Méthode itérative de Trefftz pour la simulation d'ondes électromagnétiques en trois dimensions.

Margot Sirdey

ACADÉMIE DE BORDEAUX |__|__|__|__|__|__|__|__|__|__|





Thèse de Doctorat

Présentée à l'UNIVERSITÉ DE PAU ET DES PAYS DE L'ADOUR

École Doctorale des Sciences Exactes et leurs Applications - ED SEA 211

par

Margot Sirdey

Équipe-Projet MAKUTU, INRIA Bordeaux Sud-Ouest, ONERA Toulouse, Université de Pau et des Pays de l'Adour, E2S-UPPA, Laboratoire de Mathématiques et de leurs Applications UMR CNRS 5142,

pour obtenir le titre de

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

en

Mathématiques

MÉTHODE ITÉRATIVE DE TREFFTZ POUR LA SIMULATION D'ONDES ÉLECTROMAGNÉTIQUES EN TROIS DIMENSIONS

soutenue le lundi 19 décembre 2022

après avis de

Bruno DESPRÉS	Professeur des universités la Sorbonne	Rapporteur
Stéphane LANTERI	Directeur de recherche INRIA ATLANTIS	Rapporteur

devant le jury composé de

Hélène BARUCQ	Directeur de recherche INRIA MAKUTU	Examinatrice
Bruno DESPRÉS	Professeur des universités la Sorbonne	Rapporteur
Luc GIRAUD	Directeur de recherche INRIA HiePACS	Examinateur
Lise-Marie IMBERT-GÉRARD	Associate professor Université Arizona	Examinatrice
Stéphane LANTERI	Directeur de recherche INRIA ATLANTIS	Rapporteur
Sébastien PERNET	Directeur de recherche ONERA	Co-directeur
Sébastien TORDEUX	Maître de Conférences Université de Pau	Directeur

Remerciements

J'adresse de chaleureux remerciements à mes encadrants de thèse : Sébastien Pernet et Sébastien Tordeux, inconditionnellement présents, encourageants et investis dans ce travail de thèse. Grâce à eux, j'ai grandi scientifiquement et humainement.

Je remercie Sébastien Pernet pour les nombreux échanges passionnants et variés que nous avons eus ensemble, en visio ou en présentiel. Son expertise dans les méthodes de GD, de GMRES et beaucoup d'autres, m'a été d'une grande aide. Évidemment, merci pour les incomptables heures de débug où Spiro donnait l'envie de s'arracher les cheveux !

Je remercie Sébastien Tordeux pour sa disponibilité et les multiples chat Teams - confinement ou non, week-end ou non - qui ont permis à la thèse d'avancer à une vitesse grand V. Merci d'avoir toléré mes créneaux footings et merci pour les nombreuses heures de tests où le travail en équipe - en meute - payait finalement : les résultats analytiques pour l'un et l'implémentation pour l'autre.

Un grand MERCI à vous deux. Vos apports scientifiques, humains et votre soutien ont permis d'atteindre le Milliard, où GoTEM3 était en fait déjà devenu ToTEM3. Les Sébastien, à quand PoTEM3 ?

Je souhaite remercier ensuite l'ensemble des membres du jury : Luc Giraud (président), pour ses remarques pragmatiques sur l'utilisation du GMRES, Bruno Després et Stéphane Lanteri (rapporteurs), pour leur expertise dans le domaine de la simulation numérique qui les a amenés à pointer judicieusement les atouts et les limites de GoTEM3, Hélène Barucq et Lise-Marie Imbert-Gérard (examinateurs) pour leurs perspectives encourageantes et pour les discussions Trefftz/Quasi-Trefftz que nous avons eues.

Mention particulière pour Hélène, chef d'équipe Makutu, pour les bons moments passés ensemble allant des réunions de travail à ECCOMAS ou encore à la formation de sauvetage en mer qui fût riche en émotions.

J'en profite pour remercier l'intégralité de l'équipe Makutu - et leurs collaborateurs - pour leur accueil et leur bonne humeur, les anciens et les nouveaux : Algiane, Arjeta, Augustin E, Augustin L, Aurélien, Chengyi, Ibrahima, Julien B, Julien D, Juliette, Justine, Lola, Marc F, Matthias, Nathan, Nicolas, Florian, Ha, Henri, Pierre, Rose, Stefano, Vinduja, et merci enfin à Sylvie pour sa patience avec nos ordres de mission !

Merci à mes parents, à ma soeur, Ninon, et à Max, qui m'ont apporté leur soutien tout au long de la thèse. Je remercie mes parents de m'avoir accueillie - et supportée - pendant ces longues semaines de confinement, où j'ai pris plaisir à travailler à leurs côtés. Je remercie Ninon pour ses précieux conseils vis à vis de la thèse ou de mes candidatures.

Bien qu'arrivé au milieu de ma thèse, je souhaite ensuite remercier Flavien. Toujours patient, attentionné et de bons conseils, il a su m'apporter du réconfort durant la dernière année de thèse. En particulier présent lors des moments les plus difficiles, il a réussi à me supporter. Chapeau ! Merci aussi pour les nombreux moments de loisirs partagés ensemble : footings, sorties rando/ski, restaurants, vacances, et qui ne sont pas prêts de s'arrêter.

Naturellement, je remercie Lucie et Pauline. La première, amie de longue date, toujours lucide et raisonnée, ses conseils m'ont permis de prendre du recul sur mon travail de thèse et sur mon quotidien. Je la remercie aussi vivement pour la première année Toulousaine, agrémentée de soirées houblonnées et de footings, et pour les deux dernières années où nous sommes restées quasi-quotidiennement en contact. La seconde, avec ses talents d'actrice et de mixeuse officielle de soupe, a été ma colocataire durant ces trois années à Toulouse. Merci pour nos multiples soirées (tisanes ou non), pour nos moments privilégiés affalées dans nos canapés, et surtout pour tous nos fous rire, qui faisaient le plus grand bien après une dure journée de travail. Après un déménagement intense, notre fine équipe est séparée, géographiquement oui, mais à jamais dans nos coeurs (ça c'est le moment festival de Cannes).

Enfin, mais pas des moindres, j'aimerais remercier tous les copains. De près ou de loin, ils ont tous contribué au bon avancement de ma thèse.

Merci aux Toulousains : Nadir qui m'a gentiment intégrée à l'ONERA. Une fois sa thèse et les confinements terminés, il a été à mes côtés dans mon quotidien Toulousain pour les marchés du dimanche matin ou pour les tapas en terrasse. - Gaëlle et Seb, rencontrés plus tardivement, toujours partants pour discuter durant de longues heures autour d'un verre. - Lucien, pour les (rares) fois où nous nous sommes croisés à Toulouse, pour lesquelles se fût toujours un plaisir. - Paul, pour ses nombreuses activités toutes plus originales les unes que les autres. - Adrien et Alexis, malgré nos chemins séparés en troisième année de thèse, pour les quelques apéros et tours de vélo.

Merci à la Team INSA et leurs $+1$: Tim, pour ses messages de soutien et les nouvelles régulières. - Val, toujours de bonne humeur et partant pour faire des footings (ou le fou). - Etienne, pour ses séjours Toulousains et sa bienveillance. - Quentin, pour nos moments dans la ville rose, qu'ils aient été sportifs ou festifs. - Corentin, pour nos vacances randos ou ski. - Maxime, Thibaut, Thomas, toujours prêts à accueillir "à la Nantaise". - Camille, Hugo, Romane, Thomas, Tom, les Rouennais (plus pour longtemps) : toujours fidèles au poste ! - Adriane et Pierre, pour votre couple de choc. Adriane pour son caractère bien trempé et Pierre pour ses blagues bien placées. - Juliette et Marine, pour nos amitiés malgré la distance.

MERCI

TABLE DES MATIÈRES

Introduction générale

La propagation des ondes électromagnétiques est un phénomène essentiel en physique. La très grande diversité des applications civiles ou militaires (télécommunication, imagerie, détection, furtivité, guerre électronique ...) exige des outils de simulation numérique utiles à la prédiction, à la conception ou à la maintenance. Dans cette thèse, nous nous focalisons sur des problèmes de diffraction d'ondes qui sont d'une part posés sur des domaines de plusieurs centaines de longueurs d'ondes, et qui d'autre part nécessitent le développement de schémas assurant une précision équivalente aux méthodes classiques dites moyennes fréquences.

Limites des solveurs classiques :

Plus précisément, de nombreuses méthodes numériques existent pour simuler les ondes électromagnétiques (Éléments Finis (EF) [51, 66, 78] , Différences Finies (DF) [96], Éléments Finis de Frontière (BEM) [16, 83, 93, 100], ...). Cependant, ces méthodes rencontrent deux problèmes importants.

1. Le phénomène de dispersion numérique apparaît dans le contexte de milieux de propagation non dissipatifs et de très grande taille devant la longueur d'onde [2, 3, 62, 63]. Le nombre de points de discrétisation par longueur d'onde doit alors être important pour espérer obtenir une solution numérique précise du problème de Maxwell, *ie* retranscrivant correctement les phases de l'onde. Une conséquence directe est une augmentation de la taille des systèmes linéaires à résoudre.
2. Le coût mémoire de la méthode est directement lié à la résolution du système linéaire de grande taille. De plus, ce coût augmente beaucoup plus vite que le nombre de degrés de liberté. En particulier, dans le cadre d'une factorisation LU 3D, il est actuellement impossible d'envisager ce type d'approche pour la simulation de grandes scènes de calcul, même sur des architectures HPC.

Les solveurs directs modernes et leurs limites :

Des méthodes numériques existent pour lutter contre la dispersion numérique et le coût mémoire.

Les méthodes d'EF d'ordre élevé [27, 82], les méthodes de Galerkin Discontinues (GD) basées sur des fonctions de base polynomiales [26, 36, 38, 43, 50, 59, 19], les méthodes d'Éléments Finis de Frontière (BEM) [16, 83, 93, 100] et les méthodes d'équations intégrales [14] limitent l'augmentation du nombre de points de discrétisation.

Des méthodes de condensation telles que des méthodes de GD Hybride (HDG en anglais) [24, 79, 84] et de Trefftz [10, 22, 76, 54, 55, 67, 88, 47, 95] ont été introduites afin d'améliorer les approches précédentes en réduisant drastiquement cette fois-ci la mémoire nécessaire au stockage de la factorisation LU tout en garantissant une précision équivalente.

Néanmoins, il nous semble que les gains importants induits par ces techniques d'optimisation permettent seulement de traiter des domaines de taille intermédiaire.

Les solveurs itératifs modernes et leurs limites :

Une alternative aux méthodes directes est de mettre en place des approches itératives pour limiter le coût mémoire. Nous pouvons approximativement les classer selon deux catégories dans le cadre des problèmes d'ondes.

La première, dite algébrique, met principalement en oeuvre des méthodes de Krylov (Gradient Conjugué, GMRES, multi-grilles algébriques, ...). À notre connaissance, ces méthodes ont des très bonnes performances pour les formulations coercives. Il est toutefois communément admis que les méthodes itératives algébriques appliquées à des systèmes pseudo-elliptiques amenant à des matrices dites indéfinies convergent lentement [41]. De plus, la construction d'un préconditionneur efficace reste une question ouverte et est encore l'objet de nombreux travaux [25, 90, 97].

La seconde catégorie, dite géométrique, repose sur des partitionnements du domaine de calcul et correspond aux Méthodes de Décomposition de Domaine (DDM) itératives [28, 29, 37, 39, 75, 99]. Elles bénéficient d'un cadre théorique très solide qui assure leur convergence. Cependant, ce formalisme classique est souvent difficile à mettre en œuvre sur des supercalculateurs.

Une autre approche est celle initiée par Cessenat-Després dans un cadre de milieux homogènes [21]. Il s'agit d'une Formulation Variationnelle Ultra-Faible (UWVF en anglais) [17, 22, 35, 60, 72, 98] dont l'inconnue est de même nature (traces entrantes et sortantes) que les conditions de transmission utilisées dans les DDM. Cette formulation étant contractante par construction, des solveurs itératifs de type "Jacobi" (relaxés ou non) sont naturellement envisageables [21]. Néanmoins, en pratique nous notons que ce solveur souffre de problèmes de convergence liés aux erreurs d'arrondis lorsque la taille du domaine devient importante.

Notre alternative, une méthode de Trefftz variationnelle itérative :

Nous pensons que les méthodes de Trefftz possèdent toutes les caractéristiques pour initier la construction d'un solveur capable de simuler des grandes scènes de calcul hétérogènes. En effet, d'une part, leur construction est basée sur des fonctions de base adaptées à la physique (solutions locales) permettant de diminuer drastiquement le phénomène de pollution

numérique et ainsi de réduire la taille du problème à résoudre. D'autre part, elles peuvent contrer le caractère indéfini inhérent aux systèmes linéaires issus de la discrétisation des équations d'ondes en proposant des formulations coercives.

Cependant, cette méthode n'est pas la seule alternative pour le traitement numérique des grands domaines ou des hautes fréquences. De manière non exhaustive, nous pouvons citer les formulations intégrales de type BEM ou de collocation, lorsqu'elles sont accélérées par une méthode de Fast Multipole [33, 34] ou par une Adaptive Cross Approximation [11, 68]. Leur implémentation reste toutefois très complexe à mettre en œuvre dans le cas des milieux hétérogènes. D'autre part, les méthodes de lancer de rayons issues de l'optique géométrique permettent de simuler des ondes électromagnétiques sur des domaines de plusieurs centaines ou milliers de longueurs d'ondes [7, 8]. Néanmoins, cette méthode exige encore des développements afin d'espérer obtenir des solutions précises au voisinage des matériaux diffractants.

Pour développer un solveur de Trefftz, plusieurs verrous doivent successivement être levés afin de passer à l'échelle en termes de capacité de simulation. Toutes les méthodes de Trefftz ne sont pas adaptées aux solveurs itératifs. La première étape consiste à réaliser un rapprochement entre les différents concepts de construction (Trefftz/UWVF) afin de dériver les formulations mieux adaptées aux enjeux de la simulation des problèmes électromagnétiques tridimensionnels en milieu hétérogène.

La seconde difficulté concerne les erreurs d'arrondis. Généralement, des ondes planes sont utilisées comme fonctions de base pour dériver les méthodes de Trefftz [46, 48, 49, 54, 77, 88]. En pratique, la mise en place d'une base d'ondes planes est facilement réalisable. Mais ces fonctions n'approchent pas avec précision les effets de pointe ou encore les ondes piégées dans des géométries complexes. De plus, ce choix de base de discrétisation induit de nombreux problèmes de conditionnement, provoquant à leur tour de nombreuses erreurs d'arrondis car les ondes planes sont linéairement dépendantes numériquement [9, 30, 49, 52, 72, 77, 87]. Cela peut conduire à une solution numérique imprécise, au sens où elle n'approche pas correctement la solution exacte du problème. Afin d'éviter cela, ces auteurs ont développé des stratégies, telles que l'emploi d'ondes planes évanescentes ou l'orthogonalisation de la base d'ondes planes donnant une solution numérique plus précise. Une autre voie consiste à ne pas se restreindre aux solutions analytiques et à considérer des solutions locales approchées des équations de Maxwell [64, 65].

Pour conclure, cette thèse ne s'intéresse pas à la totalité des méthodes de Trefftz. Le concept de méthode de Trefftz, introduit par le mathématicien Allemand Erich Trefftz, propose une méthode numérique reposant sur des solutions locales des équations aux dérivées partielles étudiées [67]. Nous nous sommes limités aux méthodes de Trefftz variationnelles alors qu'il est aussi possible d'écrire des méthodes de collocation. Dans notre cas, elles sont

généralement considérées comme des méthodes de GD, aussi appelées Trefftz-GD. Elles sont en effet dotées d'un cadre théorique rigoureux qui permet d'obtenir des algorithmes numériques performants et convergents à coup sûr.

L'objectif de la thèse est de **développer une méthode de type Trefftz itérative pour la simulation d'ondes électromagnétiques sur de grands domaines de calcul en trois dimensions**.

Le solveur de Trefftz itératif développé durant la thèse repose sur un code Fortran. Il a été créé de toutes pièces, hormis les librairies de résolution telles que :

— MUMPS© [4], librairie de résolution de systèmes linéaires,

— l'algorithme de GMRES du CERFACS© [44], développé par Luc Giraud, Serge Gratton, Valérie Fraysse et Julien Langou, librairie performante de résolution itérative de systèmes linéaires.

De plus, notre code s'appelle **GoTEM3**, où cet acronyme a été choisi pour

— **Go**, comme nous développons un solveur à faible coût mémoire : **G**iga-**o**ctet,

— **T**, comme nous développons une méthode de type **T**refftz,

— **EM**, comme nous simulons des ondes **É**lectro**M**agnétiques,

— **3**, comme nous étudions un problème défini sur un domaine en **3** dimensions.

Nous répondons à l'objectif de la thèse en passant par cinq étapes, où chacune est associée à un chapitre de ce manuscrit.

Dans un premier temps, nous étudions deux méthodes numériques représentatives des caractéristiques des méthodes classiques existantes pour simuler les ondes électromagnétiques. En particulier, nous choisissons d'expliquer une méthode d'EF de Nédélec d'ordre élevé et une méthode de GD. Nous suivrons des étapes similaires pour leurs présentations respectives, à savoir : la dérivation de la formulation variationnelle, puis sa discrétisation pour enfin obtenir le système matriciel associé. Pour conclure ce chapitre, nous mettrons en évidence les limites de ce type de solveurs, au sens où ils consomment trop de mémoire pour espérer les utiliser pour traiter des grandes scènes de calcul.

Dans le second chapitre, nous proposons une alternative aux méthodes d'EF de Nédélec ou de GD du Chapitre 1 : une méthode de type Trefftz. Nous construirons des formes générales de formulations variationnelles Trefftz pour les cas des milieux hétérogènes. Des critères sur les paramètres de pénalisation mèneront à la coercivité des formulations et ainsi au caractère bien posé du problème variationnel de Trefftz. Une des particularités des formulations Trefftz est qu'elles sont définies sur la frontière des éléments du maillage. En effet,

elles font intervenir la composante tangentielle du champ électrique et la trace tangentielle du champ magnétique. Nous introduirons deux types de traces numériques dans ce chapitre : celles obtenues par un solveur de Riemann dans le cas d'un milieu homogène, et celles de type upwind dans le cas hétérogène. Ces deux points de vue s'avèrent être équivalents entre eux, et mènent à des cas particuliers des formulations variationnelles Trefftz bien posées. Nous résoudrons finalement le système matriciel associé grâce à une factorisation LU. Le coût mémoire de la méthode se révèle être moins important que celui des méthodes classiques, et la méthode de Trefftz directe est alors une bonne alternative. En effet, nous doublons la taille des domaines qu'il est possible de traiter. Toutefois, le stockage de la factorisation LU est inévitable et est un frein pour considérer de très grands domaines de calcul.

Pour remédier à cela, nous formulons dans le Chapitre 3 une méthode de Trefftz itérative, basée sur des algorithmes de Krylov pour la résolution du système linéaire. Nous étudions en particulier l'UWVF obtenue à partir des traces numériques de Cessenat-Després, que nous présentons pour le cas des milieux hétérogènes. Nous verrons que ces traces numériques sont équivalentes aux traces numériques introduites dans le Chapitre 2. Ainsi, nous obtenons des formulations de Riemann, upwind ou de Cessenat-Després équivalentes, au sens où la solution d'une est la solution d'une autre à une bijection près. Deux méthodes reposant sur l'emploi de bases de Krylov sont exploitées : la méthode de GMRES et la méthode que nous nommons Krylov Galerkin (KG). En pratique, elles utilisent une stratégie de *restart* que nous expliquerons. En considérant un nombre réduit de vecteurs dans l'espace de Krylov, cette technique permet de considérablement réduire le coût mémoire de la méthode et répond ainsi à l'objectif principal de notre démarche scientifique. Ensuite, nous introduirons une méthode de KG qui s'inspire de la méthode de GMRES comme elle est aussi basée sur un espace de Krylov. La construction de ce dernier sera explicitée et son intérêt mis en avant. Cette méthode de KG utilise la propriété de coercivité des formulations Trefftz, au sens où le spectre de la matrice se situe dans le demi-plan complexe, et la solution numérique est alors supposée converger rapidement vers la solution exacte. Néanmoins, nous proposerons d'accélérer la convergence en mettant en jeu le préconditionneur de Cessenat-Després. Nous verrons qu'il rend la matrice contractante, de telle sorte que son spectre est contenu dans le cercle unité quelle que soit la taille du domaine considéré. Cela témoigne de l'amélioration du conditionnement du système. Enfin, nous présenterons des résultats numériques dans le cas particulier d'un algorithme de GMRES. Nous mettrons en valeur les faibles coûts mémoire de la méthode de GMRES UWVF face aux méthodes directes des Chapitres 1 et 2. De plus, nous analyserons la convergence de la solution, accélérée dans le cas d'un système préconditionné.

Ensuite, nous proposons trois stratégies qui scinderont le quatrième chapitre. Premièrement, nous élaborerons une technique de désassemblage de la matrice Trefftz. Cette dernière

est fondée sur le caractère cartésien du maillage que nous utiliserons. Elle apportera des gains considérables en mémoire, menant à la considération de plus d'un milliard de degrés de liberté dans le domaine de calcul. Elle diminuera aussi les temps d'exécution grâce à son adaptation HPC (utilisation d'OpenMP). Deuxièmement, nous développerons une réduction de la base d'ondes planes. Cette stratégie a pour but d'améliorer le conditionnement du problème tout en diminuant le coût mémoire de sa résolution. En se basant sur la construction d'une base réduite, nous éliminerons les fonctions de base donnant peu d'information à la description de la solution numérique. Toutefois, nous verrons par des études numériques qu'il est important de ne pas trop réduire la base Trefftz au risque d'obtenir une solution numérique erronée. Troisièmement, nous introduirons un préconditionneur global, au sens où il implique les trois directions de propagation du domaine. Nous expliquerons sa construction et son implémentation informatique, et nous mettrons en avant son efficacité face à un préconditionneur de Cessenat-Després.

Dans le dernier chapitre [1], nous souhaitons remédier aux problèmes de conditionnements liés aux ondes planes et ainsi aux phénomènes de dispersion numérique. Une alternative est par exemple de développer une méthode Quasi-Trefftz. Le mot-clé "quasi" correspond à l'emploi de solutions numériques approchées comme fonctions de base Trefftz. Nous expliquerons notre approche dans un cas en deux dimensions qui exploite déjà les capacités d'une telle méthode. En premier, nous passons en revue les équations de Maxwell, le problème du second ordre associé et les hypothèses nécessaires à la dérivation de la méthode. Ensuite, nous élaborerons la formulation variationnelle Trefftz donnée sur un maillage dont les éléments peuvent prendre différentes configurations géométriques (pas seulement des triangles et des quadrangles). Puis, nous donnerons des approximations à la fois de l'espace de Trefftz et des fonctions de base. Plus précisément, il s'agira d'utiliser une méthode d'EF de Nédélec d'ordre élevé (similaire au Chapitre 1) pour résoudre localement les équations de Maxwell. Finalement, le problème de dispersion numérique est atténué grâce au phénomène de super-convergence observé pour des cas numériques dans lesquels le domaine est géométriquement complexe (frontières ou obstacles). Ce dernier chapitre ouvre des perspectives d'évolution évidentes pour notre solveur.

1. Il s'agit d'une transcription (en anglais) de l'article [45] qui a été rédigé durant la thèse.

Lexique mathématique

Voici quelques notations mathématiques (liste non exhaustive) utiles à la compréhension du manuscrit. Ce lexique permet de recenser les plus importantes, mais certaines d'entre elles sont définies plus en détail dans le document.

Notations :

— $\mathcal{T}$ est l'ensemble des éléments du maillage,

— $\mathcal{F}$ est l'ensemble des faces du maillage,

— $\mathcal{F}_{\mathrm{int}}$ est l'ensemble des faces intérieures du maillage,

— $\mathcal{F}_{\mathrm{ext}}$ est l'ensemble des faces extérieures du maillage,

— $\mathcal{D}_\Omega$ est le côté du domaine en longueur d'onde (λ),

— La composante tangentielle sur la surface ∂T d'un élément $T \in \mathcal{T}$ est notée $\gamma_t \mathbf{E}^T := \left(\mathbf{n}_T \times \mathbf{E}^T\right) \times \mathbf{n}_T$,

— La trace tangentielle (ou trace "tournée") sur la surface ∂T d'un élément $T \in \mathcal{T}$ est notée $\gamma_\times^T \mathbf{H}^T := \mathbf{n}_T \times \mathbf{H}^T$,

— La notation $\#\mathrm{ddl}$ est le nombre de degrés de liberté global

— La notation $\#\mathrm{ddlelem}$ est le nombre de degrés de liberté par élément,

— $R_{\partial\Omega} := (1 - Z_{\partial\Omega})/(1 + Z_{\partial\Omega})$ est le coefficient de réflexion sur $\partial\Omega$,

— $Z_{\partial\Omega} := (1 - R_{\partial\Omega})/(1 + R_{\partial\Omega})$ est l'impédance relative à la frontière $\partial\Omega$,

— Sur un élément $T \in \mathcal{T}$, la trace entrante d'une onde électromagnétique est définie par $\gamma_{\mathrm{in}}^T \mathbb{E}^T := \gamma_t \mathbf{E}^T + Z_T \gamma_\times^T \mathbf{H}^T$,

— Sur un élément $T \in \mathcal{T}$, la trace sortante d'une onde électromagnétique est définie par $\gamma_{\mathrm{out}}^T \mathbb{E}^T := \gamma_t \mathbf{E}^T - Z_T \gamma_\times^T \mathbf{H}^T$,

— La trace numérique électrique est notée $\left(\widehat{\gamma_t \mathbf{E}}\right)_{|F}$,

— La trace numérique magnétique est notée $\left(\widehat{\gamma_t \mathbf{H}}\right)_{|F}$,

— L'opérateur bijectif $\mathcal{S}_{\mathrm{out}}^T : \mathrm{L}_t^2(\partial T) \to \mathbb{X}_T, \quad \mathbf{x}^T \mapsto \mathbb{E}^T$ associe une trace sortante $\mathbf{x}^T$ à un champ volumique $\mathbb{E}^T$,

— $\mathrm{MEM}^{\mathrm{GMRES}}$: coût mémoire du solveur GMRES,

— $\mathrm{MEM}_{\mathrm{prec}}^{\mathrm{GMRES}}$: coût mémoire du solveur GMRES préconditionné,

— ε est le seuil de troncature pour réduire la base d'ondes planes.

Accronymes :

— MUMPS : MUltifrontal Massively Parallel Solver,

— GMRES : General Minimal RESidual,

— EF : Éléments Finis,

— GD : Galerkin Discontinu,

— HDG : Galerkin Discontinu Hybride (ou Hybrid Discontinuous Galerkin en anglais),

— KG : Krylov Galerkin,

— UWVF : Formulation Variationnelle Ultra-Faible (ou Ultra-Weak Variational Formulation en anglais),

— FV : Formulation Variationnelle,

— CD : Cessenat-Després,

— HPC : High Performance Computing,

— PMV : Produit Matrice-Vecteur,

— To : TeraOctet, Go : GigaOctet.

Cadre de développement des solveurs numériques basés sur des formulations faibles

Différentes mises en œuvre interviennent dans la construction des méthodes numériques. En effet, des méthodes de type Différences Finies (DF) ou encore de type Équations Intégrales par exemple, n'utilisent pas les mêmes stratégies que des méthodes d'Élements Finis (EF) ou de Galerkin Discontinues (GD). Dans le cadre de cette thèse, uniquement des techniques d'EF ou de GD (de GD d'ordre élevé et de Trefftz) sont présentées. Leurs développements reposent sur les quatre points suivants.

Modélisation du problème par un système d'équations aux dérivées partielles bien posé

Dans cette thèse, nous étudions un problème de Maxwell adimensionné afin de simplifier les calculs. Pour construire ce problème, nous rappelons dans un premier temps les formules générales de Maxwell en absence de charges et de courants pour un milieu linéaire et isotropique

$$\begin{cases} \nabla \cdot \mathbf{d} = 0, & \nabla \times \mathbf{e} = -\dfrac{\partial \mathbf{b}}{\partial t}, & \mathbf{d} = \varepsilon_0 \varepsilon_r \mathbf{e}, \\ \nabla \cdot \mathbf{b} = 0, & \nabla \times \mathbf{h} = \dfrac{\partial \mathbf{d}}{\partial t}, & \mathbf{b} = \mu_0 \mu_r \mathbf{h}, \end{cases}$$

où ε_0 (*resp.* ε_r) et μ_0 (*resp.* μ_r) sont la permittivité (*resp.* permittivité relative) et la perméabilité (*resp.* perméabilité relative) du vide (*resp.* du milieu). De plus, $\mathbf{e}$, $\mathbf{h}$, $\mathbf{d}$ et $\mathbf{b}$ sont respectivement : le champ électrique, le champ magnétique, le déplacement électrique et l'induction magnétique. Nous nous plaçons en régime harmonique. Ils peuvent alors être représentés par quatre fonctions complexes normalisées $\mathbf{E}$, $\mathbf{D}$, $\mathbf{H}$ et $\mathbf{B}$, associées à leurs amplitudes normalisées

$$e_0, \quad d_0 = \varepsilon_0 e_0, \quad h_0 = \sqrt{\frac{\varepsilon_0}{\mu_0}}\, e_0 \quad \text{et} \quad b_0 = \sqrt{\varepsilon_0 \mu_0}\, e_0,$$

menant à

$$\begin{cases} \mathbf{e}(\mathbf{x},t) = e_0 \mathcal{R}\left(exp(i\omega t)\mathbf{E}(\mathbf{x})\right), & \mathbf{h}(\mathbf{x},t) = h_0 \mathcal{R}\left(exp(i\omega t)\mathbf{H}(\mathbf{x})\right), \\ \mathbf{d}(\mathbf{x},t) = d_0 \mathcal{R}\left(exp(i\omega t)\mathbf{D}(\mathbf{x})\right), & \mathbf{b}(\mathbf{x},t) = b_0 \mathcal{R}\left(exp(i\omega t)\mathbf{B}(\mathbf{x})\right), \end{cases}$$

où ω est la fréquence angulaire. Celle-ci est représentative de la dépendance fréquentielle du problème. Nous définissons le nombre d'onde $k_0 := \dfrac{\omega}{c_0}$, où $c_0 = (\varepsilon_0 \mu_0)^{-\frac{1}{2}}$ est la vitesse dans le vide. Nous obtenons le problème de Maxwell normalisé suivant, défini sur un domaine Lipschitz connexe borné $\Omega \subset \mathbb{R}^3$.

Problème 1 (Problème de Maxwell du premier ordre)**.** *Le problème de Maxwell hétérogène du premier ordre s'écrit*

$$\left\{ \begin{array}{rcll} \nabla \times \mathbf{H} & = & ik_0\, \varepsilon_r \mathbf{E}, & \text{sur } \Omega, \\ \nabla \times \mathbf{E} & = & -ik_0\, \mu_r \mathbf{H}, & \text{sur } \Omega, \end{array} \right.$$

muni de la condition de bord d'impédance

$$\gamma_t \mathbf{E} + Z_{\partial\Omega}\, \mathbf{n}_{\partial\Omega} \times \gamma_t \mathbf{H} = \mathbf{g} \quad \text{sur } \partial\Omega, \tag{1}$$

où les champs $\mathbf{E}$ *et* $\mathbf{H}$ *sont tous les deux dans l'espace* $H(\mathrm{rot}, \Omega)$ *défini par*

$$H(\mathrm{rot}, \Omega) := \left\{ \mathbf{u} : \Omega \to \mathbb{C}^3, \int_\Omega |\mathbf{u}|^2 \, \mathrm{d}\mathbf{x} < \infty, \int_\Omega |\nabla \times \mathbf{u}|^2 \, \mathrm{d}\mathbf{x} < \infty \right\},$$

et où nous posons différentes définitions et notations :

- $\mathbf{n}_{\partial\Omega} \in \mathbb{R}^3$ *est la normale sortante unitaire de* $\partial\Omega$*,*
- $Z_{\partial\Omega} : \partial\Omega \to \mathbb{R}^+$ *est une fonction constante par morceaux, avec une partie réelle strictement non positive ; elle sera même supposée réelle dans tout le manuscrit,*
- $\gamma_t \mathbf{E} = \mathbf{E} - (\mathbf{E} \cdot \mathbf{n}_{\partial\Omega})\, \mathbf{n}_{\partial\Omega} = (\mathbf{n}_{\partial\Omega} \times \mathbf{E}) \times \mathbf{n}_{\partial\Omega}$ *est la composante tangentielle du champ électrique,*
- $\gamma_t \mathbf{H} = \mathbf{H} - (\mathbf{H} \cdot \mathbf{n}_{\partial\Omega})\, \mathbf{n}_{\partial\Omega} = (\mathbf{n}_{\partial\Omega} \times \mathbf{H}) \times \mathbf{n}_{\partial\Omega}$ *est la composante tangentielle du champ magnétique,*
- ε_r *et* μ_r *sont des fonctions constantes par morceaux définies sur une partition de* Ω*, qui est*

  $$\overline{\Omega} = \bigcup_{p=1,\ldots,P} \overline{\Omega}_p,$$

  avec P *sous-domaines polyédriques non dégénérés notés* Ω_p*, et* $\Omega_p \cap \Omega_q = \emptyset$*, si* $p \neq q$*.*
- $\mathbf{g} : \partial\Omega \to \mathbb{C}^3$ *est un champ tangentiel appartenant à l'espace fonctionnel suivant*

  $$\mathrm{L}^2_t(\partial\Omega) := \left\{ \mathbf{u} \in \left(\mathrm{L}^2(\partial\Omega)\right)^3, \quad \mathbf{u} \cdot \mathbf{n}_{\partial\Omega} = 0 \right\}.$$

Le Problème 1 est bien posé lorsqu'il est complété par une condition de bord d'impédance imposée sur la frontière du domaine $\partial\Omega$, voir [78]. Ce problème est d'ordre 1. Il peut aussi s'écrire sous la forme d'un problème du second ordre, muni de la même condition de bord d'impédance.

Problème 2 (Problème de Maxwell du second ordre)**.** *Le problème de Maxwell hétérogène du*

second ordre s'écrit

$$\left\{\begin{array}{rcl} -k_0^2\varepsilon_r\mathbf{E} + \nabla\times(\mu_r^{-1}\nabla\times\mathbf{E}) & = & 0, \quad \textit{sur } \Omega, \\ \gamma_t\mathbf{E} - \dfrac{Z_{\partial\Omega}}{ik_0\mu_r}\,\mathbf{n}_{\partial\Omega}\times(\nabla\times\mathbf{E}) & = & \mathbf{g}, \quad \textit{sur } \partial\Omega. \end{array}\right. \tag{2}$$

Maillage du domaine en éléments géométriques

Pour mailler le domaine de calcul Ω, nous considérons un maillage en trois dimensions composé d'éléments polyédriques T ne se chevauchant pas. En effet, le maillage est choisi selon les partitions Ω_p, $1 \leq p \leq P$, de Ω, dans lesquelles ε_r et μ_r sont constants. Autrement dit, il existe un unique $1 \leq p_0 \leq P$ tel que $T \subset \Omega_{p_0}$. Les fonctions ε_r et μ_r sont alors constantes par morceaux. De plus, nous notons $\mathcal{T}$ l'ensemble des éléments T, et $\mathcal{F}$ l'ensemble des faces F de Ω. Nous comptons au total $\#\mathrm{elem} := \mathsf{card}(\mathcal{T})$ éléments. Nous définissons aussi les ensembles de faces suivants :

— l'ensemble $\mathcal{F}_{\mathrm{int}}$ des faces intérieures

$$\mathcal{F}_{\mathrm{int}} := \{\partial T \cap \partial K \ : \ T, K \in \mathcal{T} \text{ avec } T \neq K \text{ et } \textit{aire}(\partial T \cap \partial K) \neq 0\},$$

où l'aire d'une face I est notée $\textit{aire}(I)$ (égale à zéro pour les arêtes et pour les sommets),

— l'ensemble $\mathcal{F}_{\mathrm{ext}}$ des faces extérieures

$$\mathcal{F}_{\mathrm{ext}} := \{\partial T \cap \partial\Omega \ : \ T \in \mathcal{T} \text{ et } \textit{aire}(\partial T \cap \partial\Omega) \neq 0\},$$

— l'ensemble $\mathcal{F}_T$ des faces associées à un élément $T \in \mathcal{T}$

$$\mathcal{F}_T := \{F \in \mathcal{F}_{\mathrm{int}} \cup \mathcal{F}_{\mathrm{ext}} \ : \ \textit{aire}(F \cap \partial T) \neq 0\}.$$

Construction de la formulation variationnelle

La formule de Stokes est l'ingrédient majeur nécessaire pour construire la formulation variationnelle. Soit $\mathbf{u} \in H(\mathrm{rot}, \Omega)$ et $\mathbf{v} \in H(\mathrm{rot}, \Omega)$, elle prend la forme générale suivante

$$\int_\Omega \nabla\times\mathbf{u}\cdot\overline{\mathbf{v}} \ - \ \mathbf{u}\cdot\overline{\nabla\times\mathbf{v}} = \left\langle \mathbf{n}_{\partial\Omega}\times\gamma_t\mathbf{u}\,,\,\gamma_t\overline{\mathbf{v}}\right\rangle_{\partial\Omega}, \tag{3}$$

où $\left\langle .\,,\,.\right\rangle_{\partial\Omega}$ peut être interprété comme un produit de dualité entre $H_t^{-\frac{1}{2}}(\mathrm{rot}_{\partial\Omega}, \partial\Omega)$ et $H_t^{-\frac{1}{2}}(\mathrm{div}_{\partial\Omega}, \partial\Omega)$ [15]. L'opérateur $\mathbf{n}_{\partial\Omega}\times$ définit alors un isomorphisme entre $H_t^{-\frac{1}{2}}(\mathrm{rot}_{\partial\Omega}, \partial\Omega)$

et $H_t^{-\frac{1}{2}}(\mathrm{div}_{\partial\Omega}, \partial\Omega)$. De plus, les composantes tangentielles $\gamma_t\mathbf{E}$ et $\gamma_t\mathbf{H}$ définissent un opérateur continu [26]

$$\gamma_t : H(\mathrm{rot}, \Omega) \to H_t^{-\frac{1}{2}}(\mathrm{rot}_{\partial\Omega}, \partial\Omega).$$

Nous pouvons alors introduire l'opérateur de trace tangentielle [26]

$$\gamma_\times : H(\mathrm{rot}, \Omega) \to H_t^{-\frac{1}{2}}(\mathrm{div}_{\partial\Omega}, \partial\Omega).$$

La condition de bord d'impédance (1) s'écrit alors aussi

$$\gamma_t\mathbf{E} + Z_{\partial\Omega}\,\gamma_\times\mathbf{H} = \mathbf{g} \quad \text{sur } \partial\Omega.$$

La formule de Stokes est appliquée aux champs $\mathbf{E}$ et $\mathbf{H}$. Dans le contexte des méthodes d'EF ou de GD, ces fonctions sont supposées suffisamment régulières, de telle sorte que le produit de dualité devient une intégrale. Plus précisément, les traces des fonctions $\mathbf{E}$ et $\mathbf{H}$ doivent être dans $\mathrm{L}_t^2(\partial\Omega)$. Nous les choisissons par conséquent dans l'espace $H_{\mathrm{imp}}(\Omega)$, défini par

$$H_{\mathrm{imp}}(\Omega) \;:=\; \Big\{\mathbf{u} \in H(\mathrm{rot}, \Omega),\ \text{tel que } \gamma_t\mathbf{u} \in \mathrm{L}_t^2(\partial\Omega)\Big\}.$$

De plus, nous définissons $\mathbb{H}_{\mathrm{imp}}(\Omega)$ pour un couple $\mathbb{E} := (\mathbf{E}, \mathbf{H})$ solution du Problème 1

$$\mathbb{H}_{\mathrm{imp}}(\Omega) \;:=\; \Big\{\mathbb{E} = (\mathbf{E}, \mathbf{H}) \in H(\mathrm{rot}, \Omega) \times H(\mathrm{rot}, \Omega),\ \text{tel que } \gamma_t\mathbf{E} \in \mathrm{L}_t^2(\partial\Omega) \text{ et } \gamma_\times\mathbf{H} \in \mathrm{L}_t^2(\partial\Omega)\Big\}.$$

En prenant $\mathbb{E}$ dans $\mathbb{H}_{\mathrm{imp}}(\Omega)$ (*resp.* $\mathbf{E}$ dans $H_{\mathrm{imp}}(\Omega)$), une hypothèse forte est imposée sur les traces des champs électrique et magnétique. D'après [78], il existe alors une unique solution au Problème 1 (*resp.* 2) muni d'une condition de bord d'impédance. Ainsi, pour le Problème 1 d'ordre 1 et pour le Problème 2 d'ordre 2, nous poserons leurs problèmes variationnels respectifs.

Obtention de la solution numérique

Les solveurs déterminent les solutions numériques des Problèmes 1 ou 2 par différents moyens selon la méthode numérique utilisée. Pour les méthodes directes, nous inversons le système matriciel associé à la méthode numérique grâce à un solveur MUMPS©, voir [4]. Pour les méthodes itératives, voir [23, 40, 71], nous utilisons un algorithme de Jacobi, une méthode de GMRES [92] ou une méthode de Krylov Galerkin (que nous expliquons en Section 3.3).

Des résultats numériques appuieront les avantages et les inconvénients des méthodes développées. **Une remarque importante valable pour l'intégralité du manuscrit est que les distances sont données en longueur d'onde λ. Comme nous posons le nombre d'onde dans le vide $k_0 = 2\pi$, alors $\lambda = 1$ mètre est fixé.** Notre objectif est de simuler l'onde électromagnétique sur des domaines de taille $(200\lambda)^3 = 8 \cdot 10^6 \lambda^3$. Cela équivaut aussi naturellement à des domaines de taille $(400\frac{\lambda}{2})^3 = 64 \cdot 10^6 (\frac{\lambda}{2})^3$ ou encore $(800\frac{\lambda}{4})^3 = 512 \cdot 10^6 \ (\frac{\lambda}{4})^3$, où λ, $\frac{\lambda}{2}$ ou $\frac{\lambda}{4}$ sont les tailles h des éléments du maillage. Ainsi, résoudre un problème avec $\mathcal{D}_\Omega = 200\lambda$ revient à considérer un maillage avec 8 millions d'éléments dans le cas où $h = 1$.

