sum_{k_1+k_2=m} B_{k_1k_2}x^{k_1}y^{k_2}\right| \leq Ch^{n+1}\|\varphi_e\|_{\mathcal{C}^{n+1}},$$

$$\left|\varphi_l(x,y) - \sum_{m=0}^{n}\sum_{k_1+k_2=m} M^l_{k_1k_2}x^{k_1}y^{k_2}\right| \leq Ch^{n+1}\|\varphi_l\|_{\mathcal{C}^{n+1}},$$

Here again, for the sake of simplicity, $M^l_{k_1k_2}$ stands for the coefficient of M_n^X that corresponds to $\partial_x^{k_1}\partial_y^{k_2}\varphi_l/(k_1!k_2!)$, namely the coefficient $(M_n^X)_{\frac{(k_2+k_1)(k_2+k_1+1)}{2}+k_2+1,l}$, and in the same way B_{k_1,k_2} stands for $(B_n)_{\frac{(k_2+k_1)(k_2+k_1+1)}{2}+k_2+1}$. The system to be solved is then

$$\begin{cases} \text{Find } (x_l)_{l\in[\![1,2n+1]\!]} \in \mathbb{C}^{2n+1} \text{ such that} \\ \sum_{l=1}^{2n+1} M^l_{k_1,k_2}x_l = B_{k_1,k_2},\ \forall m \in [\![0,n]\!],\ \forall (k_1,k_2)\in[\![0,n]\!]^2 \text{ such that } k_1+k_2=m. \end{cases}$$

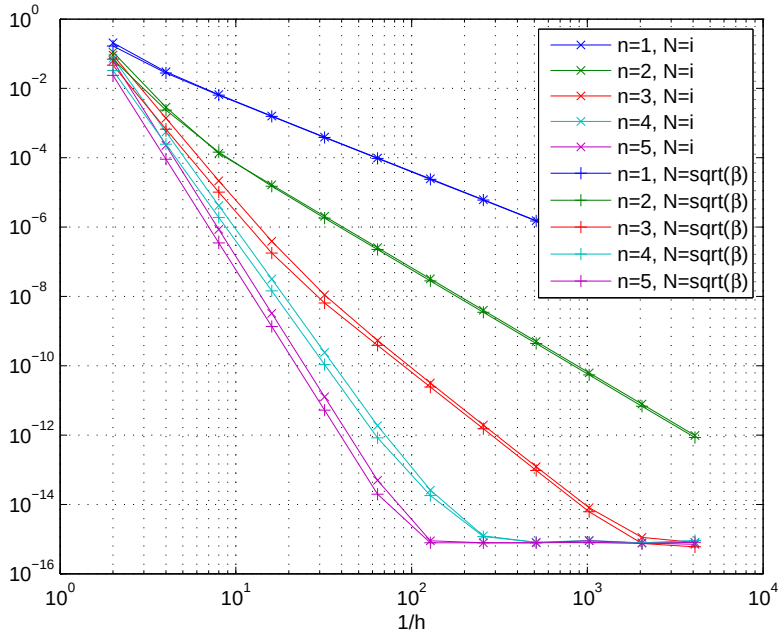
Since the rank of the matrix is known, the next step is to identify a subset $K^X \subset \mathbb{C}^{\frac{(n+1)(n+2)}{2}}$ such that $Im(M_n^X) \subset K^X$ and $B_n \in K^X$. This subspace K^X is built from the fact that the shape functions are designed to fit the Taylor expansion of equation (6.14) :

$$\begin{aligned} K^X := \Bigg\{ & (C_{k_1,k_2}) \in \mathbb{C}^{\frac{(n+1)(n+2)}{2}}, \forall (k_1,k_2)\in\mathbb{N}^2, k_1+k_2 \leq n-2, \\ & \sum_{k=0}^{i}\sum_{l=0}^{j} A_{i-k,j-l}\Big((k+2)(k+1)C_{k+2,l} + (l+2)(l+1)C_{k,l+2} \\ & \qquad + \sum_{n_i=0}^{k}\sum_{n_j=0}^{l}(n_i+1)(k-n_i+1)\lambda_{n_i+1,n_j}C_{k-n_i+1,l-n_j} \\ & \qquad + \sum_{n_i=0}^{k}\sum_{n_j=0}^{l}(n_j+1)(l-n_j+1)\lambda_{n_i,n_j+1}C_{k-n_i,l-n_j+1}\Big) \\ & + ((i-k+1)A_{i-k+1,j-l} + (j-l+1)B_{i-k,j-l+1})(k+1)C_{k+1,l} \\ & + ((j-l+1)A_{i-k,j-l+1} - (i-k+1)B_{i-k+1,j-l})(l+1)C_{k,l+1} + \mathbf{1}_{(j=0,i=0)} = 0\Bigg\}, \end{aligned}$$

It is straightforward to see that $Im(M_n^X) \subset K^X$. The fact that $B_n^X \in K^X$ simply stems from plugging the Taylor expansions of α, γ and φ_e into equation (6.14). □

Figure 6.22 provides numerical convergence results at the resonance, approximating the regular solution of equation (6.10). It shows that the generalized plane waves are able to approximate the regular solution with a high order convergence with respect to h even if the point considered here is lying along the resonance.

The second inequality of (6.14) then directly provides an interpolation result on the electric field E solution of the homogeneous adjoint equation of the X mode system, namely (6.10), in a vicinity $\mathcal{V}_G$ of $G \in \mathbb{R}^2$, h denoting the size of $\mathcal{V}_G$.



	$n=1$		$n=2$		$n=3$		$n=4$		$n=5$	
h	error	order	error	order	error	order	error	order	error	order
$1/2^1$	1.7e-01	-	7.0e-02	-	4.6e-02	-	3.3e-02	-	2.4e-02	-
$1/2^2$	2.8e-02	2.59	2.4e-03	4.89	6.7e-04	6.12	2.5e-04	7.06	9.1e-05	8.03
$1/2^3$	6.4e-03	2.12	1.4e-04	4.03	1.0e-05	6.04	1.9e-06	7.03	3.5e-07	8.02
$1/2^4$	1.6e-03	2.04	1.5e-05	3.27	1.8e-07	5.85	1.4e-08	7.03	1.4e-09	8.01
$1/2^5$	3.9e-04	2.01	1.8e-06	3.05	6.4e-09	4.78	1.1e-10	7.06	5.2e-12	8.01

	$n=1$		$n=2$		$n=3$		$n=4$		$n=5$	
h	error	order	error	order	error	order	error	order	error	order
$1/2^1$	2.1e-01	-	1.1e-01	-	9.1e-02	-	7.5e-02	-	6.3e-02	-
$1/2^2$	3.0e-02	2.78	2.9e-03	5.29	1.4e-03	6.05	5.5e-04	7.08	2.3e-04	8.13
$1/2^3$	6.6e-03	2.19	1.3e-04	4.41	2.2e-05	6.00	4.2e-06	7.06	8.5e-07	8.05
$1/2^4$	1.6e-03	2.05	1.6e-05	3.03	3.9e-07	5.79	3.2e-08	7.03	3.3e-09	8.03
$1/2^5$	4.0e-04	2.01	2.0e-06	3.01	1.1e-08	5.14	2.4e-10	7.04	1.3e-11	8.01

FIGURE 6.22 – Convergence results computed at $G = [0, 1]$ for the X mode interpolation. Both errors and corresponding orders of convergence are provided, for the normalization $N = \sqrt{\beta}$ on top, and $N = i$ at the bottom.