CHAPITRE 1

PERFORMANCES DES MÉTHODES CLASSIQUES POUR LA SIMULATION DE GRANDES SCÈNES DE CALCUL

Sommaire

Nous rappelons premièrement l'objectif de cette thèse : simuler une onde électromagnétique sur de grandes scènes de calcul. L'augmentation de la taille du domaine étudié implique cependant de nombreuses difficultés : stabilité, précision, temps de calcul ou encore coût mémoire de la méthode numérique. Ce dernier point est un problème majeur pour

les méthodes classiques, et nous le mettons en avant dans le premier chapitre en étudiant des méthodes numériques qui nous semblent être représentatives de celles existantes :

— une méthode d'EF de Nédélec sur un maillage structuré,

— une méthode de GD sur un maillage non structuré.

Chacune d'entre elles constitue une partie du Chapitre 1. Tout d'abord, nous dérivons une formulation variationnelle quasi-elliptique basée sur le problème de Maxwell d'ordre 2 pour la méthode de Nédélec (Section 1.1.1) ; puis une seconde hyperbolique basée sur la problème de Maxwell d'ordre 1 pour la méthode de GD (Section 1.2.1). Ces points de départ très différents pour les formulations nous permettent de balayer un large spectre de méthodes. C'est en ce sens que nous jugeons étudier deux exemples de méthodes qui sont à l'image de celles que nous pouvons trouver dans la littérature. De plus, nous définissons les espaces EF (Sous-sections 1.1.2 et 1.2.2) sur lesquels nous discrétisons chaque formulation (Sous-sections 1.1.3 et 1.2.3) menant à leur problème matriciel respectif. Finalement, afin de vérifier le bon développement des deux solveurs, nous présentons des courbes de convergence pour chacun d'entre eux (Sous-sections 1.1.4 et 1.2.4).

La dernière section porte sur une étude du coût mémoire de ces méthodes, où nous pourrons confirmer qu'elles ne sont pas appropriées pour simuler des ondes électromagnétiques sur de grands domaines de calcul.

1.1 Éléments finis de type Nédélec

Nous débutons par la dérivation d'une méthode d'EF de Nédélec d'ordre élevé [81, 83]. Il existe plusieurs familles d'EF de Nédélec construites à partir de simplexes ou d'hexaèdres. En particulier, nous employons celle que nous jugeons capable d'offrir les meilleurs résultats dans notre contexte : la première famille des EF de Nédélec. Plus précisément, nous la considérons pour des cellules hexaédriques [82]. Dans cette section, nous nous restreignons au cas des cubes qui sont suffisants pour mettre en avant les limites des solveurs classiques. Dans un premier temps, nous explicitons les formes sesquilinéaire et antilinéaire de la formulation variationnelle. Puis, en choisissant des points d'interpolation adaptés, nous discrétisons la formulation. Une fois la convergence du schéma montrée numériquement, le coût mémoire du solveur de Nédélec est mis en évidence.

1.1.1 Construction de la formulation variationnelle

Le Problème 2 de Maxwell du second ordre est utilisé pour la méthode d'EF de Nédélec. Nous l'étudions dans le cas où Ω est associé à un domaine homogène. Autrement dit, nous imposons $\varepsilon_r = 1$ et $\mu_r = 1$.

Remarque 1.1. *Le Problème 2 du second ordre est lié uniquement à l'inconnue* $\mathbf{E}$*. Il permet de limiter le nombre total d'inconnues du problème matriciel final.*

Nous rappelons que ce problème est muni d'une condition de bord d'impédance, voir le système (2). Nous construisons la formulation variationnelle en prenant une fonction test $\mathbf{E}' \in H_{\text{imp}}(\text{rot}, \Omega)$. En multipliant et en intégrant sur Ω, nous avons

$$-k_0^2 \int_\Omega \mathbf{E} \cdot \overline{\mathbf{E}'} + \int_\Omega \nabla \times (\nabla \times \mathbf{E}) \cdot \overline{\mathbf{E}'} = 0.$$

Puis en utilisant la formule de Stokes (3), nous obtenons

$$-k_0^2 \int_\Omega \mathbf{E} \cdot \overline{\mathbf{E}'} + \int_\Omega \nabla \times \mathbf{E} \cdot \overline{\nabla \times \mathbf{E}'} + \int_{\partial\Omega} \mathbf{n}_{\partial\Omega} \times \nabla \times \mathbf{E} \cdot \overline{\gamma_t \mathbf{E}'} = 0.$$

Nous introduisons ensuite la condition de bord

$$-k_0^2 \int_\Omega \mathbf{E} \cdot \overline{\mathbf{E}'} + \int_\Omega \nabla \times \mathbf{E} \cdot \overline{\nabla \times \mathbf{E}'} + \int_{\partial\Omega} \frac{ik_0}{Z_{\partial\Omega}} (\gamma_t \mathbf{E} - \mathbf{g}) \cdot \overline{\gamma_t \mathbf{E}'} = 0.$$

Ainsi, le problème variationnel associé au Problème 2 est le suivant.

Problème 3. *[Problème variationnel du problème d'ordre 2] Trouver* $\mathbf{E} \in H_{\text{imp}}(\text{rot}, \Omega)$*, tel que pour tout* $\mathbf{E}' \in H_{\text{imp}}(\text{rot}, \Omega)$

$$a(\mathbf{E}, \mathbf{E}') = \ell(\mathbf{E}'),$$

avec

$$a(\mathbf{E}, \mathbf{E}') = -k_0^2 \int_\Omega \mathbf{E} \cdot \overline{\mathbf{E}'} + \int_\Omega \nabla \times \mathbf{E} \cdot \overline{\nabla \times \mathbf{E}'} + \int_{\partial\Omega} \frac{ik_0}{Z_{\partial\Omega}} \gamma_t \mathbf{E} \cdot \overline{\gamma_t \mathbf{E}'},$$

et

$$\ell(\mathbf{E}') = \int_\Omega \frac{ik_0}{Z_{\partial\Omega}} \mathbf{g} \cdot \overline{\gamma_t \mathbf{E}'}.$$

Ce problème admet une unique solution d'après [78].

1.1.2 Définition de l'espace Élément Fini de Nédélec

Nous choisissons de créer un espace Élément Fini de la première famille de Nédélec [81]. Nous rappelons sa forme générale. Pour cela, nous utilisons un élément fini de référence $\left(\widehat{T}, \widehat{\mathcal{N}}_r, \widehat{\Sigma}_r\right)$, où $r \in \mathbb{N}$ est l'ordre de la méthode. L'élément de référence $\widehat{T}$ est le cube unité défini par $\widehat{T} = [0,1]^3$. L'ensemble des fonctions de base polynomiales de référence sont dans l'espace $\widehat{\mathcal{N}}_r$ défini par

$$\widehat{\mathcal{N}}_r := \mathbb{Q}_{r,r+1,r+1}\left(\widehat{T}\right) \times \mathbb{Q}_{r+1,r,r+1}(\widehat{T}) \times \mathbb{Q}_{r+1,r+1,r}(\widehat{T}),$$

où $\mathbb{Q}_{n1,n2,n3}(\widehat{T}) := \mathbb{P}^{n1}([0,1]) \otimes \mathbb{P}^{n2}([0,1]) \otimes \mathbb{P}^{n3}([0,1])$.

L'ensemble des degrés de liberté de l'élément de référence est $\widehat{\Sigma}_r$. Pour un élément quelconque du maillage, nous notons le nombre total de degrés de liberté $\#\mathrm{ddlelem}$. Dans notre cas, pour définir les degrés de liberté et les fonctions de base associées, nous introduisons les points de Gauss $\left(\hat{x}_k^G\right)_{k=1,\cdots,N^G}$ et de Gauss-Lobatto $\left(\hat{x}_k^{GL}\right)_{k=1,\cdots,N^{GL}}$ associés au segment unité $[0,1]$ [82]. Ce choix induit un élément fini unisolvant et permet d'assurer aisément la conformité dans $H(\mathrm{rot},\Omega)$ nécessaire à cette méthode d'EF. Plus précisément, l'utilisation d'une telle discrétisation permet de raccorder facilement les traces (tangentielles dans notre contexte) entre les éléments du maillage. Ainsi, nous assurons la continuité de la solution numérique.

D'une part, les points de Gauss sont les zéros du polynôme de Legendre d'ordre N^G-1, voir les Tableaux 1.1, 1.2, 1.3 et 1.4. D'autre part, les points de Gauss-Lobatto sont les zéros de la dérivée du polynôme de Legendre d'ordre $r := N^{GL}-1$ auxquels nous ajoutons les points $\{0\}$ et $\{1\}$, voir les Tableaux 1.5, 1.6, 1.7 et 1.8. Leurs poids de quadrature sont respectivement $\left(\hat{\omega}_k^G\right)_{k=1,\cdots,N^G}$ et $\left(\hat{\omega}_k^{GL}\right)_{k=1,\cdots,N^{GL}}$.

k	1
$\hat{x}_k^G$	$\frac{1}{2}$
$\hat{\omega}_k^G$	1

TABLE 1.1 – Point et poids de Gauss sur le segment $[0,1]$ pour $N^G=1$.

k	1	2
$\hat{x}_k^G$	$\frac{1-\frac{1}{\sqrt{3}}}{2}$	$\frac{1+\frac{1}{\sqrt{3}}}{2}$
$\hat{\omega}_k^G$	$\frac{1}{2}$	$\frac{1}{2}$

TABLE 1.2 – Points et poids de Gauss sur le segment $[0,1]$ pour $N^G=2$.

Nous définissons ensuite par tensorisation 3 familles de points sur $\widehat{\mathrm{T}}$:

k	1	2	3
$\hat{x}_k^G$	$\frac{1-\frac{3}{\sqrt{3}}}{2}$	$\frac{1}{2}$	$\frac{1+\frac{3}{\sqrt{3}}}{2}$
$\hat{\omega}_k^G$	$\frac{5}{18}$	$\frac{4}{9}$	$\frac{5}{18}$

TABLE 1.3 – Points et poids de Gauss sur le segment $[0,1]$ pour $N^G = 3$.

k	1	2	3	4
$\hat{x}_k^G$	$1-\hat{x}_4^G$	$1-\hat{x}_3^G$	$\frac{1}{2}-\frac{\nu}{2}$	$\frac{1}{2}+\frac{\eta}{2}$
$\hat{\omega}_k^G$	$\frac{1}{2}\frac{\frac{1}{3}-\nu^2}{\eta^2-\nu^2}$	$\frac{1}{2}-\hat{w}_1^G$	$\hat{w}_2^G$	$\hat{w}_1^G$

TABLE 1.4 – Points et poids de Gauss sur le segment $[0,1]$ pour $N^G = 4$, avec $\eta := 2^{-\frac{1}{2}}\sqrt{\frac{6}{7}+\frac{\sqrt{96}}{\sqrt{245}}}$ et $\nu := 2^{-\frac{1}{2}}\sqrt{\frac{6}{7}-\frac{\sqrt{96}}{\sqrt{245}}}$.

k	1	2
$\hat{x}_k^{GL}$	0	1
$\hat{\omega}_k^{GL}$	0.5	0.5

TABLE 1.5 – Points et poids de Gauss-Lobatto sur le segment $[0,1]$ pour $N^{GL} = 2$.

— La famille 1 : $\hat{x}^1_{\mathrm{ix,iy,iz}} := (\hat{x}^G_{\mathrm{ix}}\,,\,\hat{x}^{GL}_{\mathrm{iy}}\,,\,\hat{x}^{GL}_{\mathrm{iz}})$ pour $\mathrm{ix} = 1, N^G$ et $\mathrm{iy}, \mathrm{iz} = 1, N^{GL}$, voir la Figure 1.1,

k	1	2	3
$\hat{x}_k^{GL}$	0	0.5	1
$\hat{\omega}_k^{GL}$	$\frac{1}{6}$	$\frac{2}{3}$	$\frac{1}{6}$

TABLE 1.6 – Points et poids de Gauss-Lobatto sur le segment $[0, 1]$ pour $N^{GL} = 3$.

k	1	2	3	4
$\hat{x}_k^{GL}$	0	$\frac{5-\sqrt{5}}{10}$	$\frac{5+\sqrt{5}}{10}$	1
$\hat{\omega}_k^{GL}$	$\frac{1}{12}$	$\frac{5}{12}$	$\frac{5}{12}$	$\frac{1}{12}$

TABLE 1.7 – Points et poids de Gauss-Lobatto sur le segment $[0, 1]$ pour $N^{GL} = 4$.

k	1	2	3	4	5
$\hat{x}_k^{GL}$	0	$\frac{7-\sqrt{21}}{14}$	0.5	$\frac{7+\sqrt{21}}{14}$	1
$\hat{\omega}_k^{GL}$	$\frac{1}{20}$	$\frac{49}{180}$	$\frac{16}{45}$	$\frac{49}{180}$	$\frac{1}{20}$

TABLE 1.8 – Points et poids de Gauss-Lobatto sur le segment $[0, 1]$ pour $N^{GL} = 5$.

— La famille 2 : $\hat{x}^2_{\mathrm{ix,iy,iz}} := (\hat{x}^{GL}_{\mathrm{ix}}, \hat{x}^{G}_{\mathrm{iy}}, \hat{x}^{GL}_{\mathrm{iz}})$ pour $\mathrm{iy} = 1, N^G$ et $\mathrm{ix}, \mathrm{iz} = 1, N^{GL}$, voir la Figure 1.2,

— La famille 3 : $\hat{x}^3_{\mathrm{ix,iy,iz}} := (\hat{x}^{GL}_{\mathrm{ix}}, \hat{x}^{GL}_{\mathrm{iy}}, \hat{x}^{G}_{\mathrm{iz}})$ pour $\mathrm{iz} = 1, N^G$ et $\mathrm{ix}, \mathrm{iy} = 1, N^{GL}$, voir la Figure 1.3.

Remarque 1.2. *Par convenance, nous utilisons le mot clé famille ci-dessus. Cependant, nous notons qu'il n'a aucun lien avec les première et deuxième familles de Nédélec que nous pouvons trouver dans la littérature.*

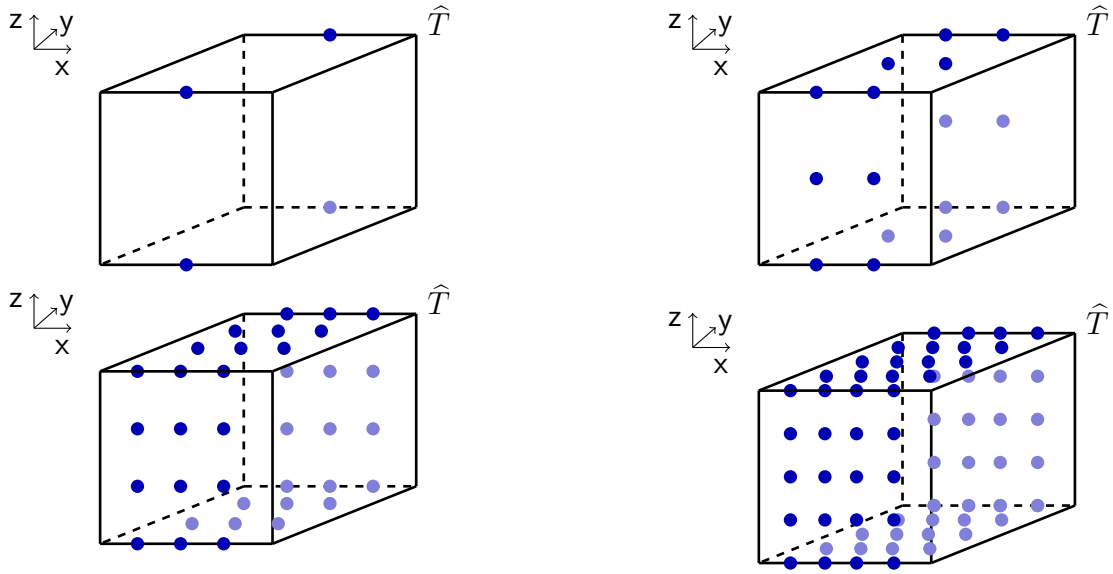

FIGURE 1.1 – Points de Gauss et de Gauss-Lobatto de la famille 1 pour les ordres $r = 1, 4$.

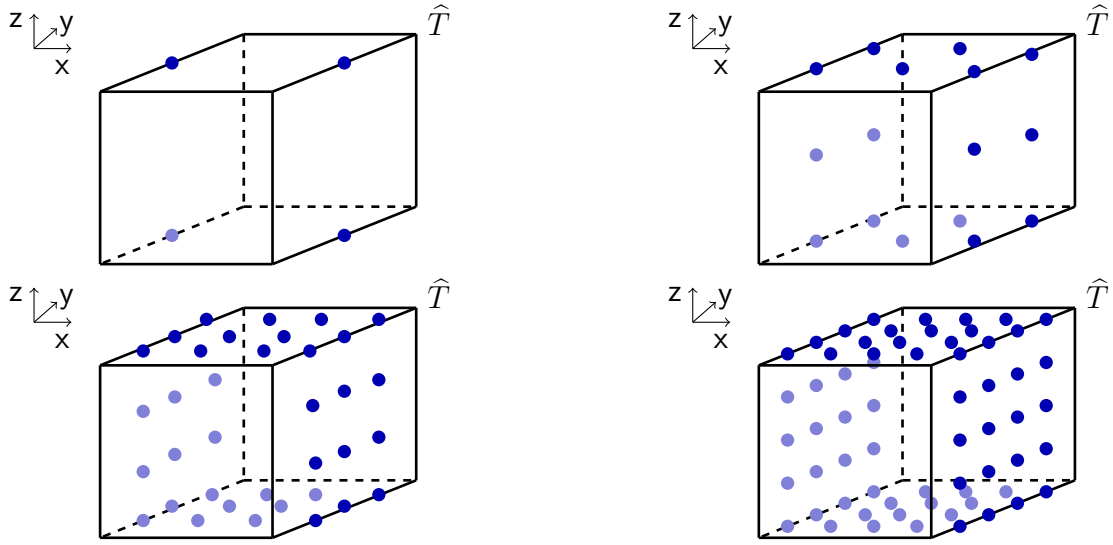

FIGURE 1.2 – Points de Gauss et de Gauss-Lobatto de la famille 2 pour les ordres $r = 1, 4$.

Nous définissons alors l'ensemble des degrés de liberté locaux à partir de ces points.

Définition 1.1 (Degrés de liberté). *Les degrés de liberté sont définis aux points $\hat{x}^f_{\mathrm{ix,iy,iz}}$, par les formes linéaires $\widehat{\Lambda}^f_{\mathrm{ix,iy,iz}}$, où $f = 1, 3$ sont les numéros des familles de points sur $\widehat{T}$. L'ensemble de degrés de liberté est défini par*

$$\widehat{\Sigma}_r := \left\{ \left(\widehat{\Lambda}^1_{\mathrm{ix,iy,iz}}\right)_{\mathrm{ix,iy,iz}} , \left(\widehat{\Lambda}^2_{\mathrm{ix,iy,iz}}\right)_{\mathrm{ix,iy,iz}} , \left(\widehat{\Lambda}^3_{\mathrm{ix,iy,iz}}\right)_{\mathrm{ix,iy,iz}} \right\},$$

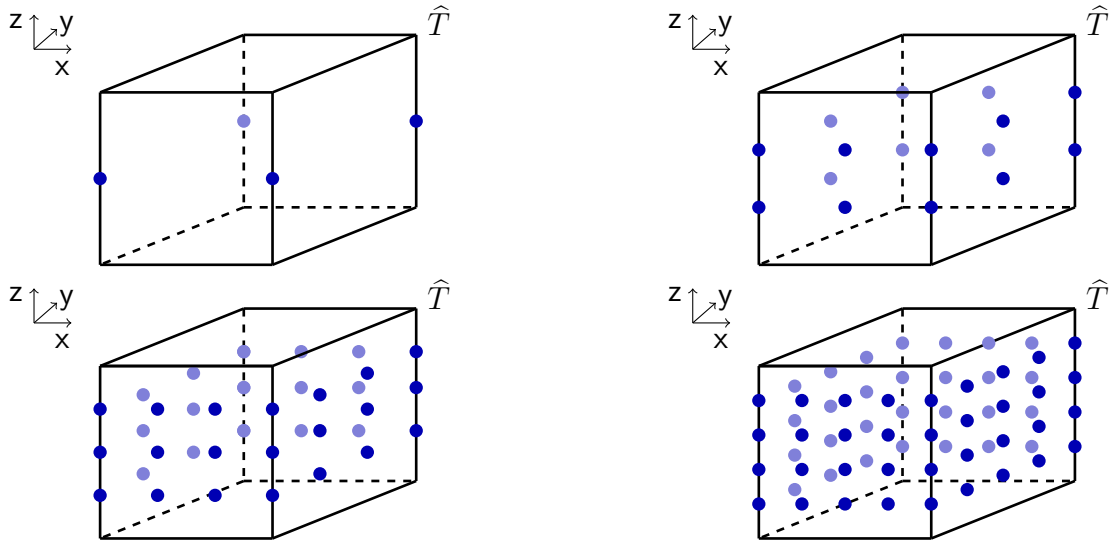

FIGURE 1.3 – Points de Gauss et de Gauss-Lobatto de la famille 3 pour les ordres $r = 1, 4$.

où nous définissons, pour tout $\mathbf{p} \in \widehat{\mathcal{N}}_r$*,*

— *Pour la famille* 1 *: pour* $\mathrm{ix} = 1, \cdots, N^G$ *et* $\mathrm{iy}, \mathrm{iz} = 1, \cdots, N^{GL}$*,*

$$\widehat{\Lambda}^1_{\mathrm{ix,iy,iz}}(\mathbf{p}) := \mathbf{p}(\hat{x}^1_{\mathrm{ix,iy,iz}}) \cdot \mathbf{e}_1,$$

— *Pour la famille* 2 *: pour* $\mathrm{iy} = 1, \cdots, N^G$ *et* $\mathrm{ix}, \mathrm{iz} = 1, \cdots, N^{GL}$*,*

$$\widehat{\Lambda}^2_{\mathrm{ix,iy,iz}}(\mathbf{p}) := \mathbf{p}(\hat{x}^2_{\mathrm{ix,iy,iz}}) \cdot \mathbf{e}_2,$$

— *Pour la famille* 3 *: pour* $\mathrm{iz} = 1, \cdots, N^G$ *et* $\mathrm{ix}, \mathrm{iy} = 1, \cdots, N^{GL}$*,*

$$\widehat{\Lambda}^3_{\mathrm{ix,iy,iz}}(\mathbf{p}) := \mathbf{p}(\hat{x}^3_{\mathrm{ix,iy,iz}}) \cdot \mathbf{e}_3,$$

où $(\mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_3)$ *est la base canonique de* $\mathbb{R}^3$ *et où nous avons choisi la loi standard*

$$N^{GL} := N^G + 1.$$

Définition 1.2 (Base de référence)**.** *Les fonctions de base associées à l'élément fini* $\left(\widehat{T}, \widehat{\mathcal{N}}_r, \widehat{\Sigma}_r\right)$ *sont définies par*

$$\widehat{\Lambda}^f_{\mathrm{ix,iy,iz}}\left(\widehat{\phi}^g_{\mathrm{jx,jy,jz}}\right) = \delta_{f,g}\, \delta_{\mathrm{ix,jx}}\, \delta_{\mathrm{iy,jy}}\, \delta_{\mathrm{iz,jz}},$$

où $f = 1, 3$ *et* $g = 1, 3$ *sont des numéros de famille. Plus précisément, nous avons*

$$\begin{aligned}
\widehat{\boldsymbol{\phi}}^1_{\mathrm{ix,iy,iz}}(\hat{x},\hat{y},\hat{z}) &= \begin{pmatrix} \hat{\phi}^G_{\mathrm{ix}}(\hat{x})\hat{\phi}^{GL}_{\mathrm{iy}}(\hat{y})\hat{\phi}^{GL}_{\mathrm{iz}}(\hat{z}) \\ 0 \\ 0 \end{pmatrix}, \\
\widehat{\boldsymbol{\phi}}^2_{\mathrm{ix,iy,iz}}(\hat{x},\hat{y},\hat{z}) &= \begin{pmatrix} 0 \\ \hat{\phi}^{GL}_{\mathrm{ix}}(\hat{x})\hat{\phi}^{G}_{\mathrm{iy}}(\hat{y})\hat{\phi}^{GL}_{\mathrm{iz}}(\hat{z}) \\ 0 \end{pmatrix}, \\
\widehat{\boldsymbol{\phi}}^3_{\mathrm{ix,iy,iz}}(\hat{x},\hat{y},\hat{z}) &= \begin{pmatrix} 0 \\ 0 \\ \hat{\phi}^{GL}_{\mathrm{ix}}(\hat{x})\hat{\phi}^{GL}_{\mathrm{iy}}(\hat{y})\hat{\phi}^{G}_{\mathrm{iz}}(\hat{z}) \end{pmatrix},
\end{aligned}$$

où les fonctions $\hat{\phi}^G_{\mathrm{k}} : [0,1] \to \mathbb{C}$ *et* $\hat{\phi}^{GL}_{\mathrm{k}} : [0,1] \to \mathbb{C}$ *sont respectivement les polynômes d'interpolation de Lagrange associés aux points de Gauss et Gauss-Lobatto défnis sur* $[0,1]$ *:*

$$\hat{\phi}^G_{\mathrm{k}}(\hat{x}) := \prod_{\mathrm{j}\neq\mathrm{k}} \frac{\hat{x} - \hat{x}^G_{\mathrm{j}}}{\hat{x}^G_{\mathrm{k}} - \hat{x}^G_{\mathrm{j}}} \quad \textit{et} \quad \hat{\phi}^{GL}_{\mathrm{k}}(\hat{x}) := \prod_{\mathrm{j}\neq\mathrm{k}} \frac{\hat{x} - \hat{x}^{GL}_{\mathrm{j}}}{\hat{x}^{GL}_{\mathrm{k}} - \hat{x}^{GL}_{\mathrm{j}}}.$$

Ainsi, dans chaque élément T, nous comptons $\#\mathrm{ddlelem} := 3 \times N^G \times (N^{GL})^2$ degrés de liberté locaux, de numéros $\mathrm{i_{loc}} = 1, \#\mathrm{ddlelem}$. Chacun d'entre eux est associé à une fonction de base vectorielle $\overrightarrow{\phi}^{\mathrm{i_{elem}},f}_{\mathrm{i_{loc}}} : T \to \mathbb{C}^3$, avec $f = 1, 3$, le numéro des familles. Par famille, nous avons $N_r := N^G(N^{GL})^2$ fonctions de base vectorielles, *ie* $\#\mathrm{ddlelem} := 3 \times N_r$, pour un ordre r donné. Ces fonctions de base sont définies comme suit.

Définition 1.3 (Base locale)**.** *Soit un cube, de numéro* $\mathrm{i_{elem}}$*, défini par*

$$T := [x^T_{min}, x^T_{min} + h] \times [y^T_{min}, y^T_{min} + h] \times [z^T_{min}, z^T_{min} + h],$$

et soit un noeud $(x, y, z) \in T$*, nous définissons*

$$\overrightarrow{\phi}^{\mathrm{i_{elem}},f}_{\mathrm{i_{loc}}}(x,y,z) := \widehat{\boldsymbol{\phi}}^f_{\mathrm{ix,iy,iz}}(\hat{x},\hat{y},\hat{z}), \tag{1.1}$$

avec $(x, y, z) = F_T(\hat{x}, \hat{y}, \hat{z})$ *où* $F_T : \widehat{T} \to T$ *est la transformation affine définie par*

$$F_T(\hat{x},\hat{y},\hat{z}) := \begin{pmatrix} x^T_{min} + h\hat{x} \\ y^T_{min} + h\hat{y} \\ z^T_{min} + h\hat{z} \end{pmatrix}, \tag{1.2}$$

et où les indices $(\mathrm{ix}, \mathrm{iy}, \mathrm{iz})$ *sont liés à* $\mathrm{i_{loc}}$ *par*

$$\mathrm{i_{loc}} = \mathrm{NoeudsGGLToLoc}(f, \mathrm{ix}, \mathrm{iy}, \mathrm{iz}),$$

voir la Remarque 1.4, et correspondent aux indices des noeuds d'interpolation de Gauss et de Gauss-Lobatto variant suivant les familles :

- $f = 1$ *:* $\mathrm{ix} = 1, N^G$ *et* $\mathrm{iy}, \mathrm{iz} = 1, N^{GL}$,
- $f = 2$ *:* $\mathrm{iy} = 1, N^G$ *et* $\mathrm{ix}, \mathrm{iz} = 1, N^{GL}$,
- $f = 3$ *:* $\mathrm{iz} = 1, N^G$ *et* $\mathrm{ix}, \mathrm{iy} = 1, N^{GL}$.

Remarque 1.3. *Dans la définition des fonctions de base (1.1), nous n'avons pas utilisé la transformation classique associée à la conformité* $H(\mathrm{rot}, \Omega)$*, ie* $(J^*_{F_T})^{-1}$*, car dans le cas des cubes elle se réduit à la matrice* $\frac{1}{h}\mathbf{I}_{3\times 3}$.

De plus, compte tenu de (1.2), les éléments du maillage sont tous de même taille et ont les mêmes points de Gauss et de Gauss-Lobatto.

Remarque 1.4. *Les degrés de liberté locaux* $\mathrm{i_{loc}}$ *sont en correspondance avec trois indices* $(\mathrm{ix}, \mathrm{iy}, \mathrm{iz})$ *et une famille* f*. Ce lien est créé grâce à l'algorithme suivant :*

$\mathrm{i_{loc}} = 0$
$f = 1$
do $\mathrm{iz} = 1, N^{GL}$
  do $\mathrm{iy} = 1, N^{GL}$
    do $\mathrm{ix} = 1, N^G$
      $\mathrm{i_{loc}} = \mathrm{i_{loc}} + 1$
      $\mathrm{NoeudsGGLToLoc}(f, \mathrm{ix}, \mathrm{iy}, \mathrm{iz}) = \mathrm{i_{loc}}$
    end do
  end do
end do
$f = 2$
do $\mathrm{iz} = 1, N^{GL}$
  do $\mathrm{ix} = 1, N^{GL}$
    do $\mathrm{iy} = 1, N^G$
      $\mathrm{i_{loc}} = \mathrm{i_{loc}} + 1$
      $\mathrm{NoeudsGGLToLoc}(f, \mathrm{ix}, \mathrm{iy}, \mathrm{iz}) = \mathrm{i_{loc}}$
    end do
  end do
end do
$f = 3$

```
do ix = 1, N^GL
  do iy = 1, N^GL
    do iz = 1, N^G
      i_loc = i_loc + 1
      NoeudsGGLToLoc(f, ix, iy, iz) = i_loc
    end do
  end do
end do
```

Définition 1.4 (Connectivité). *L'image des noeuds* $\hat{x}^f_{\mathrm{ix,iy,iz}}$ *par l'ensemble des transformations* F_T *définit naturellement la position des degrés de liberté globaux. Nous introduisons une table de connectivité* NoeudsGGLToGlob *liant la numérotation locale à une numérotation globale par*

$$\mathrm{i_{glob}} = \mathrm{NoeudsGGLToGlob}(\mathrm{i_{elem}}, f, \mathrm{ix}, \mathrm{iy}, \mathrm{iz}).$$

Définition 1.5 (Fonctions de base globales). *À chaque degré de liberté global* $\mathrm{i_{glob}}$*, nous associons une fonction de base* $\overrightarrow{\phi}_{\mathrm{i_{glob}}}$*. Pour* T *de numéro* $\mathrm{i_{elem}}$*, cette fonction de base globale est définie par*

$$\begin{cases} \overrightarrow{\phi}_{\mathrm{i_{glob}}}(x,y,z) & := & \overrightarrow{\phi}^{\mathrm{i_{elem}},f}_{\mathrm{i_{loc}}}(x,y,z), & \text{si } (x,y,z) \in T, \text{ où } \mathrm{i_{glob}} \text{ vérifie la Déf. 1.4,} \\ \overrightarrow{\phi}_{\mathrm{i_{glob}}}(x,y,z) & \equiv & 0, & \text{si } (x,y,z) \notin T. \end{cases}$$

1.1.3 Discrétisation de la formulation

Nous discrétisons le Problème 3 sur l'espace $\mathcal{N}_r$ de dimension finie, défini par

$$\mathcal{N}_r := \underset{\mathrm{i_{glob}}=1,\#\mathrm{ddl}}{\mathrm{span}} \left(\left\{ \overrightarrow{\phi}_{\mathrm{i_{glob}}} \right\} \right).$$

Cela nous mène naturellement au système linéaire suivant.

Problème 4 (Problème matriciel). *Le Problème 3 revient à chercher les coefficients* $[\mathbf{E}^h]_{\mathrm{i_{glob}}}$ *du vecteur solution* $[\mathbf{E}^h]$ *du système matriciel*

$$\mathbf{A}[\mathbf{E}^h] = \mathbf{F},$$

où $\mathbf{A} := (\mathbf{A}_{\mathrm{i_{glob}},\mathrm{j_{glob}}})_{\mathrm{i_{glob}},\mathrm{j_{glob}}=1,\#\mathrm{ddl}} := (-k_0^2\,\mathbf{M}_{\mathrm{i_{glob}},\mathrm{j_{glob}}} + \mathbf{R}_{\mathrm{i_{glob}},\mathrm{j_{glob}}} + \mathbf{B}_{\mathrm{i_{glob}},\mathrm{j_{glob}}})_{\mathrm{i_{glob}},\mathrm{j_{glob}}=1,\#\mathrm{ddl}}$

et $\mathbf{F} := (\mathbf{F}_{\mathrm{i_{glob}}})_{\mathrm{i_{glob}}=1,\#\mathrm{ddl}}$ *sont définis par*

$$\begin{aligned}
\mathbf{M}_{\mathrm{i_{glob}},\mathrm{j_{glob}}} &:= \int_\Omega \overrightarrow{\phi}_{\mathrm{j_{glob}}} \cdot \overline{\overrightarrow{\phi}_{\mathrm{i_{glob}}}}, \\
\mathbf{R}_{\mathrm{i_{glob}},\mathrm{j_{glob}}} &:= \int_\Omega \nabla \times \overrightarrow{\phi}_{\mathrm{j_{glob}}} \cdot \overline{\nabla \times \overrightarrow{\phi}_{\mathrm{i_{glob}}}}, \\
\mathbf{B}_{\mathrm{i_{glob}},\mathrm{j_{glob}}} &:= \int_{\partial\Omega} \frac{ik_0}{Z_{\partial\Omega}} \gamma_t \overrightarrow{\phi}_{\mathrm{j_{glob}}} \cdot \overline{\gamma_t \overrightarrow{\phi}_{\mathrm{i_{glob}}}},
\end{aligned} \tag{1.3}$$

et

$$\mathbf{F}_{\mathrm{i_{glob}}} := \int_\Omega \frac{ik_0}{Z_{\partial\Omega}} \mathbf{g} \cdot \overline{\overrightarrow{\phi}_{\mathrm{i_{glob}}}}. \tag{1.4}$$

Remarque 1.5. *La matrice* $\mathbf{M}$ *est la matrice de masse. La matrice* $\mathbf{R}$ *est la matrice de rigidité. La matrice* $\mathbf{B}$ *est la matrice de bord.*

Les intégrales définies dans (1.3) et (1.4) sont calculables numériquement. Nous choisissons une loi basée sur les points de Gauss et de Gauss-Lobatto.

Remarque 1.6. *D'une part,* N^G *points d'intégration de Gauss permettent d'intégrer exactement les polynômes d'ordre* $2N^G - 1$*. D'autre part,* N^{GL} *points d'intégration de Gauss-Lobatto permettent d'intégrer exactement les polynômes de degré* $2N^{GL} - 3$*. Ainsi, en fonction du nombre de noeuds de Gauss et de Gauss-Lobatto, nous intégrons de manière exacte différents ordres de polynômes, voir le Tableau 1.9.*

N^G	N^{GL}	r	Ordre du polynôme
1	2	1	1
2	3	2	3
3	4	3	5
4	5	4	7

TABLE 1.9 – Ordres des polynômes intégrables de manière exacte en fonction du nombre de points de Gauss N^G et du nombre de points de Gauss-Lobatto N^{GL}.

Dans le cadre de cette méthode, nous choisissons un maillage cartésien. De cette manière, pour calculer les matrices $\mathbf{M}$, $\mathbf{R}$ et $\mathbf{B}$ de (1.3), nous construisons des matrices dites élémentaires, *resp.* $\mathbf{M}^{\mathrm{elem}}$, $\mathbf{R}^{\mathrm{elem}}$ et $\mathbf{B}^{\mathrm{elem}}$, indépendantes de l'élément considéré. Les degrés de liberté locaux $\mathrm{i_{loc}}, \mathrm{j_{loc}} = 1, \#\mathrm{ddlelem}$ servent à la description des matrices. Nous rappelons qu'ils sont respectivement en bijection avec une famille f (*resp.* g) et les indices $(\mathrm{ix}, \mathrm{iy}, \mathrm{iz})$ (*resp.* $(\mathrm{jx}, \mathrm{jy}, \mathrm{jz})$), voir la Remarque 1.4. La matrice $\mathbf{B}^{\mathrm{elem}}$ est une matrice élémentaire définie sur les faces du cube unité. Nous avons alors $\mathrm{i_{loc}}, \mathrm{j_{loc}} = 1, 2\, N_r^{\mathrm{bord}}$, où $N_r^{\mathrm{bord}} := N^G\, N^{GL}$.

Ces matrices élémentaires sont elles-mêmes construites grâce à des matrices de référence $\widehat{\mathbf{M}}$, $\widehat{\mathbf{R}}$ et $\widehat{\mathbf{C}}$. Nous pouvons ainsi définir les trois matrices élémentaires $\mathbf{M}^{\mathrm{elem}}$, $\mathbf{R}^{\mathrm{elem}}$ et $\mathbf{B}^{\mathrm{elem}}$.