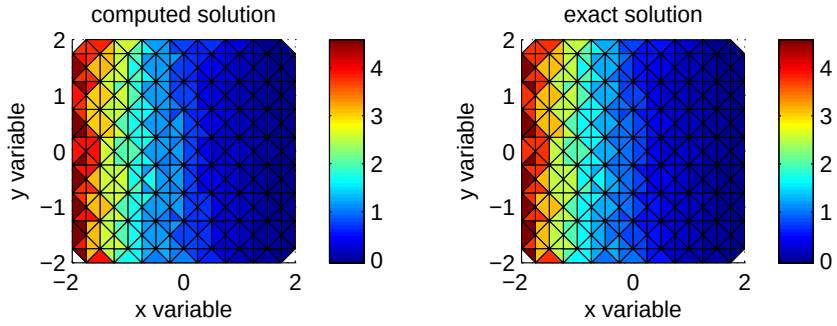


FIGURE 6.23 – Regularized solution of the X mode problem (6.8) computed for $\mu = 0.8$, $p = 7$ and $q = 4$ over the domain $\Omega =]-2, 2[^2$. Left : computed approximation of the *curl* of the electric field. The resonance occurs for $x = 0$. Right : *curl* of the analytic solution. Each vertex of the plotted solution is the center of an edge from the original mesh.

Proposition 6.3. *Suppose that* $E \in C^{n+1}$ *there are a function* $E_a \in Span\left(curl\left(\mathcal{E}^X(G,p)\right)\right)$ *depending on* α, γ *and* n, *and a constant* C *depending on* α, γ *and* n *such that for all* $\overrightarrow{x} \in \mathcal{V}_G$

$$|E(\overrightarrow{x}) - E_a(\overrightarrow{x})| \leq Ch^n \|E\|_{\mathcal{C}^{n+1}} .$$

Proof. It is straightforward considering φ_a defined in Proposition 6.2 and E_a such that

$$E_a = \varepsilon_\perp^{-*} \begin{pmatrix} \partial_y \varphi_a \\ -\partial_x \varphi_a \end{pmatrix}.$$

□

6.6.5 A first UWVF computation for the X mode

Figure 6.23 displays the first numerical result that I computed with the UWVF and the generalized plane waves for the regularized X mode problem. The solution approximated is the regularized regular solution defined in (6.18) which corresponds to the coefficients α and γ set in (6.17). Since the basis functions are potential functions for the X mode system, the *curl* of the solution can be reconstructed at the middle point of every edge of the mesh. The numerical solution displayed in Figure 6.23 is the *curl* of the electric field. The traces of $(\nabla \wedge E)_h$ are expressed with respect to the solution of the discrete system

$X_h = \sum_{k=1}^{N_h} \sum_{l=1}^{p_k} x_{k,l} Z_{k,l} \in V$ on an edge e

$$\begin{cases} 2(\nabla \wedge E)_h = (I + \Pi) X_h + g & \text{if } e \subset \partial\Omega, \\ 2(\nabla \wedge E)_h = (I + \Pi) X_h & \text{if } e \not\subset \partial\Omega. \end{cases}$$

So $(\nabla \wedge E)_h$ is explicitly reconstructed at the $\mathbb{R}^2$ points that are centers x_e of the mesh edges e thanks to the formulas

$$\begin{cases} (\nabla \wedge E)_h(x_e) = \dfrac{1}{2} \left((1+Q) \displaystyle\sum_{l=1}^{p_k} x_{k,l} (\nabla \wedge E_{k,l} - i\sigma E_{k,l} \wedge \nu_k) + g \right)(x_e), & e = \Gamma_k, \\ (\nabla \wedge E)_h(x_e) = \dfrac{1}{2} \left(\displaystyle\sum_{\tilde{k} \in \{k,j\}} \sum_{l=1}^{p_{\tilde{k}}} x_{\tilde{k},l} (\nabla \wedge E_{\tilde{k},l} - i\sigma E_{\tilde{k},l} \wedge \nu_{\tilde{k}}) \right)(x_e), & e = \Sigma_{kj}. \end{cases}$$

The accuracy can be reported using a discrete l^2 norm

$$\sqrt{\frac{\sum_e |(\nabla \wedge E)_{ex}(x_e) - (\nabla \wedge E)_h(x_e)|^2}{\sum_e |(\nabla \wedge E)_{ex}(x_e)|^2}},$$

where the sum is computed over the edges e of the initial mesh, x_e standing for the center of the edge e. First convergence computations tend to show that in such a norm the expected orders of convergence are reached on this test case : as suggested by the interpolation result, if $p = 2n + 1$, $q = n + 1$ are set for some $n \in \mathbb{N}^*$, then the order of convergence on the electric field is close to n. Further investigation is required in this direction.

The definition of parameters to compute an approximation of the singular solution (6.19) is more intricate.

Part III

Appendices

Appendix A

Addendum for the X mode

A.1 Approximation of Airy functions

Suppose that A and B are the two fundamental solutions of the equation $-u''-\alpha u = 0$ satisfying the normalization conditions

$$\begin{cases} A(0) = 1,\ A'(0) = 0, \\ B(0) = 0,\ B'(0) = 1, \end{cases} \tag{A.1}$$

such that the corresponding Wronskian is equal to 1. Let A_μ be an approximation of A in the following sense

$$\begin{cases} -A_\mu'' - \alpha A_\mu = f_{A_\mu}, \qquad \text{with } f_{A_\mu} := i\mu A_\mu, \\ \\ A_\mu(0) = 1,\ A_\mu'(0) = 0. \end{cases} \tag{A.2}$$

From the variation of constants one gets

$$A_\mu(x) = A(x)\left(c_A + \int_0^x f_{A_\mu}(t)B(t)\ dt\right) + B(x)\left(c_B - \int_0^x f_{A_\mu}(t)A(t)\ dt\right).$$

The initial values (A.2) yield $A_\mu(x) = A(x) + \int_0^x f_{A_\mu}(t)k(x,t)\ dt$, where

$$k(x,y) = A(x)B(y) - A(y)B(x).$$

So A_μ satisfies a classical Volterra integral equation

$$A_\mu(x) - i\mu \int_0^x A_\mu(t)k(x,t)\ dt = A(x). \tag{A.3}$$

Define the series of integral kernels

$$\begin{cases} K_1(x,y) = k(x,y), \\ K_{n+1}(x,y) = \displaystyle\int_0^x k(x,x_n)K_n(x_n,y)\ dx_n. \end{cases}$$

The solution of the integral equation (A.3) is

$$A_\mu(x) = A(x) + \int_0^x \left(\sum_{n=0}^{\infty} (i\mu)^{n+1} K_{n+1}(x,y)\right) A(y)\ dy.$$

For $n > 1$

$$K_{n+1}(x,y) = \int_{0<x_1<\cdots<x_n<x} k(x,x_n) \prod_{1\leq i\leq n-1} k(x_{i+1},x_i)\ dx_{i+1} k(x_1,y)\ dx_1,$$

and

$$\begin{array}{rl} I_n(x) & = \int_{(x_1,\ldots,x_n)\in\{0<x_1<\cdots<x_n<x\}} \prod_{1\leq i\leq n}\ dx_i, \\ & = \int_0^x I_{n-1}(x_n)\ dx_n, \\ & = \frac{x^n}{n!} \text{ since } I_1(x) = x. \end{array}$$

So the iterated kernels satisfy $\forall n \geq 0$

$$|K_{n+1}(x,y)| \leq (2\,\|A\|_\infty \|B\|_\infty)^{n+1} x^n/n!.$$

On the compact interval $]0, L_+[$ the sum and integral symbols can be inverted, which gives with a shift of the index n

$$A_\mu(x) = A(x) + \sum_{n=1}^{\infty} \left(\int_0^x K_n(x,y) A(y)\ dy \right) (i\mu)^n .$$