Définition 1.6 (Matrice de masse élémentaire)**.** *La matrice de masse élémentaire* $\mathbf{M}^{\mathrm{elem}}$*, de dimension* $\#\mathrm{ddlelem}$*, a la structure suivante*

$$\mathbf{M}^{\mathrm{elem}} = \begin{pmatrix} \left(\mathbf{M}^{\mathrm{elem},1,1}\right) & 0_{N_r\times N_r} & 0_{N_r\times N_r} \\ 0_{N_r\times N_r} & \left(\mathbf{M}^{\mathrm{elem},2,2}\right) & 0_{N_r\times N_r} \\ 0_{N_r\times N_r} & 0_{N_r\times N_r} & \left(\mathbf{M}^{\mathrm{elem},3,3}\right) \end{pmatrix},$$

où les matrices $\mathbf{M}^{\mathrm{elem},1,1} \in \mathbb{C}^{N_r\times N_r}$*,* $\mathbf{M}^{\mathrm{elem},2,2} \in \mathbb{C}^{N_r\times N_r}$ *et* $\mathbf{M}^{\mathrm{elem},3,3} \in \mathbb{C}^{N_r\times N_r}$*, sont définies par*

$$\mathbf{M}^{\mathrm{elem},1,1}_{\mathrm{i_{loc}},\mathrm{j_{loc}}} := h^3\, \widehat{\mathbf{M}}^{G}_{\mathrm{ix},\mathrm{jx}} \widehat{\mathbf{M}}^{GL}_{\mathrm{iy},\mathrm{jy}} \widehat{\mathbf{M}}^{GL}_{\mathrm{iz},\mathrm{jz}},$$

$$\mathbf{M}^{\mathrm{elem},2,2}_{\mathrm{i_{loc}},\mathrm{j_{loc}}} := h^3\, \widehat{\mathbf{M}}^{GL}_{\mathrm{ix},\mathrm{jx}} \widehat{\mathbf{M}}^{G}_{\mathrm{iy},\mathrm{jy}} \widehat{\mathbf{M}}^{GL}_{\mathrm{iz},\mathrm{jz}},$$

$$\mathbf{M}^{\mathrm{elem},3,3}_{\mathrm{i_{loc}},\mathrm{j_{loc}}} := h^3\, \widehat{\mathbf{M}}^{GL}_{\mathrm{ix},\mathrm{jx}} \widehat{\mathbf{M}}^{GL}_{\mathrm{iy},\mathrm{jy}} \widehat{\mathbf{M}}^{G}_{\mathrm{iz},\mathrm{jz}},$$

où la matrice de masse $1D$ *de Gauss, construite sur le segment de référence* $[0,1]$*, est définie par*

$$\widehat{\mathbf{M}}^{G}_{\mathrm{i,j}} := \int_0^1 \widehat{\phi}^{G}_{\mathrm{j}}(\hat{x})\overline{\widehat{\phi}^{G}_{\mathrm{i}}(\hat{x})},$$

et où la matrice de masse $1D$ *de Gauss-Lobatto est définie de la même manière par*

$$\widehat{\mathbf{M}}^{GL}_{\mathrm{i,j}} := \int_0^1 \widehat{\phi}^{GL}_{\mathrm{j}}(\hat{x})\overline{\widehat{\phi}^{GL}_{\mathrm{i}}(\hat{x})}.$$

Définition 1.7 (Matrice de rigidité élémentaire)**.** *La matrice de rigidité élémentaire* $\mathbf{R}^{\mathrm{elem}}$*, de dimension* $\#\mathrm{ddlelem}$*, a la structure suivante*

$$\mathbf{R}^{\mathrm{elem}} := \begin{pmatrix} \left(\mathbf{R}^{\mathrm{elem},1,1}\right) & \left(\mathbf{R}^{\mathrm{elem},1,2}\right) & \left(\mathbf{R}^{\mathrm{elem},1,3}\right) \\ \left(\mathbf{R}^{\mathrm{elem},2,1}\right) & \left(\mathbf{R}^{\mathrm{elem},2,2}\right) & \left(\mathbf{R}^{\mathrm{elem},2,3}\right) \\ \left(\mathbf{R}^{\mathrm{elem},3,1}\right) & \left(\mathbf{R}^{\mathrm{elem},3,2}\right) & \left(\mathbf{R}^{\mathrm{elem},3,3}\right) \end{pmatrix},$$

où les matrices $\mathbf{R}^{\mathrm{elem},\mathbf{1},\mathbf{1}}$*,* $\mathbf{R}^{\mathrm{elem},\mathbf{2},\mathbf{2}}$*,* $\mathbf{R}^{\mathrm{elem},\mathbf{3},\mathbf{3}}$*,* $\mathbf{R}^{\mathrm{elem},\mathbf{1},\mathbf{2}}$*,* $\mathbf{R}^{\mathrm{elem},\mathbf{1},\mathbf{3}}$*,* $\mathbf{R}^{\mathrm{elem},\mathbf{2},\mathbf{1}}$*,* $\mathbf{R}^{\mathrm{elem},\mathbf{2},\mathbf{3}}$*,* $\mathbf{R}^{\mathrm{elem},\mathbf{3},\mathbf{1}}$ *et* $\mathbf{R}^{\mathrm{elem},\mathbf{3},\mathbf{2}}$*, dans* $\mathbb{C}^{N_r\times N_r}$*, sont définies par*

$$\mathbf{R}^{\mathrm{elem},1,1}_{\mathrm{i_{loc}},\mathrm{j_{loc}}} := h\, \widehat{\mathbf{M}}^{G}_{\mathrm{ix},\mathrm{jx}} \widehat{\mathbf{M}}^{GL}_{\mathrm{iy},\mathrm{jy}} \widehat{\mathbf{R}}^{GL}_{\mathrm{iz},\mathrm{jz}} + h\, \widehat{\mathbf{M}}^{G}_{\mathrm{ix},\mathrm{jx}} \widehat{\mathbf{M}}^{GL}_{\mathrm{iz},\mathrm{jz}} \widehat{\mathbf{R}}^{GL}_{\mathrm{iy},\mathrm{jy}},$$

$$\mathbf{R}^{\mathrm{elem},2,2}_{\mathrm{i_{loc}},\mathrm{j_{loc}}} := h\, \widehat{\mathbf{M}}^{G}_{\mathrm{iy},\mathrm{jy}} \widehat{\mathbf{M}}^{GL}_{\mathrm{ix},\mathrm{jx}} \widehat{\mathbf{R}}^{GL}_{\mathrm{iz},\mathrm{jz}} + h\, \widehat{\mathbf{M}}^{G}_{\mathrm{iy},\mathrm{jy}} \widehat{\mathbf{M}}^{GL}_{\mathrm{iz},\mathrm{jz}} \widehat{\mathbf{R}}^{GL}_{\mathrm{ix},\mathrm{jx}},$$

$$\mathbf{R}^{\mathrm{elem},3,3}_{\mathrm{i_{loc}},\mathrm{j_{loc}}} := h\, \widehat{\mathbf{M}}^{G}_{\mathrm{iz},\mathrm{jz}} \widehat{\mathbf{M}}^{GL}_{\mathrm{ix},\mathrm{jx}} \widehat{\mathbf{R}}^{GL}_{\mathrm{iy},\mathrm{jy}} + h\, \widehat{\mathbf{M}}^{G}_{\mathrm{iz},\mathrm{jz}} \widehat{\mathbf{M}}^{GL}_{\mathrm{iy},\mathrm{jy}} \widehat{\mathbf{R}}^{GL}_{\mathrm{ix},\mathrm{jx}},$$

$$\mathbf{R}^{\mathrm{elem},1,2}_{\mathrm{i_{loc}},\mathrm{j_{loc}}} := -h\, \widehat{\mathbf{C}}_{\mathrm{ix},\mathrm{jx}} \widehat{\mathbf{C}}_{\mathrm{jy},\mathrm{iy}} \widehat{\mathbf{M}}^{GL}_{\mathrm{iz},\mathrm{jz}},$$

$$\mathbf{R}^{\mathrm{elem},1,3}_{\mathrm{i_{loc}},\mathrm{j_{loc}}} := -h\, \widehat{\mathbf{C}}_{\mathrm{ix},\mathrm{jx}} \widehat{\mathbf{C}}_{\mathrm{jz},\mathrm{iz}} \widehat{\mathbf{M}}^{GL}_{\mathrm{iy},\mathrm{jy}},$$

$$\mathbf{R}^{\mathrm{elem},2,1}_{\mathrm{i_{loc},j_{loc}}} := -h\,\widehat{\mathbf{C}}_{\mathrm{iy,jy}}\widehat{\mathbf{C}}_{\mathrm{jx,ix}}\widehat{\mathbf{M}}^{GL}_{\mathrm{iz,jz}},$$

$$\mathbf{R}^{\mathrm{elem},2,3}_{\mathrm{i_{loc},j_{loc}}} := -h\,\widehat{\mathbf{C}}_{\mathrm{iy,jy}}\widehat{\mathbf{C}}_{\mathrm{jz,iz}}\widehat{\mathbf{M}}^{GL}_{\mathrm{ix,jx}},$$

$$\mathbf{R}^{\mathrm{elem},3,1}_{\mathrm{i_{loc},j_{loc}}} := -h\,\widehat{\mathbf{C}}_{\mathrm{iz,jz}}\widehat{\mathbf{C}}_{\mathrm{jx,ix}}\widehat{\mathbf{M}}^{GL}_{\mathrm{iy,jy}},$$

$$\mathbf{R}^{\mathrm{elem},3,2}_{\mathrm{i_{loc},j_{loc}}} := -h\,\widehat{\mathbf{C}}_{\mathrm{iz,jz}}\widehat{\mathbf{C}}_{\mathrm{jy,iy}}\widehat{\mathbf{M}}^{GL}_{\mathrm{ix,jx}},$$

où la matrice de rigidté 1*D de Gauss, construite sur le segment de référence* $[0,1]$*, est définie par*

$$\widehat{\mathbf{R}}^{G}_{\mathrm{i,j}} := \int_0^1 \frac{\partial\phi^G_{\mathrm{j}}(x)}{\partial x}\overline{\frac{\partial\phi^G_{\mathrm{j}}(x)}{\partial x}},$$

la matrice de rigidité 1*D de Gauss-Lobatto est définie de la même manière par*

$$\widehat{\mathbf{R}}^{GL}_{\mathrm{i,j}} := \int_0^1 \frac{\partial\phi^{GL}_{\mathrm{j}}(x)}{\partial x}\overline{\frac{\partial\phi^{GL}_{\mathrm{j}}(x)}{\partial x}},$$

et où la matrice mixte 1*D de Gauss est définie par*

$$\widehat{\mathbf{C}}^{G}_{\mathrm{i,j}} := \int_0^1 \phi^G_{\mathrm{j}}(x)\overline{\frac{\partial\phi^{GL}_{\mathrm{j}}(x)}{\partial x}}.$$

Définition 1.8 (Matrice de bord élémentaire)**.** *La matrice de bord élémentaire* $\mathbf{B}^{\mathrm{elem}}$*, de dimension* $2\times N_r^{\mathrm{bord}}$ *(nombre de degrés de liberté sur une face), a la structure suivante*

$$\mathbf{B}^{\mathrm{elem}} = \begin{pmatrix} \left(\mathbf{B}^{1,1}\right) & 0_{N_r^{\mathrm{bord}}} \\ 0_{N_r^{\mathrm{bord}}} & \left(\mathbf{B}^{2,2}\right) \end{pmatrix},$$

où les matrices $\mathbf{B}^{1,1}$ *et* $\mathbf{B}^{2,2}$*, de dimension* N_r^{bord}*, sont définies par*

$$\mathbf{B}^{1,1}_{\mathrm{i,j}} := \widehat{\mathbf{M}}^{G}_{\mathrm{iz,jz}}\widehat{\mathbf{M}}^{GL}_{\mathrm{ix,jx}},$$

$$\mathbf{B}^{2,2}_{\mathrm{i,j}} := \widehat{\mathbf{M}}^{G}_{\mathrm{ix,jx}}\widehat{\mathbf{M}}^{GL}_{\mathrm{iz,jz}}.$$

Finalement, nous avons intégralement décrit les matrices du Problème matriciel 4.

1.1.4 Convergence du solveur

La solution numérique $\mathbf{E}^h \in \mathcal{N}_r$ du problème variationnel discret est représentée par le vecteur solution $[\mathbf{E}^h] \in \mathbb{C}^{\#\mathrm{ddl}}$. Ce dernier est obtenu grâce à la résolution du Problème matriciel 4 à l'aide d'une factorisation LU de la matrice $\mathbf{A}$, réalisée par le solveur MUMPS© [4]. Nous souhaitons vérifier si le solveur de Nédélec développé converge vers la solution du problème. Pour cela nous cherchons à simuler une onde électromagnétique dont nous

connaissons analytiquement la solution exacte $\mathbf{E}^{\text{ex}}$. Nous étudions l'erreur relative e^h

$$e^h := \frac{\|\mathbf{E}^h - \mathbf{E}^{\text{ex}}\|_{\mathrm{L}^2}}{\|\mathbf{E}^{\text{ex}}\|_{\mathrm{L}^2}}, \tag{1.8}$$

avec la norme L^2 définie par

$$\|\mathbf{E}\|^2_{\mathrm{L}^2} = \sum_{T\in\mathcal{T}} \int_T \mathbf{E}^T \cdot \overline{\mathbf{E}^T}\,, \quad \text{pour tout } \mathbf{E} \in \mathrm{L}^2.$$

Remarque 1.7. *Soit* $[\mathbf{E}^{\text{ex}}]$ *(resp.* $[\mathbf{E}^h]$*) le vecteur des valeurs des degrés de liberté de l'interpolation de la solution exacte* $\mathbf{E}^{\text{ex}}$ *(resp. de la solution numérique* $\mathbf{E}^h$*) aux noeuds de Gauss et de Gauss-Lobatto. En pratique, l'erreur relative* L^2 *de* (1.8) *est donnée par*

$$e^h = \frac{\sqrt{([\mathbf{E}^h] - [\mathbf{E}^{\text{ex}}])^* \mathbf{M} ([\mathbf{E}^h] - [\mathbf{E}^{\text{ex}}])}}{\sqrt{[\mathbf{E}^{\text{ex}}]^* \mathbf{M} [\mathbf{E}^{\text{ex}}]}},$$

où $\mathbf{M}$ *est la matrice de masse définie par* (1.3).

Nous étudions le cas où $\Omega = [0,1]^3$ et $\mathbf{g} = \gamma_t \mathbf{E}_{inc} + Z_{\partial\Omega}\gamma_\times \mathbf{H}_{inc}$, où $Z_{\partial\Omega} = 1$ et avec une onde plane électromagnétique incidente sur la frontière du domaine. Ainsi, la solution numérique $\mathbf{E}^h$ doit retranscrire l'onde plane à l'intérieur du domaine et l'erreur relative e^h doit être proche de 0. La solution exacte de l'onde plane est connue analytiquement. Pour cet exemple, elle est de direction $\mathbf{d} = (1,0,0)^\mathsf{T}$ et de polarisation $\mathbf{p} = (0,1,0)^\mathsf{T}$, et nous cherchons son champ électrique défini par

$$\mathbf{E}^{\text{ex}}(\mathbf{x}) := \mathbf{p}\, e^{ik_0\, \mathbf{d}\cdot\mathbf{x}}, \text{ pour } \mathbf{x} \in \mathbb{C}^3.$$

Nous rappelons que le côté d'un cube $T \in \mathcal{T}$ est noté h. Si le schéma converge, la solution numérique $\mathbf{E}^h$ s'approche de la solution exacte $\mathbf{E}^{\text{ex}}$ lorsque h diminue. Les courbes sur la Figure 1.4 montrent la convergence du solveur d'EF de Nédélec d'ordre élevé pour les différents ordres d'approximation r allant de 1 à 4. Les pentes associées à chacune des courbes donnent un taux de convergence de $r+1$, ce qui est cohérent avec la théorie.

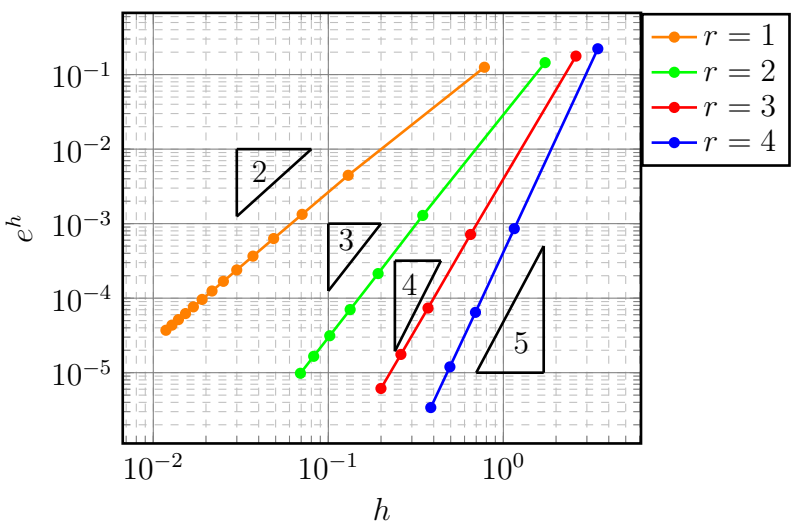

FIGURE 1.4 – Courbes représentant la norme L^2 en fonction de la taille h du côté des éléments T du maillage, pour différents ordres d'approximation $r = 1, 4$.

1.2 Méthode de Galerkin Discontinu

La seconde méthode classique étudiée est une méthode de GD appliquée à un maillage tétraédrique [26, 50]. Plus précisément, nous dérivons la formulation variationnelle associée à une méthode de GD pour le Problème 1 de Maxwell écrit sous sa forme hyperbolique en régime harmonique. Son caractère bien posé nous mène à sa discrétisation à l'aide d'un espace élément fini bien choisi. Le système matriciel obtenu est résolu grâce au solveur MUMPS©. Enfin, nous montrons que la méthode de GD est convergente.

1.2.1 Construction de la formulation variationnelle

Dans le cadre d'une méthode de GD, le schéma numérique met en jeu des fonctions de base discontinues. Nous devons alors mettre en place des stratégies pour assurer la continuité de la solution entre les éléments $T \in \mathcal{T}$. Pour cela, le principe est d'ajouter des termes de pénalisation. Ces derniers peuvent être des traces numériques ou des termes judicieusement choisis assurant le caractère bien posé du problème.

Les formulations du premier ordre permettent d'écrire naturellement des formulations de GD bien posées. Elles utilisent des traces numériques de nature physique de type centrés, upwind, ou encore de Riemann. Il existe aussi des formulations de GD pour les équations du second ordre [18]. Elles sont basées sur des pénalisations intérieures de type naturelles, symétriques ou asymétriques. Mais leurs conditions d'utilisation sont plus délicates. En effet, certains termes de pénalisation introduisent des formulations variationnelles mal posées ou mal conditionnées et provoquent des problèmes de convergence. Par exemple, la version symétrique pour un problème d'ordre 2 nécessite l'ajout d'un terme de pénalisation très grand pour assurer le caractère bien posé du problème. Cependant, la valeur optimale de ce paramètre n'est pas connue, de telle sorte qu'une valeur non adaptée tend plutôt à détériorer le conditionnement du problème. De cette façon, les traces employées dans le cadre d'une formulation hyperbolique contrent ces inconvénients. Elles sont alors de bonnes candidates pour espérer simuler l'onde électromagnétique sur de grands domaines. Pour construire la formulation variationnelle associée à la méthode de GD, nous considérons le Problème 1 de Maxwell adimensionné du premier ordre. En particulier, nous choisissons des traces numériques de Riemann. Nous détaillerons leur construction dans le Chapitre 2.

À nouveau, nous supposons que le domaine Ω est associé à un milieu homogène, *ie* $\varepsilon_r = 1$ et $\mu_r = 1$. De plus, nous introduisons le cadre fonctionnel suivant pour la méthode de GD

$$\mathbb{X} := \prod_{T \in \mathcal{T}} \mathbb{X}_T, \quad \text{avec} \quad \mathbb{X}_T := \mathbb{H}_{\text{imp}}(T).$$

Nous considérons un élément T et sa frontière ∂T. La restriction à l'élément T de la solution

$\mathbb{E} = (\mathbf{E}, \mathbf{H})$ du Problème 1 s'écrit

$$\mathbb{E}^T := (\mathbf{E}^T, \mathbf{H}^T) \in \mathbb{X}_T.$$

De la même façon, la restriction à l'élément T de la fonction test $\mathbb{E}'$ s'écrit

$$\mathbb{E}'^T := (\mathbf{E}'^T, \mathbf{H}'^T) \in \mathbb{X}_T.$$

Tout d'abord, pour un élément T, nous remarquons que pour $\mathbf{E} \in H(\mathrm{rot}, T)$ et $\mathbf{H} \in H(\mathrm{rot}, T)$, tels que $\gamma_t \mathbf{E} \in \mathrm{L}_t^2(\partial\Omega)$ et $\gamma_t \mathbf{H} \in \mathrm{L}_t^2(\partial\Omega)$, la formule de Stokes s'écrit

$$\int_T \nabla \times \mathbf{E} \cdot \mathbf{H} - \int_T \mathbf{E} \cdot \nabla \times \mathbf{H} = -\int_{\partial T} \gamma_t \mathbf{E} \cdot \gamma_\times \mathbf{H}, \tag{1.9}$$

ou de manière équivalente

$$\int_T \nabla \times \mathbf{E} \cdot \mathbf{H} - \int_T \mathbf{E} \cdot \nabla \times \mathbf{H} = \int_{\partial T} \gamma_\times \mathbf{E} \cdot \gamma_t \mathbf{H}, \tag{1.10}$$

où γ_t désigne la composante tangentielle et $\gamma_\times^T$ désigne la trace tangentielle définie par

$$\gamma_\times^T \mathbf{u} = \mathbf{n}_T \times \mathbf{u},$$

avec $\mathbf{u} \in \mathbb{C}^3$ et $\mathbf{n}_T$ la normale unitaire sortante à ∂T. Nous multiplions par $\mathbb{E}'^T \in \mathbb{X}_T$ les équations du système de Maxwell restreintes à un élément T

$$\begin{cases} \displaystyle\int_T \nabla \times \mathbf{E}^T \cdot \overline{\mathbf{H}'^T} + ik_0 \mathbf{H}^T \cdot \overline{\mathbf{H}'^T} &= 0, \\ \displaystyle\int_T -\nabla \times \mathbf{H}^T \cdot \overline{\mathbf{E}'^T} + ik_0 \mathbf{E}^T \cdot \overline{\mathbf{E}'^T} &= 0. \end{cases}$$

En sommant sur les éléments $T \in \mathcal{T}$, nous obtenons

$$\begin{cases} \displaystyle\sum_{T\in\mathcal{T}} \int_T \nabla \times \mathbf{E}^T \cdot \overline{\mathbf{H}'^T} + ik_0 \mathbf{H}^T \cdot \overline{\mathbf{H}'^T} &= 0, \\ \displaystyle\sum_{T\in\mathcal{T}} \int_T -\nabla \times \mathbf{H}^T \cdot \overline{\mathbf{E}'^T} + ik_0 \mathbf{E}^T \cdot \overline{\mathbf{E}'^T} &= 0. \end{cases}$$

Puis nous utilisons les formules de Stokes (1.9) et (1.10)

$$\begin{cases} \displaystyle\sum_{T\in\mathcal{T}} \int_T \mathbf{E}^T \cdot \overline{\nabla \times \mathbf{H}'^T} + ik_0 \mathbf{H}^T \cdot \overline{\mathbf{H}'^T} &= \displaystyle\sum_{T\in\mathcal{T}} \int_{\partial T} \gamma_t \mathbf{E}^T \cdot \gamma_\times^T \overline{\mathbf{H}'^T}, \\ \displaystyle\sum_{T\in\mathcal{T}} \int_T -\mathbf{H}^T \cdot \overline{\nabla \times \mathbf{E}'^T} + ik_0 \mathbf{E}^T \cdot \overline{\mathbf{E}'^T} &= -\displaystyle\sum_{T\in\mathcal{T}} \int_{\partial T} \gamma_t \mathbf{H}^T \cdot \gamma_\times^T \overline{\mathbf{E}'^T}. \end{cases} \tag{1.11}$$

Nous rappelons que la frontière de l'élément T est constituée de faces $F \in \mathcal{F}_T$. Sur chacune de ces faces, nous imposons des traces numériques. Ces dernières nous permettent de rétablir la continuité de la solution numérique et de prendre en compte la condition de bord d'impédance. Les traces sont différentes selon si la face est intérieure ou de bord. Nous choisissons des traces numériques obtenues par un solveur de Riemann, voir [13, 80, 86]. Soit $F \in \mathcal{F}_{\text{int}}$ séparant deux éléments $T \in \mathcal{T}$ et $K \in \mathcal{T}$, nous introduisons les définitions suivantes.

Définition 1.9 (Moyenne et saut)**.** *La moyenne de la composante tangentielle et le saut de la trace tangentielle de $\mathbf{u} \in \mathbb{C}^3$ sur $F \in \mathcal{F}_{\text{int}}$, sont respectivement donnés par*

$$\begin{cases} \{\gamma_t \mathbf{u}\}_F & := & \dfrac{\gamma_t \mathbf{u}^T + \gamma_t \mathbf{u}^K}{2}, \\ [\![\gamma_\times \mathbf{u}]\!]_F & := & \mathbf{n}_T \times \mathbf{u}^T + \mathbf{n}_K \times \mathbf{u}^K, \end{cases}$$

où $\mathbf{n}_T$ (resp. $\mathbf{n}_K$) est la normale unitaire sortante à ∂T (resp. ∂K).

Définition 1.10 (Traces numériques intérieures)**.** *Les traces numériques intérieures de $(\mathbf{E}, \mathbf{H}) \in \mathbb{X}$ notées $\widehat{\gamma_t \mathbf{E}}$ et $\widehat{\gamma_t \mathbf{H}}$ sont définies sur chaque face intérieure $F \in \mathcal{F}_{\text{int}}$ par*

$$(\widehat{\gamma_t \mathbf{E}})_{|F} := \{\gamma_t \mathbf{E}\}_F - \frac{[\![\gamma_\times \mathbf{H}]\!]_F}{2} \quad \textit{et} \quad (\widehat{\gamma_t \mathbf{H}})_{|F} := \{\gamma_t \mathbf{H}\}_F + \frac{[\![\gamma_\times \mathbf{E}]\!]_F}{2},$$

ou de manière équivalente par

$$\begin{cases} (\widehat{\gamma_t \mathbf{E}})_{|F} = \gamma_t \mathbf{E}^T + \dfrac{\mathbf{n}_T \times [\![\gamma_\times \mathbf{E}]\!]_F}{2} - \dfrac{[\![\gamma_\times \mathbf{H}]\!]_F}{2}, \\ (\widehat{\gamma_t \mathbf{H}})_{|F} = \gamma_t \mathbf{H}^T + \dfrac{\mathbf{n}_T \times [\![\gamma_\times \mathbf{H}]\!]_F}{2} + \dfrac{[\![\gamma_\times \mathbf{E}]\!]_F}{2}. \end{cases} \tag{1.12}$$

Remarque 1.8. *Pour une face intérieure $F \in \mathcal{F}_{\text{int}}$, la solution exacte vérifie*

$$(\widehat{\gamma_t \mathbf{E}})_{|F} = (\gamma_t \mathbf{E}^T)_{|F} = (\gamma_t \mathbf{E}^K)_{|F} \quad \textit{et} \quad (\widehat{\gamma_t \mathbf{H}})_{|F} = (\gamma_t \mathbf{H}^T)_{|F} = (\gamma_t \mathbf{H}^K)_{|F}.$$

Nous disons alors que les traces numériques intérieures sont consistantes.

Définition 1.11 (Traces numériques de bord)**.** *Les traces numériques de $\mathbf{E}$ et de $\mathbf{H}$ sur une face de bord $F \in \mathcal{F}_T \cap \mathcal{F}_{\text{ext}}$ sont définies par*

$$\begin{cases} (\widehat{\gamma_t \mathbf{E}})_{|F} & = & \gamma_t \mathbf{E}^T - \dfrac{1}{1 + Z_{\partial\Omega}} (\gamma_t \mathbf{E}^T + Z_{\partial\Omega} \gamma_\times^T \mathbf{H}^T - \mathbf{g}), \\ (\widehat{\gamma_\times^T \mathbf{H}})_{|F} & = & \gamma_\times^T \mathbf{H}^T - \dfrac{1}{1 + Z_{\partial\Omega}} (\gamma_t \mathbf{E}^T + Z_{\partial\Omega} \gamma_\times^T \mathbf{H}^T - \mathbf{g}), \end{cases} \tag{1.13}$$

où $\mathbf{g}$ *est défini dans la condition de bord d'impédance* (1). *De manière équivalente, nous avons*

$$\begin{cases} (\widehat{\gamma_t \mathbf{E}})_{|F} &= \dfrac{1+R_{\partial\Omega}}{2}(\gamma_t \mathbf{E}^T - \gamma_\times^T \mathbf{H}^T) + \dfrac{1-R_{\partial\Omega}}{2}\mathbf{g}, \\ (\widehat{\gamma_t \mathbf{H}})_{|F} &= \dfrac{1-R_{\partial\Omega}}{2}(\gamma_t \mathbf{H}^T + \gamma_\times^T \mathbf{E}^T) - \dfrac{1-R_{\partial\Omega}}{2}(\mathbf{n}_{\partial\Omega} \times \mathbf{g}), \end{cases}$$

où la seconde équation s'écrit aussi

$$(\widehat{\gamma_\times^T \mathbf{H}})_{|F} = \frac{1-R_{\partial\Omega}}{2}(\gamma_\times^T \mathbf{H}^T - \gamma_t \mathbf{E}^T) + \frac{1-R_{\partial\Omega}}{2}\mathbf{g},$$

avec $R_{\partial\Omega}$ *le coefficient de réflexion sur la frontière du domaine. De plus, nous avons*

$$R_{\partial\Omega} = \frac{Z_{\partial\Omega}-1}{Z_{\partial\Omega}+1} \quad \textit{ou de manière équivalente} \quad Z_{\partial\Omega} = \frac{R_{\partial\Omega}-1}{R_{\partial\Omega}+1}.$$

Remarque 1.9. *Pour une face extérieure* $F \in \mathcal{F}_{\text{ext}} \cap \mathcal{F}_T$*, la solution exacte vérifie*

$$(\widehat{\gamma_t \mathbf{E}})_{|F} = (\gamma_t \mathbf{E}^T)_{|F} \quad \textit{et} \quad (\widehat{\gamma_t \mathbf{H}})_{|F} = (\gamma_t \mathbf{H}^T)_{|F}.$$

Nous disons alors que les traces numériques de bord sont consistantes.

Remarque 1.10. *L'obtention des traces numériques définies par* (1.12) *et* (1.13) *est détaillée dans la suite du document grâce à un solveur de Riemann, voir le Chapitre 2.*

Dans le membre de droite de l'équation (1.11), nous décomposons l'intégrale sur ∂T en une somme sur les faces de T. Puis, nous remplaçons les traces $\gamma_t \mathbf{E}^T$ et $\gamma_t \mathbf{H}^T$ par leurs traces numériques, voir les Remarques 1.8 et 1.9,

$$\begin{cases} \displaystyle\sum_{T\in\mathcal{T}} \int_T \mathbf{E}^T \cdot \overline{\nabla \times \mathbf{H}'^T} + ik_0 \mathbf{H}^T \cdot \overline{\mathbf{H}'^T} &= \displaystyle\sum_{T\in\mathcal{T}} \sum_{F\in\mathcal{F}_T} \int_F \widehat{\gamma_t \mathbf{E}} \cdot \gamma_\times^T \overline{\mathbf{H}'^K}, \\ \displaystyle\sum_{T\in\mathcal{T}} \int_T -\mathbf{H}^T \cdot \overline{\nabla \times \mathbf{E}'^T} + ik_0 \mathbf{E}^T \cdot \overline{\mathbf{E}'^T} &= -\displaystyle\sum_{T\in\mathcal{T}} \sum_{F\in\mathcal{F}_T} \int_F \widehat{\gamma_t \mathbf{H}} \cdot \gamma_\times^T \overline{\mathbf{E}'^T}. \end{cases}$$

La dernière étape consiste à appliquer de nouveau la formule de Stokes

$$\begin{cases} \displaystyle\sum_{T\in\mathcal{T}} \int_T \nabla \times \mathbf{E}^T \cdot \overline{\mathbf{H}'^T} + ik_0 \mathbf{H}^T \cdot \overline{\mathbf{H}'^T} &= \mathcal{I}_1, \\ \displaystyle\sum_{T\in\mathcal{T}} \int_T -\nabla \times \mathbf{H}^T \cdot \overline{\mathbf{E}'^T} + ik_0 \mathbf{E}^T \cdot \overline{\mathbf{E}'^T} &= \mathcal{I}_2, \end{cases}$$

où

$$\mathcal{I}_1 := \sum_{T\in\mathcal{T}} \sum_{F\in\mathcal{F}_T} \int_F \left(\widehat{\gamma_t \mathbf{E}} - \gamma_t \mathbf{E}^T\right) \cdot \overline{\gamma_\times^T \mathbf{H}'^T},$$

et

$$\mathcal{I}_2 := -\sum_{T\in\mathcal{T}}\sum_{F\in\mathcal{F}_T}\int_F \left(\widehat{\gamma_t\mathbf{H}} - \gamma_t\mathbf{H}^T\right)\cdot\overline{\gamma_\times^T\mathbf{E}'^T}.$$

Nous rappelons que $\mathbf{n}_T\times\gamma_\times^T\mathbf{u} = -\gamma_t\mathbf{u}$, pour $\mathbf{u}\in\mathbb{C}^3$. En séparant les faces intérieures et extérieures, nous avons

$$\begin{aligned}\mathcal{I}_1 \;=\; & \sum_{T\in\mathcal{T}}\sum_{F\in\mathcal{F}_{\rm int}\cap\mathcal{F}_T}\int_F \left(\widehat{\gamma_t\mathbf{E}} - \gamma_t\mathbf{E}^T\right)\cdot\overline{\gamma_\times^T\mathbf{H}'^T}\\ & + \sum_{T\in\mathcal{T}}\sum_{F\in\mathcal{F}_{\rm ext}\cap\mathcal{F}_T}\int_F \left(\widehat{\gamma_t\mathbf{E}} - \gamma_t\mathbf{E}^T\right)\cdot\overline{\gamma_\times^T\mathbf{H}'^T},\end{aligned} \tag{1.14}$$

et

$$\begin{aligned}\mathcal{I}_2 \;=\; & -\sum_{T\in\mathcal{T}}\sum_{F\in\mathcal{F}_{\rm int}\cap\mathcal{F}_T}\int_F \left(\widehat{\gamma_t\mathbf{H}} - \gamma_t\mathbf{H}^T\right)\cdot\overline{\gamma_\times^T\mathbf{E}'^T}\\ & + \sum_{T\in\mathcal{T}}\sum_{F\in\mathcal{F}_{\rm ext}\cap\mathcal{F}_T}\int_F \left(\widehat{\gamma_\times^T\mathbf{H}} - \gamma_\times^T\mathbf{H}^T\right)\cdot\overline{\gamma_t\mathbf{E}'^T}.\end{aligned} \tag{1.15}$$

Nous introduisons les traces numériques (1.12) et (1.13) dans (1.14)

$$\begin{aligned}\mathcal{I}_1 \;=\; & \sum_{T\in\mathcal{T}}\sum_{F\in\mathcal{F}_{\rm int}\cap\mathcal{F}_T}\int_F \Big(\frac{\mathbf{n}_T\times[\![\gamma_\times\mathbf{E}]\!]_F}{2} - \frac{[\![\gamma_\times\mathbf{H}]\!]_F}{2}\Big)\cdot\overline{\gamma_\times^T\mathbf{H}'^T}\\ & - \sum_{T\in\mathcal{T}}\sum_{F\in\mathcal{F}_{\rm ext}\cap\mathcal{F}_T}\int_F \Big(\frac{1-R_{\partial\Omega}}{2}\gamma_t\mathbf{E}^T + \frac{1+R_{\partial\Omega}}{2}\gamma_\times^T\mathbf{H}^T\Big)\cdot\overline{\gamma_\times^T\mathbf{H}'^T},\\ & + \sum_{T\in\mathcal{T}}\sum_{F\in\mathcal{F}_{\rm ext}\cap\mathcal{F}_T}\int_F \frac{1-R_{\partial\Omega}}{2}\mathbf{g}\cdot\overline{\gamma_\times^T\mathbf{H}'^T},\end{aligned} \tag{1.16a}$$

puis dans (1.15)

$$\begin{aligned}\mathcal{I}_2 \;=\; & -\sum_{T\in\mathcal{T}}\sum_{F\in\mathcal{F}_{\rm int}\cap\mathcal{F}_T}\int_F \Big(\frac{\mathbf{n}_T\times[\![\gamma_\times\mathbf{H}]\!]_F}{2} + \frac{[\![\gamma_\times\mathbf{E}]\!]_F}{2}\Big)\cdot\overline{\gamma_\times^T\mathbf{E}'^T}\\ & - \sum_{T\in\mathcal{T}}\sum_{F\in\mathcal{F}_{\rm ext}\cap\mathcal{F}_T}\int_F \Big(\frac{R_{\partial\Omega}+1}{2}\gamma_\times^T\mathbf{H}^T + \frac{1-R_{\partial\Omega}}{2}\gamma_t\mathbf{E}^T\Big)\cdot\overline{\gamma_t\mathbf{E}'^T}\\ & + \sum_{T\in\mathcal{T}}\sum_{F\in\mathcal{F}_{\rm ext}\cap\mathcal{F}_T}\int_F \frac{1-R_{\partial\Omega}}{2}\mathbf{g}\cdot\overline{\gamma_t\mathbf{E}'^T}.\end{aligned} \tag{1.16b}$$

Finalement, nous utilisons que $(\mathbf{n}_T\times\gamma_\times^T\mathbf{u})\cdot(\mathbf{n}_T\times\gamma_t\mathbf{u}) = \gamma_\times^T\mathbf{u}\cdot\gamma_t\mathbf{u}$, et nous simplifions les intégrales en sommant sur toutes les faces intérieures puis sur toutes les faces extérieures. Nous obtenons alors la formulation suivante.

Problème 5 (Formulation variationnelle de GD)**.** *La formulation variationnelle de la méthode de GD associée au problème de Maxwell du premier ordre est*

$$\textit{Trouver } \mathbb{E}\in\mathbb{X} \textit{ tel que pour tout } \mathbb{E}'\in\mathbb{X}\quad a(\mathbb{E},\mathbb{E}') = \ell(\mathbb{E}'),$$

avec

$$\left\{\begin{array}{rcl} a(\mathbb{E},\mathbb{E}') & = & \displaystyle\sum_{T\in\mathcal{T}} a_T(\mathbb{E},\mathbb{E}') \;+\; \sum_{F\in\mathcal{F}_{\rm int}} b_F^{\rm int}(\mathbb{E},\mathbb{E}') \;+\; \sum_{F\in\mathcal{F}_{\rm ext}} b_F^{\rm ext}(\mathbb{E},\mathbb{E}'), \\ \ell(\mathbb{E}') & = & \displaystyle\int_{\partial\Omega} \frac{1-R_{\partial\Omega}}{2}\Big(\mathbf{g}\cdot\overline{\gamma_t\mathbf{E}'} + \mathbf{g}\cdot\overline{\gamma_\times^T\mathbf{H}'}\Big), \end{array}\right. \tag{1.17}$$

et

$$a_T(\mathbb{E},\mathbb{E}') = \int_T \nabla\times\mathbf{E}^T\cdot\overline{\mathbf{H}'^T} - \nabla\times\mathbf{H}^T\cdot\overline{\mathbf{E}'^T} - ik_0\int_T \mathbf{H}^T\cdot\overline{\mathbf{H}'^T} + \mathbf{E}^T\cdot\overline{\mathbf{E}'^T},$$

$$\begin{array}{rcl} b_F^{\rm int}(\mathbb{E},\mathbb{E}') & = & \displaystyle\int_F [\![\gamma_\times\mathbf{H}]\!]_F\cdot\overline{\{\gamma_t\mathbf{E}'^T\}_F} - [\![\gamma_\times\mathbf{E}]\!]_F\cdot\overline{\{\gamma_t\mathbf{H}'^T\}_F} \\ & + & \displaystyle\int_F \frac{[\![\gamma_\times\mathbf{H}]\!]_F\cdot\overline{[\![\gamma_\times\mathbf{H}']\!]_F} + [\![\gamma_\times\mathbf{E}]\!]_F\cdot\overline{[\![\gamma_\times\mathbf{E}']\!]_F}}{2}, \\ b_F^{\rm ext}(\mathbb{E},\mathbb{E}') & = & \displaystyle\int_F \frac{1-R_{\partial\Omega}}{2}\gamma_t\mathbf{E}^T\cdot\overline{\gamma_t\mathbf{E}'^T} + \frac{1+R_{\partial\Omega}}{2}\gamma_\times^T\mathbf{H}^T\cdot\overline{\gamma_\times^T\mathbf{H}'^T} \\ & + & \displaystyle\int_F \frac{1-R_{\partial\Omega}}{2}\gamma_t\mathbf{E}^T\cdot\overline{\gamma_\times^T\mathbf{H}'^T} + \frac{1+R_{\partial\Omega}}{2}\gamma_\times^T\mathbf{H}^T\cdot\overline{\gamma_t\mathbf{E}'^T}. \end{array} \tag{1.18}$$

Remarque 1.11. *Nous détaillons ci-dessous les calculs pour la forme intérieure $b_F^{\rm int}$, où nous ne considérons que les sommes sur les faces intérieures dans $\mathcal{I}_1$ et $\mathcal{I}_2$. À partir de* (1.16a) *et* (1.16b)*, il suit*

$$\begin{array}{rcl} b_F^{\rm int}(\mathbb{E},\mathbb{E}') & = & \displaystyle\sum_{T\in\mathcal{T}}\sum_{F\in\mathcal{F}_{\rm int}\cap\mathcal{F}_T}\int_F \frac{[\![\gamma_\times\mathbf{H}]\!]_F}{2}\cdot\overline{\gamma_\times^T\mathbf{H}'^T} + \frac{[\![\gamma_\times\mathbf{E}]\!]_F}{2}\cdot\overline{\gamma_\times^T\mathbf{E}'^T} \\ & + & \displaystyle\sum_{T\in\mathcal{T}}\sum_{F\in\mathcal{F}_{\rm int}\cap\mathcal{F}_T}\int_F \frac{\mathbf{n}_T\times[\![\gamma_\times\mathbf{H}]\!]_F}{2}\cdot\overline{\gamma_\times^T\mathbf{E}'^T} - \frac{\mathbf{n}_T\times[\![\gamma_\times\mathbf{E}]\!]_F}{2}\cdot\overline{\gamma_\times^T\mathbf{H}'^T}. \end{array}$$

En appliquant l'opérateur $\mathbf{n}_T\times$ à la deuxième ligne de la formule ci-dessus, nous obtenons

$$\begin{array}{rcl} b_F^{\rm int}(\mathbb{E},\mathbb{E}') & = & \displaystyle\sum_{T\in\mathcal{T}}\sum_{F\in\mathcal{F}_{\rm int}\cap\mathcal{F}_T}\int_F \frac{[\![\gamma_\times\mathbf{H}]\!]_F}{2}\cdot\overline{\gamma_\times^T\mathbf{H}'^T} + \frac{[\![\gamma_\times\mathbf{E}]\!]_F}{2}\cdot\overline{\gamma_\times^T\mathbf{E}'^T} \\ & + & \displaystyle\sum_{T\in\mathcal{T}}\sum_{F\in\mathcal{F}_{\rm int}\cap\mathcal{F}_T}\int_F \frac{[\![\gamma_\times\mathbf{H}]\!]_F\cdot\overline{\gamma_t\mathbf{E}'^T}}{2} - \frac{[\![\gamma_\times\mathbf{E}]\!]_F\cdot\overline{\gamma_t\mathbf{H}'^T}}{2}. \end{array}$$

Lors du parcours des faces intérieures, la face F qui sépare deux éléments T et K est considérée

deux fois (une fois vis à vis de T, une fois vis à vis de K). Ainsi, nous avons

$$\begin{aligned} b_F^{\text{int}}(\mathbb{E},\mathbb{E}') \;=\; & \sum_{F\in\mathcal{F}_{\text{int}}} \int_F \frac{[\![\gamma_\times \mathbf{H}]\!]_F}{2} \cdot \left(\overline{\gamma_\times^T \mathbf{H}'^T} + \overline{\gamma_\times^T \mathbf{H}'^K}\right) \\ + & \sum_{F\in\mathcal{F}_{\text{int}}} \int_F \frac{[\![\gamma_\times \mathbf{E}]\!]_F}{2} \cdot \left(\overline{\gamma_\times^T \mathbf{E}'^T} + \overline{\gamma_\times^T \mathbf{E}'^K}\right) \\ + & \sum_{F\in\mathcal{F}_{\text{int}}} \int_F [\![\gamma_\times \mathbf{H}]\!]_F \cdot \frac{\overline{\gamma_t \mathbf{E}'^T} + \overline{\gamma_t \mathbf{E}'^K}}{2} \\ - & \sum_{F\in\mathcal{F}_{\text{int}}} \int_F [\![\gamma_\times \mathbf{E}]\!]_F \cdot \frac{\overline{\gamma_t \mathbf{H}'^T} + \overline{\gamma_t \mathbf{H}'^K}}{2}. \end{aligned}$$

Cela mène finalement à

$$\begin{aligned} b_F^{\text{int}}(\mathbb{E},\mathbb{E}') \;=\; & \sum_{F\in\mathcal{F}_{\text{int}}} \int_F \frac{[\![\gamma_\times \mathbf{H}]\!]_F \cdot \overline{[\![\gamma_\times \mathbf{H}']\!]_F}}{2} + \frac{[\![\gamma_\times \mathbf{E}]\!]_F \cdot \overline{[\![\gamma_\times \mathbf{E}']\!]_F}}{2} \\ + & \sum_{F\in\mathcal{F}_{\text{int}}} \int_F [\![\gamma_\times \mathbf{H}]\!]_F \cdot \overline{\{\gamma_t \mathbf{E}'^T\}_F} - [\![\gamma_\times \mathbf{E}]\!]_F \cdot \overline{\{\gamma_t \mathbf{H}'^T\}_F}. \end{aligned}$$

Le Problème 5 associé à la méthode de GD est défini. Maintenant, nous nous appuyons sur une base d'EF pour le discrétiser.