Assuming μ is bounded positive for the simplicity of notations, A_μ is indeed bounded independently for $0 < \mu \leq 1$

$$|A_\mu(x)| \leq \|A\|_\infty \left(1 + \sum_{n=1}^{\infty} \frac{(\mu C_0)^n}{n!} \right) = \|A\|_\infty \left(1 + e^{\mu C_0} - 1 \right) = \|A\|_\infty e^{\mu C_0},$$

with $C_0 = 2L_+ \|A\|_\infty \|B\|_\infty$. From (A.3) it yields

$$|A_\mu(x) - A(x)| \leq \mu C_0\, \|A\|_\infty\, e^{\mu C_0}.$$

Similarly if B_μ approximates B in the following sense

$$\begin{cases} -B_\mu'' - \alpha B_\mu = i\mu B_\mu, \\ \\ B_\mu(0) = 0,\ B_\mu'(0) = 1, \end{cases}$$

then one has the inequality $|B_\mu(x)| \leq \|B\|_\infty\, e^{\mu C_0}$ together with

$$|B_\mu(x) - B(x)| \leq \mu C_0\, \|B\|_\infty\, e^{\mu C_0}.$$

Both $A_\mu(x)$ and $B_\mu(x)$ are $\mathcal{C}^\infty$ functions with respect to μ and x. Since any H^1 function f is $1/2$ Hölder thanks to the inequality $|f(x) - f(y)| \leq \|f'\|_{L^2} |x-y|^{1/2}$, then $A_\mu(x)$, $B_\mu(x)$ as well as all their derivatives with respect to μ and x are also $1/2$ Hölder, with constants bounded independently of μ as far as μ is bounded.

A.2 Adapted Privalov theorem

This section provides two versions of this complex analysis theorem. Regarding the literature about Cauchy principal value and other singular integrals topics useful in this perspective, my personal favorite reference is [Mus92].

A.2.1 A classical version

Lemma A.1. *Suppose φ is 1/2 Hölder in the vicinity of 0 and L^1 on $]L_-, L_+[$. Put $\varphi = 0$ on $\mathbb{C}\backslash]L_-, L_+[$. Suppose L^* is a smooth closed contour of integration including $[L_-, L_+]$ and included in the upper half complex plan. Then*

$$\int_{L^*} \frac{\varphi(z) - \varphi(0)}{z \pm i\mu}\, dz \to_{\mu\to 0} \int_{L^*} \frac{\varphi(z) - \varphi(0)}{z}\, dz.$$

Proof. Let ϵ be a positive number.

$$\int_{L^*} \frac{\varphi(z) - \varphi(0)}{z}\, dz - \int_{L^*} \frac{\varphi(z) - \varphi(0)}{z \pm i\mu}\, dz = \pm i\mu \int_{L^*} \frac{\varphi(z) - \varphi(0)}{z(z \pm i\mu)}\, dz$$

Split this integral into two parts separating the vicinity $]-\rho, \rho[$ of 0 from the rest of L^* such that $\forall z \in L^*\backslash]-\rho, \rho[$, $|z| > \rho$.

Using the Hölder condition $|\varphi(z) - \varphi(0)| \le C|z|^{1/2}$ one obtains, since $|z + i\mu| \ge \mu$,

$$\left| i\mu \int_{-\rho}^{\rho} \frac{\varphi(z) - \varphi(0)}{z(z \pm i\mu)}\, dz \right| \le C \int_{-\rho}^{\rho} |z|^{-1/2}\, dz = 4C\rho^{1/2}.$$

Take ρ so small that $\left| i\mu \int_{L^*} \frac{\varphi(z)-\varphi(0)}{z(z\pm i\mu)}\, dz \right| \le \epsilon/2$; the choice of ρ may obviously be made independently of the position of μ.

Further for z on $L^*\backslash]-\rho, \rho[$, i.e. not close to 0, $|z| \ge \rho$, $|z \pm i\mu| \ge \rho$ and therefore

$$\left| i\mu \int_{L^*\backslash]-\rho,\rho[} \frac{\varphi(z) - \varphi(0)}{z(z \pm i\mu)}\, dz \right| \le \frac{\mu}{\rho^2} \int_{L^*\backslash]-\rho,\rho[} |\varphi(z) - \varphi(0)|\, dz \le \frac{\mu}{\rho^2} \|\varphi - \varphi(0)\|_{L^1}.$$

Thus for sufficiently small μ, $\left| i\mu \int_{L^*\backslash]-\rho,\rho[} \frac{\varphi(z)-\varphi(0)}{z(z\pm i\mu)}\, dz \right| \le \epsilon/2$, and the lemma is proved. □

Theorem A.2.1. Under the assumptions of Lemma A.1, both functions $\phi_\pm(\mu) = \int_{L_-}^{L_+} \frac{\varphi(z)}{z \pm i\mu}\, dz$ admit a continuous limit when μ goes to zero, and therefore are continuous for $\mu \in [0, \infty[$.

Proof. Let μ be any non zero positive parameter. Then

$$\phi_\pm(\mu) = \int_{L^*} \frac{\varphi(z)}{z \pm i\mu}\, dz = \int_{L^*} \frac{\varphi(z) - \varphi(0)}{z \pm i\mu}\, dz + \varphi(0) \int_{L^*} \frac{1}{z \pm i\mu}\, dz,$$

whence

$$\begin{cases} \phi_+(\mu) = \int_{L^*} \frac{\varphi(z)-\varphi(0)}{z+i\mu}\, dz, \\ \phi_-(\mu) = \int_{L^*} \frac{\varphi(z)-\varphi(0)}{z-i\mu}\, dz + 2i\pi\varphi(0). \end{cases}$$

Then, by the lemma proved above, $\phi_\pm$ tends to the limits

$$\begin{cases} \phi_+(0) = \int_{L^*} \frac{\varphi(z)-\varphi(0)}{z}\, dz, \\ \phi_-(0) = \int_{L^*} \frac{\varphi(z)-\varphi(0)}{z}\, dz + 2i\pi\varphi(0). \end{cases}$$

Since $PV \int_{L^*} \frac{1}{z}\, dz = i\pi$ it yields

$$\begin{cases} \phi_+(0) = \mathrm{PV} \int_{L_-}^{L_+} \frac{\varphi(z)}{z}\, dz - i\pi\varphi(0), \\ \phi_-(0) = \mathrm{PV} \int_{L_-}^{L_+} \frac{\varphi(z)}{z}\, dz + i\pi\varphi(0). \end{cases}$$

□

Remark 17. These two proofs essentially come from [Mus92]. Moreover the Cauchy principal value of an integral is also defined in [Mus92] in the following sense

$$\mathrm{PV}\int_{L_-}^{L_+}\frac{\varphi(z)}{z}\,dz = i\pi\varphi(0)+\varphi(0)\log\frac{L_+}{L_-}+\int_{L_-}^{L_+}\frac{\varphi(z)-\varphi(0)}{z}\,dz,$$

where log is the principal value of the complex logarithm. Then

$$\mathrm{PV}\int_{L_-}^{L_+}\frac{\varphi(z)}{z}\,dz = \varphi(0)\log\frac{L_+}{|L_-|}+\int_{L_-}^{L_+}\frac{\varphi(z)-\varphi(0)}{z}\,dz.$$

A.2.2 An adapted version

Since all the numerators of the kernels appearing in Section 4.3.2 do also depend on the parameter μ, here is a result specifying one case in which the result of the previous section can be extended. The following theorem states that if the Hölder condition of a derivative of the numerator is **uniform with respect to** μ, then one also has a continuity when μ goes to zero.

Theorem A.2.2. Suppose φ is defined on $]L_-,L_+[\times\mathbb{R}^+$, and suppose $\partial_2\varphi$ exists and is $1/2$ Hölder with respect to the first variable uniformly with respect to the second one, i.e. there exists a constant C independent on μ such that $|\partial_2\varphi(z_1,\mu)-\partial_2\varphi(z_2,\mu)|\leq C|z_1-z_2|^{1/2}$ for all $z\in]L_-,L_+[$. Suppose that $\varphi(.,0)$ and $\partial_2\varphi(.,\mu)$ for each value of μ belong to $L^1(]L_-,L_+[)$. Suppose that $\partial_2\varphi(z,.)$ is continuous. Then both functions $\tilde{\phi}_\pm(\mu)=\int_{L_-}^{L_+}\frac{\varphi(z,\mu)}{z\pm i\mu}\,dz$ admit a continuous limit when μ goes to zero, and therefore are continuous for $\mu\in[0,\infty[$.