1.2.2 Définition de l'espace Élément Fini

Pour discrétiser la formulation variationnelle de GD, nous nous basons sur un espace éléments finis. Nous considérons un maillage tétraédrique, où $\mathcal{T}$ est l'ensemble des tétraèdres. Chaque tétraèdre $T \in \mathcal{T}$ est défini par ses quatre sommets $(\mathbf{x}_0, \mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3)$, voir la Figure 1.5.

Tout d'abord, nous définissons un élément fini de référence $(\widehat{T}, \mathcal{P}_q(\widehat{T}), \Sigma_q(\widehat{T}))$, où $q \in \mathbb{N}$ est l'ordre de la méthode. Plus précisément, une fonction polynomiale $p : \widehat{T} \to \mathbb{C}$ est dans l'ensemble $\mathcal{P}_q(\widehat{T})$ si et seulement si elle vérifie

$$p(\hat{\mathbf{x}}) = \sum_{n_x=0}^{q} \sum_{n_y=0}^{q-n_x} \sum_{n_z=0}^{q-n_y-n_x} a_{n_x,n_y,n_z}\ \hat{x}^{n_x}\ \hat{y}^{n_y}\ \hat{z}^{n_z}, \quad \text{pour } \hat{\mathbf{x}} := (\hat{x},\hat{y},\hat{z}) \in \widehat{T},$$

où p est un polynôme de degré total inférieur ou égal à q et $a_{n_x,n_y,n_z} \in \mathbb{C}$ sont les poids associés à chaque monôme du polynôme.

Définition 1.12 (Élément de référence)**.** *Soit $\widehat{T}$ l'élément de référence. Il est défini par*

$$\widehat{T} := \left\{ \hat{\mathbf{x}} = (\hat{x},\hat{y},\hat{z}) \in \mathbb{R}^3 \;\middle|\; \hat{x}+\hat{y}+\hat{z} \le 1,\ \hat{x} \ge 0,\ \hat{y} \ge 0,\ \hat{z} \ge 0 \right\},$$

avec ses sommets $\hat{\mathbf{x}}_{\mathrm{i}} = \mathbf{e}_{\mathrm{i}}$ pour $1 \le \mathrm{i} \le 3$ et $\hat{\mathbf{x}}_0 = (0,0,0)^{\mathsf{T}}$, où les $\mathbf{e}_{\mathrm{i}}$ sont les vecteurs de la base

canonique.

Nous associons à $\mathcal{P}_q(\widehat{T})$ un ensemble $\Sigma_q(\widehat{T})$ de degrés de liberté. Pour définir les degrés de liberté et les fonctions de base associées, nous introduisons les noeuds de référence.

Définition 1.13 (Noeuds de référence)**.** *Les noeuds locaux de Lagrange* $\hat{\mathbf{x}}_{\mathrm{i}_{\mathrm{loc}1},\mathrm{i}_{\mathrm{loc}2},\mathrm{i}_{\mathrm{loc}3}}$*, indicés par* $(\mathrm{i}_{\mathrm{loc}k})_{k=1,3}$*, sont définis dans l'élément de référence par*

$$\hat{\mathbf{x}}_{\mathrm{i}_{\mathrm{loc}1},\mathrm{i}_{\mathrm{loc}2},\mathrm{i}_{\mathrm{loc}3}} := \Big(\frac{\mathrm{i}_{\mathrm{loc}1}}{q}, \frac{\mathrm{i}_{\mathrm{loc}2}}{q}, \frac{\mathrm{i}_{\mathrm{loc}3}}{q}\Big).$$

Ces points d'interpolation de référence $\hat{\mathbf{x}}_{\mathrm{i}_{\mathrm{loc}1},\mathrm{i}_{\mathrm{loc}2},\mathrm{i}_{\mathrm{loc}3}}$ *satisfont*

$$0 \leq \mathrm{i}_{\mathrm{loc}k} \leq q, \quad \textit{avec} \quad 0 \leq k \leq 3, \quad \textit{et} \quad \mathrm{i}_{\mathrm{loc}0} = q - \mathrm{i}_{\mathrm{loc}1} - \mathrm{i}_{\mathrm{loc}2} - \mathrm{i}_{\mathrm{loc}3},$$

où $q \in \mathbb{N}^*$*, et le cas* $q = 0$ *n'est pas considéré.*

Définition 1.14 (Degrés de liberté)**.** *Les degrés de liberté sont définis aux points* $\hat{\mathbf{x}}_{\mathrm{i}_{\mathrm{loc}1},\mathrm{i}_{\mathrm{loc}2},\mathrm{i}_{\mathrm{loc}3}}$ *par les formes linéaires* $\widehat{\Lambda}_{\mathrm{i}_{\mathrm{loc}1},\mathrm{i}_{\mathrm{loc}2},\mathrm{i}_{\mathrm{loc}3}}$*. L'ensemble des degrés de liberté de* $\widehat{T}$ *est défini par*

$$\Sigma_q(\widehat{T}) := \Big\{(\widehat{\Lambda}_{\mathrm{i}_{\mathrm{loc}1},\mathrm{i}_{\mathrm{loc}2},\mathrm{i}_{\mathrm{loc}3}})_{\mathrm{i}_{\mathrm{loc}1},\mathrm{i}_{\mathrm{loc}2},\mathrm{i}_{\mathrm{loc}3}}, \ \textit{tel que } \widehat{\Lambda}_{\mathrm{i}_{\mathrm{loc}1},\mathrm{i}_{\mathrm{loc}2},\mathrm{i}_{\mathrm{loc}3}}(p) := p(\hat{\mathbf{x}}_{\mathrm{i}_{\mathrm{loc}1},\mathrm{i}_{\mathrm{loc}2},\mathrm{i}_{\mathrm{loc}3}}), \\ \forall p \in \mathcal{P}_q(\widehat{T}) \textit{ et } 0 \leq \mathrm{i}_{\mathrm{loc}k} \leq q, \textit{ pour } 0 \leq k \leq 3\Big\}.$$

Remarque 1.12. *[Ensemble des degrés de liberté d'un élément quelconque] On définit aussi plus généralement l'ensemble des degrés de liberté d'un élément quelconque* T *par*

$$\Sigma_q(T) := \Big\{(\Lambda_{\mathrm{i}_{\mathrm{loc}1},\mathrm{i}_{\mathrm{loc}2},\mathrm{i}_{\mathrm{loc}3}})_{\mathrm{i}_{\mathrm{loc}1},\mathrm{i}_{\mathrm{loc}2},\mathrm{i}_{\mathrm{loc}3}}, \quad \textit{tel que } \Lambda_{\mathrm{i}_{\mathrm{loc}1},\mathrm{i}_{\mathrm{loc}2},\mathrm{i}_{\mathrm{loc}3}}(p) := p(\hat{\mathbf{x}}_{\mathrm{i}_{\mathrm{loc}1},\mathrm{i}_{\mathrm{loc}2},\mathrm{i}_{\mathrm{loc}3}}), \\ \forall p \in \mathcal{P}_q(T) \textit{ et } 0 \leq \mathrm{i}_{\mathrm{loc}k} \leq q, \textit{ pour } 0 \leq k \leq 3\Big\}.$$

Définition 1.15 (Base de référence)**.** *Les fonctions de base scalaires associées à l'élément fini* $\Big(\widehat{T}, \mathcal{P}_q(\widehat{T}), \Sigma_q(\widehat{T})\Big)$ *sont définies par*

$$\widehat{\Lambda}_{\mathrm{i}_{\mathrm{loc}1},\mathrm{i}_{\mathrm{loc}2},\mathrm{i}_{\mathrm{loc}3}}(\hat{\phi}_{\mathrm{j}_{\mathrm{loc}1},\mathrm{j}_{\mathrm{loc}2},\mathrm{j}_{\mathrm{loc}3}}) := \delta_{\mathrm{i}_{\mathrm{loc}1},\mathrm{j}_{\mathrm{loc}1}}\delta_{\mathrm{i}_{\mathrm{loc}2},\mathrm{j}_{\mathrm{loc}2}}\delta_{\mathrm{i}_{\mathrm{loc}3},\mathrm{j}_{\mathrm{loc}3}}.$$

Plus précisément, les fonctions $\hat{\phi}_{\mathrm{j}_{\mathrm{loc}1},\mathrm{j}_{\mathrm{loc}2},\mathrm{j}_{\mathrm{loc}3}}$ *sont les polynômes de Lagrange scalaires définis par*

$$\hat{\phi}_{\mathrm{j}_{\mathrm{loc}1},\mathrm{j}_{\mathrm{loc}2},\mathrm{j}_{\mathrm{loc}3}}(\hat{\mathbf{x}}_{\mathrm{j}_{\mathrm{loc}1},\mathrm{j}_{\mathrm{loc}2},\mathrm{j}_{\mathrm{loc}3}}) \ = \ \hat{\phi}_{\mathrm{j}_{\mathrm{loc}0}}(1-\hat{x}-\hat{y}-\hat{z})\hat{\phi}_{\mathrm{j}_{\mathrm{loc}1}}(\hat{x})\hat{\phi}_{\mathrm{j}_{\mathrm{loc}2}}(\hat{y})\hat{\phi}_{\mathrm{j}_{\mathrm{loc}3}}(\hat{z}),$$

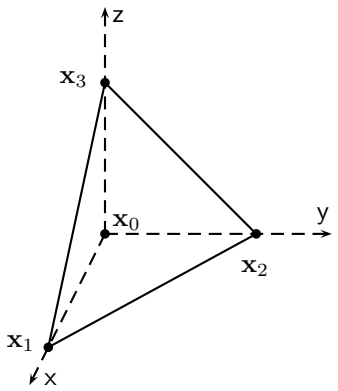

FIGURE 1.5 – Élément quelconque et ses sommets $(\mathbf{x}_0, \mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3)$.

avec les fonctions polynomiales de Lagrange 1D

$$\hat{\phi}_{\mathrm{k}}(\hat{x}) = \prod_{\mathrm{i}=0}^{\mathrm{k}-1} \frac{\hat{x} - \dfrac{\mathrm{i}}{q}}{\dfrac{\mathrm{k}}{q} - \dfrac{\mathrm{i}}{q}}, \quad \textit{pour} \quad \mathrm{k} \geq 1 \quad \textit{et} \quad \hat{\phi}_0(\hat{x}) = 1.$$

Remarque 1.13. *La fonction polynomiale de Lagrange $\hat{\phi}_{\mathrm{i_{loc1}},\mathrm{i_{loc2}},\mathrm{i_{loc3}}}$ est égale à 1 au point d'interpolation $\hat{\mathbf{x}}_{\mathrm{i_{loc1}},\mathrm{i_{loc2}},\mathrm{i_{loc3}}}$ et est égale à 0 à tous les autres points de référence.*

Le lien entre l'élément de référence et un élément arbitraire, de numéro $\mathrm{i_{elem}}$, est défini comme suit.

Définition 1.16. *Chaque élément quelconque T, de sommets $(\mathbf{x}_0, \mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3)$, est l'image de $\hat{T}$ par l'application affine $\overrightarrow{\psi}_T$ définie par*

$$\overrightarrow{\psi}_T(\hat{\mathbf{x}}_{\mathrm{i}}) = \mathbf{x}_{\mathrm{i}} \quad \textit{avec} \quad \mathbf{x}_{\mathrm{i}} := (x_i, y_i, z_i) \quad \textit{et} \quad \mathbf{x} = \overrightarrow{\psi}_T(\hat{\mathbf{x}}).$$

Nous calculons aussi la matrice jacobienne de cette transformation géométrique

$$J_{\overrightarrow{\psi}_T} := \begin{pmatrix} \dfrac{\partial x}{\partial \hat{x}} & \dfrac{\partial x}{\partial \hat{y}} & \dfrac{\partial x}{\partial \hat{z}} \\ \dfrac{\partial y}{\partial \hat{x}} & \dfrac{\partial y}{\partial \hat{y}} & \dfrac{\partial y}{\partial \hat{z}} \\ \dfrac{\partial z}{\partial \hat{x}} & \dfrac{\partial z}{\partial \hat{y}} & \dfrac{\partial z}{\partial \hat{z}} \end{pmatrix} = \begin{pmatrix} x_1 - x_0 & x_2 - x_0 & x_3 - x_0 \\ y_1 - y_0 & y_2 - y_0 & y_3 - y_0 \\ z_1 - z_0 & z_2 - z_0 & z_3 - z_0 \end{pmatrix}.$$

Ces fonctions vectorielles sont associées à chaque degré de liberté numéroté (localement ou globalement). Nous utilisons donc deux espaces éléments finis. Le premier est local, relatif

à un élément T. Le second est global, relatif au maillage. Ils sont respectivement définis par

$$\mathbb{X}^{\mathrm{loc}}_{q,\mathrm{i_{elem}}}(T) := (\mathcal{P}_q(T))^6 \quad \text{et} \quad \mathbb{X}^{\mathrm{glob}}_{q} := \prod_{T\in\mathcal{T}} \mathbb{X}^{\mathrm{loc}}_{q,\mathrm{i_{elem}}}(T).$$

Plus précisémment, nous définissons $\mathbb{X}^{\mathrm{loc}}_{q,\mathrm{i_{elem}}}$ (*resp.* $\mathbb{X}^{\mathrm{glob}}_{q}$), composé de toutes les fonctions de base locales (*resp.* globales), pour $\mathrm{k} = 1, 6$.

Définition 1.17. *L'espace de fonctions de base locales* $\mathbb{X}^{\mathrm{loc}}_{q,\mathrm{i_{elem}}}$ *prend la forme*

$$\mathbb{X}^{\mathrm{loc}}_{q,\mathrm{i_{elem}}} := \mathrm{span}\Big(\{\overrightarrow{\phi}^{\mathrm{i_{elem}},\mathrm{k}}_{\mathrm{i_{loc}}} \ : \ \mathrm{k} = 1,6 \textit{ et } \mathrm{i_{loc}} = 1, N_q\}\Big),$$

où N_q *est le nombre de degrés de liberté scalaires de l'élément* T.

Définition 1.18. *L'espace de fonctions de base globales* $\mathbb{X}^{\mathrm{glob}}_{q}$ *prend la forme*

$$\mathbb{X}^{\mathrm{glob}}_{q} := \mathrm{span}\Big(\{\overrightarrow{\phi}_{\mathrm{i_{glob}}} \in \Big[\mathrm{L}^2(\Omega)\Big]^6 \ : \ \forall \mathrm{i_{glob}} = \mathrm{LocToGlob}(\mathrm{i_{elem}}, \mathrm{i_{loc}}, \mathrm{k}),$$
$$\textit{supp}\left(\overrightarrow{\phi}_{\mathrm{i_{glob}}}\right) = T \textit{ et } \left(\overrightarrow{\phi}_{\mathrm{i_{glob}}}\right)_{|_T} = \overrightarrow{\phi}^{\mathrm{i_{elem}},\mathrm{k}}_{\mathrm{i_{loc}}}\}\Big),$$

où $\mathrm{i_{glob}}$*, le numéro global du degré de liberté, est associé à : l'élément* T *de numéro* $\mathrm{i_{elem}}$*, au numéro local du degré de liberté* $\mathrm{i_{loc}}$*, pour* $\mathrm{i_{loc}} = 1, N_q$*, et aux composantes* $k = 1, 6$.

La bijection entre $(\mathrm{i_{elem}}, \mathrm{i_{loc}}, k)$ et le numéro global du degré de liberté $\mathrm{i_{glob}}$ est représentée par la table de connectivité construite à l'aide de l'algorithme suivant

$\mathrm{i_{glob}} = 1$
do $\mathrm{elem} = 1, \#\mathrm{elem}$
  do $\mathrm{i_{loc}} = 1, N_q$
    do $k = 1, 6$
      $\mathrm{LocToGlob}(\mathrm{i_{elem}}, \mathrm{i_{loc}}, k) = \mathrm{i_{glob}}$
      $\mathrm{i_{glob}} = \mathrm{i_{glob}} + 1$
    end do
  end do
end do

L'approximation numérique de la solution dans chaque élément T, notée $\mathbb{E}^{h,T} = (\mathbf{E}^{h,T}, \mathbf{H}^{h,T})$,

est alors définie par :

$$\mathbb{E}^{h,T} = \sum_{\mathrm{k}=1}^{6} \sum_{\mathrm{i_{loc}}=1}^{N_q} \mathbb{E}^{\mathrm{i_{elem},k}}_{\mathrm{i_{loc}}} \overrightarrow{\phi}^{\mathrm{i_{elem},k}}_{\mathrm{i_{loc}}} = \sum_{\mathrm{i_{loc}}=1}^{N_q} \begin{pmatrix} \mathbf{E}^{\mathrm{i_{elem},x}}_{\mathrm{i_{loc}}} \\ \mathbf{E}^{\mathrm{i_{elem},y}}_{\mathrm{i_{loc}}} \\ \mathbf{E}^{\mathrm{i_{elem},z}}_{\mathrm{i_{loc}}} \\ \mathbf{H}^{\mathrm{i_{elem},x}}_{\mathrm{i_{loc}}} \\ \mathbf{H}^{\mathrm{i_{elem},y}}_{\mathrm{i_{loc}}} \\ \mathbf{H}^{\mathrm{i_{elem},z}}_{\mathrm{i_{loc}}} \end{pmatrix} \phi^{\mathrm{i_{elem}}}_{\mathrm{i_{loc}}},$$

où nous avons

$$\begin{cases} \mathbb{E}^{\mathrm{i_{elem},1}}_{\mathrm{i_{loc}}} = \mathbf{E}^{\mathrm{i_{elem},x}}_{\mathrm{i_{loc}}}, \\ \mathbb{E}^{\mathrm{i_{elem},2}}_{\mathrm{i_{loc}}} = \mathbf{E}^{\mathrm{i_{elem},y}}_{\mathrm{i_{loc}}}, \\ \mathbb{E}^{\mathrm{i_{elem},3}}_{\mathrm{i_{loc}}} = \mathbf{E}^{\mathrm{i_{elem},z}}_{\mathrm{i_{loc}}}, \\ \mathbb{E}^{\mathrm{i_{elem},4}}_{\mathrm{i_{loc}}} = \mathbf{H}^{\mathrm{i_{elem},x}}_{\mathrm{i_{loc}}}, \\ \mathbb{E}^{\mathrm{i_{elem},5}}_{\mathrm{i_{loc}}} = \mathbf{H}^{\mathrm{i_{elem},y}}_{\mathrm{i_{loc}}}, \\ \mathbb{E}^{\mathrm{i_{elem},6}}_{\mathrm{i_{loc}}} = \mathbf{H}^{\mathrm{i_{elem},z}}_{\mathrm{i_{loc}}}, \end{cases}$$

et où les composantes selon l'axe x (*resp.* y ou z) de $\mathbb{R}^3$ sont notées $\mathbf{E}^{\mathrm{i_{elem},x}}_{\mathrm{i_{loc}}}$ et $\mathbf{H}^{\mathrm{i_{elem},x}}_{\mathrm{i_{loc}}}$ (*resp.* $\mathbf{E}^{\mathrm{i_{elem},y}}_{\mathrm{i_{loc}}}$ et $\mathbf{H}^{\mathrm{i_{elem},y}}_{\mathrm{i_{loc}}}$, ou $\mathbf{E}^{\mathrm{i_{elem},z}}_{\mathrm{i_{loc}}}$ et $\mathbf{H}^{\mathrm{i_{elem},z}}_{\mathrm{i_{loc}}}$). Nous avons aussi pour les restrictions au champ électrique et au champ magnétique

$$\mathbf{E}^{h,T}(\mathbf{x}) = \sum_{\mathrm{i_{loc}}=1}^{N_q} \begin{pmatrix} \mathbf{E}^{\mathrm{i_{elem},x}}_{\mathrm{i_{loc}}} \\ \mathbf{E}^{\mathrm{i_{elem},y}}_{\mathrm{i_{loc}}} \\ \mathbf{E}^{\mathrm{i_{elem},z}}_{\mathrm{i_{loc}}} \end{pmatrix} \phi^{\mathrm{i_{elem}}}_{\mathrm{i_{loc}}}(\mathbf{x}) \quad \text{pour } \mathbf{x} \in T,$$

$$\mathbf{H}^{h,T}(\mathbf{x}) = \sum_{\mathrm{i_{loc}}=1}^{N_q} \begin{pmatrix} \mathbf{H}^{\mathrm{i_{elem},x}}_{\mathrm{i_{loc}}} \\ \mathbf{H}^{\mathrm{i_{elem},y}}_{\mathrm{i_{loc}}} \\ \mathbf{H}^{\mathrm{i_{elem},z}}_{\mathrm{i_{loc}}} \end{pmatrix} \phi^{\mathrm{i_{elem}}}_{\mathrm{i_{loc}}}(\mathbf{x}) \quad \text{pour } \mathbf{x} \in T.$$

1.2.3 Discrétisation de la formulation

Nous cherchons dans cette partie à décrire la matrice $\mathbf{A}$, le vecteur $\mathbf{F}$ et le vecteur solution $[\mathbb{E}^h]$. Ce dernier prend la forme

$$[\mathbb{E}^h] := \begin{pmatrix} \mathbb{E}_1 \\ \mathbb{E}_2 \\ \vdots \\ \vdots \\ \vdots \\ \mathbb{E}_{\#\text{ddl}} \end{pmatrix} := \begin{pmatrix} \mathbb{E}^{T_1} \\ \mathbb{E}^{T_2} \\ \vdots \\ \mathbb{E}^{T_{\#\text{elem}}} \end{pmatrix}, \quad \text{avec} \quad \mathbb{E}^{T_{\mathrm{i}_{\text{elem}}}} := \begin{pmatrix} \mathbf{E}^{T_{\mathrm{i}_{\text{elem}}},\mathsf{x}} \\ \mathbf{E}^{T_{\mathrm{i}_{\text{elem}}},\mathsf{y}} \\ \mathbf{E}^{T_{\mathrm{i}_{\text{elem}}},\mathsf{z}} \\ \mathbf{H}^{T_{\mathrm{i}_{\text{elem}}},\mathsf{x}} \\ \mathbf{H}^{T_{\mathrm{i}_{\text{elem}}},\mathsf{y}} \\ \mathbf{H}^{T_{\mathrm{i}_{\text{elem}}},\mathsf{z}} \end{pmatrix} \in \mathbb{C}^{\#\text{ddlelem}},$$

où les composantes x, y et z du champ électrique

$$\mathbf{E}^{T_{\mathrm{i}_{\text{elem}}}} := \left(\mathbf{E}^{T_{\mathrm{i}_{\text{elem}}},\mathsf{x}}, \mathbf{E}^{T_{\mathrm{i}_{\text{elem}}},\mathsf{y}}, \mathbf{E}^{T_{\mathrm{i}_{\text{elem}}},\mathsf{z}}\right)^{\mathsf{T}} \in \mathbb{C}^{3\,N_q},$$

sont définies par

$$\begin{aligned} \mathbf{E}^{T_{\mathrm{i}_{\text{elem}}},\mathsf{x}} &:= \left(\mathbf{E}_{\mathrm{i}_{\text{loc}}}^{T_{\mathrm{i}_{\text{elem}}},\mathsf{x}}\right)_{\mathrm{i}_{\text{loc}}=1,N_q}, \\ \mathbf{E}^{T_{\mathrm{i}_{\text{elem}}},\mathsf{y}} &:= \left(\mathbf{E}_{\mathrm{i}_{\text{loc}}}^{T_{\mathrm{i}_{\text{elem}}},\mathsf{y}}\right)_{\mathrm{i}_{\text{loc}}=1,N_q}, \\ \mathbf{E}^{T_{\mathrm{i}_{\text{elem}}},\mathsf{z}} &:= \left(\mathbf{E}_{\mathrm{i}_{\text{loc}}}^{T_{\mathrm{i}_{\text{elem}}},\mathsf{z}}\right)_{\mathrm{i}_{\text{loc}}=1,N_q}, \end{aligned}$$

et celles du champ magnétique

$$\mathbf{H}^{T_{\mathrm{i}_{\text{elem}}}} := \left(\mathbf{H}^{T_{\mathrm{i}_{\text{elem}}},\mathsf{x}}, \mathbf{H}^{T_{\mathrm{i}_{\text{elem}}},\mathsf{y}}, \mathbf{H}^{T_{\mathrm{i}_{\text{elem}}},\mathsf{z}}\right)^{\mathsf{T}} \in \mathbb{C}^{3\,N_q},$$

sont définies par

$$\begin{aligned} \mathbf{H}^{T_{\mathrm{i}_{\text{elem}}},\mathsf{x}} &:= \left(\mathbf{H}_{\mathrm{i}_{\text{loc}}}^{T_{\mathrm{i}_{\text{elem}}},\mathsf{x}}\right)_{\mathrm{i}_{\text{loc}}=1,N_q}, \\ \mathbf{H}^{T_{\mathrm{i}_{\text{elem}}},\mathsf{y}} &:= \left(\mathbf{H}_{\mathrm{i}_{\text{loc}}}^{T_{\mathrm{i}_{\text{elem}}},\mathsf{y}}\right)_{\mathrm{i}_{\text{loc}}=1,N_q}, \\ \mathbf{H}^{T_{\mathrm{i}_{\text{elem}}},\mathsf{z}} &:= \left(\mathbf{H}_{\mathrm{i}_{\text{loc}}}^{T_{\mathrm{i}_{\text{elem}}},\mathsf{z}}\right)_{\mathrm{i}_{\text{loc}}=1,N_q}, \end{aligned}$$

où nous donnons un exemple détaillé pour la composante x

$$\mathbf{E}^{T_{\mathrm{i}_{\text{elem}}},\mathsf{x}} = \begin{pmatrix} \left(\mathbf{E}^{T_{\mathrm{i}_{\text{elem}}},\mathsf{x}}\right)_1 \\ \left(\mathbf{E}^{T_{\mathrm{i}_{\text{elem}}},\mathsf{x}}\right)_2 \\ \vdots \\ \left(\mathbf{E}^{T_{\mathrm{i}_{\text{elem}}},\mathsf{x}}\right)_{N_q} \end{pmatrix} := \begin{pmatrix} E_1^{T_{\mathrm{i}_{\text{elem}}},\mathsf{x}} \\ E_2^{T_{\mathrm{i}_{\text{elem}}},\mathsf{x}} \\ \vdots \\ E_{N_q}^{T_{\mathrm{i}_{\text{elem}}},\mathsf{x}} \end{pmatrix},$$

et

$$\mathbf{H}^{T_{\mathrm{i_{elem}}},\times} = \begin{pmatrix} (\mathbf{H}^{T_{\mathrm{i_{elem}}},\times})_1 \\ (\mathbf{H}^{T_{\mathrm{i_{elem}}},\times})_2 \\ \vdots \\ (\mathbf{H}^{T_{\mathrm{i_{elem}}},\times})_{N_q} \end{pmatrix} := \begin{pmatrix} H_1^{T_{\mathrm{i_{elem}}},\times} \\ H_2^{T_{\mathrm{i_{elem}}},\times} \\ \vdots \\ H_{N_q}^{T_{\mathrm{i_{elem}}},\times} \end{pmatrix}.$$

Remarque 1.14. *Par convention, nous considérons que nous avons toujours l'indice* $\mathrm{i_{elem}}$ *associé à l'élément* T*, sauf mention contraire.*

Nous prenons comme fonction test $\overrightarrow{\phi}_{\mathrm{i_{glob}}}$. Nous avons

$$\sum_{\mathrm{j_{glob}}=1}^{\#\mathrm{ddl}} \mathbb{E}_{\mathrm{j_{glob}}}\, a(\overrightarrow{\phi}_{\mathrm{j_{glob}}}, \overrightarrow{\phi}_{\mathrm{i_{glob}}}) = \ell(\overrightarrow{\phi}_{\mathrm{i_{glob}}}), \quad \text{pour } \mathrm{i_{glob}} = 1, \#\mathrm{ddl}.$$

Ces équations induisent un système linéaire de la forme $\mathbf{A}[\mathbb{E}^h] = \mathbf{F}$ défini par :

$$\mathbf{A}_{\mathrm{i_{glob}},\mathrm{j_{glob}}} = a(\overrightarrow{\phi}_{\mathrm{j_{glob}}}, \overrightarrow{\phi}_{\mathrm{i_{glob}}}) \quad \text{et} \quad \mathbf{F}_{\mathrm{i_{glob}}} = \ell(\overrightarrow{\phi}_{\mathrm{i_{glob}}}) \quad \text{avec } \mathrm{i_{glob}}, \mathrm{j_{glob}} = 1, \#\mathrm{ddl}.$$

Nous rappelons la forme sesquilinéaire définie en (1.17)

$$a(\mathbb{E}, \mathbb{E}') = \sum_{T\in\mathcal{T}} a_T(\mathbb{E}, \mathbb{E}') + \sum_{F\in\mathcal{F}_{\mathrm{int}}} b_F^{\mathrm{int}}(\mathbb{E}, \mathbb{E}') + \sum_{F\in\mathcal{F}_{\mathrm{ext}}} b_F^{\mathrm{ext}}(\mathbb{E}, \mathbb{E}').$$

La matrice $\mathbf{A}$ est décomposée de la même façon.

Définition 1.19 (Matrice globale)**.** *La matrice globale* $\mathbf{A}$ *est définie par*

$$\mathbf{A} := \sum_{T\in\mathcal{T}} \mathbf{A}_T + \sum_{F\in\mathcal{F}_{\mathrm{int}}} \mathbf{B}_F^{\mathrm{int}} + \sum_{F\in\mathcal{F}_{\mathrm{ext}}} \mathbf{B}_F^{\mathrm{ext}}, \tag{1.19}$$

où les différentes matrices sont définies par

$$\begin{aligned}
(\mathbf{A}_T)_{\mathrm{i_{glob}},\mathrm{j_{glob}}} &:= a_T(\overrightarrow{\phi}_{\mathrm{j_{glob}}}, \overrightarrow{\phi}_{\mathrm{i_{glob}}}) = a_T(\overrightarrow{\phi}^{\mathrm{j_{elem}},\mathrm{k_j}}_{\mathrm{j_{loc}}}, \overrightarrow{\phi}^{\mathrm{i_{elem}},\mathrm{k_i}}_{\mathrm{i_{loc}}}),\\
(\mathbf{B}_F^{\mathrm{int}})_{\mathrm{i_{glob}},\mathrm{j_{glob}}} &:= b_F^{\mathrm{int}}(\overrightarrow{\phi}_{\mathrm{j_{glob}}}, \overrightarrow{\phi}_{\mathrm{i_{glob}}}) = b_F^{\mathrm{int}}(\overrightarrow{\phi}^{\mathrm{j_{elem}},\mathrm{k_j}}_{\mathrm{j_{loc}}}, \overrightarrow{\phi}^{\mathrm{i_{elem}},\mathrm{k_i}}_{\mathrm{i_{loc}}}),\\
(\mathbf{B}_F^{\mathrm{ext}})_{\mathrm{i_{glob}},\mathrm{j_{glob}}} &:= b_F^{\mathrm{ext}}(\overrightarrow{\phi}_{\mathrm{j_{glob}}}, \overrightarrow{\phi}_{\mathrm{i_{glob}}}) = b_F^{\mathrm{ext}}(\overrightarrow{\phi}^{\mathrm{j_{elem}},\mathrm{k_j}}_{\mathrm{j_{loc}}}, \overrightarrow{\phi}^{\mathrm{i_{elem}},\mathrm{k_i}}_{\mathrm{i_{loc}}}),
\end{aligned}$$

avec

$$\mathrm{i_{glob}} = \mathrm{LocToGlob}(\mathrm{i_{elem}}, \mathrm{i_{loc}}, \mathrm{k_i}) \quad \textit{et} \quad \mathrm{j_{glob}} = \mathrm{LocToGlob}(\mathrm{j_{elem}}, \mathrm{j_{loc}}, \mathrm{k_j}),$$

où $\mathrm{i_{elem}}, \mathrm{j_{elem}} = 1, \#\mathrm{elem}$*,* $\mathrm{i_{loc}}, \mathrm{j_{loc}} = 1, N_q$ *et* $\mathrm{k_i}, \mathrm{k_j} = 1, 6$*.*

Dans le but de définir les matrices globales de (1.19), nous introduisons des matrices élémentaires de masse et de rigidité associées à un élément T.

Définition 1.20 (Matrice de masse scalaire)**.** *La matrice de masse scalaire associée à T est*

$$(\mathbf{M}^{\mathrm{i_{elem}}})_{\mathrm{i_{loc}},\mathrm{j_{loc}}} := \mathbf{M}^{\mathrm{i_{elem}}}_{\mathrm{i_{loc}},\mathrm{j_{loc}}} := \int_T \phi^{\mathrm{i_{elem}}}_{\mathrm{i_{loc}}}\ \phi^{\mathrm{i_{elem}}}_{\mathrm{j_{loc}}} \quad \textit{pour } \mathrm{i_{loc}},\mathrm{j_{loc}} = 1, N_q.$$

Remarque 1.15. *Pour $\mathrm{i_{glob}},\mathrm{j_{glob}} = 1, \#\mathrm{ddl}$, la matrice de masse scalaire permet de définir une matrice de masse globale $\mathbf{M}$ qui est bloc diagonale. Ce constat se déduit rapidement grâce au support local des fonctions de base globales. Ainsi, la méthode de GD peut s'avérer attractive pour l'étude de problèmes temporels [26].*

Définition 1.21 (Matrices de rigidité scalaires)**.** *Les matrices de rigidité scalaires, de dimension $N_q \times N_q$, sont*

$$\begin{cases} (\mathbf{R}^{\mathrm{i_{elem}},1})_{\mathrm{i_{loc}},\mathrm{j_{loc}}} = \mathbf{R}^{\mathrm{i_{elem}},1}_{\mathrm{i_{loc}},\mathrm{j_{loc}}} = \displaystyle\int_T \phi^{\mathrm{i_{elem}}}_{\mathrm{i_{loc}}} \frac{\partial\, \phi^{\mathrm{i_{elem}}}_{\mathrm{j_{loc}}}}{\partial \mathsf{x}}, & \textit{pour } \mathrm{i_{loc}},\mathrm{j_{loc}} = 1, N_q, \\ (\mathbf{R}^{\mathrm{i_{elem}},2})_{\mathrm{i_{loc}},\mathrm{j_{loc}}} = \mathbf{R}^{\mathrm{i_{elem}},2}_{\mathrm{i_{loc}},\mathrm{j_{loc}}} = \displaystyle\int_T \phi^{\mathrm{i_{elem}}}_{\mathrm{i_{loc}}} \frac{\partial\, \phi^{\mathrm{i_{elem}}}_{\mathrm{j_{loc}}}}{\partial \mathsf{y}}, & \textit{pour } \mathrm{i_{loc}},\mathrm{j_{loc}} = 1, N_q, \\ (\mathbf{R}^{\mathrm{i_{elem}},3})_{\mathrm{i_{loc}},\mathrm{j_{loc}}} = \mathbf{R}^{\mathrm{i_{elem}},3}_{\mathrm{i_{loc}},\mathrm{j_{loc}}} = \displaystyle\int_T \phi^{\mathrm{i_{elem}}}_{\mathrm{i_{loc}}} \frac{\partial\, \phi^{\mathrm{i_{elem}}}_{\mathrm{j_{loc}}}}{\partial \mathsf{z}}, & \textit{pour } \mathrm{i_{loc}},\mathrm{j_{loc}} = 1, N_q. \end{cases}$$

À partir de ces matrices élémentaires, nous pouvons plus facilement expliciter les termes de la matrice $\mathbf{A}_T$.

Définition 1.22 (Matrice des termes volumiques)**.** *La matrice $\mathbf{A}_T$ est donnée par*

$$a_T(\overrightarrow{\phi}_{\mathrm{j_{glob}}}, \overrightarrow{\phi}_{\mathrm{i_{glob}}}) = a_T(\overrightarrow{\phi}^{\mathrm{j_{elem}},\mathrm{k_j}}_{\mathrm{j_{loc}}}, \overrightarrow{\phi}^{\mathrm{i_{elem}},\mathrm{k_i}}_{\mathrm{i_{loc}}}) = 0,$$

sauf si $T_{\mathrm{i_{elem}}} = T_{\mathrm{j_{elem}}} = T$, voir la Figure 1.6. Autrement dit, tous les termes de la matrice sont

nuls, à l'exception des termes suivants

$$\left\{\begin{array}{rcll}
a_T(\overrightarrow{\phi}^{\mathrm{j_{elem}},\mathrm{k_j}}_{\mathrm{j_{loc}}}, \overrightarrow{\phi}^{\mathrm{i_{elem}},\mathrm{k_i}}_{\mathrm{i_{loc}}}) &=& -ik_0\mathbf{M}^{\mathrm{i_{elem}}}_{\mathrm{i_{loc}},\mathrm{j_{loc}}} & \textit{pour } 1 \leq \mathrm{k_j}, \mathrm{k_i} \leq 6,\\
a_T(\overrightarrow{\phi}^{\mathrm{j_{elem}},1}_{\mathrm{j_{loc}}}, \overrightarrow{\phi}^{\mathrm{i_{elem}},5}_{\mathrm{i_{loc}}}) &=& \mathbf{R}^{\mathrm{i_{elem}},3}_{\mathrm{i_{loc}},\mathrm{j_{loc}}},\\
a_T(\overrightarrow{\phi}^{\mathrm{j_{elem}},1}_{\mathrm{j_{loc}}}, \overrightarrow{\phi}^{\mathrm{i_{elem}},6}_{\mathrm{i_{loc}}}) &=& -\mathbf{R}^{\mathrm{i_{elem}},2}_{\mathrm{i_{loc}},\mathrm{j_{loc}}},\\
a_T(\overrightarrow{\phi}^{\mathrm{j_{elem}},2}_{\mathrm{j_{loc}}}, \overrightarrow{\phi}^{\mathrm{i_{elem}},4}_{\mathrm{i_{loc}}}) &=& -\mathbf{R}^{\mathrm{i_{elem}},3}_{\mathrm{i_{loc}},\mathrm{j_{loc}}},\\
a_T(\overrightarrow{\phi}^{\mathrm{j_{elem}},2}_{\mathrm{j_{loc}}}, \overrightarrow{\phi}^{\mathrm{i_{elem}},6}_{\mathrm{i_{loc}}}) &=& \mathbf{R}^{\mathrm{i_{elem}},1}_{\mathrm{i_{loc}},\mathrm{j_{loc}}},\\
a_T(\overrightarrow{\phi}^{\mathrm{j_{elem}},3}_{\mathrm{j_{loc}}}, \overrightarrow{\phi}^{\mathrm{i_{elem}},4}_{\mathrm{i_{loc}}}) &=& \mathbf{R}^{\mathrm{i_{elem}},2}_{\mathrm{i_{loc}},\mathrm{j_{loc}}},\\
a_T(\overrightarrow{\phi}^{\mathrm{j_{elem}},3}_{\mathrm{j_{loc}}}, \overrightarrow{\phi}^{\mathrm{i_{elem}},5}_{\mathrm{i_{loc}}}) &=& -\mathbf{R}^{\mathrm{i_{elem}},1}_{\mathrm{i_{loc}},\mathrm{j_{loc}}},\\
a_T(\overrightarrow{\phi}^{\mathrm{j_{elem}},4}_{\mathrm{j_{loc}}}, \overrightarrow{\phi}^{\mathrm{i_{elem}},2}_{\mathrm{i_{loc}}}) &=& -\mathbf{R}^{\mathrm{i_{elem}},3}_{\mathrm{i_{loc}},\mathrm{j_{loc}}},\\
a_T(\overrightarrow{\phi}^{\mathrm{j_{elem}},4}_{\mathrm{j_{loc}}}, \overrightarrow{\phi}^{\mathrm{i_{elem}},3}_{\mathrm{i_{loc}}}) &=& \mathbf{R}^{\mathrm{i_{elem}},2}_{\mathrm{i_{loc}},\mathrm{j_{loc}}},\\
a_T(\overrightarrow{\phi}^{\mathrm{j_{elem}},5}_{\mathrm{j_{loc}}}, \overrightarrow{\phi}^{\mathrm{i_{elem}},1}_{\mathrm{i_{loc}}}) &=& \mathbf{R}^{\mathrm{i_{elem}},3}_{\mathrm{i_{loc}},\mathrm{j_{loc}}},\\
a_T(\overrightarrow{\phi}^{\mathrm{j_{elem}},5}_{\mathrm{j_{loc}}}, \overrightarrow{\phi}^{\mathrm{i_{elem}},3}_{\mathrm{i_{loc}}}) &=& -\mathbf{R}^{\mathrm{i_{elem}},1}_{\mathrm{i_{loc}},\mathrm{j_{loc}}},\\
a_T(\overrightarrow{\phi}^{\mathrm{j_{elem}},6}_{\mathrm{j_{loc}}}, \overrightarrow{\phi}^{\mathrm{i_{elem}},1}_{\mathrm{i_{loc}}}) &=& -\mathbf{R}^{\mathrm{i_{elem}},2}_{\mathrm{i_{loc}},\mathrm{j_{loc}}},\\
a_T(\overrightarrow{\phi}^{\mathrm{j_{elem}},6}_{\mathrm{j_{loc}}}, \overrightarrow{\phi}^{\mathrm{i_{elem}},2}_{\mathrm{i_{loc}}}) &=& \mathbf{R}^{\mathrm{i_{elem}},1}_{\mathrm{i_{loc}},\mathrm{j_{loc}}}.
\end{array}\right.$$

Pour chaque élément T, associé au numéro $\mathrm{i_{elem}}$, la matrice $\mathbf{A}_T$ a la structure suivante

$$\mathbf{A}_T = \begin{bmatrix}
-ik_0\mathbf{M}^{\mathrm{i_{elem}}} & 0 & 0 & 0 & \mathbf{R}^{\mathrm{i_{elem}},3} & \mathbf{R}^{\mathrm{i_{elem}},2}\\
0 & -ik_0\mathbf{M}^{\mathrm{i_{elem}}} & 0 & -\mathbf{R}^{\mathrm{i_{elem}},3} & 0 & \mathbf{R}^{\mathrm{i_{elem}},1}\\
0 & 0 & -ik_0\mathbf{M}^{\mathrm{i_{elem}}} & \mathbf{R}^{\mathrm{i_{elem}},2} & -\mathbf{R}^{\mathrm{i_{elem}},1} & 0\\
0 & -\mathbf{R}^{\mathrm{i_{elem}},3} & \mathbf{R}^{\mathrm{i_{elem}},2} & -ik_0\mathbf{M}^{\mathrm{i_{elem}}} & 0 & 0\\
\mathbf{R}^{\mathrm{i_{elem}},3} & 0 & -\mathbf{R}^{\mathrm{i_{elem}},1} & 0 & -ik_0\mathbf{M}^{\mathrm{i_{elem}}} & 0\\
-\mathbf{R}^{\mathrm{i_{elem}},2} & \mathbf{R}^{\mathrm{i_{elem}},1} & 0 & 0 & 0 & -ik_0\mathbf{M}^{\mathrm{i_{elem}}}
\end{bmatrix}.$$

Maintenant, nous cherchons à déterminer la matrice $\mathbf{B}_F^{\mathrm{int}}$ représentant la forme b_F^{int} de la formulation variationnelle. Nous la rappelons

$$\left\{\begin{array}{rcl}
b_F^{\mathrm{int}}(\mathbb{E},\mathbb{E}') &=& \displaystyle\int_F [\![\gamma_\times \mathbf{H}]\!]_F \cdot \overline{\{\gamma_t \mathbf{E}'^T\}_F} - [\![\gamma_\times \mathbf{E}]\!]_F \cdot \overline{\{\gamma_t \mathbf{H}'^T\}_F}\\
&+& \displaystyle\int_F \frac{[\![\gamma_\times \mathbf{H}]\!]_F \cdot \overline{[\![\gamma_\times \mathbf{H}']\!]_F} + [\![\gamma_\times \mathbf{E}]\!]_F \cdot \overline{[\![\gamma_\times \mathbf{E}']\!]_F}}{2}.
\end{array}\right.$$

Nous remarquons que

$$b_F^{\mathrm{int}}(\overrightarrow{\phi}_{\mathrm{j_{glob}}}, \overrightarrow{\phi}_{\mathrm{i_{glob}}}) = 0 \quad \text{sauf si } F \in \mathcal{F}_{\mathrm{int}} \text{ et } F \subset \{T, K\}.$$

Définition 1.23 (Matrice des faces intérieures)**.** *La matrice* $\mathbf{B}_F^{\mathrm{int}}$ *représentant la forme inté-*

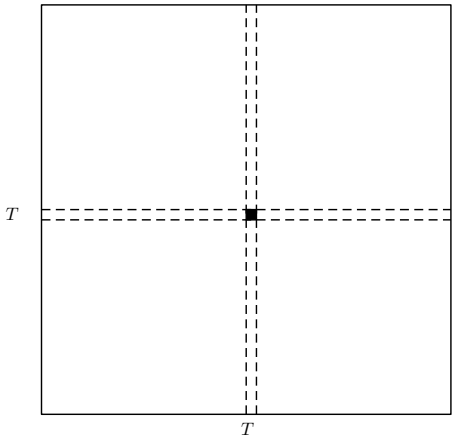

FIGURE 1.6 – Emplacements des blocs non nuls de la matrice $\mathbf{A}$.

rieure b_F^{int} *est donnée par*

$$(\mathbf{B}_F^{\mathrm{int}})_{\mathrm{i_{glob}},\mathrm{j_{glob}}} := \frac{1}{2}\,\varepsilon^F_{\mathrm{i_{elem}},\mathrm{j_{elem}}}\left((\mathbf{C}^{\mathrm{j_{elem}}})_{\mathrm{k_i},\mathrm{k_j}} + (\mathbf{D}^{\mathrm{j_{elem}}})_{\mathrm{k_i},\mathrm{k_j}}\right)(\mathbf{M}_F^{\mathrm{i_{elem}},\mathrm{j_{elem}}})_{\mathrm{i_{loc}},\mathrm{j_{loc}}},$$

avec $\varepsilon^F_{\mathrm{i_{elem}},\mathrm{j_{elem}}} = 0$ *sauf si* $F \subset \{T, K\}$*, où nous avons*

$$\begin{cases} \varepsilon^F_{\mathrm{i_{elem}},\mathrm{j_{elem}}} = 1 & \textit{pour } T = K, \\ \varepsilon^F_{\mathrm{i_{elem}},\mathrm{j_{elem}}} = -1 & \textit{pour } T \neq K, \end{cases}$$

et les matrices $\mathbf{C}^{\mathrm{i_{elem}}}$ *et* $\mathbf{D}^{\mathrm{i_{elem}}}$ *données par*

$$\mathbf{C}^{\mathrm{i_{elem}}} = \begin{pmatrix} 1-n_1^2 & -n_1n_2 & -n_1n_3 & 0 & 0 & 0 \\ -n_1n_2 & 1-n_2^2 & -n_2n_3 & 0 & 0 & 0 \\ -n_1n_3 & -n_2n_3 & 1-n_3^2 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1-n_1^2 & -n_1n_2 & -n_1n_3 \\ 0 & 0 & 0 & -n_1n_2 & 1-n_2^2 & -n_2n_3 \\ 0 & 0 & 0 & -n_1n_3 & -n_2n_3 & 1-n_3^2 \end{pmatrix},$$

et

$$\mathbf{D}^{\mathrm{i_{elem}}} = \begin{pmatrix} 0 & 0 & 0 & 0 & -n_3 & n_2 \\ 0 & 0 & 0 & n_3 & 0 & -n_1 \\ 0 & 0 & 0 & -n_2 & n_1 & 0 \\ 0 & n_3 & -n_2 & 0 & 0 & 0 \\ -n_3 & 0 & n_1 & 0 & 0 & 0 \\ n_2 & -n_1 & 0 & 0 & 0 & 0 \end{pmatrix},$$

où $\mathbf{n}_T := (n_1, n_2, n_3)^\mathsf{T}$ *est la normale sortante unitaire à l'élément* T.