Proof. Define $\psi(\mu)=\int_{L^*}\frac{\varphi(z,\mu)-\varphi(0,0)}{z\pm i\mu}\,dz$ and consider the difference

$$\begin{aligned}\psi(\mu)-\psi(0) &= \int_{L^*}\frac{\varphi(z,\mu)-\varphi(z,0)}{z\pm i\mu}\,dz \pm i\mu\int_{L^*}\frac{\varphi(0,0)-\varphi(z,0)}{z(z\pm i\mu)}\,dz,\\ &= \int_0^\mu\int_{L^*}\frac{\partial_2\varphi(z,v)}{z\pm i\mu}\,dz\,dv \pm i\mu\int_{L^*}\frac{\varphi(0,0)-\varphi(z,0)}{z(z\pm i\mu)}\,dz.\end{aligned}$$

It has been shown in the proof of lemma A.1 that the second term tends to zero. Moreover, from lemma A.1 and from the uniform Hölder condition on $\partial_2\varphi$ one has when μ tends to zero

$$\int_{L^*}\frac{\partial_2\varphi(z,v)}{z\pm i\mu}\,dz \to PV\int_{L_-}^{L_+}\frac{\partial_2\varphi(z,v)}{z}\,dz \mp i\pi\partial_2\varphi(0,v).$$

Since $\partial_2\varphi(z,.)$ is continuous, a classical argument of continuity under the integral sign shows that $\psi(\mu)-\psi(0)$ goes to zero when μ goes to zero.

As $\tilde{\phi}_\pm(\mu)=\psi(\mu)+\varphi(0,0)\int_{L^*}\frac{dz}{z\pm i\mu}$ it yields that $\tilde{\phi}$ is indeed continuous and converges to

$$\begin{cases}\tilde{\phi}_+(0)=\psi(0),\\ \tilde{\phi}_-(0)=\psi(0)+2i\pi\varphi(0,0).\end{cases}$$

in other words

$$\begin{cases}\tilde{\phi}_+(0)=\mathrm{PV}\int_{L_-}^{L_+}\frac{\varphi(z,0)}{z}\,dz-i\pi\varphi(0,0),\\ \tilde{\phi}_-(0)=\mathrm{PV}\int_{L_-}^{L_+}\frac{\varphi(z,0)}{z}\,dz+i\pi\varphi(0,0).\end{cases}$$

□

Appendix B

Additional illustrations

UWVF matrices profile

The structure of the sparse matrix to be inverted to solve the UWVF discrete problem is linked to the numbering of the mesh elements. The kth $p_k \times p_k$ diaganonal blocks are always non zero because of matrix D blocks and the terms from the possible boundary Γ_k of matrix C. On the other hand, the $p_j \times p_k$ off-diagonal block corresponding to the term from the interface Σ_{kj} of matrix C are non zero if and only if the kth and jth elemnts of the mesh have a common an edge. As a result, in dimension one, because each cell k is at most the neighbor of cells $k+1$ and $k-1$, the matrix has a block tridiagonal structure. In dimension two, these non zero off-diagonal blocks can be anywhere in the matrix, but every non zero (j,k) block corresponds to a non zero (k,j) block. See Figure B.1.

A look at the generalized plane waves

In dimension two, the basis functions are designed locally at a point $G \in \mathbb{R}^2$. Consider here $G = (-2,2)$, which is lying in the propagative zone for $\beta(x,y) = x - 1$. Figure B.3 displays the classical plane waves designed at G. The only difference between these five plane waves is their own direction. Figure B.2 displays some generalized plane waves designed at G. They also seem to have a main direction, however they have different shapes.

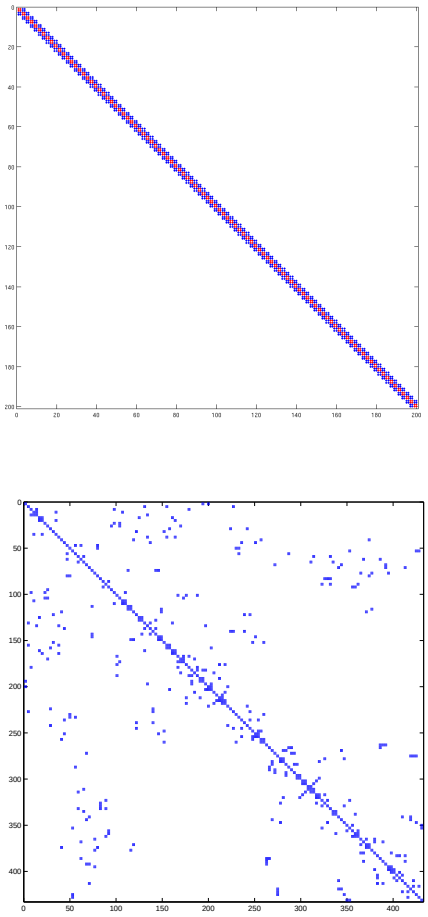


FIGURE B.1 – Sparse profile of the matrices of the UWVF discrete problem. Left : one dimensional case. Right : two dimensional case.

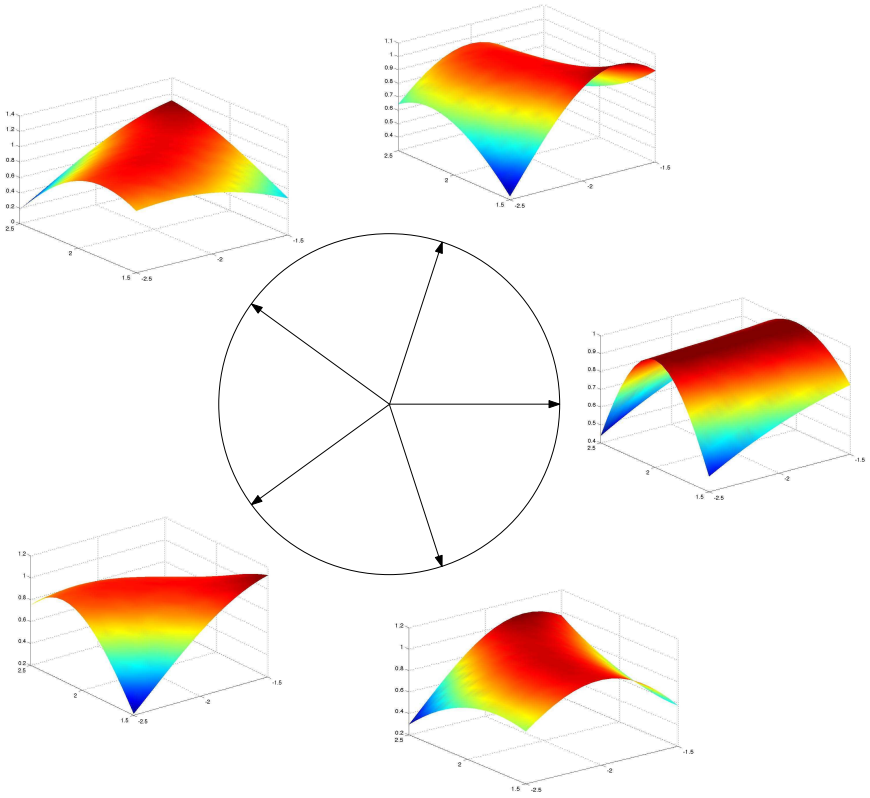


FIGURE B.2 – For $p = 5$ and $q = 5$, generalized plane waves designed at $G = (-2, 2)$ with the β-normalization, $\beta(x, y) = x - 1$.