Remarque 1.16. *Nous considérons alternativement les éléments* T *et* K *pour construire les matrices globales. Ainsi, nous remarquons que les blocs non nuls de la matrice* $\mathbf{B}_F^{\text{int}}$*, voir la Figure 1.7, sont*

$$\begin{aligned}
&\left[\ \frac{1}{2}(\mathbf{C}^{\mathrm{i_{elem}}} + \mathbf{D}^{\mathrm{i_{elem}}})\mathbf{M}_F^{\mathrm{i_{elem}},\mathrm{i_{elem}}}\right]_{\mathrm{i_{elem}},\mathrm{i_{elem}}}, && \textit{pour } \mathrm{i_{elem}} = 1, \#\mathrm{ddlelem},\\
&\left[\ -\frac{1}{2}(\mathbf{C}^{\mathrm{i_{elem}}} + \mathbf{D}^{\mathrm{i_{elem}}})\mathbf{M}_F^{\mathrm{i_{elem}},\mathrm{j_{elem}}}\right]_{\mathrm{i_{elem}},\mathrm{j_{elem}}}, && \textit{pour } \mathrm{i_{elem}}, \mathrm{j_{elem}} = 1, \#\mathrm{ddlelem},\\
&\left[\ -\frac{1}{2}(\mathbf{C}^{\mathrm{j_{elem}}} + \mathbf{D}^{\mathrm{j_{elem}}})\mathbf{M}_F^{\mathrm{j_{elem}},\mathrm{i_{elem}}}\right]_{\mathrm{j_{elem}},\mathrm{i_{elem}}}, && \textit{pour } \mathrm{j_{elem}}, \mathrm{i_{elem}} = 1, \#\mathrm{ddlelem},\\
&\left[\ \frac{1}{2}(\mathbf{C}^{\mathrm{j_{elem}}} + \mathbf{D}^{\mathrm{j_{elem}}})\mathbf{M}_F^{\mathrm{j_{elem}},\mathrm{j_{elem}}}\right]_{\mathrm{j_{elem}},\mathrm{j_{elem}}}, && \textit{pour } \mathrm{j_{elem}} = 1, \#\mathrm{ddlelem}.
\end{aligned}$$

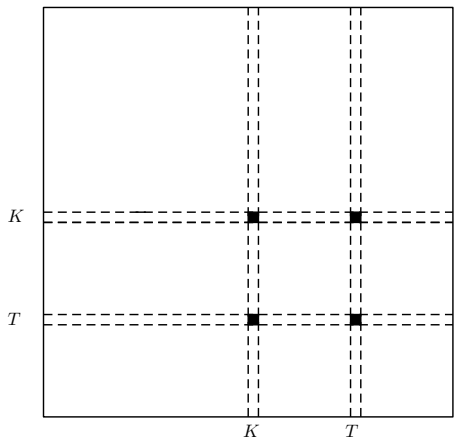

FIGURE 1.7 – Emplacements des blocs non nuls de la matrice $\mathbf{B}_F^{\text{int}}$.

Nous définissons ensuite la matrice $\mathbf{B}_F^{\text{ext}}$ associée à la forme sesquilinéaire (1.18) que nous rappelons ici

$$\begin{aligned}
b_F^{\text{ext}}(\mathbb{E}, \mathbb{E}') \;=\; & \int_F \frac{1 - R_{\partial\Omega}}{2}\gamma_t\mathbf{E}^T \cdot \overline{\gamma_t\mathbf{E}'^T} + \frac{1 + R_{\partial\Omega}}{2}\gamma_\times^T\mathbf{H}^T \cdot \overline{\gamma_\times\mathbf{H}'^T}\\
+ \; & \int_F \frac{1 - R_{\partial\Omega}}{2}\gamma_t\mathbf{E}^T \cdot \overline{\gamma_\times\mathbf{H}'^T} + \frac{1 + R_{\partial\Omega}}{2}\gamma_\times^T\mathbf{H}^T \cdot \overline{\gamma_t\mathbf{E}'^T}.
\end{aligned}$$

Nous remarquons que

$$b_F^{\text{ext}}(\overrightarrow{\phi}_{\mathrm{j_{glob}}}, \overrightarrow{\phi}_{\mathrm{i_{glob}}}) = 0 \quad \text{sauf si } F \in \mathcal{F}_{\text{ext}} \cap \mathcal{F}_{T_{\mathrm{i_{elem}}}} \text{ et } T_{\mathrm{i_{elem}}} = T_{\mathrm{j_{elem}}}.$$

Les valeurs non nulles sont calculées à partir des actions des opérateurs de composante

tangentielle et de trace tangentielle sur les fonctions de base globales, à savoir

$$\gamma_t \mathbf{E}^T\big[\overrightarrow{\phi}^{\mathrm{i_{elem}},\mathrm{k_i}}_{\mathrm{i_{loc}}}\big] \quad \text{et} \quad \gamma_\times^T \mathbf{E}^T\big[\overrightarrow{\phi}^{\mathrm{i_{elem}},\mathrm{k_i}}_{\mathrm{i_{loc}}}\big], \quad \text{pour } \mathrm{i_{loc}} = 1, N_q \text{ et } \mathrm{k_i} = 1, 6.$$

Nous notons dans la suite : le numéro $\mathrm{i_{elem}}$ associé à l'élément T et le numéro $\mathrm{j_{elem}}$ associé à l'élément K. Nous définissons des matrices de masse de frontière pour chaque face $F \in \mathcal{F}_T$.

Définition 1.24 (Matrice de masse de frontière)**.** *La matrice de masse sur une face $F \in \mathcal{F}_T$ est définie par*

$$(\mathbf{M}_F^{\mathrm{i_{elem}},\mathrm{j_{elem}}})_{\mathrm{i_{loc}},\mathrm{j_{loc}}} = \int_F \phi^{\mathrm{j_{elem}}}_{\mathrm{j_{loc}}} \phi^{\mathrm{i_{elem}}}_{\mathrm{i_{loc}}} \quad \textit{pour } \mathrm{i_{loc}}, \mathrm{j_{loc}} = 1, N_q. \tag{1.20}$$

Dans le cas où $F \in \mathcal{F}_T$ est une face extérieure, l'élément T n'a pas de voisin K à travers F, de telle sorte que nous prenons $K = T$ dans (1.20). Ainsi, les coefficients de la matrice deviennent, pour $F \in \mathcal{F}_{\mathrm{ext}}$,

$$(\mathbf{M}_F^{\mathrm{i_{elem}},\mathrm{i_{elem}}})_{\mathrm{i_{loc}},\mathrm{j_{loc}}} = \int_F \phi^{\mathrm{i_{elem}}}_{\mathrm{j_{loc}}} \phi^{\mathrm{i_{elem}}}_{\mathrm{i_{loc}}} \quad \text{pour } \mathrm{i_{loc}}, \mathrm{j_{loc}} = 1, N_q.$$

Grâce aux outils et aux définitions ci-dessus, nous pouvons désormais introduire la matrice des bords extérieurs.

Définition 1.25 (Matrice des bords extérieurs)**.** *Soit $F \in \mathcal{F}_{\mathrm{ext}} \cap \mathcal{F}_T$, où $T \in \mathcal{T}$, la matrice associée à la forme sesquilinéaire b_F^{ext} est définie par*

$$(\mathbf{B}_F^{\mathrm{ext}})_{\mathrm{i_{glob}},\mathrm{j_{glob}}} := \zeta^F_{\mathrm{i_{elem}},\mathrm{j_{elem}}} \left(D_1^{\mathrm{ext}} \mathbf{C}_1^{\mathrm{i_{elem}}} + D_2^{\mathrm{ext}} \mathbf{C}_2^{\mathrm{i_{elem}}}\right)_{\mathrm{k_i},\mathrm{k_j}} (\mathbf{M}_F^{\mathrm{i_{elem}},\mathrm{i_{elem}}})_{\mathrm{i_{loc}},\mathrm{j_{loc}}},$$

où $\zeta^F_{\mathrm{i_{elem}},\mathrm{j_{elem}}} = 0$, sauf si $T_{\mathrm{i_{elem}}} = T_{\mathrm{j_{elem}}} = T$ (Figure 1.8), où nous avons $\zeta^F_{\mathrm{i_{elem}},\mathrm{j_{elem}}} = 1$,

$$\mathrm{i_{glob}} = \mathrm{LocToGlob}(\mathrm{i_{elem}}, \mathrm{i_{loc}}, \mathrm{k_i}) \quad \textit{et} \quad \mathrm{j_{glob}} = \mathrm{LocToGlob}(\mathrm{j_{elem}}, \mathrm{j_{loc}}, \mathrm{k_j}),$$

et

$$D_1^{\mathrm{ext}} := \mathrm{diag}\left(\frac{1-R}{2} \quad \frac{1-R}{2} \quad \frac{1-R}{2} \quad \frac{1+R}{2} \quad \frac{1+R}{2} \quad \frac{1+R}{2}\right),$$

$$D_2^{\mathrm{ext}} := \mathrm{diag}\left(\frac{1+R}{2} \quad \frac{1+R}{2} \quad \frac{1+R}{2} \quad \frac{1-R}{2} \quad \frac{1-R}{2} \quad \frac{1-R}{2}\right),$$

$$\mathbf{C}_1^{\mathrm{i_{elem}}} = \begin{pmatrix} 1-n_1^2 & -n_1n_2 & -n_1n_3 & 0 & 0 & 0 \\ -n_1n_2 & 1-n_2^2 & -n_2n_3 & 0 & 0 & 0 \\ -n_1n_3 & -n_2n_3 & 1-n_3^2 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1-n_1^2 & -n_1n_2 & -n_1n_3 \\ 0 & 0 & 0 & -n_1n_2 & 1-n_2^2 & -n_2n_3 \\ 0 & 0 & 0 & -n_1n_3 & -n_2n_3 & 1-n_3^2 \end{pmatrix},$$

$$\mathbf{C}_2^{\mathrm{i_{elem}}} = \begin{pmatrix} 0 & 0 & 0 & 0 & -n_3 & n_2 \\ 0 & 0 & 0 & n_3 & 0 & -n_1 \\ 0 & 0 & 0 & -n_2 & n_1 & 0 \\ 0 & n_3 & -n_2 & 0 & 0 & 0 \\ -n_3 & 0 & n_1 & 0 & 0 & 0 \\ n_2 & -n_1 & 0 & 0 & 0 & 0 \end{pmatrix},$$

où $\mathbf{n}_T = (n_1, n_2, n_3)^\mathsf{T}$ *est la normale sortante unitaire à l'élément* T.

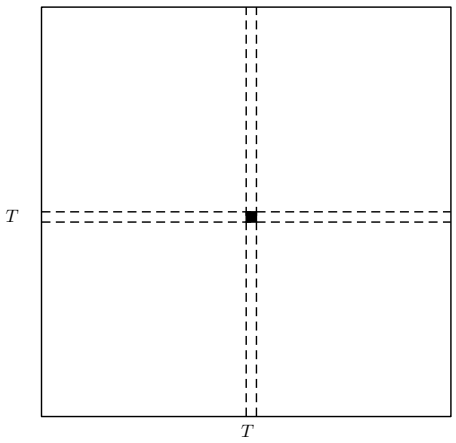

FIGURE 1.8 – Emplacements des blocs non nuls de la matrice $\mathbf{B}_F^{\mathrm{ext}}$.

Pour conclure, le problème matriciel associé à la méthode de GD est le suivant.

Problème 6 (Problème matriciel)**.** *Le système matriciel de la méthode GD est*

$$\mathbf{A}[\mathbb{E}^h] = \mathbf{F},$$

avec

$$\mathbf{A} = \sum_{T\in\mathcal{T}} \mathbf{A}_T + \sum_{F\in\mathcal{F}_{\mathrm{int}}} \mathbf{B}_F^{\mathrm{int}} + \sum_{F\in\mathcal{F}_{\mathrm{ext}}} \mathbf{B}_F^{\mathrm{ext}},$$

où $\mathbf{A}_T$, $\mathbf{B}_F^{\mathrm{int}}$, $\mathbf{B}_F^{\mathrm{ext}}$ *sont définis par les Définitions 1.22, 1.23 et 1.25, et avec* $\mathbf{F}$ *défini par*

$$(\mathbf{F})_{\mathrm{i_{glob}}} := \ell(\overrightarrow{\phi}_{\mathrm{i_{glob}}}) = \int_{\partial\Omega} \frac{1 - R_{\partial\Omega}}{2} \Big(\mathbf{g} \cdot \overline{\overrightarrow{\phi}_{\mathrm{i_{glob}}}} + \mathbf{g} \cdot \overline{\overrightarrow{\phi}_{\mathrm{i_{glob}}}} \Big).$$

1.2.4 Convergence du solveur

Le solveur trouve la solution numérique $[\mathbb{E}^h]$ du Problème 6 grâce à une factorisation LU de la matrice $\mathbf{A}$. Afin de vérifier que notre solveur de GD simule correctement une onde élec-

tromagnétique, nous mettons en place une étude de convergence. Comme pour la méthode d'EF de Nédélec, nous étudions le cube unité $\Omega = [0,1]^3$ où une onde plane électromagnétique $\mathbb{E}_{inc}$, de direction $\mathbf{d} = (1,0,0)^\mathsf{T}$ et de polarisation $\mathbf{p} = (0,1,0)^\mathsf{T}$, est imposée sur sa frontière. En particulier, nous choisissons $\mathbf{g} = \gamma_t \mathbf{E}_{inc} + Z_{\partial\Omega}\gamma_\times \mathbf{H}_{inc}$, avec $Z_{\partial\Omega} = 1$.

Nous connaissons la valeur analytique $\mathbb{E}^{\text{ex}} := (\mathbf{E}^{\text{ex}}, \mathbf{H}^{\text{ex}})$ de l'onde plane que nous souhaitons obtenir dans le domaine, avec

$$\mathbf{E}^{\text{ex}}(\mathbf{x}) := \mathbf{p}\, e^{ik_0\, \mathbf{d}\cdot\mathbf{x}} \quad \text{et} \quad \mathbf{H}^{\text{ex}}(\mathbf{x}) := (\mathbf{p}\times\mathbf{d})\, e^{ik_0\, \mathbf{d}\cdot\mathbf{x}}, \quad \text{pour } \mathbf{x}\in\mathbb{C}^3.$$

Nous étudions la convergence L^2 du schéma. L'erreur associée s'écrit

$$e^h := \|\mathbb{E}^h - \mathbb{E}^{\text{ex}}\|_{\mathrm{L}^2},$$

avec la norme L^2, définie par

$$\|\mathbb{E}\|_{\mathrm{L}^2}^2 = \sum_{T\in\mathcal{T}} \int_T |\mathbf{E}|^2 + |\mathbf{H}|^2, \quad \text{pour tout} \quad \mathbb{E} = (\mathbf{E},\mathbf{H}) \in \mathbb{X}.$$

Nous notons $h := \max_T(h_T)$ où h_T est la hauteur du tétraèdre $T \in \mathcal{T}$. Nous rappelons que nous utilisons des fonctions de base polynomiales scalaires de $P_q(T)$, où q est le degré maximal du polynôme. Nous avons développé l'algorithme pour $q = 1, 3$. Ainsi, trois ordres d'approximation sont possibles. Pour chacun d'entre eux, la méthode de GD converge, voir la Figure 1.9. En effet, la norme e^h, en ordonnées, devient petite lorsque h, en abscisses, diminue. Les oscillations sur les courbes proviennent de l'utilisation d'un maillage irrégulier généré automatiquement par le logiciel GMSH©. À taille de domaine fixée, $\mathcal{D}_\Omega = 1\lambda$, et en donnant une taille d'élément h au logiciel, il est difficile d'obtenir des éléments tétraédriques de taille homogène. Par conséquent, bien que nous diminuons de manière régulière le pas de maillage, ce n'est pas le cas du nombre d'éléments.

1.3 Incapacités à traiter de grands domaines de calcul

Dans cette section, nous analysons les coûts mémoire des méthodes présentées. Pour cela, nous augmentons la taille $\mathcal{D}_\Omega$ du domaine à précision fixée et à pas de maillage h fixé pour chaque ordre d'approximation. Pour la méthode d'EF de Nédélec, nous construisons des courbes à précision donnée. Nous définissons l'erreur relative associée aux méthodes

$$e_r = \frac{\|\mathbf{E}^h - \mathbf{E}^{\text{ex}}\|_{\mathrm{L}^2}}{\|\mathbf{E}^{\text{ex}}\|_{\mathrm{L}^2}}.$$

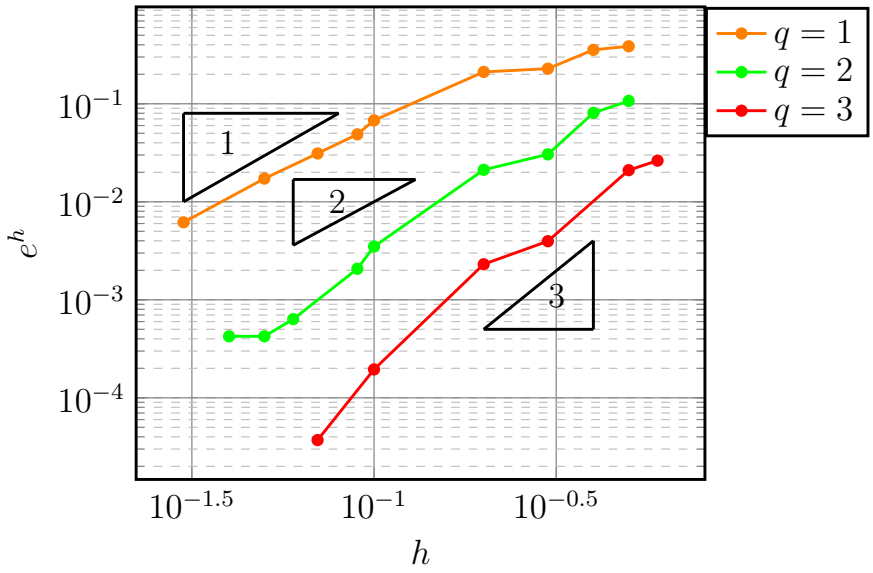

FIGURE 1.9 – Courbes représentant la norme L^2 en fonction de la taille h des éléments $T \in \mathcal{T}$ du maillage, pour différents ordres d'approximation q.

Nous utilisons des cubes de côté h pour mailler Ω. Les pas de maillage h sont choisis, pour chaque ordre d'approximation possible $r = 1, 4$, de telle sorte à obtenir $e_r = 10\%$ pour $\mathcal{D}_\Omega = 5\lambda$, voir le Tableau 1.10. Puis, ils sont fixés pour toute l'étude du coût mémoire, *ie* lorsque nous augmentons $\mathcal{D}_\Omega$, pour chaque ordre d'approximation.

Remarque 1.17. *L'erreur est choisie relativement grande et ne produit pas une solution numérique précise. En effet, pour obtenir une meilleure précision sur la solution numérique, ie $e_r = 1\%$ par exemple, nous aurions dû utiliser davantage de mémoire (plus de points de discrétisation ou plus d'éléments). De cette façon, nous sommes même optimistes vis à vis du coût mémoire de la méthode. En effet, si nous prouvons son coût mémoire trop important pour $e_r = 10\%$ alors cela ne pourra que être pire pour $e_r = 1\%$.*

La Figure 1.10 représente les coûts de la factorisation LU du solveur d'EF pour chaque résolution du système matriciel. Le coût mémoire est proportionnel à $(\mathcal{D}_\Omega)^4$, comme le montre les pentes des courbes sur la Figure 1.10. Ainsi, nous retrouvons bien le coût mémoire théorique d'une factorisation LU réalisée avec la librairie MUMPS©, voir [5].

r	1	2	3	4
$h\,(\lambda)$	$\frac{1}{50} = 0.02$	$\frac{1}{10} = 0.1$	$\frac{5}{18} \approx 0.27$	$\frac{1}{2} = 0.5$

TABLE 1.10 – Taille h du côté d'un cube pour les différents ordres de la méthode d'EF de Nédélec $r = 1, 4$.

Nous procédons de la même façon pour la méthode de GD. Nous fixons les hauteurs des tétraèdres h pour chaque ordre d'approximation q allant de 1 à 3, voir le Tableau 1.11. De nouveau, nous retrouvons le coût mémoire théorique d'une factorisation LU réalisée avec MUMPS©, voir la Figure 1.11. En impliquant à la fois le champ électrique $\mathbf{E}$ et le champ magnétique $\mathbf{H}$, l'utilisation d'un problème d'ordre 1 augmente le nombre d'inconnues et limite la capacité à traiter de grands domaines de calcul. De plus, avec notre implémentation

q	1	2	3
$h\,(\lambda)$	$\frac{1}{10} = 0.1$	$\frac{1}{5} = 0.2$	$\frac{2}{5} = 0.4$

TABLE 1.11 – Taille de la hauteur d'un tétraèdre h pour les différents ordres de la méthode de GD d'ordre $q = 1, 3$.

de ces méthodes et à coût mémoire égal, *ie* 1 Tera Octet (To), nous pouvons simuler une onde électromagnétique sur un domaine allant jusqu'à $\mathcal{D}_\Omega^{\max} = 21\lambda$ avec la méthode de Nédélec, contrairement à la méthode de GD qui ne traite que des domaines de taille maximale $\mathcal{D}_\Omega^{\max} = 9\lambda$, voir les Figures 1.10 et 1.11. Cette différence s'explique principalement à cause des six composantes présentes dans la méthode de GD et du caractère discontinu des fonctions de base.

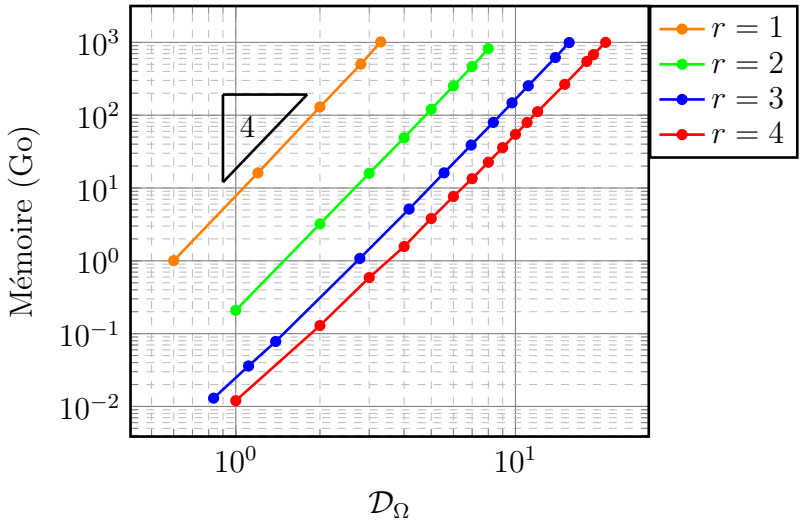

FIGURE 1.10 – Coût mémoire de la résolution du problème de Nédélec pour différentes tailles de domaine $\mathcal{D}_\Omega$.

Remarque 1.18. *Nous avons choisi de ne pas investiguer dans cette thèse les méthodes de compression de type Block-Low Rank (BLR), car ce type d'approche nous semble moins approprié à notre contexte haute fréquence.*

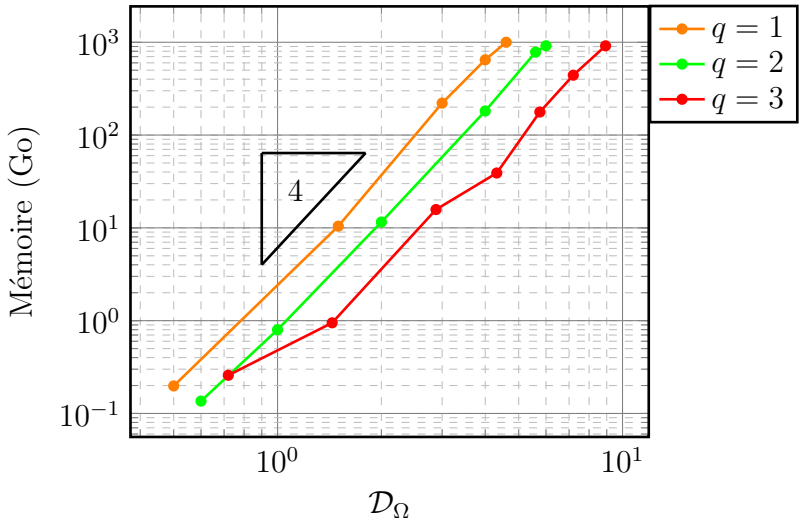

FIGURE 1.11 – Coût mémoire de la résolution du problème de GD pour différentes tailles de domaine $\mathcal{D}_\Omega$.

Une alternative peut être d'utiliser une méthode itérative. Nous donnons l'exemple d'une méthode de Generalised Minimal RESidual (GMRES) ici. Comme le montre la Figure 1.12, la méthode de GD ne converge pas pour $\mathcal{D}_\Omega = 5\lambda$, au sens où le résidu GMRES ne diminue pas au fur et à mesure des itérations. De plus, pour la méthode d'EF, nous observons que plus de 10^3 itérations sont nécessaires pour atteindre un résidu GMRES de 1%. Le code d'EF met du temps à converger, pour une taille de domaine pourtant extrêmement raisonnable : $\mathcal{D}_\Omega = 10\lambda$. Cet exemple montre les limites de ces méthodes même en utilisant un solveur itératif de type GMRES. Ces taux de convergence pourraient être améliorés par l'utilisation d'un préconditionneur. Néanmoins, sa construction est difficile dans le cadre des systèmes indéfinis. Nous traiterons ce problème avec la méthode de Trefftz développée dans cette thèse.

Pour conclure, il est donc nécessaire de développer de nouvelles méthodes à faible coût mémoire.

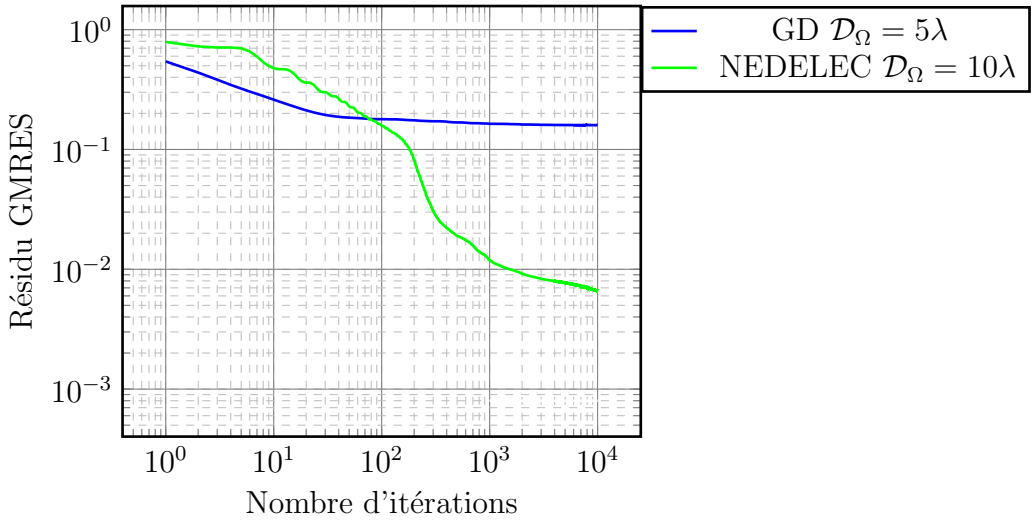

FIGURE 1.12 – Résidu GMRES pour la méthode d'EF de Nédélec et pour la méthode de GD en fonction du nombre d'itérations de la méthode itérative, pour $R_{\partial\Omega} = \dfrac{1 - Z_{\partial\Omega}}{1 + Z_{\partial\Omega}} = 0.9$.

1.4 Conclusion

Dans le chapitre 1, nous avons introduit deux méthodes numériques représentatives des méthodes classiques existantes pour la simulation d'ondes électromagnétiques.

La première, la méthode d'EF de Nédélec d'ordre élevé, repose sur un maillage cubique dont la discrétisation est réalisée grâce à des degrés de liberté localisés à des noeuds de Gauss et de Gauss-Lobatto, voir la Section 1.1. La formulation ainsi que de la solution sont interpolées à l'aide de fonctions de base de Lagrange vectorielles. Ce choix introduit des produits tensoriels de matrices, ce qui facilite son implémentation et induit une matrice de masse diagonale.

La seconde, la méthode de GD d'ordre élevé, est basée sur une formulation d'ordre 2 du problème de Maxwell, voir la Section 1.2. Nous employons un maillage tétraédrique et des fonctions de base de type Lagrange pour assembler le système matriciel. Nous maillons le domaine de calcul en tétraèdres et nous interpolons les quantités intégrables grâce aux fonctions de Lagrange scalaires. La complexité de ce solveur provient de la forme que prend sa solution numérique. L'évaluation des champs $\mathbf{E}$ et $\mathbf{H}$ en trois dimensions implique le calcul de 6 composantes à chaque degré de liberté du domaine. Le coût mémoire pour la résolution s'en voit alors considérablement augmenté.

Plus particulièrement, les résolutions directes des systèmes associés à ces deux méthodes présentent clairement des problèmes de coût mémoire, voir la Section 1.3. Très rapidement, lorsque la taille du domaine augmente, le coût mémoire des deux méthodes dépasse 1To.

De plus, même en utilisant une résolution itérative, les résultats numériques ne sont pas concluants au sens où l'algorithme itératif ne converge pas ou devient trop coûteux en temps pour obtenir la solution.

La conclusion est, qu'en l'état, ces méthodes ne permettent pas de simuler avec précision une onde électromagnétique sur de grands domaines. Il existe de multiples solutions pour pallier ce problème. Les idées les plus connues sont l'utilisation de méthodes de Décomposition De Domaine (DDM) [37, 99] ou encore HDG [24, 79, 84] qui réduisent le coût mémoire de stockage de la factorisation LU. Dans le cadre de cette thèse, nous étudions une méthode de GD particulière, nommée méthode de Trefftz [54, 88]. Nous proposons une méthode de Trefftz capable de traiter de plus grandes scènes de calcul que les méthodes classiques. Dans le chapitre suivant, nous mettons en place le solveur de Trefftz direct, *ie* utilisant une factorisation LU pour sa résolution.

CHAPITRE 2

SOLVEUR DIRECT DE TYPE TREFFTZ

Sommaire

Dans le Chapitre 1, nous avons présenté deux méthodes numériques classiques (d'EF et de GD) que nous jugeons représentatives de celles existantes dans la littérature. Il est clair qu'elles font face toutes les deux à un problème de coût mémoire, voir les Figures 1.10 et 1.11. Une méthode alternative doit être mise en place afin de pouvoir traiter de très grandes scènes de calcul. Nous explorons dans cette section la capacité des méthodes de Trefftz [46, 48, 49, 77] à pallier ces difficultés mémoire comme elles impliquent moins de degrés de liberté.

L'objectif du Chapitre 2 est de construire un solveur Trefftz direct. La méthode de Trefftz met en jeu deux types d'espaces : les espaces de Trefftz continus et discrets dont les éléments sont des solutions locales des équations de Maxwell. Le point de départ de toutes ces constructions est une formule de réciprocité sur ces espaces. Nous investiguons alors deux points de vue afin de dériver des formulations bien posées et coercives. Le premier consiste à utiliser des perturbations basées sur des formes consistantes. Le caractère bien posé de ces formulations est prouvé en assurant la coercivité grâce à des critères portant sur les paramètres de pénalisation. Le second utilise la notion de traces numériques. Celles-ci sont obtenues par un solveur de Riemann, dans le cas homogène, et par un schéma upwind, dans le cas hétérogène. Elles sont donc étroitement liées à la nature hyperbolique du système de Maxwell. Nous verrons qu'elles s'avèrent équivalentes. Bien que leur présentation soit moins concise, les traces de Riemann relient les choix des paramètres de pénalisation de la méthode de GD à la physique du phénomène étudié. Nous prouvons ensuite le caractère bien posé de la méthode de Trefftz. Dans le cas hétérogène, la formulation variationnelle upwind obtenue admet une propriété de coercivité faible que nous démontrons. Ceci nous permet d'assurer le caractère bien posé de la méthode de Trefftz dans le cas général.

Enfin, nous présentons la structure du problème matriciel pour un maillage cartésien. Le solveur Trefftz direct basé sur des traces numériques upwind dans le cas de milieux hétérogènes est implémenté. Cela mène au code GoTEM3, dans lequel la résolution du système matriciel est réalisée grâce à une factorisation LU de la matrice $\mathbf{A}$. Des résultats numériques témoignent de la convergence de ce solveur et mettent en avant son faible coût mémoire face aux méthodes du Chapitre 1. Toutefois, le fait d'avoir recours à une factorisation LU ne fait que retarder l'explosion mémoire et ne permet pas de traiter des domaines de plusieurs centaines de longueurs d'onde.

2.1 Espaces continus et discrets de type Trefftz

Dans cette partie, nous allons tout d'abord définir l'espace de Trefftz continu. Celui-ci est l'espace fonctionnel nécessaire à la dérivation et à l'étude des formulations Trefftz présentées

dans ce chapitre. Nous présentons ensuite l'espace de Trefftz discret, basé sur des espaces d'approximation d'ondes planes. Ceci permet la discrétisation des formulations. Enfin, nous introduisons la formule de réciprocité, liée à la formule des travaux virtuels et associée à ces espaces. Elle est le point de départ de la construction des formulations Trefftz.

L'objectif de cette thèse est de simuler des ondes électromagnétiques avec précision sur des domaines de grande taille et complexes, tant géométriquement que dans leur composition matérielle. En effet, les industriels manipulent des géométries constituées de différents composants. C'est pourquoi il est indispensable de développer un solveur Trefftz adapté aux hétérogénéités des matériaux. Dans les cas hétérogènes, les caractéristiques physiques du milieu varient. Toutefois, les paramètres physiques sont supposés constants sur chaque élément T. Cette restriction ne semble pas poser de difficultés en pratique et est couramment réalisée dans les librairies de simulation. Nous introduisons l'impédance normalisée relative à un élément T notée

$$Z_T := \sqrt{\frac{\mu_r^T}{\varepsilon_r^T}},$$

où nous rappelons que ε_r et μ_r sont respectivement la permitivité relative et la perméabilité relative, dont les restrictions à un élément T sont respectivement notées ε_r^T et μ_r^T.