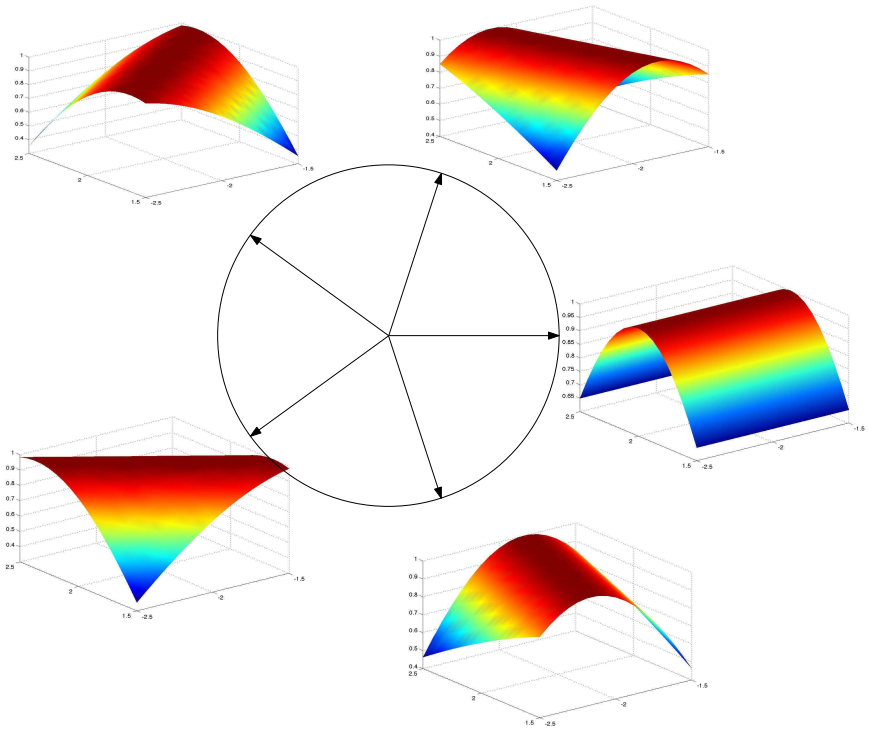


FIGURE B.3 – For $p = 5$, classical plane waves designed at $G = (-2, 2)$ for $\beta(x, y) = x - 1$.

Appendix C

Some images of ITER

All the following images come from the official ITER website [Org]. It is specified in the multimedia section of the website that the images from the Iter image galleries may be freely downloaded for non-commercial, scientific, news and educational purposes.

Figure C.1 represents a cutaway of the ITER tokamak, highlighting the important size of the diagnostic system with respect to the reactor. Figure C.2 represents a cyclotron antenna which is used to heat the plasma.

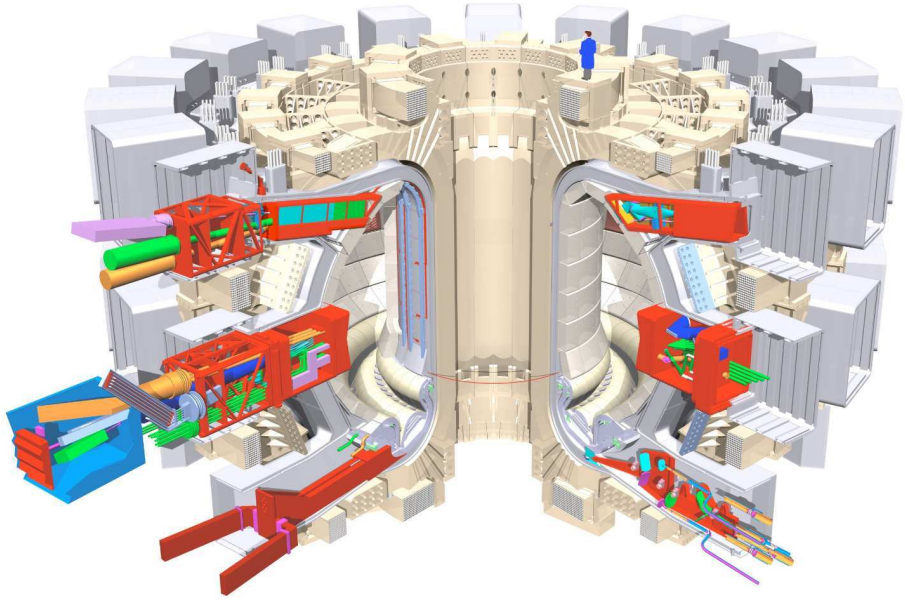


FIGURE C.1 – Diagnostic systems. About 50 individual measurement systems will help to control, evaluate and optimize plasma performance in ITER and to further understanding of plasma physics.

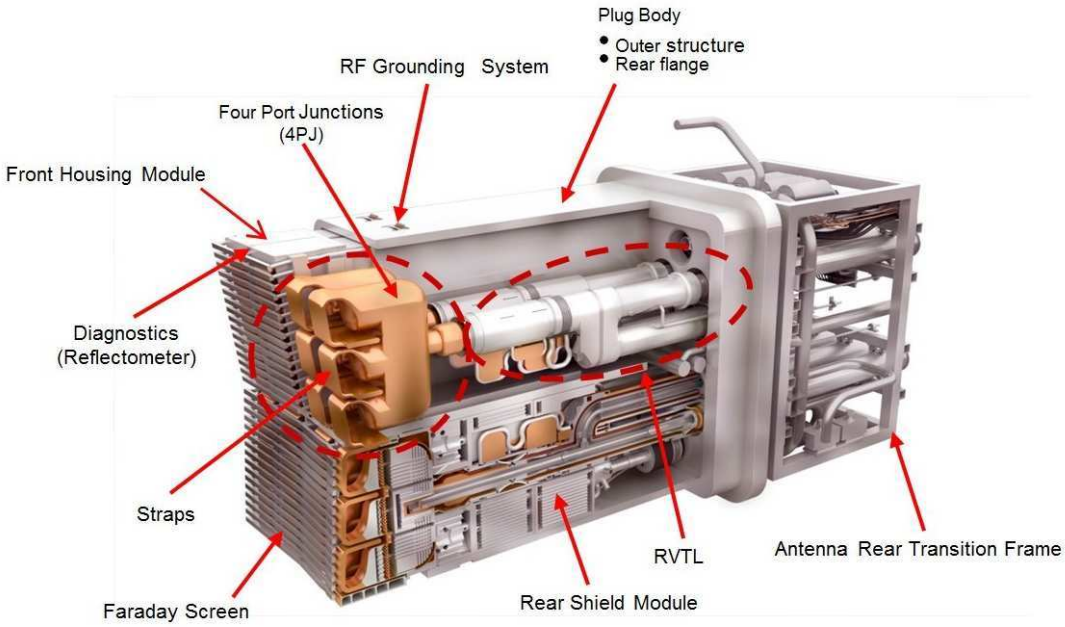


FIGURE C.2 – ITER's ion cyclotron antenna. One of the two 45-ton Ion Cyclotron Resonant Heating antenna systems that will deliver 10 MW of heating power each into the ITER machine.

References

[ACR09] M. Amara, F. Charbel, and D. Rabia. Convergence analysis of a discontinuous Galerkin method with plane waves and Lagrange multipliers for the solution of Helmholtz problems. *Ann. of Math. (2) SIAM J. Numer. Anal.*, 47(2) :1038–1066, 2009.

[Ami78] C. J. Amick. Some remarks on Rellich's theorem and the Poincaré inequality. *J. London Math. Soc. (2)*, 18(1) :81–93, 1978.

[AS72] M. Abramowitz and I. Stegun. *Handbook of Mathematical. Functions with Formulas, Graphs, and Mathematical Tables*. New York : Dover Publications, 1972.