Nous notons $\mathbb{X}_T$ l'espace de Trefftz local continu défini pour $T \in \mathcal{T}$ comme l'ensemble des fonctions

$$\mathbb{E}^T := (\mathbf{E}^T, \mathbf{H}^T) \in H(\mathrm{rot}, T) \times H(\mathrm{rot}, T) \text{ vérifiant} \tag{2.1a}$$

$$\nabla \times \mathbf{H}^T = ik_0\, \varepsilon_r \mathbf{E}^T \quad \text{et} \quad \nabla \times \mathbf{E}^T = -ik_0\, \mu_r \mathbf{H}^T, \quad \text{sur } T, \tag{2.1b}$$

$$\gamma_t \mathbf{E}^T \in \mathrm{L}_t^2(\partial T) \quad \text{et} \quad \gamma_t \mathbf{H}^T \in \mathrm{L}_t^2(\partial T). \tag{2.1c}$$

Remarque 2.1. *La condition (2.1c) ne joue pas en faveur des méthodes de Trefftz. En effet, elle nécessite une régularité $H^{\frac{3}{2}}(T)$ de la solution du problème de Maxwell. Cette hypothèse est le plus souvent satisfaite même dans le cas hétérogène, bien qu'il soit possible de construire des domaines à faible régularité. Nous pouvons nous référer aux travaux [12, 31, 32, 53] qui traitent partiellement cette problématique.*

L'espace $\mathbb{X}_T$ est équipé d'un produit scalaire hermitien, pondéré par $\alpha > 0$ et $\beta > 0$,

$$(\mathbb{E}^T, \mathbb{E}'^T)_{\mathbb{X}_T} := \int_{\partial T} \left(\alpha\, \gamma_t \mathbf{E}^T \cdot \overline{\gamma_t \mathbf{E}'^T} + \beta\, \gamma_t \mathbf{H}^T \cdot \overline{\gamma_t \mathbf{H}'^T} \right). \tag{2.2}$$

Remarque 2.2. *Ce résultat se prouve facilement en utilisant les propriétés du produit scalaire L^2 et le théorème du prolongement unique [58, 89]. Nous avons*

$$(\mathbb{E}^T, \mathbb{E}^T)_{\mathbb{X}_T} = 0 \implies \gamma_t \mathbf{E}^T = 0 \textit{ et } \gamma_t \mathbf{H}^T = 0 \textit{ sur } \partial T \implies \mathbb{E}^T = (\mathbf{E}^T, \mathbf{H}^T) = 0 \textit{ dans } T.$$

Ainsi, l'espace de Trefftz global continu $\mathbb{X}_{\mathcal{T}}$ est défini élément par élément par

$$\mathbb{X}_{\mathcal{T}} := \prod_{T\in\mathcal{T}} \mathbb{X}_T. \tag{2.3}$$

Il est équipé du produit scalaire

$$(\mathbb{E},\mathbb{E}')_{\mathbb{X}_{\mathcal{T}}} := \sum_{T\in\mathcal{T}} (\mathbb{E}^T,\mathbb{E}'^T)_{\mathbb{X}_T}.$$

Remarque 2.3. *Chaque fonction $\mathbb{E}\in\mathbb{X}_{\mathcal{T}}$ est associée à une fonction définie sur Ω dont la restriction à T est $\mathbb{E}^T$. Chaque $\mathbb{E}^T=(\mathbf{E}^T,\mathbf{H}^T)$ satisfait (2.1b) et $\mathbb{E}$ est généralement discontinue à travers les faces.*

Maintenant, nous définissons l'espace de Trefftz global discret $\mathbb{X}^h_{\mathcal{T}}$. C'est un sous-espace vectoriel de dimension finie de $\mathbb{X}_{\mathcal{T}}$. Il est défini comme

$$\mathbb{X}^h_{\mathcal{T}} := \prod_{T\in\mathcal{T}} \mathbb{X}^h_T \subset \mathbb{X}_{\mathcal{T}}, \tag{2.4}$$

où l'espace de Trefftz local discret $\mathbb{X}^h_T$ est défini par

$$\mathbb{X}^h_T := \operatorname*{span}_{\ell=1,N_T}\Big\{\mathbf{v}^\ell_T\in\mathbb{X}_T\Big\},$$

où les fonctions $\mathbf{v}^\ell_T$ sont des ondes planes électromagnétiques de direction $\mathbf{d}^\ell_T$ et de polarisation $\mathbf{p}^\ell_T$, et où N_T est le nombre de degrés de liberté par élément. L'espace global discret de Trefftz $\mathbb{X}^h_{\mathcal{T}}$, défini par (2.4), implique des fonctions de base hétérogènes. Plus précisément, l'espace de Trefftz local discret $\mathbb{X}^h_T$ est la base engendrée par l'ensemble des ondes planes hétérogènes définies pour $\ell=1,N_T$ par

$$\mathbf{v}^\ell_T := (\mathbf{E}^T,\mathbf{H}^T)\in\mathbb{X}_T \quad \text{avec} \quad \left\{\begin{array}{rcl} \mathbf{E}^T &:=& \mathbf{p}^\ell_T\, e^{ik_0\sqrt{\varepsilon_r^T\mu_r^T}\mathbf{d}^\ell_T\cdot\mathbf{x}},\\ \mathbf{H}^T &:=& Z_T(\mathbf{p}^\ell_T\times\mathbf{d}^\ell_T)\, e^{ik_0\sqrt{\varepsilon_r^T\mu_r^T}\mathbf{d}^\ell_T\cdot\mathbf{x}}, \end{array}\right. \tag{2.5}$$

avec $\mathbf{x}\in T$ et où $Z_T=\sqrt{\mu_r^T/\varepsilon_r^T}$ est une impédance normalisée définie sur T. Plus précisément, la direction de propagation $\mathbf{d}^\ell_T$ appartient à $\mathcal{D}$, où $\mathcal{D}$ est un sous-espace discret de la sphère unité, et la polarisation $\mathbf{p}^\ell_T$ appartient à $S_{\mathbf{d}}$, où $S_{\mathbf{d}}$ est une base orthonormale du sous-espace vectoriel de deux dimensions $(\mathbf{d}^\ell_T)^\perp\subset\mathbb{R}^3$.

Différents choix pour $\mathcal{D}$ existent dans le cas d'une base discrète d'ondes planes. Par exemple, les directions peuvent être données par les sommets d'un maillage surfacique de la sphère unité [77] ou du cube unité. Dans cette thèse, la seconde approche est considérée en

utilisant un maillage cartésien de la surface du cube. Grâce à ce dernier, nous définissons les directions de propagation discrètes à partir des vecteurs orientés du centre du cube vers l'ensemble des noeuds du maillage. Par exemple, la discrétisation de la surface d'un cube avec 4 carrés par face correspond à 9 points par face, et conduit à 26 directions différentes, comme le montre la Figure 2.1. Nous considérons des ondes planes en 3 dimensions et associons donc deux polarisations à chaque direction. De cette façon, nous avons une base de $N_T = 52$ ondes planes pour cet exemple.

Bien souvent, les espaces d'ondes planes induisent des problèmes de stabilité de la méthode numérique. De nombreuses recherches ont étudié l'erreur entre la solution numérique et la solution exacte en fonction du nombre d'ondes planes dans l'espace discret, voir [9, 49, 52, 77]. Il s'avère qu'une augmentation du nombre de fonctions de base tend à mal conditionner le système étudié. Pour éviter ce phénomène, nous développons dans cette thèse une stratégie de réduction du nombre de fonctions de base pour espérer obtenir un nombre optimal d'ondes planes, voir la Section 4.2.

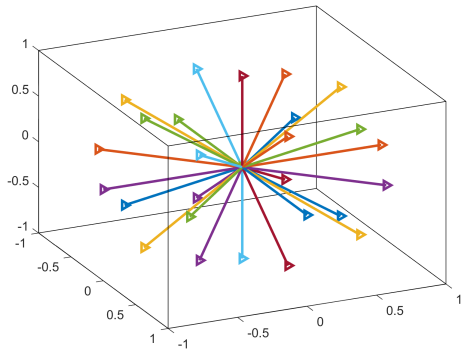

FIGURE 2.1 – 26 directions d'ondes planes dans le cube unité.

D'autres fonctions de base auraient pu être mises en jeu. Par exemple, nous aurions pu employer des ondes évanescentes qui favorisent la stabilité de la méthode, voir [54, 85], en particulier pour les méthodes UWVF [72]. Un autre exemple est celui des fonctions de type Bessel [54]. En présence de singularités, elles permettent de pallier le mauvais conditionnement induit par les ondes planes et évitent un raffinement local du maillage qui pourrait provoquer des coûts de calcul importants. Cependant, contrairement aux ondes planes, ces fonctions ne peuvent pas être intégrées précisément par une simple quadrature et des méthodes spécifiques doivent être utilisées. Nous aurions aussi pu avoir recours à des fonctions qui satisfont les équations de Maxwell de manière approchée. Plus précisément, cela revient par exemple à déterminer des solutions numériques en résolvant sur chaque élément des problèmes de Maxwell, puis de les choisir comme fonctions de base. Cette méthode s'appelle quasi-Trefftz [64, 65], et nous en donnons un exemple dans ce manuscrit, voir le Chapitre 5.

Un important avantage de l'utilisation d'ondes planes est leur facilité d'implémentation

et de dérivation. C'est pourquoi les autres possibilités n'ont pas été investiguées dans cette thèse. Néanmoins, elles pourraient être étudiées dans le futur pour améliorer les performances de notre solveur.

Nous associons aux espaces de Trefftz (discrets ou non) une formule de réciprocité. Elle est basée sur la formule des travaux virtuels. Ainsi, la formule de réciprocité établit un lien entre la physique du phénomène et sa modélisation mathématique. De plus, elle s'appuie sur les restrictions à $T \in \mathcal{T}$ de la composante tangentielle du champ électrique et de la trace tangentielle du champ magnétique, respectivement définies comme

$$\gamma_t \mathbf{E}^T := \left(\mathbf{n}_T \times \mathbf{E}^T\right) \times \mathbf{n}_T \quad \text{et} \quad \gamma_\times^T \mathbf{H}^T := \mathbf{n}_T \times \mathbf{H}^T, \tag{2.6}$$

où $\mathbf{n}_T \in \mathbb{R}^3$ est la normale sortante unitaire de ∂T.

Proposition 2.1 (Formule de réciprocité)**.** *Sur $\mathbb{X}_\mathcal{T}$, nous avons la formule de réciprocité suivante : pour tout $\mathbb{E} \in \mathbb{X}_\mathcal{T}$ et $\mathbb{E}' \in \mathbb{X}_\mathcal{T}$,*

$$\begin{aligned} r(\mathbb{E}, \mathbb{E}') &:= \sum_{T \in \mathcal{T}} r_T(\mathbb{E}, \mathbb{E}') = 0, \\ \textit{avec} \quad r_T(\mathbb{E}, \mathbb{E}') &:= \int_{\partial T} \left(\gamma_\times^T \mathbf{H}^T \cdot \overline{\gamma_t \mathbf{E}'^T} + \gamma_t \mathbf{E}^T \cdot \gamma_\times^T \overline{\mathbf{H}'^T}\right). \end{aligned} \tag{2.7}$$

Démonstration. Nous prenons $\mathbb{E}^T = \left(\mathbf{E}^T, \mathbf{H}^T\right)$ et $\mathbb{E}'^T = \left(\mathbf{E}'^T, \mathbf{H}'^T\right)$ dans $\mathbb{X}_T$. À partir de l'aspect physique du phénomène électromagnétique, nous écrivons la formule des travaux virtuels

$$W_T = ik_0 \int_T \varepsilon_r^T \mathbf{E}^T \cdot \overline{\mathbf{E}'^T} + \mu_r^T \mathbf{H}^T \cdot \overline{\mathbf{H}'^T}, \quad \text{pour } T \in \mathcal{T}.$$

Puisque $\mathbb{E}^T \in \mathbb{X}_T$ et $\mathbb{E}'^T \in \mathbb{X}_T$, et donc satisfont (2.1), nous avons

$$W_T = \int_T \nabla \times \mathbf{H}^T \cdot \overline{\mathbf{E}'^T} - \nabla \times \mathbf{E}^T \cdot \overline{\mathbf{H}'^T} = \int_T \mathbf{H}^T \cdot \overline{\nabla \times \mathbf{E}'^T} - \mathbf{E}^T \cdot \overline{\nabla \times \mathbf{H}'^T}.$$

En soustrayant ces deux dernières expressions de W_T, nous avons

$$\int_T \nabla \times \mathbf{H}^T \cdot \overline{\mathbf{E}'^T} - \mathbf{H}^T \cdot \overline{\nabla \times \mathbf{E}'^T} + \int_T \mathbf{E}^T \cdot \overline{\nabla \times \mathbf{H}'^T} - \nabla \times \mathbf{E}^T \cdot \overline{\mathbf{H}'^T} = 0.$$

Grâce à la formule de Stokes [78], nous obtenons la formule de réciprocité locale

$$r_T(\mathbb{E}, \mathbb{E}') = \int_{\partial T} \left((\mathbf{n}_T \times \mathbf{H}^T) \cdot \overline{\mathbf{E}'^T} + \mathbf{E}^T \cdot (\mathbf{n}_T \times \overline{\mathbf{H}'^T})\right) = 0, \quad \text{pour } \mathbb{E} \text{ et } \mathbb{E}' \text{ dans } \mathbb{X}_\mathcal{T}.$$

La preuve se termine en utilisant les définitions (2.6). □

Remarque 2.4. *La formule de réciprocité* (2.7) *est alors équivalente à*

$$r_T(\mathbb{E},\mathbb{E}') := \sum_{F\in\mathcal{F}_T}\int_F \left(\gamma_\times^T\mathbf{H}^T\cdot\overline{\gamma_t\mathbf{E}'^T}+\gamma_t\mathbf{E}^T\cdot\gamma_\times^T\overline{\mathbf{H}'^T}\right)=0.$$

2.2 Formulations variationnelles Trefftz vues par les formes consistantes

Une formulation variationnelle Trefftz est construite à partir de la formule de réciprocité (2.7). Cette formule n'assure cependant aucun contrôle sur le niveau de continuité de la solution numérique, au sens où nous ne pouvons rien garantir sur les sauts de la solution numérique entre les éléments du maillage. Nous perturbons alors la formule de réciprocité en ajoutant des formes consistantes élémentaires bien choisies qui assurent, à convergence, la continuité de la solution aux interfaces du maillage.

2.2.1 Formes consistantes intérieures et de bord

Avant de déterminer les formes consistantes intérieures et de bord, nous introduisons tout d'abord les notations générales du problème variationnel de Trefftz.

Problème 7 (Problème variationnel de Trefftz)**.** *Le problème variationnel de Trefftz est :*

Trouver $\mathbb{E}\in\mathbb{X}_\mathcal{T}$ *tel que pour tout* $\mathbb{E}'\in\mathbb{X}_\mathcal{T}$

$$a(\mathbb{E},\mathbb{E}')=\ell(\mathbb{E}'), \tag{2.8}$$

avec a *et* ℓ *définies selon deux conventions possibles. D'une part, un assemblage par éléments conduit à*

$$a(\mathbb{E},\mathbb{E}')=\sum_{T\in\mathcal{T}}\sum_{F\in\mathcal{F}_T}a_{T,F}(\mathbb{E},\mathbb{E}'), \tag{2.9}$$

et

$$\ell(\mathbb{E}')=\sum_{T\in\mathcal{T}}\sum_{F\in\mathcal{F}_T}\ell_{T,F}(\mathbb{E}'). \tag{2.10}$$

D'autre part, un assemblage par faces conduit à

$$a(\mathbb{E},\mathbb{E}')=\sum_{F\in\mathcal{F}}\mathsf{a}_F(\mathbb{E},\mathbb{E}'), \tag{2.11}$$

et

$$\ell(\mathbb{E}')=\sum_{F\in\mathcal{F}}\mathsf{l}_F(\mathbb{E}').$$

Dans le but de déterminer la forme sesquilinéaire a et la forme antilinéaire ℓ, nous construisons les formes consistantes élémentaires pour les faces intérieures et extérieures.

Nous prenons $\mathbb{E} \in \mathbb{X}_{\mathcal{T}}$ et $\mathbb{E}' \in \mathbb{X}_{\mathcal{T}}$. La solution exacte $\mathbb{E}$ vérifie sur $F \in \mathcal{F}_{\text{int}}$ séparant deux éléments T et K

$$(\gamma_t \mathbf{E})_{|F} = (\gamma_t \mathbf{E}^T)_{|F} = (\gamma_t \mathbf{E}^K)_{|F} \quad \text{et} \quad (\gamma_\times^T \mathbf{H})_{|F} = (\gamma_\times^T \mathbf{H}^T)_{|F} = (\gamma_\times^T \mathbf{H}^K)_{|F}. \tag{2.12}$$

L'existence des deux opérateurs de trace γ_t et $\gamma_\times^T$ nous permettent de construire quatre formules faibles équivalentes à (2.12) pour $F \in \mathcal{F}$:

$$\underbrace{\int_F \gamma_t \mathbf{E}^T \cdot \gamma_t \overline{\mathbf{E}'^T}}_{:=a_{T,F}^{1,1}(\mathbb{E},\mathbb{E}')} - \underbrace{\int_F \gamma_t \mathbf{E}^K \cdot \gamma_t \overline{\mathbf{E}'^T}}_{:=a_{T,F}^{1,2}(\mathbb{E},\mathbb{E}')} = \int_F \left(\gamma_t \mathbf{E}^T - \gamma_t \mathbf{E}^K\right) \cdot \gamma_t \overline{\mathbf{E}'^T} = 0, \tag{2.13a}$$

$$\underbrace{\int_F \gamma_t \mathbf{E}^T \cdot \gamma_\times^T \overline{\mathbf{H}'^T}}_{:=a_{T,F}^{2,1}(\mathbb{E},\mathbb{E}')} - \underbrace{\int_F \gamma_t \mathbf{E}^K \cdot \gamma_\times^T \overline{\mathbf{H}'^T}}_{:=a_{T,F}^{2,2}(\mathbb{E},\mathbb{E}')} = \int_F \left(\gamma_t \mathbf{E}^T - \gamma_t \mathbf{E}^K\right) \cdot \gamma_\times^T \overline{\mathbf{H}'^T} = 0, \tag{2.13b}$$

$$\underbrace{\int_F \gamma_\times^T \mathbf{H}^T \cdot \gamma_t \overline{\mathbf{E}'^T}}_{:=a_{T,F}^{3,1}(\mathbb{E},\mathbb{E}')} - \underbrace{\int_F \gamma_\times^T \mathbf{H}^K \cdot \gamma_t \overline{\mathbf{E}'^T}}_{:=a_{T,F}^{3,2}(\mathbb{E},\mathbb{E}')} = \int_F \left(\gamma_\times^T \mathbf{H}^T - \gamma_\times^T \mathbf{H}^K\right) \cdot \gamma_t \overline{\mathbf{E}'^T} = 0, \tag{2.13c}$$

$$\underbrace{\int_F \gamma_\times^T \mathbf{H}^T \cdot \gamma_\times^T \overline{\mathbf{H}'^T}}_{:=a_{T,F}^{4,1}(\mathbb{E},\mathbb{E}')} - \underbrace{\int_F \gamma_\times^T \mathbf{H}^K \cdot \gamma_\times^T \overline{\mathbf{H}'^T}}_{:=a_{T,F}^{4,2}(\mathbb{E},\mathbb{E}')} = \int_F \left(\gamma_\times^T \mathbf{H}^T - \gamma_\times^T \mathbf{H}^K\right) \cdot \gamma_\times^T \overline{\mathbf{H}'^T} = 0, \tag{2.13d}$$

où nous définissons huit formes consistantes élémentaires $a_{T,F}^{i,1}$ et $a_{T,F}^{i,2}$, pour $i = 1, 4$.

Remarque 2.5. *Ces formes sont appelées consistantes par abus, comme ce sont en réalité leurs soustractions (pour chaque $i = 1, 4$) qui sont nulles pour la solution exacte. Nous avons choisi de distinguer $a_{T,F}^{i,1}$ et $a_{T,F}^{i,2}$ pour nous permettre d'avoir des notations adaptées à la fois au cas des faces intérieures et à celui des faces extérieures.*

Les formes $a_{T,F}^{i,1}$ (*resp.* $a_{T,F}^{i,2}$) représentent une interaction de l'élément T sur lui-même (*resp.* de l'élément T avec son voisin K). Il est important de remarquer ici que ces formes consistantes dites élémentaires sont bien définies relativement à un élément T. De plus, la notion d'interactions fait référence aux intégrales impliquant les produits scalaires de

1. deux traces de fonctions provenant de l'élément T : interaction sur lui-même,
2. une trace de fonction provenant de l'élément T et une autre de l'élément K : interaction de T avec son voisin.

Pour construire la formulation Trefftz, l'idée est de perturber la formule de réciprocité en utilisant ces formes. Pour les faces $F \in \mathcal{F}_{\text{int}} \cap \mathcal{F}_T$ partageant deux éléments T et K, nous

associons à chaque couple de formes $a^{i,1}_{T,F}$ et $a^{i,2}_{T,F}$, un coefficient $\alpha^i_{T,F} \in \mathbb{R}$, $i = 1, 4$. Comme les équations (2.13a), (2.13b), (2.13c) et (2.13d), sont toutes égales à zéro, elles peuvent être ajoutées $\alpha^i_{T,F}$ fois à la formule de réciprocité. En effet, nous avons

$$\alpha^1_{T,F} a^{1,1}_{T,F}(\mathbb{E}, \mathbb{E}') - \alpha^1_{T,F} a^{1,2}_{T,F}(\mathbb{E}, \mathbb{E}') = 0, \tag{2.14a}$$

$$\alpha^2_{T,F} a^{2,1}_{T,F}(\mathbb{E}, \mathbb{E}') - \alpha^2_{T,F} a^{2,2}_{T,F}(\mathbb{E}, \mathbb{E}') = 0, \tag{2.14b}$$

$$\alpha^3_{T,F} a^{3,1}_{T,F}(\mathbb{E}, \mathbb{E}') - \alpha^3_{T,F} a^{3,2}_{T,F}(\mathbb{E}, \mathbb{E}') = 0, \tag{2.14c}$$

$$\alpha^4_{T,F} a^{4,1}_{T,F}(\mathbb{E}, \mathbb{E}') - \alpha^4_{T,F} a^{4,2}_{T,F}(\mathbb{E}, \mathbb{E}') = 0. \tag{2.14d}$$

Finalement, avec un assemblage par éléments, nous définissons $a_{T,F}$ et $\ell_{T,F}$ pour les faces intérieures du maillage, voir (2.9) et (2.10).

Définition 2.1. *Soient $T \in \mathcal{T}$ et $F \in \mathcal{F}_{\text{int}} \cap \mathcal{F}_T$. La forme sesquilinéaire $a_{T,F} : \mathbb{X}_{\mathcal{T}} \times \mathbb{X}_{\mathcal{T}} \to \mathbb{C}$ est définie par*

$$a_{T,F}(\mathbb{E}, \mathbb{E}') := \sum_{i=1}^{4} \alpha^i_{T,F} a^i_{T,F}(\mathbb{E}, \mathbb{E}'), \quad \textit{avec } a^i_{T,F}(\mathbb{E}, \mathbb{E}') := a^{i,1}_{T,F}(\mathbb{E}, \mathbb{E}') - a^{i,2}_{T,F}(\mathbb{E}, \mathbb{E}'),$$

où les formes $a^i_{T,F} : \mathbb{X}_{\mathcal{T}} \times \mathbb{X}_{\mathcal{T}} \to \mathbb{C}$ sont définies en (2.13a), (2.13b), (2.13c) et (2.13d).

De plus, nous définissons la forme antilinéaire $\ell_{T,F} : \mathbb{X}_{\mathcal{T}} \to \mathbb{C}$ par

$$\ell_{T,F}(\mathbb{E}') = 0.$$

Chacun des coefficients correspond à des interactions de l'élément T sur lui-même ou avec un voisin K. La formulation variationnelle pour $F \in \mathcal{F}_{\text{int}}$, s'écrit comme un tableau de coefficients, voir le Tableau 2.1. Dans ce tableau, nous symbolisons les types d'interactions par

- $a^{\mathbf{E},\mathbf{E}}$ pour les intégrales impliquant deux traces de type champ électrique $\gamma_t \mathbf{E}^{T \text{ou} K}$ et $\gamma_t \mathbf{E}'^{T}$, voir (2.14a),
- $a^{\mathbf{H},\mathbf{E}}$ pour les intégrales impliquant une trace de type champ électrique $\gamma_t \mathbf{E}^{T \text{ou} K}$ et une trace de type champ magnétique $\gamma^T_\times \mathbf{H}'^{T}$, voir (2.14b),
- $a^{\mathbf{E},\mathbf{H}}$ pour les intégrales impliquant une trace de type champ magnétique $\gamma^T_\times \mathbf{H}^{T \text{ou} K}$ et une trace de type champ électrique $\gamma_t \mathbf{E}'^{T}$, voir (2.14c),
- $a^{\mathbf{H},\mathbf{H}}$ pour les intégrales impliquant deux traces de type champ magnétique $\gamma^T_\times \mathbf{H}^{T \text{ou} K}$ et $\gamma^T_\times \mathbf{H}'^{T}$, voir (2.14d).

Interactions sur T	$a^{\mathbf{E},\mathbf{E}}$	$a^{\mathbf{E},\mathbf{H}}$	$a^{\mathbf{H},\mathbf{E}}$	$a^{\mathbf{H},\mathbf{H}}$
Lui-même	$\alpha^1_{T,F}$	$\alpha^2_{T,F}$	$\alpha^3_{T,F}$	$\alpha^4_{T,F}$
Voisin	$-\alpha^1_{T,F}$	$-\alpha^2_{T,F}$	$-\alpha^3_{T,F}$	$-\alpha^4_{T,F}$

TABLE 2.1 – Tableau de coefficients d'une formulation Trefftz pour $F \in \mathcal{F}_{\text{int}}$.

Maintenant, nous définissons les formes consistantes élémentaires pour les faces de bord $F \in \mathcal{F}_{\text{ext}}$. Dans ce cas, l'élément K, voisin de T, peut être assimilé à un élément virtuel K_{ext} extérieur au domaine Ω. Sur la face commune F séparant T et K_{ext}, différentes conditions aux limites peuvent être imposées : condition de Dirichlet, condition de Neumann ou condition d'impédance. Une condition de bord implique la présence d'un champ incident $\mathbb{E}_{inc}$ sur $F \in \mathcal{F}_{\text{ext}} \cap \mathcal{F}_T$. Nous allons voir par la suite que cela induit des formes sesquilinéaires associées au bord et des formes linéaires associées au second membre du Problème variationnel 7. Ainsi, les formes consistantes associées à une face extérieure correspondent à des interactions de bord (interaction entre T et K_{ext}) et à des interactions du second membre. De nouveau, ces formes sont utiles pour la perturbation de la formule de réciprocité afin de créer la formulation. L'ajout est réalisé grâce aux quatre coefficients réels β^1, β^2, β^3 et β^4.

Une condition aux limites de Dirichlet s'écrit

$$\gamma_t \mathbf{E} = \gamma_t \mathbf{E}_{inc} \text{ sur } \partial\Omega, \quad \text{où} \quad \mathbf{E}_{inc} : \mathbb{R}^3 \to \mathbb{C}^3. \tag{2.15}$$

Nous considérons une face $F \in \mathcal{F}_{\text{ext}} \cap \mathcal{F}_T$ d'un élément T et située sur le bord du domaine $\partial\Omega$. Nous pouvons reformuler la restriction de (2.15) à la face F en utilisant les deux identités suivantes : $\forall \mathbb{E}' = (\mathbf{E}', \mathbf{H}') \in \mathbb{X}_{\mathcal{T}}$,

$$\underbrace{\int_F \gamma_t \mathbf{E}^T \cdot \gamma_t \overline{\mathbf{E}'^T}}_{a^{1,1}_{T,F}(\mathbb{E},\mathbb{E}')} = \underbrace{\int_F \gamma_t \mathbf{E}^T_{inc} \cdot \gamma_t \overline{\mathbf{E}'^T}}_{:=\ell^1_{T,F}(\mathbb{E}')}, \tag{2.16a}$$

$$\underbrace{\int_F \gamma_t \mathbf{E}^T \cdot \gamma^T_\times \overline{\mathbf{H}'^T}}_{a^{2,1}_{T,F}(\mathbb{E},\mathbb{E}')} = \underbrace{\int_F \gamma_t \mathbf{E}^T_{inc} \cdot \gamma^T_\times \overline{\mathbf{H}'^T}}_{:=\ell^2_{T,F}(\mathbb{E}')}, \tag{2.16b}$$

où nous reconnaissons les formes consistantes élémentaires $a^{1,1}_{T,F}$ et $a^{2,1}_{T,F}$ précédentes. Ces formes peuvent être ajoutées autant de fois que nous le souhaitons à la formule de réciprocité (2.7). La première, (2.16a), est ajoutée β^1_F fois. La seconde, (2.16b), est ajoutée β^2_F fois. Cela s'écrit :

$$\beta^1_F a^{1,1}_{T,F}(\mathbb{E}, \mathbb{E}') = \beta^1_F \ell^1_{T,F}(\mathbb{E}'),$$

$$\beta_F^2 a_{T,F}^{2,1}(\mathbb{E},\mathbb{E}') = \beta_F^2 \ell_{T,F}^2(\mathbb{E}').$$

Nous remarquons que les formes $\ell_{T,F}^1$ et $\ell_{T,F}^2$ peuvent s'exprimer à l'aide de $a_{T,F}^{1,1}$ et $a_{T,F}^{2,1}$ de (2.13a). Ainsi, nous avons

$$\beta_F^1 a_{T,F}^{1,1}(\mathbb{E},\mathbb{E}') = \beta_F^1 a_{T,F}^{1,1}(\mathbb{E}_{inc},\mathbb{E}'),$$

$$\beta_F^2 a_{T,F}^{2,1}(\mathbb{E},\mathbb{E}') = \beta_F^2 a_{T,F}^{2,1}(\mathbb{E}_{inc},\mathbb{E}')\,.$$

Finalement, nous définissons les formes $a_{T,F}$ et $\ell_{T,F}$ pour une face $F \in \mathcal{F}_{\text{ext}} \cap \mathcal{F}_T$ dans le cas d'une condition de bord de Dirichlet.

Définition 2.2. *Soit une face $F \in \mathcal{F}_{\text{ext}} \cap \mathcal{F}_T$ associée à un élément $T \in \mathcal{T}$, sur laquelle nous imposons une condition aux limites de Dirichlet. Nous définissons les formes sesquilinéaire $a_{T,F}$ et antilinéaire $\ell_{T,F}$ sur $\mathbb{X}_{\mathcal{T}}$ par*

$$a_{T,F}(\mathbb{E},\mathbb{E}') = \beta_F^1 a_{T,F}^{1,1}(\mathbb{E},\mathbb{E}') + \beta_F^2 a_{T,F}^{2,1}(\mathbb{E},\mathbb{E}'),$$

et

$$\ell_{T,F}(\mathbb{E}') = \beta_F^1 a_{T,F}^{1,1}(\mathbb{E}_{inc},\mathbb{E}') + \beta_F^2 a_{T,F}^{2,1}(\mathbb{E}_{inc},\mathbb{E}').$$

Pour une condition de bord de Dirichlet, nous avons finalement le Tableau 2.2.

Interactions sur T	$a^{\mathbf{E},\mathbf{E}}$	$a^{\mathbf{E},\mathbf{H}}$	$a^{\mathbf{H},\mathbf{E}}$	$a^{\mathbf{H},\mathbf{H}}$
Bord	β_F^1	β_F^2	0	0
2^{nd} membre	β_F^1	β_F^2	0	0

TABLE 2.2 – Tableau de coefficients d'une formulation Trefftz pour $F \in \mathcal{F}_{\text{ext}} \cap \mathcal{F}_T$ avec une condition de bord de Dirichlet.

Une condition aux limites de Neumann prend la forme

$$\gamma_\times \mathbf{H} = \gamma_\times \mathbf{H}_{inc} \text{ sur } \partial\Omega, \quad \text{où} \quad \mathbf{H}_{inc} : \mathbb{R}^3 \to \mathbb{C}^3. \tag{2.19}$$

Soit une face $F \in \mathcal{F}_{\text{ext}} \cap \mathcal{F}_T$. Nous pouvons restreindre (2.19) à la face F en utilisant les deux identités suivantes :

$$\underbrace{\int_F \gamma_\times^T \mathbf{H}^T \cdot \gamma_t \overline{\mathbf{E}'^T}}_{a_{T,F}^{3,1}(\mathbb{E},\mathbb{E}')} = \underbrace{\int_F \gamma_\times^T \mathbf{H}_{inc} \cdot \gamma_t \overline{\mathbf{E}'^T}}_{:=\ell_{T,F}^{3,1}(\mathbb{E}')} \tag{2.20}$$

$$\underbrace{\int_F \gamma_\times^T \mathbf{H}^T \cdot \gamma_\times^T \overline{\mathbf{H}'^T}}_{a_{T,F}^{4,1}(\mathbb{E},\mathbb{E}')} = \underbrace{\int_F \gamma_\times^T \mathbf{H}_{inc} \cdot \gamma_\times^T \overline{\mathbf{H}'^T}}_{:=\ell_{T,F}^{4,1}(\mathbb{E}')}; \tag{2.21}$$

où nous reconnaissons dans les membres de droite les formes consistantes $a_{T,F}^{3,1}$ et $a_{K,F}^{4,1}$ appliquées au champ incident $\mathbb{E}_{inc}$. Les équations (2.20) et (2.21), peuvent être ajoutées β_F^3 et β_F^4 fois à la formule de réciprocité

$$\beta_F^3 a_{T,F}^{3,1}(\mathbb{E},\mathbb{E}') = \beta_F^3 a_{T,F}^{3,1}(\mathbb{E}_{inc},\mathbb{E}'),$$

$$\beta_F^4 a_{T,F}^{4,1}(\mathbb{E},\mathbb{E}') = \beta_F^4 a_{T,F}^{4,1}(\mathbb{E}_{inc},\mathbb{E}').$$

Définition 2.3. *Soit $F \in \mathcal{F}_{\text{ext}} \cap \mathcal{F}_T$ associée à un élément $T \in \mathcal{T}$ sur laquelle nous imposons une condition aux limites de Neumann. Nous définissons les formes sesquilinéaire $a_{T,F}$ et antilinéaire $\ell_{T,F}$ sur $\mathbb{X}_\mathcal{T}$ par*

$$a_{T,F}(\mathbb{E},\mathbb{E}') = \beta_F^3 a_{T,F}^{3,1}(\mathbb{E},\mathbb{E}') + \beta_F^4 a_{T,F}^{4,1}(\mathbb{E},\mathbb{E}'),$$

et

$$\ell_{T,F}(\mathbb{E}') = \beta_F^3 a_{T,F}^{3,1}(\mathbb{E}_{inc},\mathbb{E}') + \beta_F^4 a_{T,F}^{4,1}(\mathbb{E}_{inc},\mathbb{E}').$$

Ainsi, sous la condition (2.19), le tableau d'interactions devient le Tableau 2.3.

Interactions sur T	$a^{\mathbf{E},\mathbf{E}}$	$a^{\mathbf{E},\mathbf{H}}$	$a^{\mathbf{H},\mathbf{E}}$	$a^{\mathbf{H},\mathbf{H}}$
Bord	0	0	β_F^3	β_F^4
2^{nd} membre	0	0	β_F^3	β_F^4

TABLE 2.3 – Tableau de coefficients d'une formulation Trefftz pour $F \in \mathcal{F}_{\text{ext}} \cap \mathcal{F}_T$ avec une condition de bord de Neumann.

Nous pouvons écrire la condition de bord d'impédance (1) sous la forme

$$\gamma_t \mathbf{E} + Z_{\partial\Omega}\, \mathbf{n}_{\partial\Omega} \times \gamma_t \mathbf{H} = \gamma_t \mathbf{E}_{inc} + Z_{\partial\Omega}\, \mathbf{n}_{\partial\Omega} \times \gamma_t \mathbf{H}_{inc} \text{ sur } \partial\Omega,$$
$$\text{où } \mathbb{E}_{inc} : \mathbb{R}^3 \times \mathbb{R}^3 \to \left(\mathbb{C}^3\right)^2. \tag{2.22}$$

Nous reconnaissons ici des formes consistantes élémentaires déjà définies précédemment. Elles sont de nouveau associées aux coefficients β_F^i, $i = 1, 4$. Cependant, pour déterminer les valeurs des β_F^i, nous introduisons de nouveaux coefficients $\kappa \in \mathbb{R}$ et $\delta \in \mathbb{R}$. Nous écrivons tout d'abord la restriction de (2.22) aux faces extérieures d'un élément T : soit $F \in \mathcal{F}_{\text{ext}} \cap \mathcal{F}_T$,

$$\gamma_t \mathbf{E}^T + Z_F\, \gamma_\times^T \mathbf{H}^T = \gamma_t \mathbf{E}_{inc} + Z_F\, \gamma_\times^T \mathbf{H}_{inc} \text{ sur } F,$$

avec $Z_F = Z_{\partial\Omega|F}$. Nous reformulons alors cette condition sous la forme faible pour chaque face $F \in \mathcal{F}_{\text{ext}} \cap \mathcal{F}_T$: pour tout $\mathbb{E}'^T = \left(\mathbf{E}'^T, \mathbf{H}'^T\right) \in \mathbb{X}_T$,

$$Z_F \underbrace{\int_F \gamma_\times^T \mathbf{H}^T \cdot \gamma_\times^T \overline{\mathbf{H}'^T}}_{a_{T,F}^{4,1}(\mathbb{E},\mathbb{E}')} + \underbrace{\int_F \gamma_t \mathbf{E}^T \cdot \gamma_\times^T \overline{\mathbf{H}'^T}}_{a_{T,F}^{2,1}(\mathbb{E},\mathbb{E}')} = Z_F \underbrace{\int_F \gamma_\times^T \mathbf{H}_{inc} \cdot \gamma_\times^T \overline{\mathbf{H}'^T}}_{\ell_{T,F}^4(\mathbb{E}')} + \underbrace{\int_F \gamma_t \mathbf{E}_{inc} \cdot \gamma_\times^T \overline{\mathbf{H}'^T}}_{\ell_{T,F}^2(\mathbb{E}')},$$

et

$$Z_F \underbrace{\int_F \gamma_\times^T \mathbf{H}^T \cdot \overline{\gamma_t \mathbf{E}'^T}}_{a_{T,F}^{3,1}(\mathbb{E},\mathbb{E}')} + \underbrace{\int_F \gamma_t \mathbf{E}^T \cdot \overline{\gamma_t \mathbf{E}'^T}}_{a_{T,F}^{1,1}(\mathbb{E},\mathbb{E}')} = Z_F \underbrace{\int_F \gamma_\times^T \mathbf{H}_{inc} \cdot \overline{\gamma_t \mathbf{E}'^T}}_{\ell_{T,F}^3(\mathbb{E}')} + \underbrace{\int_F \gamma_t \mathbf{E}_{inc} \cdot \overline{\gamma_t \mathbf{E}'^T}}_{\ell_{T,F}^1(\mathbb{E}')}.$$

Les combinaisons de formes consistantes élémentaires ci-dessus peuvent être ajoutées κ ou δ fois à la formule de réciprocité (2.7) :

$$Z_F\, \delta\, a_{T,F}^{4,1}(\mathbb{E},\mathbb{E}') + \delta a_{T,F}^{2,1}(\mathbb{E},\mathbb{E}') = Z_F\, \delta\, \ell_{T,F}^4(\mathbb{E}') + \delta \ell_{T,F}^2(\mathbb{E}'),$$

$$Z_F\, \kappa\, a_{T,F}^{3,1}(\mathbb{E},\mathbb{E}') + \kappa a_{T,F}^{1,1}(\mathbb{E},\mathbb{E}') = Z_F\, \kappa\, \ell_{T,F}^3(\mathbb{E}') + \kappa \ell_{T,F}^1(\mathbb{E}').$$

De nouveau, nous reconnaissons ici les formes consistantes élémentaires $a_{T,F}^{1,1}$, $a_{T,F}^{2,1}$, $a_{T,F}^{3,1}$ et $a_{T,F}^{4,1}$ considérées pour un champ incident $\mathbb{E}_{inc}$

$$Z_F\, \delta\, a_{T,F}^{4,1}(\mathbb{E},\mathbb{E}') + \delta\, a_{T,F}^{2,1}(\mathbb{E},\mathbb{E}') = Z_F\, \delta\, a_{T,F}^{4,1}(\mathbb{E}_{inc},\mathbb{E}') + \delta\, a_{T,F}^{2,1}(\mathbb{E}_{inc},\mathbb{E}'),$$

$$Z_F\, \kappa\, a_{T,F}^{3,1}(\mathbb{E},\mathbb{E}') + \kappa a_{T,F}^{1,1}(\mathbb{E},\mathbb{E}') = Z_F\, \kappa\, a_{T,F}^{3,1}(\mathbb{E}_{inc},\mathbb{E}') + \kappa\, a_{T,F}^{1,1}(\mathbb{E}_{inc},\mathbb{E}').$$

Définition 2.4. *Soit $F \in \mathcal{F}_{\text{ext}} \cap \mathcal{F}_T$ associée à un élément $T \in \mathcal{T}$ sur laquelle nous imposons une condition de bord d'impédance. Nous définissons les formes sesquilinéaire $a_{T,F}$ et antilinéaire $\ell_{T,F}$ sur $\mathbb{X}_\mathcal{T}$ par*

$$a_{T,F}(\mathbb{E},\mathbb{E}') = \beta_F^1 a_{T,F}^{1,1}(\mathbb{E},\mathbb{E}') + \beta_F^2 a_{T,F}^{2,1}(\mathbb{E},\mathbb{E}') + \beta_F^3 a_{T,F}^{3,1}(\mathbb{E},\mathbb{E}') + \beta_F^4 a_{T,F}^{4,1}(\mathbb{E},\mathbb{E}'),$$

et

$$\ell_{T,F}(\mathbb{E}') = \beta_F^1 a_{T,F}^{1,1}(\mathbb{E}_{inc},\mathbb{E}') + \beta_F^2 a_{T,F}^{2,1}(\mathbb{E}_{inc},\mathbb{E}') + \beta_F^3 a_{T,F}^{3,1}(\mathbb{E}_{inc},\mathbb{E}') + \beta_F^4 a_{T,F}^{4,1}(\mathbb{E}_{inc},\mathbb{E}'),$$

où $\beta_F^1 = \kappa$, $\beta_F^2 = \delta$, $\beta_F^3 = Z_F\, \kappa$ et $\beta_F^4 = Z_F\, \delta$.

Finalement, nous écrivons $a_{T,F}$ et $\ell_{T,F}$ sous la forme d'un tableau de coefficients, voir le Tableau 2.4.

Les formes consistantes élémentaires sont maintenant définies. Il reste à exprimer les formulations variationnelles Trefftz générales en fonction de celles-ci.

Interactions sur T	$a^{\mathbf{E},\mathbf{E}}$	$a^{\mathbf{E},\mathbf{H}}$	$a^{\mathbf{H},\mathbf{E}}$	$a^{\mathbf{H},\mathbf{H}}$
Bord	$\beta_F^1 = \kappa$	$\beta_F^2 = \delta$	$\beta_F^3 = Z_F\,\kappa$	$\beta_F^4 = Z_F\,\delta$
2^{nd} membre	$\beta_F^1 = \kappa$	$\beta_F^2 = \delta$	$\beta_F^3 = Z_F\,\kappa$	$\beta_F^4 = Z_F\,\delta$

TABLE 2.4 – Tableau de coefficients d'une formulation Trefftz pour $F \in \mathcal{F}_{\text{ext}} \cap \mathcal{F}_T$ avec une condition de bord d'impédance.

2.2.2 Construction des formulations

Comme nous l'avons mentionné précédemment, les formulations variationnelles Trefftz définies sur toutes les faces $F \in \mathcal{F}$ s'écrivent grâce à une perturbation de la formule de réciprocité (2.7) par une combinaison de formes consistantes.

Dans un premier temps, nous exprimons la formule de réciprocité (2.7) à l'aide des formes consistantes définies en (2.13b) et (2.13c).

Définition 2.5. *Soient $\mathbb{E} \in \mathbb{X}_{\mathcal{T}}$ et $\mathbb{E}' \in \mathbb{X}_{\mathcal{T}}$, la formule de réciprocité (2.7) se décompose en formes consistantes comme*

$$r_T(\mathbb{E},\mathbb{E}') := \sum_{F\in\mathcal{F}_T} r_{T,F}(\mathbb{E},\mathbb{E}') := \sum_{F\in\mathcal{F}_T} a^{2,1}_{T,F}(\mathbb{E},\mathbb{E}') + a^{3,1}_{T,F}(\mathbb{E},\mathbb{E}') = 0.$$

Nous prenons un coefficient $\xi_T \in \mathbb{R}$ associé à la formule de réciprocité. En effet, étant nulle, elle peut aussi être ajoutée autant de fois que nous le souhaitons à la formulation Trefftz.

La forme sesquilinéaire a du Problème 7 s'écrit, au choix, sous la forme d'une somme sur les éléments $T \in \mathcal{T}$, voir (2.9), ou d'une somme sur les faces $F \in \mathcal{F}$, voir (2.11). Nous prenons l'exemple d'un assemblage par faces.

Problème 8 (Formulation variationnelle Trefftz avec formes consistantes)**.** *Trouver $\mathbb{E} \in \mathbb{X}_{\mathcal{T}}$ tel que pour tout $\mathbb{E}' \in \mathbb{X}_{\mathcal{T}}$*

$$a(\mathbb{E},\mathbb{E}') = \ell(\mathbb{E}'),$$

avec

$$a(\mathbb{E},\mathbb{E}') = \sum_{F\in\mathcal{F}} \mathsf{a}_F(\mathbb{E},\mathbb{E}') \quad \textit{et} \quad \ell(\mathbb{E}') = \sum_{F\in\mathcal{F}} \mathsf{l}_F(\mathbb{E}'),$$

où

$$\mathsf{a}_F(\mathbb{E},\mathbb{E}') \quad := \quad \begin{cases} r_F(\mathbb{E},\mathbb{E}') + \displaystyle\sum_{i=1}^{4} \mathsf{a}_F^i(\mathbb{E},\mathbb{E}') & \textit{pour } F \in \mathcal{F}_{\text{int}}, \\ r_F(\mathbb{E},\mathbb{E}') + \displaystyle\sum_{i=1}^{4} \mathsf{a}_F^i(\mathbb{E},\mathbb{E}') & \textit{pour } F \in \mathcal{F}_{\text{ext}}, \end{cases}$$

et

$$\mathsf{l}_F(\mathbb{E}') \quad := \quad \begin{cases} 0 & \text{pour } F \in \mathcal{F}_{\text{int}}, \\ \displaystyle\sum_{i=1}^{4} \mathsf{a}_F^i(\mathbb{E}_{inc}, \mathbb{E}') & \text{pour } F \in \mathcal{F}_{\text{ext}}, \end{cases}$$

avec

$$\mathsf{r}_F(\mathbb{E}, \mathbb{E}') := \xi_T \, r_{T,F}(\mathbb{E}, \mathbb{E}') + \xi_K \, r_{K,F}(\mathbb{E}, \mathbb{E}'), \text{ pour } F \in \mathcal{F},$$

où nous rappelons que pour $F \in \mathcal{F}_T$

$$r_{T,F}(\mathbb{E}, \mathbb{E}') = a_{T,F}^{2,1}(\mathbb{E}, \mathbb{E}') + a_{T,F}^{3,1}(\mathbb{E}, \mathbb{E}').$$

Dans ce problème, nous définissons aussi

$$\mathsf{a}_F^i(\mathbb{E}, \mathbb{E}') := \begin{cases} \mathsf{a}_F^{i,1}(\mathbb{E}, \mathbb{E}') - \mathsf{a}_F^{i,2}(\mathbb{E}, \mathbb{E}'), & \text{pour } F \in \mathcal{F}_{\text{int}} \text{ et } i = 1, 4, \\ \beta_F^i a_{T,F}^{i,1}(\mathbb{E}, \mathbb{E}'), & \text{pour } F \in \mathcal{F}_{\text{ext}} \cap \mathcal{F}_T \text{ et } i = 1, 4, \end{cases}$$

où nous avons pour F *séparant deux éléments* T *et* K

$$\begin{aligned} \mathsf{a}_F^{i,1}(\mathbb{E}, \mathbb{E}') &= \alpha_{T,F}^i \, a_{T,F}^{i,1}(\mathbb{E}, \mathbb{E}') + \alpha_{K,F}^i \, a_{K,F}^{i,1}(\mathbb{E}, \mathbb{E}'), \quad \text{pour } F \in \mathcal{F}_T \cap \mathcal{F}_K \text{ et } i = 1, 4, \\ \mathsf{a}_F^{i,2}(\mathbb{E}, \mathbb{E}') &= \alpha_{T,F}^i \, a_{T,F}^{i,2}(\mathbb{E}, \mathbb{E}') + \alpha_{K,F}^i \, a_{K,F}^{i,2}(\mathbb{E}, \mathbb{E}'), \quad \text{pour } F \in \mathcal{F}_T \cap \mathcal{F}_K \text{ et } i = 1, 4. \end{aligned}$$

Finalement, pour une condition de bord d'impédance par exemple, la formulation variationnelle Trefftz est résumée grâce au Tableau 2.5, où nous avons assemblé les Tableaux 2.1 et 2.4 et où $\beta_F^1 = \kappa$, $\beta_F^2 = \delta$, $\beta_F^3 = Z_F \, \kappa$ et $\beta_F^4 = Z_F \, \delta$ (voir le Tableau 2.4).

Interactions sur T	$a^{\mathbf{E},\mathbf{E}}$	$a^{\mathbf{E},\mathbf{H}}$	$a^{\mathbf{H},\mathbf{E}}$	$a^{\mathbf{H},\mathbf{H}}$
Lui-même	$\alpha_{T,F}^1$	$\alpha_{T,F}^2 + \xi_T$	$\alpha_{T,F}^3 + \xi_T$	$\alpha_{T,F}^4$
Voisin	$-\alpha_{T,F}^1$	$-\alpha_{T,F}^2$	$-\alpha_{T,F}^3$	$-\alpha_{T,F}^4$
Bord	β_F^1	$\beta_F^2 + \xi_T$	$\beta_F^3 + \xi_T$	β_F^4
2^{nd} membre	β_F^1	β_F^2	β_F^3	β_F^4

TABLE 2.5 – Tableau de coefficients d'une formulation Trefftz pour $F \in \mathcal{F}_T$ avec une condition de bord d'impédance.

2.2.3 Caractère bien posé des formulations Trefftz

D'après le théorème de Lax-Milgram, la convergence du schéma numérique dépend de la coercivité de la forme sesquilinéaire associée à la formulation variationnelle. Dans cette

section, nous prouvons, sous certaines conditions, la coercivité de la forme sesquilinéaire associée à une formulation Trefftz.

Le théorème suivant énonce les conditions sur les coefficients $\alpha^i_{T,F}$ et β^i_F pour assurer la coercivité de la forme sesquilinéaire a.

Théorème 2.1 (Coercivité faible d'une formulation Trefftz)**.** *Sous les conditions suivantes, pour toute face intérieure $F = T \cap K \in \mathcal{F}_{\text{int}}$:*

$$\begin{cases} \alpha^1_{T,F} = \alpha^1_{K,F} = \alpha^1_F > 0, \\ \alpha^4_{T,F} = \alpha^4_{K,F} = \alpha^4_F > 0, \\ \alpha^2_{T,F} = \alpha^3_{K,F}, \end{cases}$$

pour toute face extérieure $F \in \mathcal{F}_{\text{ext}}$: $\beta^1_F,\ \beta^4_F > 0$, et il existe une constante fixe $\gamma \in \mathbb{R}$ telle que :

$$\begin{cases} \alpha^2_{T,F} + \alpha^3_{T,F} = \gamma, \quad \forall T \in \mathcal{T} \text{ et } \forall F \in \mathcal{F}_T \cap \mathcal{F}_{\text{int}}, \\ \beta^2_F + \beta^3_F = \gamma, \quad \forall F \in \mathcal{F}_{\text{ext}}. \end{cases}$$

la forme sesquilinéaire $a : \mathbb{X}_{\mathcal{T}} \times \mathbb{X}_{\mathcal{T}} \to \mathbb{C}$ vérifie la propriété de coercivité suivante :

$$\Re(a(\mathbb{E}, \mathbb{E})) = \|\mathbb{E}\|^2_{\text{GD}},$$

où

$$\|\mathbb{E}\|^2_{\text{GD}} = \|\mathbb{E}\|^2_{\text{int}} + \|\mathbb{E}\|^2_{\text{ext}}, \tag{2.23}$$

avec

$$\|\mathbb{E}\|^2_{\text{int}} := \sum_{F \in \mathcal{F}_{\text{int}}} \int_F (\alpha^1_F \, |[\![\gamma_t \mathbf{E}]\!]_{T,F}|^2 + \alpha^4_F \, |[\![\gamma_\times \mathbf{H}]\!]_F|^2),$$

et

$$\|\mathbb{E}\|^2_{\text{ext}} := \sum_{F \in \mathcal{F}_{\text{ext}}} \int_F (\beta^1_F \, |\gamma_t \mathbf{E}|^2 + \beta^4_F \, |\gamma_\times \mathbf{H}|^2),$$

où

$$[\![\gamma_t \mathbf{E}]\!]_{T,F} := \gamma_t \mathbf{E}^T - \gamma_t \mathbf{E}^K \quad \text{et} \quad [\![\gamma_\times \mathbf{H}]\!]_F := \gamma^T_\times \mathbf{H}^T - \gamma^T_\times \mathbf{H}^K. \tag{2.25}$$

Démonstration. Tout d'abord, nous ajoutons $\xi = -\gamma/2 - \xi_T$ fois la réciprocité. La forme sesquilinéaire a se décompose sous la forme

$$a(\mathbb{E}, \mathbb{E}) = \sum_{F \in \mathcal{F}} C_F(\mathbb{E}, \mathbb{E}), \quad \text{pour } \mathbb{E} \in \mathbb{X}_{\mathcal{T}},$$

avec pour une face $F = T \cap K \in \mathcal{F}_{\text{int}}$ intérieure

$$\begin{aligned}
C_F(\mathbb{E},\mathbb{E}) &= \xi\left(\int_F \gamma_t\mathbf{E}^T\cdot\overline{\gamma_\times^T\mathbf{H}^T} + \gamma_\times^T\mathbf{H}^T\cdot\overline{\gamma_t\mathbf{E}^T}\right)\\
&+ \xi\left(\int_F \gamma_t\mathbf{E}^K\cdot\overline{\gamma_\times^K\mathbf{H}^K} + \gamma_\times^K\mathbf{H}^K\cdot\overline{\gamma_t\mathbf{E}^K}\right)\\
&+ \alpha^1_{T,F}\int_F\left(\gamma_t\mathbf{E}^T-\gamma_t\mathbf{E}^K\right)\cdot\overline{\gamma_t\mathbf{E}^T} + \alpha^1_{K,F}\int_F\left(\gamma_t\mathbf{E}^K-\gamma_t\mathbf{E}^T\right)\cdot\overline{\gamma_t\mathbf{E}^K}\\
&+ \alpha^2_{T,F}\int_F\left(\gamma_t\mathbf{E}^T-\gamma_t\mathbf{E}^K\right)\cdot\overline{\gamma_\times^T\mathbf{H}^T} + \alpha^2_{K,F}\int_F\left(\gamma_t\mathbf{E}^K-\gamma_t\mathbf{E}^T\right)\cdot\overline{\gamma_\times^K\mathbf{H}^K}\\
&+ \alpha^3_{T,F}\int_F\left(\gamma_\times^T\mathbf{H}^T-\gamma_\times^T\mathbf{H}^K\right)\cdot\overline{\gamma_t\mathbf{E}^T} + \alpha^3_{K,F}\int_F\left(\gamma_\times^K\mathbf{H}^K-\gamma_\times^K\mathbf{H}^T\right)\cdot\overline{\gamma_t\mathbf{E}^K}\\
&+ \alpha^4_{T,F}\int_F\left(\gamma_\times^T\mathbf{H}^T-\gamma_\times^T\mathbf{H}^K\right)\cdot\overline{\gamma_\times^T\mathbf{H}^T} + \alpha^4_{K,F}\int_F\left(\gamma_\times^K\mathbf{H}^K-\gamma_\times^K\mathbf{H}^T\right)\cdot\overline{\gamma_\times^K\mathbf{H}^K}.
\end{aligned}$$

Nous regroupons judicieusement les termes et tenons compte des hypothèses du théorème :

$$\begin{aligned}
C_F(\mathbb{E},\mathbb{E}) &= \left(\xi+\alpha^2_{T,F}\right)\Big[\int_F \gamma_t\mathbf{E}^T\cdot\overline{\gamma_\times^T\mathbf{H}^T} - \overline{\int_F \gamma_t\mathbf{E}^T\cdot\overline{\gamma_\times^T\mathbf{H}^T}}\Big]\\
&+ \left(\xi+\alpha^2_{K,F}\right)\Big[\int_F \gamma_t\mathbf{E}^K\cdot\overline{\gamma_\times^K\mathbf{H}^K} - \overline{\int_F \gamma_t\mathbf{E}^K\cdot\overline{\gamma_\times^K\mathbf{H}^K}}\Big]\\
&- \alpha^2_{T,F}\Big[\int_F \gamma_t\mathbf{E}^K\cdot\overline{\gamma_\times^T\mathbf{H}^T} - \overline{\int_F \gamma_t\mathbf{E}^K\cdot\overline{\gamma_\times^T\mathbf{H}^T}}\Big]\\
&- \alpha^2_{K,F}\Big[\int_F \gamma_t\mathbf{E}^T\cdot\overline{\gamma_\times^K\mathbf{H}^K} - \overline{\int_F \gamma_t\mathbf{E}^T\cdot\overline{\gamma_\times^K\mathbf{H}^K}}\Big]\\
&+ \alpha^1_{T,F}\Big[\int_F\left(\gamma_t\mathbf{E}^T-\gamma_t\mathbf{E}^K\right)\cdot\overline{\gamma_t\mathbf{E}^T} + \int_F\left(\gamma_t\mathbf{E}^K-\gamma_t\mathbf{E}^T\right)\cdot\overline{\gamma_t\mathbf{E}^K}\Big]\\
&+ \alpha^4_{T,F}\Big[\int_F\left(\gamma_\times^T\mathbf{H}^T-\gamma_\times^T\mathbf{H}^K\right)\cdot\overline{\gamma_\times^T\mathbf{H}^T} + \int_F\left(\gamma_\times^K\mathbf{H}^K-\gamma_\times^K\mathbf{H}^T\right)\cdot\overline{\gamma_\times^K\mathbf{H}^K}\Big].
\end{aligned}$$

Nous remarquons que les quatre premières lignes sont des nombres complexes imaginaires purs. Les deux dernières lignes nous donnent ensuite la définition de la semi-norme $\|\cdot\|^2_{\text{int}}$ pour une face intérieure F. Les faces extérieures se traitent de manière similaire. □

Sous ces conditions, la forme a du Problème 8 est coercive sur $\mathbb{X}_\mathcal{T}$. Nous montrons maintenant qu'elle est injective.

Théorème 2.2 (Injectivité des formulations)**.** *Sous les hypothèses du Théorème 2.1, la forme a associée au Problème 8, est injective. En effet, nous avons*

$$a(\mathbb{E},\mathbb{E}) = 0 \quad\Longrightarrow\quad \mathbb{E}\equiv 0.$$

Démonstration. Comme $a(\mathbb{E},\mathbb{E})=0$, nous avons $\|\mathbb{E}\|^2_{\text{GD}}=0$ et alors

$$[\![\gamma_t\mathbf{E}]\!]_{T,F} = [\![\gamma_\times\mathbf{H}]\!]_F = 0,\ \text{sur } F\in\mathcal{F}_{\text{int}} \quad\text{et}\quad \gamma_t\mathbf{E} = \gamma_\times^T\mathbf{H} = 0,\ \text{sur } F\in\mathcal{F}_{\text{ext}}. \tag{2.26}$$

Nous rappelons que si $\mathbb{E}^T$ vérifie (2.1) et s'il existe une face $F\in\mathcal{F}_T$ telle que $\gamma_t\mathbf{E}^T=0$ et

$\gamma_\times^T \mathbf{H}^T = 0$ sur F, alors grâce au théorème de prolongement unique $\mathbb{E}^T \equiv 0$.

Soit $\widetilde{\mathcal{T}} \subset \mathcal{T}$ l'ensemble des éléments défini par

$$\widetilde{\mathcal{T}} := \Big\{ T \in \mathcal{T} \;\Big|\; \mathbb{E}^T \equiv 0 \Big\}.$$

Nous devons maintenant montrer que $\widetilde{\mathcal{T}} = \mathcal{T}$.
(*a*) L'ensemble $\widetilde{\mathcal{T}}$ est non vide comme $\mathbb{E}^T$ s'annule dans tout élément T avec une face $F \subset \mathcal{F}_T$ incluse dans la frontière extérieure $\mathcal{F}_{\text{ext}}$, voir (2.26).
(*b*) Soit $T \in \mathcal{T}$ et $K \in \widetilde{\mathcal{T}}$ avec une face commune $F \in \mathcal{F}_T \cap \mathcal{F}_K$. Nous cherchons à montrer que $T \in \widetilde{\mathcal{T}}$. Avec (2.26), nous avons

$$(\gamma_t \mathbf{E}^T)_{|F} = (\gamma_t \mathbf{E}^K)_{|F} = 0 \quad \text{et} \quad (\gamma_\times^T \mathbf{H}^T)_{|F} = (-\gamma_\times^K \mathbf{H}^K)_{|F} = 0 \quad \text{sur } F \in \mathcal{F}_T \cap \mathcal{F}_K,$$

ce qui implique, par le théorème de prolongement unique,

$$\mathbb{E}^T = 0 \text{ dans T} \quad \Longrightarrow \quad T \in \widetilde{\mathcal{T}}.$$

(*c*) D'après le caractère complet du graphe du voisinage, il suit que $\mathcal{T} = \widetilde{\mathcal{T}}$. Ainsi, nous avons $\mathbb{E}^T = 0$. □

De cette façon, l'unicité de la solution du Problème 8 est prouvée. Son existence est assurée grâce au fait que la solution de (2.1), voir [78], est aussi une solution du Problème 8. Cependant, cela ne veut pas dire qu'une formulation variationnelle (2.8) est bien posée pour tout $\ell \in (\mathbb{X}_{\mathcal{T}})^*$. Ce dernier point reste une question ouverte.

Finalement, à condition d'ajouter judicieusement des formes consistantes à la formule de réciprocité (Théorème 2.1), la coercivité de la formulation est assurée, et il existe une unique solution au Problème 8. Une bonne perturbation, au sens où elle assure la coercivité, peut consister à utiliser des traces numériques obtenues par un solveur de Riemann ou upwind. C'est l'objet de la section suivante.

2.3 Formulations variationnelles Trefftz vues par les traces numériques

Nous avons construit dans les parties précédentes des formulations Trefftz par l'approche des formes consistantes. Bien que cette dernière nous donne des critères pour construire des formulations coercives, elle ne nous fournit pas un choix de coefficients effectifs. Cela peut entraîner des défauts d'approximation. Dans cette section, nous allons utiliser l'approche des solveurs de Riemann dans le cas homogène et des schémas upwind dans le cas hétérogène

pour réaliser ce choix. Tout d'abord, nous tenons à expliquer brièvement les démarches de construction d'une formulation variationnelle par l'utilisation de traces numériques. Ces dernières sont ensuite dérivées grâce à un solveur de Riemann ou à un schéma upwind. Enfin, nous prouvons la coercivité faible de la formulation variationnelle Trefftz upwind. Il s'avérera que ces deux approches sont équivalentes, voir le Chapitre 3.

2.3.1 Démarche de construction

La mise en place des formulations variationnelles Trefftz est précédemment passée par une perturbation de la formule de réciprocité. Dans la suite du manuscrit, l'idée reste la même. Cependant, il ne s'agit plus d'ajouter des formes consistantes mais de définir une trace numérique généralisée au sens où elle peut s'appliquer aux fonctions discontinues. Cette trace numérique est une extension de la définition de la trace d'une fonction continue pour les fonctions discontinues. En adoptant une approche similaire aux méthodes de GD mixtes polynomiales [43, 55], la formulation Trefftz est déduite en insérant les traces numériques [86], dans la formule de réciprocité (2.7).

Pour une fonction continue les composantes tangentielles des champs électromagnétiques s'écrivent

$$\left\{\begin{array}{rcccl} (\gamma_t\mathbf{E})_{|F} & = & (\gamma_t\mathbf{E}^T)_{|F} & = & (\gamma_t\mathbf{E}^K)_{|F}, \\ (\gamma_t\mathbf{H})_{|F} & = & (\gamma_t\mathbf{H}^T)_{|F} & = & (\gamma_t\mathbf{H}^K)_{|F}. \end{array}\right.$$

Cette identité n'est plus valable pour les fonctions discontinues et est remplacée par la notion de traces numériques

$$(\widehat{\gamma_t\mathbf{E}})_{|F} := \alpha_1^F\gamma_t\mathbf{E}^T + (1-\alpha_1^F)\gamma_t\mathbf{E}^K + \alpha_2^F(\gamma_\times^T\mathbf{H}^T + \gamma_\times^K\mathbf{H}^K), \tag{2.27a}$$

$$(\widehat{\gamma_t\mathbf{H}})_{|F} := \alpha_3^F\gamma_t\mathbf{H}^T + (1-\alpha_3^F)\gamma_t\mathbf{H}^K + \alpha_4^F(\gamma_\times^T\mathbf{E}^T + \gamma_\times^K\mathbf{E}^K), \tag{2.27b}$$

où $\alpha_i^F \in \mathbb{R}$ pour $i = 1, 4$ et où F est une face intérieure. Cette écriture assure que la trace numérique coïncide avec la trace classique quand elles sont appliquées aux fonctions continues. Toutefois, en fixant arbitrairement ces quatre coefficients nous obtenons le plus souvent des problèmes mal posés ne vérifiant pas les conditions du Théorème 2.1.

Nous introduisons aussi la notion de trace numérique de bord qui coïncide avec la notion de trace tangentielle pour les solutions vérifiant la condition d'impédance

$$(\widehat{\gamma_t\mathbf{E}})_{|F} := \gamma_t\mathbf{E} + \beta_1^F\Big(\gamma_t\mathbf{E} + Z_{\partial\Omega}\gamma_\times\mathbf{H} - \gamma_t\mathbf{g}\Big),$$

$$(\widehat{\gamma_t\mathbf{H}})_{|F} := \gamma_t\mathbf{H} + \beta_2^F\mathbf{n}_{\partial\Omega}\times\Big(\gamma_t\mathbf{E} + Z_{\partial\Omega}\gamma_\times\mathbf{H} - \gamma_t\mathbf{g}\Big),$$

où $\beta_i^F \in \mathbb{R}$ pour $i = 1, 2$ et où F est une face extérieure. La construction d'une formulation variationnelle Trefftz passe par deux étapes :

1. le choix de ces quatre paramètres par des arguments physiques,
2. l'insertion des traces numériques obtenues dans la formule de réciprocité locale (2.7) et la définition de la forme sesquilinéaire a comme

$$a(\mathbb{E}, \mathbb{E}') := \sum_{T \in \mathfrak{T}} \sum_{F \in \partial T} \int_F \left(\widehat{\gamma_\times^T \mathbf{H}} \cdot \overline{\gamma_t \mathbf{E}'^T} + \widehat{\gamma_t \mathbf{E}} \cdot \gamma_\times^T \overline{\mathbf{H}'^T} \right).$$

Dans cette thèse, les traces numériques sont obtenues grâce à un solveur de Riemann dans le cas homogène et grâce à un schéma upwind ou à une méthode UWVF dans le cas hétérogène. Nous expliquons les deux premières possibilités dans la présente section. L'obtention de traces numériques par une méthode UWVF sera présentée dans le Chapitre 3.

La seconde étape permet d'établir une formulation variationnelle Trefftz dont les coefficients sont déterminés avec les définitions des traces numériques obtenues. La coercivité sera alors démontrée grâce à une preuve directe bien que nous aurions pu aussi avoir recours au Théorème 2.1. Ce dernier point est implicitement attaché aux choix des paramètres de pénalisation liés à la physique associée au système hyperbolique de Maxwell.

Il s'avèrera que cette approche est donc équivalente à la méthode de perturbations par les formes consistantes.

2.3.2 Détermination des traces numériques en utilisant un problème de Riemann pour les milieux homogènes

Tout d'abord, le principe sera de considérer un problème hyperbolique ne dépendant que d'une seule des trois dimensions de $\mathbb{R}^3$. Ce choix permet une résolution analytique du système hyperbolique et d'en déduire les traces numériques de Riemann. En effet, elles prennent une forme nettement moins compliquée lorsque la solution ne dépend plus des trois coordonnées spatiales du domaine.

Par la suite, nous définissons les traces numériques de Riemann pour le problème initial en distinguant les faces intérieures et de bord. Elles sont obtenues en considérant des ondes à incidence normale aux frontières des éléments. Ces traces numériques ne sont déterminées que pour une solution dépendant d'une variable spatiale. Malgré cela, elles sont une valeur ajoutée pour la compréhension de la méthode de Trefftz aux interfaces entre les éléments du maillage.

Système hyperbolique de Maxwell

Tout problème de propagation peut être vu comme un système hyperbolique. En particulier, la solution $\mathbb{E}$ du Problème 1 satisfait aussi le problème hyperbolique général suivant

$$\frac{\partial \mathbb{E}}{\partial t}(\mathbf{x},t) = \mathbf{A}_1 \frac{\partial \mathbb{E}}{\partial \mathsf{x}}(\mathbf{x},t) + \mathbf{A}_2 \frac{\partial \mathbb{E}}{\partial \mathsf{y}}(\mathbf{x},t) + \mathbf{A}_3 \frac{\partial \mathbb{E}}{\partial \mathsf{z}}(\mathbf{x},t),$$

où $\mathbf{x} \in \mathbb{R}^3$, $t \in \mathbb{R}^+$, $\mathsf{x},\mathsf{y},\mathsf{z}$ sont les composantes de $\mathbb{R}^3$, et où nous adoptons la notation (comme dans la Section 1.2)

$$\mathbb{E} = \begin{bmatrix} \mathbf{E} \\ \mathbf{H} \end{bmatrix} = \begin{bmatrix} E_1 \\ E_2 \\ E_3 \\ H_1 \\ H_2 \\ H_3 \end{bmatrix}, \quad \text{où } \mathbf{E} = [E_1,E_2,E_3]^\mathsf{T} \text{ et } \mathbf{H} = [H_1,H_2,H_3]^\mathsf{T}.$$

Afin de définir les traces numériques de Riemann ayant des expressions simples, nous considérons des champs électromagnétiques à incidence normale par rapport à l'interface considérée. Nous supposons que les conditions initiales sont indépendantes de y et de z. Autrement dit, nous admettons que nous avons

$$\mathbf{E}(\mathsf{x},\mathsf{y},\mathsf{z},t) = \mathbf{E}(\mathsf{x},t) \quad \text{et} \quad \mathbf{H}(\mathsf{x},\mathsf{y},\mathsf{z},t) = \mathbf{H}(\mathsf{x},t).$$

Le système hyperbolique devient finalement

$$\frac{\partial \mathbb{E}}{\partial t}(\mathsf{x},t) = \mathbf{A}_1 \frac{\partial \mathbb{E}}{\partial \mathsf{x}}(\mathsf{x},t), \quad \text{où} \quad \mathbf{A}_1 : \mathbb{R}^6 \to \mathbb{R}^6. \tag{2.29}$$

Bien qu'associé à une solution qui ne dépend que de x et de t, ce problème reste bien posé. Sa solution est donnée par le théorème suivant.

Théorème 2.3 (Solution du système hyperbolique de Maxwell)**.** *La solution du système* (2.29) *est donnée par*

$$\begin{aligned} \mathbb{E}(\mathsf{x},t) \;=\; & E_1(\mathsf{x},0)\overrightarrow{\Phi}_1 + H_1(\mathsf{x},0)\overrightarrow{\Phi}_2 \\ & + \frac{E_2+H_3}{2}(\mathsf{x}-c_0t,0)\overrightarrow{\Phi}_3 + \frac{E_3-H_2}{2}(\mathsf{x}-c_0t,0)\overrightarrow{\Phi}_4 \\ & + \frac{E_2-H_3}{2}(\mathsf{x}+c_0t,0)\overrightarrow{\Phi}_5 + \frac{E_3+H_2}{2}(\mathsf{x}+c_0t,0)\overrightarrow{\Phi}_6, \end{aligned}$$

où c_0 est la vitesse d'une onde dans le vide, et où les vecteurs $\overrightarrow{\Phi}_i$ sont donnés, pour $i = 1, 6$, par

$$\overrightarrow{\Phi}_1 := \begin{bmatrix} 1 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}, \overrightarrow{\Phi}_2 := \begin{bmatrix} 0 \\ 0 \\ 0 \\ 1 \\ 0 \\ 0 \end{bmatrix}, \overrightarrow{\Phi}_3 := \begin{bmatrix} 0 \\ 1 \\ 0 \\ 0 \\ 0 \\ 1 \end{bmatrix}, \overrightarrow{\Phi}_4 := \begin{bmatrix} 0 \\ 0 \\ 1 \\ 0 \\ -1 \\ 0 \end{bmatrix}, \overrightarrow{\Phi}_5 := \begin{bmatrix} 0 \\ 1 \\ 0 \\ 0 \\ 0 \\ -1 \end{bmatrix}, \overrightarrow{\Phi}_6 := \begin{bmatrix} 0 \\ 0 \\ 1 \\ 0 \\ 1 \\ 0 \end{bmatrix}.$$

Démonstration. Le système de Maxwell peut s'écrire comme

$$\left\{ \begin{array}{l} \dfrac{\partial E_3}{\partial \mathsf{y}} - \dfrac{\partial E_2}{\partial \mathsf{z}} = -\dfrac{1}{c_0}\dfrac{\partial H_1}{\partial t}, \\ \dfrac{\partial E_1}{\partial \mathsf{z}} - \dfrac{\partial E_3}{\partial \mathsf{x}} = -\dfrac{1}{c_0}\dfrac{\partial H_2}{\partial t}, \\ \dfrac{\partial E_2}{\partial \mathsf{x}} - \dfrac{\partial E_1}{\partial \mathsf{y}} = -\dfrac{1}{c_0}\dfrac{\partial H_3}{\partial t}, \end{array} \right. \qquad \left\{ \begin{array}{l} \dfrac{\partial H_3}{\partial \mathsf{y}} - \dfrac{\partial H_2}{\partial \mathsf{z}} = \dfrac{1}{c_0}\dfrac{\partial E_1}{\partial t}, \\ \dfrac{\partial H_1}{\partial \mathsf{z}} - \dfrac{\partial H_3}{\partial \mathsf{x}} = \dfrac{1}{c_0}\dfrac{\partial E_2}{\partial t}, \\ \dfrac{\partial H_2}{\partial \mathsf{x}} - \dfrac{\partial H_1}{\partial \mathsf{y}} = \dfrac{1}{c_0}\dfrac{\partial E_3}{\partial t}. \end{array} \right.$$

Comme la solution est indépendante de y et de z, nous simplifions ces équations

$$\left\{ \begin{array}{rcl} 0 & = & -\dfrac{1}{c_0}\dfrac{\partial H_1}{\partial t}, \\ -\dfrac{\partial E_3}{\partial \mathsf{x}} & = & -\dfrac{1}{c_0}\dfrac{\partial H_2}{\partial t}, \\ \dfrac{\partial E_2}{\partial \mathsf{x}} & = & -\dfrac{1}{c_0}\dfrac{\partial H_3}{\partial t}, \end{array} \right. \qquad \left\{ \begin{array}{rcl} 0 & = & \dfrac{1}{c_0}\dfrac{\partial E_1}{\partial t}, \\ -\dfrac{\partial H_3}{\partial \mathsf{x}} & = & \dfrac{1}{c_0}\dfrac{\partial E_2}{\partial t}, \\ \dfrac{\partial H_2}{\partial \mathsf{x}} & = & \dfrac{1}{c_0}\dfrac{\partial E_3}{\partial t}. \end{array} \right.$$

Il suit que E_1 et H_1 sont indépendants du temps. Ils sont donc constants et égaux à la condition initiale pour tout $t \in \mathbb{R}^+$

$$E_1(\mathsf{x}, t) = E_1(\mathsf{x}, 0) \quad \text{et} \quad H_1(\mathsf{x}, t) = H_1(\mathsf{x}, 0).$$

Les autres coordonnées vérifient le système matriciel hyperbolique suivant

$$\frac{\partial \mathbf{U}}{\partial t}(\mathsf{x}, t) \; = \; c_0 \, \mathbf{A}_1 \frac{\partial \mathbf{U}}{\partial \mathsf{x}}(\mathsf{x}, t),$$

avec

$$\mathbf{U} = \begin{bmatrix} E_2 \\ E_3 \\ H_2 \\ H_3 \end{bmatrix} \quad \text{et} \quad \mathbf{A}_1 = \begin{bmatrix} 0 & 0 & 0 & -1 \\ 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 \\ -1 & 0 & 0 & 0 \end{bmatrix}.$$

Afin de déterminer $\mathbf{U}$, nous diagonalisons la matrice $\mathbf{A}_1$

$$\mathbf{A}_1\mathbf{P} = \mathbf{P}\mathbf{\Lambda} \quad \text{avec} \quad \mathbf{P} = \begin{bmatrix} 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 \\ 0 & -1 & 0 & 1 \\ 1 & 0 & -1 & 0 \end{bmatrix} \text{ et } \mathbf{\Lambda} = \begin{bmatrix} -1 & 0 & 0 & 0 \\ 0 & -1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}.$$

Cela nous conduit à une base orthogonale de vecteurs propres $\overrightarrow{\phi}_i$

$$\mathcal{B} = \text{span}\Big(\{\overrightarrow{\phi}_i, \text{ pour } i = 1, 4\}\Big),$$

où les vecteurs $\overrightarrow{\phi}_i$ sont de la forme

$$\overrightarrow{\phi}_1 = \begin{bmatrix} 1 \\ 0 \\ 0 \\ 1 \end{bmatrix}, \quad \overrightarrow{\phi}_2 = \begin{bmatrix} 0 \\ 1 \\ -1 \\ 0 \end{bmatrix}, \quad \overrightarrow{\phi}_3 = \begin{bmatrix} 1 \\ 0 \\ 0 \\ -1 \end{bmatrix}, \quad \overrightarrow{\phi}_4 = \begin{bmatrix} 0 \\ 1 \\ 1 \\ 0 \end{bmatrix}. \tag{2.30}$$

De par les valeurs de la matrice diagonale $\mathbf{\Lambda}$, nous avons

$$\mathbf{A}_1\overrightarrow{\phi}_1 = -\overrightarrow{\phi}_1, \quad \mathbf{A}_1\overrightarrow{\phi}_2 = -\overrightarrow{\phi}_2, \quad \mathbf{A}_1\overrightarrow{\phi}_3 = \overrightarrow{\phi}_3, \text{ et } \mathbf{A}_1\overrightarrow{\phi}_4 = \overrightarrow{\phi}_4.$$

Le vecteur solution $\mathbf{U}$ peut être décomposé dans la base $\mathcal{B}$ grâce aux coefficients $(\varepsilon_i)_{i=1,4}$: $\mathbb{R} \times \mathbb{R}^+ \to \mathbb{C}$

$$\mathbf{U}(\mathsf{x}, t) = \sum_{i=1}^{4} \varepsilon_i(\mathsf{x}, t)\overrightarrow{\phi}_i.$$

En prenant en compte l'expression de $\mathbf{U}$ dans le système hyperbolique, nous avons

$$\sum_{i=1}^{4} \frac{\partial \varepsilon_i}{\partial t}(\mathsf{x}, t)\overrightarrow{\phi}_i \;=\; c_0 \sum_{i=1}^{4} \lambda_i \frac{\partial \varepsilon_i}{\partial \mathsf{x}}(\mathsf{x}, t)\overrightarrow{\phi}_i.$$

Par conséquent, les coordonnées $\varepsilon_i(\mathsf{x}, t)$ satisfont

$$\frac{\partial \varepsilon_i}{\partial t}(\mathsf{x}, t) = c_0\, \lambda_i \frac{\partial \varepsilon_i}{\partial \mathsf{x}}(\mathsf{x}, t). \tag{2.31}$$

L'équation (2.31) est une équation de transport dont la solution s'écrit

$$\varepsilon_i(\mathsf{x}, t) = \varepsilon_i(\mathsf{x} + \lambda_i c_0 t, 0).$$

Il suit que

$$\begin{cases} \varepsilon_1(\mathsf{x},t) = \varepsilon_1(\mathsf{x}-c_0t,0), \quad \varepsilon_2(\mathsf{x},t) = \varepsilon_2(\mathsf{x}-c_0t,0), \\ \varepsilon_3(\mathsf{x},t) = \varepsilon_3(\mathsf{x}+c_0t,0), \quad \varepsilon_4(\mathsf{x},t) = \varepsilon_4(\mathsf{x}+c_0t,0). \end{cases}$$

De manière plus concrète, ces fonctions décrivent des modes électromagnétiques se propageant à la vitesse c_0, vers la droite (*resp.* gauche) pour ε_1 et ε_2 (*resp.* ε_3 et ε_4), voir la Figure 2.2. Comme $\|\overrightarrow{\phi}_i\|^2 = 2$ pour tout i allant de 1 à 4, nous pouvons normaliser la base $\mathcal{B}$ et les modes ont pour valeur

$$\varepsilon_i(\mathsf{x},t) = \frac{\mathbf{U}(\mathsf{x},t)\cdot\overrightarrow{\phi}_i}{2} \quad \text{pour } i = 1,4.$$

Ainsi, nous avons

$$\begin{cases} \varepsilon_1(\mathsf{x},t) = \dfrac{\mathbf{U}(\mathsf{x},t)\cdot\overrightarrow{\phi}_1}{2} = \dfrac{E_2+H_3}{2}(\mathsf{x},t) = \dfrac{E_2+H_3}{2}(\mathsf{x}-c_0t,0), \\ \varepsilon_2(\mathsf{x},t) = \dfrac{\mathbf{U}(\mathsf{x},t)\cdot\overrightarrow{\phi}_2}{2} = \dfrac{E_3-H_2}{2}(\mathsf{x},t) = \dfrac{E_3-H_2}{2}(\mathsf{x}-c_0t,0), \\ \varepsilon_3(\mathsf{x},t) = \dfrac{\mathbf{U}(\mathsf{x},t)\cdot\overrightarrow{\phi}_3}{2} = \dfrac{E_2-H_3}{2}(\mathsf{x},t) = \dfrac{E_2-H_3}{2}(\mathsf{x}+c_0t,0), \\ \varepsilon_4(\mathsf{x},t) = \dfrac{\mathbf{U}(\mathsf{x},t)\cdot\overrightarrow{\phi}_4}{2} = \dfrac{E_3+H_2}{2}(\mathsf{x},t) = \dfrac{E_3+H_2}{2}(\mathsf{x}+c_0t,0). \end{cases} \tag{2.32}$$

La solution prend la forme suivante

$$\begin{aligned} \mathbf{U}(\mathsf{x},t) = \varepsilon_1(\mathsf{x}-c_0t,0)\overrightarrow{\phi}_1 + \varepsilon_2(\mathsf{x}-c_0t,0)\overrightarrow{\phi}_2 \\ + \varepsilon_3(\mathsf{x}+c_0t,0)\overrightarrow{\phi}_3 + \varepsilon_4(\mathsf{x}+c_0t,0)\overrightarrow{\phi}_4, \end{aligned} \tag{2.33}$$

où les coefficients $(\varepsilon_i)_{i=1,4}$ peuvent être déterminés grâce aux valeurs trouvées en (2.32). Ainsi, la solution au point (x,t) est donnée par les valeurs des fonctions ε_i évaluées sur les lignes caractéristiques

— $\mathsf{x}-c_0t = cte$ pour ε_1 et ε_2 (de couleur rouges sur la Figure 2.2),

— $\mathsf{x}+c_0t = cte$ pour ε_3 et ε_4 (de couleur bleues sur la Figure 2.2).

Le résultat du théorème suit naturellement avec la forme de la solution $\mathbb{E}$. □

Remarque 2.6. *Le fait que la solution ne dépende que de* x *peut être vu non pas comme une hypothèse mais comme une conséquence du caractère isotropique du système de Maxwell et de la nature invariante de la condition initiale suivant* y *et* z.

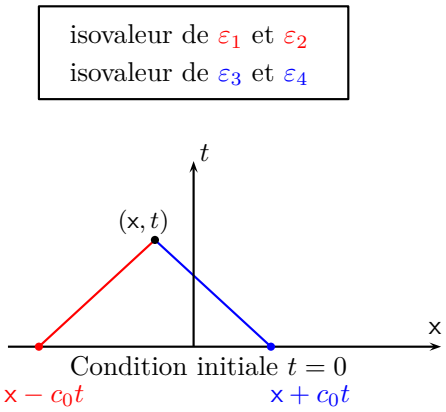

FIGURE 2.2 – Lignes des caractéristiques impliquées dans le calcul de $\mathbf{U}(\mathrm{x}, t)$.

Dérivation des traces numériques intérieures

Dans cette section, nous déterminons les coefficients définissant les traces numériques (2.27a) et (2.27b) sur les faces intérieures $F \in \mathcal{F}_{\text{int}}$, séparant deux éléments T^+ et T^-.