[Bat53] H. Bateman. *Higher Transcendental Functions Volumes 1, 2, 3*. A. Erdelyi, McGraw-Hill Book company, 1953.

[BBCD97] F. Ben Belgacem, C. Bernardi, M. Costabel, and M. Dauge. Un résultat de densité pour les équations de Maxwell. (French. English, French summary) [A denseness result for Maxwell's equations] . *C. R. Acad. Sci. Paris Sér. I Math.*, 324(6) :731–736, 1997.

[BCC12] A.S. Bonnet-Ben Dhia, L. Chesnel, and P. Ciarlet. T -coercivity for scalar interface problems between dielectrics and metamaterials. *Math. Mod. Num. Anal.*, 18 :1363–1387, 2012.

[BCZ08] A.S. Bonnet-Ben Dhia, P. Ciarlet, and C. Zwölf. A new compactness result for electromagnetic waves. Application to the transmission problem between dielectrics and metamaterials. *Math. Models Meth. App. Sci.*, 18 :1605–1631, 2008.

[Bee69] P.R. Beesack. Comparison theorems and integral inequalities for Volterra integral equations . *Proc. Amer. Math. Soc.*, 204, 1969.

[BLSS03] J.-D. Benamou, O. Lafitte, R. Sentis, and I. Solliec. A geometrical optics-based numerical method for high frequency electromagnetic fields computations near fold caustics—Part I. *Journal of Computational and Applied Mathematics*, 156(1) :93–125, 2003.

[BLSS04] J.-D. Benamou, O. Lafitte, I. Solliec, and R. Sentis. A geometric optics method for high-frequency electromagnetic fields computations near fold caustics—Part II. The energy. *Journal of Computational and Applied Mathematics*, 167(1) :91–134, 2004.

[BM96] D. Bouche and F. Molinet. *Asymptotic Methods in Electromagnetics*. Springer-Verlag Berlin and Heidelberg GmbH & Co. K, 1996.

[BM08] A. Buffa and P. Monk. Error estimates for the Ultra Weak Variational Formulation of the Helmholtz equation. *ESAIM : Mathematical Modelling and Numerical Analysis*, 42 :925–940, 2008.

[Bra98] M. Branbilla. *Kinetic Theory of Plasma Waves- Homogeneous Plasmas*. International Series of Monographs on Physics. Clarendon Press, 1998.

[Bré83] H. Brézis. *Analyse Fonctionnelle, Théorie et applications*. Masson, Paris, 1983.

[BS97] I. M. Babuska and S. A. Sauter. Is the pollution effect of the FEM avoidable for the Helmholtz equation considering high wave numbers ? *Reprint of SIAM J. Numer. Anal.*, 34(6) :2392–2423, 1997.

[BS00] J.-D. Benamou and I. Solliec. An Eulerian Method for Capturing Caustics. *Journal of Computational Physics*, 162(1) :132–163, 2000.

[BW73] G.R. Bart and R.L. Warnock. Linear integral equations of the third kind. *Siam J. of Math. Anal*, 40 :609–622, 1973.

[Cas59] K.M. Case. Plasma oscillations. *Annals of physics*, 78 :349–364, 1959.

[CD98] O. Cessenat and B. Després. Application of an ultra weak variational formulation of elliptic PDEs to the two dimensional Helmholtz problem. *SIAM J. Numer. Anal.*, 55(1) :255–299, 1998.

[Ces96a] M. Cessenat. *Mathematical Methods in Electromagnetism : Linear Theory and Applications*. World Scientific Publishing Company, 1996.

[Ces96b] O. Cessenat. *Application d'une nouvelle formulation variationnelle aux équations d'ondes harmoniques. Problèmes de Helmholtz 2D et de Maxwell 3D.* PhD thesis, Université Paris IX Dauphine, 1996.

[Che12] L. Chesnel. *Etude de quelques problèmes de transmission avec changement de signe. Application aux métamatériaux.* PhD thesis, Ecole Polytechnique, 2012.

[Cia12] P. Jr. Ciarlet. T-coercivity : application to the discretization of Helmholtz-like problems. *Comput. Math. Appl.*, 64(1) :22–34, 2012.

[CL55] E.A. Coddington and N. Levinson. *Theory of Ordinary Differential Equations*. McGraw-Hill Book Company, Inc., 1955.

[Cra05] A. D. D. Craik. Prehistory of Faà di Bruno's formula. *Amer. Math. Monthly*, 112(2) :119–130, 2005.

[CS96] G. M. H. Constantine and T. Savits. A multivariate Faà di Bruno formula with applications. *Trans. Amer. Math. Soc.*, 348(2) :503–520, 1996.

[CS12] B. Cockburn and M. Solano. Solving Dirichlet boundary-value problems on curved domains by extensions from subdomains. *SIAM J. Sci. Comput.*, 34(1) :A497–A519, 2012.

[CSS12] B. Cockburn, F.-J. Sayas, and M. Solano. Coupling at a distance HDG and BEM. *SIAM J. Sci. Comput.*, 34(1) :A28–A47, 2012.

[CW74] F.F. Chen and R.B. White. Amplification and Absorption of Electromagnetic Waves in Overdense Plasmas. *Plasma Phys. ; anthologized in Laser Interaction with Matter, Series of Selected Papers in Physics, ed. by C. Yamanaka, Phys. Soc. Japan, 1984*, 1974.

[DDF+11] E. Deriaz, B Despres, G. Faccanoni, K. P. Gostaf, L.-M. Imbert-Gerard, G. Sadaka, and R. Sart. Magnetic equations with FreeFem ++ : the Grad-Shafranov equation & the current hole. *ESAIM Proc*, CEMRACS'10 research achievements : numerical modeling of fusion(32), 2011.

[Des] B. Despres. Waves in Magnetic Plasmas.

[Des94] B. Després. Sur une formulation variationnelle de type ultra-faible. *C. R. Acad. Sci. Paris Sér. I Math.*, 318(10) :939–944, 1994.

[DHM06] F. Da Silva, S. Heuraux, and M. Manso. Developments on reflectometry simulations for fusion plasmas : applications to ITER position reflectometry. *J. Plasma Physics*, 72(1205), 2006.

[DL84] R. Dautray and J.-L. Lions. *Mathematical Analysis and Numerical Methods for Science and Technology, volume 3, chapitre 8.* . Masson, 1984.

[DL85] R. Dautray and J.-L. Lions. *Analyse mathématique et calcul numérique pour les sciences et les techniques*, volume 2. Masson, Paris, 1985.

[DM07] E. Darrigrand and P. Monk. Coupling of the ultra-weak variational formulation and an integral representation using a fast multipole method in electromagnetism. *J. Comput. Appl. Math.*, 204(2) :400–407, 2007.

[DM12] E. Darrigrand and P. Monk. Combining the ultra-weak variational formulation and the multilevel fast multipole method. *Appl. Numer. Math.*, 62(6) :709–719, 2012.

[DNS08] S. Desroziers, F. Nataf, and R. Sentis. Simulation of laser propagation in a plasma with a frequency wave equation. *Journal of Computational Physics*, 227(4) :2610–2625, 2008.

[DPS05] R.J. Dumont, C.K. Phillips, and D.N. Smithe. Effects of non-Maxwellian species on ion cyclotron waves propagation and absorption in magnetically confined plasmas. *Phys. Plasmas*, 12(042508), 2005.

[Dum09] R.J Dumont. Variational approach to radiofrequency waves in magnetic fusion devices . *Nuclear Fusion*, 49(075033), 2009.

[Eva10] L.C. Evans. *Partial differential equations. Second edition.* Graduate Studies in Mathematics, 19. American Mathematical Society, Providence, RI, 2010.

[FHF01] C. Farhat, I. Harari, and L. Franca. The discontinuous enrichment method. *Computer Methods in Applied Mechanics and Engineering*, 190 :6455–6479, 2001.

[Fre07] J.P. Freidberg. *Plasma physics and fusion energy.* Cambridge university press, 2007.

[FTWG04] C. Farhat, R. Tezaur, and P. Wiedemann-Goiran. Higher-order extensions of a discontinuous Galerkin method for mid-frequency Helmholtz problems. *International Journal for Numerical Methods in Engineering*, 61 :1938–1956, 2004.

[FW63] G. Frye and R.L. Warnock. Analysis of partial-wave dispersion relations. *Physical review*, 130 :478–494, 1963.

[GGH11] G. Gabard, P. Gamallo, and T. Huttunen. A comparison of wave-based discontinuous Galerkin, ultra-weak and least-square methods for wave problems. *International Journal for Numerical Methods in Engineering*, 85 :380–402, 2011.

[GHP09] C. J. Gittelson, R. Hiptmair, and I. Perugia. Plane wave discontinuous Galerkin methods : Analysis of the h-version. *ESAIM : Mathematical Modelling and Numerical Analysis*, 43 :297–331, 2009.

[GJ99] A.D. Greenwood and J.-M. Jin. A field picture of wave propagation in inhomogeneous dielectric lenses. *Antennas and Propagation Magazine, IEEE*, 41(5), 1999.

[GR65] I.S. Gradshteyn and I.M. Ryzhik. *Tables of integrals and products.* Academic Press, New York, 1965.

[Gri85] P. Grisvard. *Elliptic problems in nonsmooth domains.* Monographs and Studies in Mathematics, 24. Pitman (Advanced Publishing Program), Boston, MA, 1985.

[Har06] M. Hardy. Combinatorics of partial derivatives. *Electron. J. Combin.*, 13(1) :13, 2006.

[HdSG+11] S. Heuraux, F. da Silva, E. Z. Gusakov, A. Yu Popov, E. Beauvier, N. Kosolapova, and K. Syisoeva. Reflectometry Simulations on Different Methods to Extract Fusion Plasma Turbulence Characteristics and Its Dynamics. *Contributions To Plasma Physics*, 51 :126–130, 2011.

[Hen57] P. Henrici. A survey of I. N. Vekua's theory of elliptic partial differential equations with analytic coefficients. . *Z. Angew. Math. Phys.*, 82 :169–203, 1957.

[HGP10] S. Heuraux, E. Z. Gusakov, and A. Popov. A Numerical Study of Forward- and Backscattering Signatures on Doppler-Reflectometry Signals. *IEEE Transactions on Plasma Science*, 38(9), 2010.

[Hil53] D. Hilbert. *Grundzüge einer allgemeinen Theorie der linear Integralgleichungen.* Chelsea, New York, 1953.

[HKM04] T. Huttunen, J. P. Kaipio, and P. Monk. The perfectly matched layer for the ultra weak variational formulation of the 3D Helmholtz equation. *Internat. J. Numer. Methods Engrg.*, 61(7) :1072–1092, 2004.

[HKM08] T. Huttunen, J. P. Kaipio, and P. Monk. An ultra-weak method for acoustic fluid-solid interaction. *J. Comput. Appl. Math.*, 213(1) :166–185, 2008.

[HMCK04] T. Huttunen, P. Monk, F. Collino, and J. P. Kaipio. The ultra-weak variational formulation for elastic wave problems. *SIAM J. Sci. Comput.*, 25(5) :1717–1742, 2004.

[HMK02] T. Huttunen, P. Monk, and J. P. Kaipio. Computational Aspects of the Ultra-Weak Variational Formulation. *Journal of Computational Physics*, 182(1) :27–46, 2002.

[HMM07] T. Huttunen, M. Malinen, and P. Monk. Solving Maxwell's equations using the ultra weak variational formulation. *Journal of Computational Physics*, 223(2) :731–758, 2007.

[HMP11] R. Hiptmair, A. Moiola, and I. Perugia. Plane wave discontinuous Galerkin methods for the 2D Helmholtz equation : analysis of the p-version. *SIAM J. Numer. Anal*, 49(1) :264–284, 2011.

[HMP13] R. Hiptmair, A. Moiola, and I. Perugia. Error analysis of Trefftz-discontinuous Galerkin methods for the time-harmonic Maxwell equations. *Math. Comp.*, 82(281) :247–268, 2013.

[Hor76] L. Hormander. *Linear partial differential operators. (4th edition).* Springer-Verlag, 1976.

[IGD11] L.-M. Imbert-Gérard and B. Després. A generalized plane wave numerical method for smooth non constant coefficients. *To appear in IMA Journal of Numerical Analysis*, 2011.

[JW10] P. Joly and R. Weder. Analysis of Acoustic Wave Propagation in a Thin Moving Fluid. *SIAM Journal on Applied Mathematics*, 70(7) :2449–2472, 2010.

[KGH09] N. V. Kosolapova, E. Z. Gusakov, and S. Heuraux. Numerical modeling of micro turbulence wave number spectra reconstruction using radial correlation reflectometry : I. O-mode reflectometry at the linear plasma density profile. In *ECA Vol. 33E*. 36th EPS Conference on Plasma Phys. Sofia, 2009.

[Lab] S. Labrunie. Full-wave simulation for plasma heating, Iter group talk, Laboratoire Jacques-Louis Lions, June 21st, 2012.

[LBA02] O. Laghrouche, P. Bettes, and R. J. Astley. Modelling of short wave diffraction problems using approximating systems of plane waves. *International Journal for Numerical Methods in Engineering*, 54(10) :1501–1533, 2002.

[LDMS96] C. Laviron, A.J.H. Donné, M.E. Manso, and J. Sanchez. Reflectometry techniques for density profile measurements on fusion plasmas . *Plasma Phys. Control. Fusion*, 38 :905–936, 1996.

[LHM12] T. Luostari, T. Huttunen, and P. Monk. The ultra weak variational formulation using Bessel basis functions. *Commun. Comput. Phys.*, 11(2) :400–414, 2012.

[LHM13] T. Luostari, T. Huttunen, and P. Monk. Error estimates for the ultra weak variational formulation in linear elasticity. *ESAIM Math. Model. Numer. Anal.*, 47(1) :183–211, 2013.

[Ma09] T.-W. Ma. Higher chain formula proved by combinatorics. *Electron. J. Combin.*, 16(1) :7, 2009.

[MB96] J.M. Melenk and I. Babuska. The partition of unity method finite element method : basic theory and applications. *Computer Methods in Applied Mechanics and Engineering*, 139 :289–314, 1996.

[MHP11] A. Moiola, R. Hiptmair, and I. Perugia. Vekua theory for the Helmholtz operator. *Zeitschrift für angewandte Mathematik und Physik*, 62(5) :779–807, 2011.

[Mon03] P. Monk. *Finite element for Maxwell's equations*. Numerical Mathematics and Scientific Computations. Clarendon Press, Oxford, 2003.

[MPS12] J.M. Melenk, A. Parsania, and S. Sauter. Generalized DG-Methods for Highly Indefinite Helmholtz Problems based on the Ultra-Weak Variational Formulation. *ASC Report 06/2012*, 2012.

[MS] A. Moiola and E. Spence. Is the Helmholtz equation really sign-indefinite ? Preprint.

[MSK+98] M. Manso, F. Serra, B. Kurzan, I. Nunes, J. Santos, A. Silva, W. Suttrop, P. Varela, and S. Vergamota. H-mode studies with microwave reflectometry on ASDEX Upgrade. *Plasma Phys. Control. Fusion*, 40 :747–752, 1998.

[Mus92] N.I. Muskhelishvili. *Singular integral equations*. Dover Publications, 1992.

[MV11] C. Mouhot and C. Villani. On Landau damping. *Acta Mathematica*, 207 :29–201, 2011.

[Nec67] J. Necas. *Les méthodes directes en théorie des équations elliptiques (French).* . Masson et Cie, Editeurs, Paris ; Academia, Editeurs, Prague, 1967.

[Oka02] T. Okaji. Strong unique continuation property for time harmonic Maxwell equations. *J. Math. Soc. Japan*, 54(1) :89–122, 2002.

[Org] ITER Organization. ITER organization web page.

[OS01] P. Ortiz and E. Sanchez. An improved partition of unity finite element model for diffraction problems. *Int. J. Numer. Meth. Engng*, 50(12) :2727–2740, 2001.

[PDLBT04] E. Perrey-Debain, O. Laghrouche, P. Bettess, and J. Trevelyan. Plane-wave basis finite elements and boundary elements for three-dimensional wave scattering. *Philos. Trans. R. Soc. Lond. Ser. A Math. Phys. Eng. Sci.*, 362 :561–577, 2004.

[PHVD07] B. Pluymers, B. Hal, D. Vandepitte, and W. Desmet. Trefftz-based methods for time-harmonic acoustics. *Archives of Computational Methods in Engineering*, 4(14) :343–381, 2007.

[Pic11] E. Picard. Sur les équations intégrales de troisième espèce. *Annales scientifiques de l'ENS*, 28 :459–472, 1911.

[Pri56] I.I. Privalov. *Randwerteigenschaften Analytischer Funktionen.* V.E.B. Deutscher Verlag der Wissenschaften, 1956.

[PT05] A.D. Piliya and E.N. Tregubova. Linear conversion of electromagnetic waves into electron Bernstein waves in an arbitrary inhomogeneous plasma slab. *Plasma Phys. Control. Fusion*, 47(143), 2005.

[Saa] P. Saadé. http ://picviz.com/en/index.html. Technical report, Picviz labs.

[Shu97] D. Shulaia. On one Fredholm integral equation of third kind. *Georgian Mathematical Journal*, 4 :461–476, 1997.

[Sti92] T.H. Stix. *Waves in plasmas.* Springer-Verlag, 1992.

[Swa03] D. G. Swanson. *Plasma Waves, 2nd Edition.* Series in Plasma Physics. Institute of Physics Publishing, 2003.

[TF06] R. Tezaur and C. Farhat. Three-dimensional discontinuous Galerkin elements with plane waves and Lagrange multipliers for the solution of mid-frequency Helmholtz problems. *Int. J. Numer. Meth. Engng.*, 66(5) :796–815, 2006.

[Tri85] F.G. Tricomi. *Integral equations.* Dover Publications, 1985.

[Van55] N.G. Van Kampen. On the theory of stationnary waves in plasmas. *Physica*, XX1 :949–963, 1955.

[Vek67a] I.N. Vekua. *New methods for solving elliptic equations. Translated from the Russian by D. E. Brown.* . John Wiley & Sons, Inc., translation edited by a. b. tayler. north-holland edition, 1967.

[Vek67b] N.P. Vekua. *Systems of singular integral equations.* Gordon and Breach science publisher, 1967.

[Vog91] V. Vogelsang. On the strong unique continuation principle for inequalities of Maxwell type. *Math. Ann.*, 289(2) :285–295, 1991.

[Vol10] V. Volterra. *Leçons sur les équations intégrales et les équations intégro-différentielles. Leçons professées à la Faculté des sciences de Rome en 1910.* Les Grands Classiques Gauthier-Villars., 1910.

[Wed91] R. Weder. *Spectral and Scattering Theory for Wave Propagation in Perturbed Stratified Media*, volume 87. Springer Verlag, New York, 1991.

[Wed08a] R. Weder. Rigorous Analysis of High-Order Electromagnetic Invisibility Cloaks. *J.o Phys. A : Mathematical and Theoretical*, 41, 2008.

[Wed08b] R. Weder. The Boundary Conditions for Point Transformed Electromagnetic Invisibility Cloaks. *J. Phys A : Mathematical and Theoretical*, 41, 2008.

[WJIG11] M. Wolff, S. Jaouen, and L.-M. Imbert-Gérard. Conservative numerical methods for a two-temperature resistive MHD model with self-generated magnetic field term. *ESAIM Proc*, CEMRACS'10 research achievements : numerical modeling of fusion(32), 2011.

[WTTF12] D. Wang, R. Tezaur, J. Toivanen, and C. Farhat. Overview of the discontinuous enrichment method, the ultra-weak variational formulation, and the partition of unity method for acoustic scattering in medium frequency regime and performance comparisons. *Int. J. Numer. Meth. Engng.*, 89(4) :403–417, 2012.

Lise-Marie Imbert-Gérard

Analyse mathématique et numérique de problèmes d'ondes apparaissant dans les plasmas magnétiques

Résumé

Cette thèse étudie les aspects mathématiques et numériques de phénomènes d'ondes dans les plasmas magnétiques. La réflectométrie, une technique de sonde des plasmas de fusion, est modélisée par les équations de Maxwell. Le tenseur de permittivité présente dans ce model des valeurs propres ainsi que des termes diagonaux qui s'annulent. La relation de dispersion met en évidence deux phénomènes cruciaux : coupures et résonances, lorsque le nombre d'onde s'annule ou tend vers l'infini.

La partie I rassemble les résultats numériques. La grande nouveauté réside dans la définition d'une solution résonante. En effet, à cause des coefficients s'annulant continuement en changeant de signe, la solution peut être singulière, i.e. avoir une composante non intégrable. Cependant, grâce au principe d'absorption limite, une solution résonante est explicitement définie comme la limite de solutions intégrables du problème régularisé. L'expression théorique de la singularité est validée par des tests numériques du passage à la limite.

La partie II concerne l'approximation numérique. Elle comprend la mise en place d'une nouvelle méthode numérique adaptée aux coefficients réguliers. Celle-ci est basée sur la Formulation Variationnelle Ultra Faible mais nécessite des fonctions de base spécifiques, construites comme approximations locales du problème adjoint. L'analyse de convergence est effectuée en dimesion un, en dimension deux la construction des fonctions de base et leur propritété d'interpolation sont détaillées. La méthode d'ordre élevé obtenue permet de simuler le phénomène de coupure tandis que simuler le phénomène de résonance en dimension deux reste un défi.

Mots-clefs équations de Maxwell ; équations intégrales singulières ; principe d'absorption limite ; simulations numériques ; méthode numérique d'ordre élevé ; coefficients réguliers.

Résumé en anglais

Mathematical and numerical analysis of wave problems for magnetic plasmas

This dissertation investigates mathematical and numerical aspects of some wave phenomena appearing in magnetic plasmas. In order to model a probing technique for fusion plasmas, called reflectometry, a particular form of Maxwell's equations is studied. In the model, the dielectric tensor presents vanishing eigenvalues and diagonal terms. The study of the dispersion relation evidences two kinds of phenomena: cut-offs and resonances if the wave number goes either to zero or to infinity.

Part I of the thesis gathers the theoretical results. The main novelty consists in the definition of a resonant solution. Indeed, because of a smooth vanishing sign-changing coefficient, the solution may be singular: one of its components may be non-integrable. However, using a limit absorption principle, a resonant solution is explicitly obtained by studying the integrable solutions of the regularized system plus a limiting process. The theoretical expression of the singularity is validated by numerical tests concerning the regularized system as the regularizing term goes to zero.

Part II focuses on the numerical results. It includes the design of a new numerical method adapted to smooth coefficients. The method is based on the Ultra Weak Variational Formulation but requires specific shape functions, designed as local approximations of the adjoint equation. The convergence analysis of the method is performed in one dimension, for two dimensions the design procedure and the interpolation property of the shapes functions are detailed. The resulting high order method numerically tackles the approximation of cut-offs while the approximation of resonant solutions is still very challenging.

Keywords Maxwell's equations ; singular integral equations ; limiting absorption principle ; numerical simulations ; high order numerical method ; smooth varying coefficients.