La résolution du système hyperbolique sur le maillage $\mathcal{T}$ ne conduirait pas à une expression simple du champ électromagnétique. Nous simplifions donc la géométrie en considérant deux demi-espaces $\mathbb{R}^{3,+}_{\mathbf{n}}$ et $\mathbb{R}^{3,-}_{\mathbf{n}}$, avec $\mathbf{n}$ la normale au plan définissant l'interface entre les demi-espaces, voir la Figure 2.3,

$$\mathbb{R}^{3,-}_{\mathbf{n}} = \Big\{ \mathbf{x} \in \mathbb{R}^3 \text{ tel que } \mathbf{x} \cdot \mathbf{n} \leq 0 \Big\},$$

$$\mathbb{R}^{3,+}_{\mathbf{n}} = \Big\{ \mathbf{x} \in \mathbb{R}^3 \text{ tel que } \mathbf{x} \cdot \mathbf{n} \geq 0 \Big\}.$$

Afin de se placer dans le cadre du Théorème 2.3, nous utilisons la coordonnée x relative à

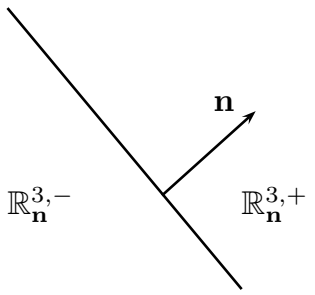

FIGURE 2.3 – Deux sous-espaces de $\mathbb{R}^3$.

la normale $\mathbf{n}$ à l'interface entre les demi-espaces,

$$\mathsf{x} = \mathbf{x} \cdot \mathbf{n}.$$

Remarque 2.7. *Nous assimilerons la normale* $\mathbf{n}$ *à la normale à une interface* F *entre deux éléments du maillage.*

Nous supposons à nouveau que les champs électromagnétiques $\mathbb{E} = (\mathbf{E}, \mathbf{H})$ ne dépendent que de la variable x

$$\mathbb{E}(\mathbf{x}, t) = \mathbb{E}(\mathsf{x}, t) \text{ pour } \mathbf{x} \in \mathbb{R}^{3,-}_{\mathbf{n}} \quad \text{et} \quad \mathbb{E}(\mathbf{x}, t) = \mathbb{E}(\mathsf{x}, t) \text{ pour } \mathbf{x} \in \mathbb{R}^{3,+}_{\mathbf{n}}.$$

D'autre part, nous supposons que les conditions initiales sont constantes par morceaux sur la partition définie par $\mathbb{R}^{3,-}_{\mathbf{n}}$ et $\mathbb{R}^{3,+}_{\mathbf{n}}$

$$\mathbb{E}(\mathbf{x}, 0) = \mathbb{E}^- \text{ pour } \mathbf{x} \in \mathbb{R}^{3,-}_{\mathbf{n}} \quad \text{et} \quad \mathbb{E}(\mathbf{x}, 0) = \mathbb{E}^+ \text{ pour } \mathbf{x} \in \mathbb{R}^{3,+}_{\mathbf{n}}.$$

Nous associons à $\mathbb{R}^3$ une base orthonormale $(\overrightarrow{e}_1 = \mathbf{n}, \overrightarrow{e}_2, \overrightarrow{e}_3)$, et nous adoptons les notations suivantes

$$E_i = \mathbf{E} \cdot \overrightarrow{e}_i \quad \text{et} \quad H_i = \mathbf{H} \cdot \overrightarrow{e}_i.$$

La détermination des traces intérieures revient à trouver la valeur du champ électromagnétique pour $\mathsf{x} = 0$ et $t > 0$ de $\mathbb{E}(\mathsf{x}, t)$. Les traces numériques de Riemann intérieures sont alors les composantes tangentielles de $\mathbb{E}(0, t)$ et sont énoncées dans le théorème suivant.

Théorème 2.4 (Traces numériques intérieures de Riemann)**.** *Pour* $F \in \mathcal{F}_{\text{int}}$*, les traces numériques de Riemann intérieures sont de la forme*

$$(\widehat{\gamma_t \mathbf{E}})_{|F} = \{\gamma_t \mathbf{E}\}_F - \frac{[\![\gamma_\times \mathbf{H}]\!]_F}{2}, \quad (\widehat{\gamma_t \mathbf{H}})_{|F} = \{\gamma_t \mathbf{H}\}_F + \frac{[\![\gamma_\times \mathbf{E}]\!]_F}{2}.$$

Remarque 2.8. *Chaque solution régulière du problème de Maxwell est continue entre les éléments. Ainsi, ses sauts sont nuls*

$$[\![\gamma_\times \mathbf{E}]\!]_F = 0 \quad \textit{et} \quad [\![\gamma_\times \mathbf{H}]\!]_F = 0.$$

Par conséquent, la solution exacte vérifie

$$(\widehat{\gamma_t \mathbf{E}})_{|F} = (\gamma_t \mathbf{E})_{|F} \quad \textit{et} \quad (\widehat{\gamma_t \mathbf{H}})_{|F} = (\gamma_t \mathbf{H})_{|F}. \tag{2.35}$$

Par ailleurs, les traces numériques (2.27a) *et* (2.27b) *sont données par*

$$(\widehat{\gamma_t \mathbf{E}})_{|F} := \frac{1}{2}\gamma_t \mathbf{E}^T + \frac{1}{2}\gamma_t \mathbf{E}^K - \frac{1}{2}(\gamma_\times^T \mathbf{H}^T + \gamma_\times^K \mathbf{H}^K),$$

$$(\widehat{\gamma_t \mathbf{H}})_{|F} := \frac{1}{2}\gamma_t \mathbf{H}^T + \frac{1}{2}\gamma_t \mathbf{H}^K + \frac{1}{2}(\gamma_\times^T \mathbf{E}^T + \gamma_\times^K \mathbf{E}^K).$$

Démonstration. Nous cherchons la valeur de la solution $\mathbb{E}(\mathsf{x}, t)$ en $\mathsf{x} = 0$, puis nous en prendrons ses traces. Le Théorème 2.3 est l'ingrédient principal pour résoudre ce problème de transmission. Nous devons trouver les valeurs des fonctions ε_i, pour $i = 1, 4$. Comme (2.32) est vraie pour $-c_0 t < \mathsf{x} < c_0 t$, nous avons alors pour $\mathsf{x} = 0$

$$\begin{cases} E_2(0,t) &= \dfrac{E_2^- + E_2^+}{2} - \dfrac{H_3^+ - H_3^-}{2}, \\ E_3(0,t) &= \dfrac{E_3^- + E_3^+}{2} + \dfrac{H_2^+ - H_2^-}{2}, \\ H_2(0,t) &= \dfrac{H_2^- + H_2^+}{2} + \dfrac{E_3^+ - E_3^-}{2}, \\ H_3(0,t) &= \dfrac{H_3^- + H_3^+}{2} - \dfrac{E_2^+ - E_2^-}{2}. \end{cases}$$

Dans ce cas, la valeur de $\mathbb{E}(0,t)$ dépend des quatre modes se propageant vers la droite ou vers la gauche, voir la Figure 2.2. Cette dernière expression peut être interprétée en tant que composantes tangentielles. En effet, nous avons

$$\begin{cases} \gamma_t \mathbf{E} = (0, E_2, E_3)^\mathsf{T}, \\ \gamma_\times^+ \mathbf{E} = \mathbf{n}^+ \times \gamma_t \mathbf{E} = (0, E_3, -E_2)^\mathsf{T}, \\ \gamma_\times^- \mathbf{E} = \mathbf{n}^- \times \gamma_t \mathbf{E} = (0, -E_3, E_2)^\mathsf{T}, \\ \mathbf{n}^+ := -\mathbf{n} \text{ et } \mathbf{n}^- := \mathbf{n}, \end{cases}$$

et nous obtenons les traces de Riemann intérieures

$$\begin{cases} (\widehat{\gamma_t \mathbf{E}})_{|F} = \dfrac{\gamma_t \mathbf{E}^+ + \gamma_t \mathbf{E}^-}{2} - \dfrac{\gamma_\times^+ \mathbf{H}^+ + \gamma_\times^- \mathbf{H}^-}{2}, \\ (\widehat{\gamma_t \mathbf{H}})_{|F} = \dfrac{\gamma_t \mathbf{H}^+ + \gamma_t \mathbf{H}^-}{2} + \dfrac{\gamma_\times^+ \mathbf{E}^+ + \gamma_\times^- \mathbf{E}^-}{2}. \end{cases}$$

Nous concluons en adoptant les notations suivantes

$$\begin{cases} \{\gamma_t \mathbf{E}\}_F &= \dfrac{\gamma_t \mathbf{E}^+ + \gamma_t \mathbf{E}^-}{2}, \\ [\![\gamma_\times \mathbf{E}]\!]_F &= \mathbf{n}^- \times \mathbf{E}^- + \mathbf{n}^+ \times \mathbf{E}^+. \end{cases}$$

□

Dérivation des traces numériques de bord

En utilisant un raisonnement similaire nous pouvons déterminer les traces numériques de bord de Riemann. Pour les faces de bord $F \in \mathcal{F}_{\text{ext}}$, nous résolvons cette fois le problème de Maxwell dans le sous-espace vérifiant $\mathsf{x} := \mathbf{x} \cdot \mathbf{n} \leq 0$, *ie* $\mathbb{R}^{3,-}$. Le problème tridimensionnel initial est

$$\begin{cases} \nabla \times \mathbf{E} = -\dfrac{1}{c_0}\dfrac{\partial \mathbf{H}}{\partial t}, & \text{pour } \mathsf{x} \leq 0, \\ \nabla \times \mathbf{H} = \quad \dfrac{1}{c_0}\dfrac{\partial \mathbf{E}}{\partial t}, & \text{pour } \mathsf{x} \leq 0. \end{cases}$$

Ce système de Maxwell est muni d'une condition de bord d'impédance

$$\gamma_t \mathbf{E} + Z_{\partial\Omega}\gamma_\times \mathbf{H} = \mathbf{g}, \quad \text{pour } \mathsf{x} = 0,$$

avec $\mathbf{g} = \gamma_t \mathbf{g}$ un champ tangent constant et d'une condition initiale constante

$$\mathbf{E}(\mathbf{x}, 0) = \mathbf{E}^- \quad \text{et} \quad \mathbf{H}(\mathbf{x}, 0) = \mathbf{H}^-.$$

Nous recherchons une solution ne dépendant que de x. En utilisant le Théorème 2.3, $E_1(\mathsf{x}, t) = E_1^-$ et $H_1(\mathsf{x}, t) = H_1^-$ sont constants.

La trace numérique de Riemann est alors définie comme la valeur de la composante tangentielle du champ électromagnétique $\mathbb{E}(\mathsf{x}, t)$ en $\mathsf{x} = 0$ et $t > 0$.

Théorème 2.5 (Traces numériques de bord de Riemann)**.** *Pour $F \in \mathcal{F}_{\text{ext}}$, les traces numériques de bord de Riemann prennent la forme*

$$\begin{cases} (\widehat{\gamma_t \mathbf{E}})_{|F} = \dfrac{1 + R_{\partial\Omega}}{2}\Big(\gamma_t \mathbf{E}^- - \gamma_\times \mathbf{H}^-\Big) + \dfrac{1 - R_{\partial\Omega}}{2}\gamma_t \mathbf{g}, \\ (\widehat{\gamma_t \mathbf{H}})_{|F} = \dfrac{1 - R_{\partial\Omega}}{2}\Big(\gamma_t \mathbf{H}^- + \gamma_\times \mathbf{E}^-\Big) - \dfrac{1 - R_{\partial\Omega}}{2}\gamma_\times \mathbf{g}, \end{cases}$$

où $R_{\partial\Omega}$ est lié à l'impédance $Z_{\partial\Omega}$ par

$$R_{\partial\Omega} := \frac{Z_{\partial\Omega} - 1}{Z_{\partial\Omega} + 1}. \tag{2.37}$$

Remarque 2.9. *Si la condition de bord d'impédance est vérifiée, alors la trace de Riemann coïncide avec la trace classique. En effet, nous avons de manière équivalente*

$$\begin{cases} (\widehat{\gamma_t \mathbf{E}})_{|F} = \gamma_t \mathbf{E}^- - \dfrac{1}{1 + Z_{\partial\Omega}}\Big(\gamma_t \mathbf{E}^- + Z_{\partial\Omega}\gamma_\times \mathbf{H}^- - \gamma_t \mathbf{g}\Big), \\ (\widehat{\gamma_\times \mathbf{H}})_{|F} = \gamma_\times \mathbf{H}^- - \dfrac{1}{1 + Z_{\partial\Omega}}\Big(\gamma_t \mathbf{E}^- + Z_{\partial\Omega}\gamma_\times \mathbf{H}^- - \gamma_t \mathbf{g}\Big). \end{cases}$$

Démonstration. Nous devons déterminer $\mathbb{E}(0, t)$, pour $t > 0$, puis en prendre sa composante

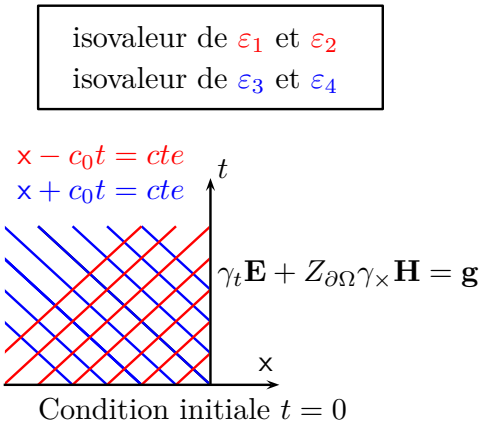

FIGURE 2.4 – Lignes des caractéristiques rencontrant à la fois la condition initiale et la condition de bord.

tangentielle pour obtenir le résultat du théorème. Nous utilisons la solution $\mathbf{U}$ du système hyperbolique, donnée par (2.33) et que nous rappelons ici

$$\mathbf{U}(\mathrm{x},t) = \varepsilon_1(\mathrm{x}-c_0t,0)\overrightarrow{\phi}_1 + \varepsilon_2(\mathrm{x}-c_0t,0)\overrightarrow{\phi}_2 + \varepsilon_3(\mathrm{x}+c_0t,0)\overrightarrow{\phi}_3 + \varepsilon_4(\mathrm{x}+c_0t,0)\overrightarrow{\phi}_4. \tag{2.38}$$

Pour une face extérieure, *ie* pour $\mathrm{x} < 0$ et $t > 0$, nous devons donc trouver les valeurs des coefficients

$$\begin{cases} \varepsilon_i^-(\mathrm{x}) &= \varepsilon_i(\mathrm{x}) \text{ pour } \mathrm{x} < 0 \text{ et } 1 \leq i \leq 4, \text{voir la Figure 2.4,} \\ \varepsilon_i^+(\mathrm{x}) &= \varepsilon_i(\mathrm{x}) \text{ pour } \mathrm{x} > 0 \text{ et } 3 \leq i \leq 4, \text{voir la Figure 2.5.} \end{cases}$$

Ces fonctions sont calculées grâce à

- la condition initiale pour ε_i^-, $1 \leq i \leq 4$,
- la condition de bord d'impédance pour ε_3^+ et ε_4^+.

En écrivant (2.38) en $t = 0$, nous avons pour $\mathrm{x} < 0$

$$\begin{pmatrix} E_2^-(\mathrm{x},0) \\ E_3^-(\mathrm{x},0) \\ H_2^-(\mathrm{x},0) \\ H_3^-(\mathrm{x},0) \end{pmatrix} = \varepsilon_1^-(\mathrm{x})\overrightarrow{\phi}_1 + \varepsilon_2^-(\mathrm{x})\overrightarrow{\phi}_2 + \varepsilon_3^-(\mathrm{x})\overrightarrow{\phi}_3 + \varepsilon_4^-(\mathrm{x})\overrightarrow{\phi}_4.$$

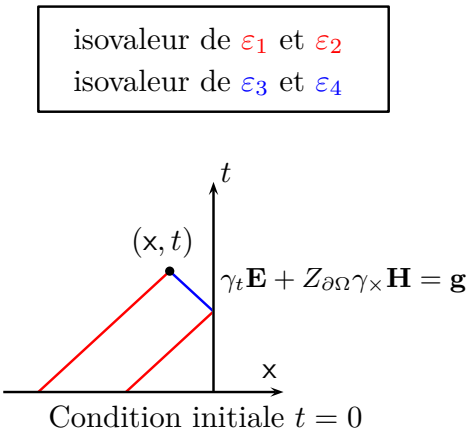

FIGURE 2.5 – Lignes des caractéristiques impliquées dans le calcul de $\mathbf{U}(\mathsf{x}, t)$ pour une face de bord.

En agissant de la même manière que pour (2.32), nous avons

$$\varepsilon_1^- = \frac{E_2^- + H_3^-}{2}, \quad \varepsilon_2^- = \frac{E_3^- - H_2^-}{2}, \quad \varepsilon_3^- = \frac{E_2^- - H_3^-}{2}, \quad \varepsilon_4^- = \frac{E_3^- + H_2^-}{2}.$$

Par conséquent, $\mathbf{U}$ a les coefficients suivants pour $\mathsf{x} < -c_0 t$

$$E_2(\mathbf{x}, t) = E_2^-, \quad E_3(\mathbf{x}, t) = E_3^-, \quad H_2(\mathbf{x}, t) = H_2^-, \quad H_3(\mathbf{x}, t) = H_3^-.$$

Nous déterminons les fonctions ε_3^+ et ε_4^+. La condition de bord d'impédance en $\mathsf{x} = 0$ et en $t > 0$ s'écrit

$$\gamma_t \mathbf{E} + Z_{\partial\Omega}\gamma_\times \mathbf{H} = \gamma_t \mathbf{g} \iff \begin{bmatrix} 0 \\ E_2(0,t) \\ E_3(0,t) \end{bmatrix} + \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix} \times \begin{bmatrix} 0 \\ Z_{\partial\Omega} H_2(0,t) \\ Z_{\partial\Omega} H_3(0,t) \end{bmatrix} = \begin{bmatrix} 0 \\ \mathbf{g}_2(t) \\ \mathbf{g}_3(t) \end{bmatrix},$$

c'est à dire

$$\begin{cases} E_2(0,t) - Z_{\partial\Omega} H_3(0,t) = \mathbf{g}_2(t), \\ E_3(0,t) + Z_{\partial\Omega} H_2(0,t) = \mathbf{g}_3(t). \end{cases}$$

En prenant $\mathsf{x} = 0$ dans (2.38), il suit que

$$\mathbf{U}(0,t) = \varepsilon_1^-(-c_0 t)\,\overrightarrow{\phi}_1 + \varepsilon_2^-(-c_0 t)\,\overrightarrow{\phi}_2 + \varepsilon_3^+(c_0 t)\,\overrightarrow{\phi}_3 + \varepsilon_4^+(c_0 t)\,\overrightarrow{\phi}_4.$$

En utilisant (2.30), nous obtenons

$$\begin{pmatrix} E_2(0,t) \\ E_3(0,t) \\ H_2(0,t) \\ H_3(0,t) \end{pmatrix} = \begin{pmatrix} \varepsilon_1^-(-c_0t) + \varepsilon_3^+(c_0t) \\ \varepsilon_2^-(-c_0t) + \varepsilon_4^+(c_0t) \\ -\varepsilon_2^-(-c_0t) + \varepsilon_4^+(c_0t) \\ \varepsilon_1^-(-c_0t) - \varepsilon_3^+(c_0t) \end{pmatrix}.$$

Ceci nous mène finalement au système linéaire suivant

$$\begin{cases} \varepsilon_1^- + \varepsilon_3^+ + Z_{\partial\Omega}\left(\varepsilon_3^+ - \varepsilon_1^-\right) = \mathbf{g}_2, \\ \varepsilon_2^- + \varepsilon_4^+ + Z_{\partial\Omega}\left(\varepsilon_4^+ - \varepsilon_2^-\right) = \mathbf{g}_3. \end{cases}$$

Nous déduisons les coefficients constants ε_i^+

$$\begin{cases} \varepsilon_3^+ = \dfrac{1}{1+Z_{\partial\Omega}}\mathbf{g}_2 + \dfrac{Z_{\partial\Omega}-1}{Z_{\partial\Omega}+1}\varepsilon_1^-, \\ \varepsilon_4^+ = \dfrac{1}{1+Z_{\partial\Omega}}\mathbf{g}_3 + \dfrac{Z_{\partial\Omega}-1}{Z_{\partial\Omega}+1}\varepsilon_2^-. \end{cases}$$

Comme nous l'avons vu en (2.37), le coefficient de réflexion est défini classiquement par

$$R_{\partial\Omega} := \frac{Z_{\partial\Omega}-1}{Z_{\partial\Omega}+1}.$$

Comme $\dfrac{1}{1+Z_{\partial\Omega}} = \dfrac{1-R_{\partial\Omega}}{2}$, nous avons

$$\begin{cases} \varepsilon_3^+ = \dfrac{1-R_{\partial\Omega}}{2}\mathbf{g}_2 + R_{\partial\Omega}\varepsilon_1^- = \dfrac{1-R_{\partial\Omega}}{2}\mathbf{g}_2 + R_{\partial\Omega}\dfrac{E_2^- + H_3^-}{2}, \\ \varepsilon_4^+ = \dfrac{1-R_{\partial\Omega}}{2}\mathbf{g}_3 + R_{\partial\Omega}\varepsilon_2^- = \dfrac{1-R_{\partial\Omega}}{2}\mathbf{g}_3 + R_{\partial\Omega}\dfrac{E_3^- - H_2^-}{2}. \end{cases}$$

Nous rappelons que pour $-c_0t < \mathrm{x} \leq 0$, $\mathbf{U}$ peut être évalué comme suit

$$\mathbf{U}(\mathrm{x},t) = \varepsilon_1^- \overrightarrow{\phi}_1 + \varepsilon_2^- \overrightarrow{\phi}_2 + \varepsilon_3^+ \overrightarrow{\phi}_3 + \varepsilon_4^+ \overrightarrow{\phi}_4.$$

De cette façon, grâce à (2.30) et (2.32), nous pouvons conclure

$$\begin{bmatrix} E_2(0,t) \\ E_3(0,t) \\ H_2(0,t) \\ H_3(0,t) \end{bmatrix} = \begin{bmatrix} \varepsilon_1^- + \varepsilon_3^+ \\ \varepsilon_2^- + \varepsilon_4^+ \\ -\varepsilon_2^- + \varepsilon_4^+ \\ \varepsilon_1^- - \varepsilon_3^+ \end{bmatrix} = \begin{bmatrix} \dfrac{E_2^- + H_3^-}{2} + R_{\partial\Omega}\dfrac{E_2^- + H_3^-}{2} + \dfrac{1-R_{\partial\Omega}}{2}\mathbf{g}_2 \\ \dfrac{E_3^- - H_2^-}{2} + R_{\partial\Omega}\dfrac{E_3^- - H_2^-}{2} + \dfrac{1-R_{\partial\Omega}}{2}\mathbf{g}_3 \\ \dfrac{H_2^- - E_3^-}{2} - R_{\partial\Omega}\dfrac{H_2^- - E_3^-}{2} + \dfrac{1-R_{\partial\Omega}}{2}\mathbf{g}_3 \\ \dfrac{H_3^- + E_2^-}{2} - R_{\partial\Omega}\dfrac{H_3^- + E_2^-}{2} - \dfrac{1-R_{\partial\Omega}}{2}\mathbf{g}_2 \end{bmatrix}.$$

Nous déduisons les valeurs des traces numériques grâce aux valeurs du champ électromagnétique

$$\begin{cases} (\widehat{(\gamma_t \mathbf{E})}_{|F})_2 = \dfrac{1+R_{\partial\Omega}}{2}\Big(E_2^- + H_3^-\Big) + \dfrac{1-R_{\partial\Omega}}{2}\mathbf{g}_2, \\ (\widehat{(\gamma_t \mathbf{E})}_{|F})_3 = \dfrac{1+R_{\partial\Omega}}{2}\Big(E_3^- - H_2^-\Big) + \dfrac{1-R_{\partial\Omega}}{2}\mathbf{g}_3, \\ (\widehat{(\gamma_t \mathbf{H})}_{|F})_2 = \dfrac{1-R_{\partial\Omega}}{2}\Big(H_2^- - E_3^-\Big) + \dfrac{1-R_{\partial\Omega}}{2}\mathbf{g}_3, \\ (\widehat{(\gamma_t \mathbf{H})}_{|F})_3 = \dfrac{1-R_{\partial\Omega}}{2}\Big(H_3^- + E_2^-\Big) - \dfrac{1-R_{\partial\Omega}}{2}\mathbf{g}_2. \end{cases}$$

En prenant en compte le fait que pour $\mathbf{u} \in \mathbb{C}^3$, nous avons

$$\begin{cases} \gamma_t \mathbf{u} = (0, \mathbf{u}_2, \mathbf{u}_3)^\mathsf{T}, \\ \gamma_\times \mathbf{u} = \mathbf{n}^- \times \gamma_t \mathbf{u} = (0, -\mathbf{u}_3, \mathbf{u}_2)^\mathsf{T}, \end{cases}$$

nous obtenons les traces numériques de Riemann extérieures. □

2.3.3 Détermination des traces numériques upwind pour des milieux hétérogènes

Nous rappelons que la formule de réciprocité peut être perturbée par des formes consistantes bien choisies (Section 2.2.1), ou par des traces numériques de Riemann (Section 2.3.2). Mais la perturbation peut aussi être réalisée grâce à d'autres traces numériques, introduites dans cette partie. Elles sont elles aussi représentatives de la physique du phénomène et sont obtenues par l'introduction de traces upwind.

Nous calculons ces traces dans le cas hétérogène. Nous rappelons tout d'abord les traces entrantes et sortantes à incidence normale $\gamma^T_{\text{in/out}} : \mathbb{X}_T \to \mathrm{L}^2_t(\partial T)$ définies classiquement par

$$\gamma^T_{\text{in}} \mathbb{E}^T := \gamma_t \mathbf{E}^T + Z_T \gamma^T_\times \mathbf{H}^T \quad \text{et} \quad \gamma^T_{\text{out}} \mathbb{E}^T := \gamma_t \mathbf{E}^T - Z_T \gamma^T_\times \mathbf{H}^T, \tag{2.39}$$

représentées dans la Figure 2.6. Soit $T \in \mathcal{T}$, les traces numériques électrique et magnétique sur une face F séparant T et l'un de ses voisins K sont respectivement notées $(\widehat{\gamma_t \mathbf{E}})_{|F}$ et $(\widehat{\gamma^T_\times \mathbf{H}})_{|F}$. Le champ magnétique est considéré à travers sa trace tangentielle qui est orientée

$$(\widehat{\gamma^T_\times \mathbf{H}})_{|F} = -(\widehat{\gamma^K_\times \mathbf{H}})_{|F} := \mathbf{n}_T \times (\widehat{\gamma_t \mathbf{H}})_{|F}.$$

Une trace numérique upwind sur une face F intérieure est une combinaison linéaire des traces sortantes des deux éléments T et K, voir la Figure 2.6 :

$$(\widehat{\gamma_t \mathbf{E}})_{|F} := \alpha^T \gamma^T_{\text{out}} \mathbb{E}^T + \alpha^K \gamma^K_{\text{out}} \mathbb{E}^K, \tag{2.40a}$$

$$\widehat{(\gamma_\times^T \mathbf{H})}_{|F} := -\beta^T \gamma_{\text{out}}^T \mathbb{E}^T + \beta^K \gamma_{\text{out}}^T \mathbb{E}^K, \tag{2.40b}$$

ou de manière équivalente

$$\widehat{(\gamma_\times^K \mathbf{H})}_{|F} := \beta^T \gamma_{\text{out}}^T \mathbb{E}^T - \beta^K \gamma_{\text{out}}^T \mathbb{E}^K.$$

Afin de définir une méthode numérique consistante, elle doit généraliser la notion de trace classique. Ainsi, comme pour les flux de Riemann, nous assurons pour des champs électromagnétiques continus que

$$\begin{array}{rcccccl} \widehat{(\gamma_t \mathbf{E})}_{|F} & = & (\gamma_t \mathbf{E}^T)_{|F} & = & (\gamma_t \mathbf{E}^K)_{|F} & = & (\gamma_t \mathbf{E})_{|F}, \\ \widehat{(\gamma_\times^T \mathbf{H})}_{|F} & = & (\gamma_\times^T \mathbf{H}^T)_{|F} & = & -(\gamma_\times^K \mathbf{H}^K)_{|F} & = & (\gamma_\times^T \mathbf{H})_{|F}. \end{array} \tag{2.41}$$

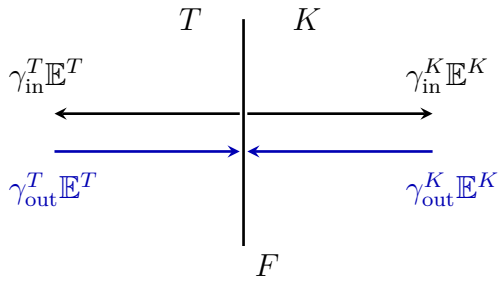

FIGURE 2.6 – Traces entrante et sortante sur une face $F \in \mathcal{F}_{\text{int}}$ séparant deux éléments T et K.

Théorème 2.6 (Traces numériques upwind intérieures)**.** *Pour deux éléments T et K séparés par une face F, l'unique combinaison de traces upwind prolongeant la notion de trace classique est donnée par* (2.40a) *et* (2.40b) *avec*

$$\alpha^T = \frac{Z_K}{Z_T + Z_K}, \quad \alpha^K = \frac{Z_T}{Z_T + Z_K}, \quad \beta^T = \frac{1}{Z_T + Z_K}, \quad \beta^K = \frac{1}{Z_T + Z_K}.$$

Démonstration. L'équation (2.40a) s'écrit

$$\gamma_t \mathbf{E} = (\alpha^T + \alpha^K)\gamma_t \mathbf{E} + (\alpha^K Z_K - \alpha^T Z_T)\gamma_\times^T \mathbf{H}.$$

De la même façon, nous avons pour (2.40b)

$$\gamma_\times^T \mathbf{H} = (\beta^K - \beta^T)\gamma_t \mathbf{E} + (\beta^T Z_T + \beta^K Z_K)\gamma_\times^T \mathbf{H}.$$

Nous en déduisons que les coefficients α^T et α^K satisfont

$$\begin{cases} \alpha^T + \alpha^K & = & 1, \\ \alpha^K Z_K - \alpha^T Z_T & = & 0, \\ \beta^T - \beta^K & = & 0, \\ \beta^T Z_T + \beta^K Z_K & = & 1. \end{cases}$$

Il ne nous reste plus qu'à résoudre ce système d'équations. □

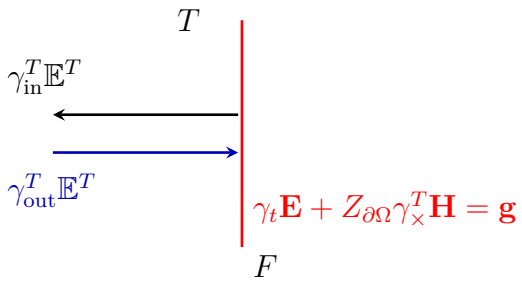

FIGURE 2.7 – Traces entrante et sortante d'un élément T dont la face F est sur le bord du domaine.

Nous traitons maintenant le cas des faces extérieures, voir la Figure 2.7. Nous adoptons une approche similaire et établissons le théorème suivant.

Théorème 2.7 (Traces numériques upwind de bord)**.** *L'unique combinaison de traces upwind prolongeant la notion de trace classique pour les champs électromagnétiques vérifiant la condition d'impédance est donnée par*

$$\widehat{(\gamma_t \mathbf{E})}_{|F} := \alpha \gamma_{\text{out}}^T \mathbb{E}^T + f_{\mathbf{E}}, \tag{2.42a}$$

$$\widehat{(\gamma_\times^T \mathbf{H})}_{|F} := \beta \gamma_{\text{out}}^T \mathbb{E}^T + f_{\mathbf{H}}, \tag{2.42b}$$

avec

$$\alpha = \frac{Z_{\partial\Omega}}{Z_T + Z_{\partial\Omega}}, \quad f_{\mathbf{E}} = \frac{Z_T}{Z_T + Z_{\partial\Omega}} \mathbf{g}, \quad \beta = -\frac{1}{Z_{\partial\Omega} + Z_T}, \quad f_{\mathbf{H}} = \frac{1}{Z_{\partial\Omega} + Z_T} \mathbf{g}.$$

Démonstration. Comme la solution exacte vérifie (2.41), et grâce à la définition d'une trace sortante (2.39), nous développons (2.42a) et (2.42b)

$$(1-\alpha)\gamma_t \mathbf{E} + \alpha Z_T \gamma_\times^T \mathbf{H} = f_{\mathbf{E}}, \tag{2.43}$$

$$-\beta \gamma_t \mathbf{E} + (1 + \beta Z_T) \gamma_\times^T \mathbf{H} = f_{\mathbf{H}}. \tag{2.44}$$

Nous rappelons la condition de bord d'impédance (1)

$$\gamma_t\mathbf{E} + Z_{\partial\Omega}\,\mathbf{n}_{\partial\Omega}\times\gamma_t\mathbf{H} = \mathbf{g} \quad \text{sur } \partial\Omega.$$

Nous identifions les équations (2.43) et (2.44) à la condition d'impédance, et il suit

$$\begin{cases} \dfrac{\alpha Z_T}{1-\alpha} &= Z_{\partial\Omega}, \\ f_{\mathbf{E}} &= (1-\alpha)\mathbf{g}, \\ \dfrac{1+\beta Z_T}{-\beta} &= Z_{\partial\Omega}, \\ f_{\mathbf{H}} &= -\beta\mathbf{g}. \end{cases}$$

Les valeurs des coefficients des traces de bord sont finalement déduites de ce système. □

Remarque 2.10. *Dans le cas homogène, les traces numériques upwind intérieures et de bord sont équivalentes aux traces numériques de Riemann. En effet, en prenant $Z_T = Z_K = 1$, nous retrouvons les Théorèmes 2.4 et 2.5.*

2.3.4 Construction de la formulation variationnelle upwind

Nous avons dérivé les traces numériques upwind pour les faces intérieures et de bord. Grâce à ces dernières, nous mettons en place la formulation variationnelle upwind pour les milieux hétérogènes. Conformément à l'étape 2 de la démarche de construction d'une formulation Trefftz, voir la Sous-section 2.3.1, nous insérons les traces numériques upwind dans la formule de réciprocité (2.7).

Problème 9 (Problème upwind)**.** *Trouver $\mathbb{E}\in\mathbb{X}_{\mathfrak{T}}^h$ tel que pour tout $\mathbb{E}'\in\mathbb{X}_{\mathfrak{T}}^h$*

$$\sum_{T\in\mathfrak{T}}\sum_{F\in\mathcal{F}_T}\int_F\left(\widehat{\gamma_t\mathbf{E}}\cdot\gamma_\times^T\overline{\mathbf{H}'^T} + \widehat{\gamma_\times^T\mathbf{H}}\cdot\overline{\gamma_t\mathbf{E}'^T}\right) = 0,$$

avec les traces upwind intérieures pour $F\in\mathcal{F}_{\text{int}}$ définies comme

$$\begin{cases} \left(\widehat{\gamma_t\mathbf{E}}\right)_{|F} &:= \dfrac{Z_K}{Z_K+Z_T}\gamma_{\text{out}}^T\mathbb{E}^T + \dfrac{Z_T}{Z_K+Z_T}\gamma_{\text{out}}^K\mathbb{E}^K, \\ \left(\widehat{\gamma_\times^T\mathbf{H}}\right)_{|F} &:= -\dfrac{1}{Z_K+Z_T}\gamma_{\text{out}}^T\mathbb{E}^T + \dfrac{1}{Z_K+Z_T}\gamma_{\text{out}}^K\mathbb{E}^K, \end{cases} \tag{2.45}$$

et les traces upwind de bord pour $F\in\mathcal{F}_{\text{ext}}$ définies comme

$$\begin{cases} \left(\widehat{\gamma_t\mathbf{E}}\right)_{|F} &:= \dfrac{Z_{\partial\Omega}}{Z_{\partial\Omega}+Z_T}\gamma_{\text{out}}^T\mathbb{E}^T + \dfrac{Z_T}{Z_{\partial\Omega}+Z_T}\mathbf{g}^T, \\ \left(\widehat{\gamma_\times^T\mathbf{H}}\right)_{|F} &:= -\dfrac{1}{Z_{\partial\Omega}+Z_T}\gamma_{\text{out}}^T\mathbb{E}^T + \dfrac{1}{Z_{\partial\Omega}+Z_T}\mathbf{g}^T. \end{cases} \tag{2.46}$$

Remarque 2.11 (Élément virtuel)**.** *Une face extérieure $F \in \mathcal{F}_{\text{ext}}$ peut être vue comme une face séparant un élément T et un élément virtuel extérieur au domaine. Plus précisément, en notant $Z_{\partial\Omega} = Z_K$ et $\mathbf{g}^T = \gamma_{\text{out}}^K \mathbb{E}^K$, l'équation (2.46) devient (2.45).*

Le Problème 9 est interprété comme un problème variationnel. Cela mène au problème suivant.

Problème 10 (Problème variationnel uwpind)**.** *Le Problème 9 s'écrit*

$$\text{Trouver } \mathbb{E} \in \mathbb{X}_{\mathcal{J}}^h \text{ tel que pour tout } \mathbb{E}' \in \mathbb{X}_{\mathcal{J}}^h \quad a(\mathbb{E}, \mathbb{E}') = \ell(\mathbb{E}'),$$

où la forme sesquilinéaire a et la forme antilinéaire ℓ sont données par

$$a(\mathbb{E}, \mathbb{E}') := \sum_{T\in\mathcal{J}} \sum_{F\in\mathcal{F}_T} \int_F \left(a_{T,F}^{EE'} + a_{T,F}^{EH'} + a_{T,F}^{HE'} + a_{T,F}^{HH'} \right), \tag{2.47a}$$

$$\ell(\mathbb{E}') := \sum_{T\in\mathcal{J}} \sum_{F\in\mathcal{F}_T\cap\mathcal{F}_{\text{ext}}} \int_F \left(\ell_{T,F}^{E'} + \ell_{T,F}^{H'} \right),$$

où nous avons, si $F \in \mathcal{F}_{\text{int}}$ interfaçant deux éléments voisins T et K,

$$\begin{cases} a_{T,F}^{EE'} &:= \dfrac{1}{Z_T + Z_K} (\gamma_t \mathbf{E}^T - \gamma_t \mathbf{E}^K) \cdot \overline{\gamma_t \mathbf{E}'^T}, \\ a_{T,F}^{EH'} &:= \dfrac{Z_T}{Z_T + Z_K} (\gamma_t \mathbf{E}^T - \gamma_t \mathbf{E}^K) \cdot \overline{\gamma_\times^T \mathbf{H}'^T}, \\ a_{T,F}^{HE'} &:= \dfrac{Z_K}{Z_T + Z_K} (\gamma_\times^T \mathbf{H}^T - \gamma_\times^T \mathbf{H}^K) \cdot \overline{\gamma_t \mathbf{E}'^T}, \\ a_{T,F}^{HH'} &:= \dfrac{Z_T\, Z_K}{Z_T + Z_K} (\gamma_\times^T \mathbf{H}^T - \gamma_\times^T \mathbf{H}^K) \cdot \overline{\gamma_\times^T \mathbf{H}'^T}, \end{cases} \tag{2.48a}$$

et si $F \in \mathcal{F}_{\text{ext}}$

$$\begin{cases} a_{T,F}^{EE'} &:= \dfrac{1}{Z_T + Z_{\partial\Omega}} \gamma_t \mathbf{E}^T \cdot \overline{\gamma_t \mathbf{E}'^T}, & a_{T,F}^{EH'} &:= \dfrac{Z_T}{Z_T + Z_{\partial\Omega}} \gamma_t \mathbf{E}^T \cdot \overline{\gamma_\times^T \mathbf{H}'^T}, \\ a_{T,F}^{HE'} &:= \dfrac{Z_{\partial\Omega}}{Z_T + Z_{\partial\Omega}} \gamma_\times^T \mathbf{H}^T \cdot \overline{\gamma_t \mathbf{E}'^T}, & a_{T,F}^{HH'} &:= \dfrac{Z_T\, Z_{\partial\Omega}}{Z_T + Z_{\partial\Omega}} \gamma_\times^T \mathbf{H}^T \cdot \overline{\gamma_\times^T \mathbf{H}'^T}, \\ \ell_{T,F}^{E'} &:= \dfrac{1}{Z_T + Z_{\partial\Omega}} \mathbf{g}^T \cdot \overline{\gamma_t \mathbf{E}'^T}, & \ell_{T,F}^{H'} &:= \dfrac{Z_T}{Z_T + Z_{\partial\Omega}} \mathbf{g}^T \cdot \overline{\gamma_\times^T \mathbf{H}'^T}. \end{cases} \tag{2.48b}$$

Remarque 2.12. *Comme la solution exacte vérifie (2.35), la formulation variationnelle du Problème 10 reflète alors la continuité de sa solution.*

Démonstration. Le Problème 9 s'écrit

$$\sum_{T\in\mathcal{J}} \sum_{F\in\mathcal{F}_T} \int_F \mathfrak{I}_{T,F} = 0, \quad \text{avec} \quad \mathfrak{I}_{T,F} := \widehat{\gamma_t \mathbf{E}} \cdot \overline{\gamma_\times^T \mathbf{H}'^T} + \widehat{\gamma_\times^T \mathbf{H}} \cdot \overline{\gamma_t \mathbf{E}'^T}.$$

(*a*) Si $F \in \mathcal{F}_{\rm int}$, en utilisant les définitions des traces numériques (2.45), nous avons

$$\begin{aligned}\mathcal{I}_{T,F} \;=\; & \Big(\frac{Z_K}{Z_K+Z_T}\gamma_{\rm out}^T\mathbb{E}^T + \frac{Z_T}{Z_T+Z_K}\gamma_{\rm out}^K\mathbb{E}^K\Big)\cdot\overline{\gamma_\times^T\mathbf{H}'^T} \\ & +\Big(-\frac{1}{Z_K+Z_T}\gamma_{\rm out}^T\mathbb{E}^T + \frac{1}{Z_K+Z_T}\gamma_{\rm out}^K\mathbb{E}^K\Big)\cdot\overline{\gamma_t\mathbf{E}'^T}.\end{aligned}$$

En prenant en compte la définition des opérateurs de traces sortante et entrante dans (2.39), nous obtenons

$$\begin{aligned}\mathcal{I}_{T,F} = & \Big[\frac{Z_K}{Z_K+Z_T}(\gamma_t\mathbf{E}^T - Z_T\gamma_\times^T\mathbf{H}^T) + \frac{Z_T}{Z_T+Z_K}(\gamma_t\mathbf{E}^K + Z_K\gamma_\times^T\mathbf{H}^K)\Big]\cdot\overline{\gamma_\times^T\mathbf{H}'^T} \\ & + \Big[-\frac{1}{Z_K+Z_T}(\gamma_t\mathbf{E}^T - Z_T\gamma_\times^T\mathbf{H}^T) + \frac{1}{Z_K+Z_T}(\gamma_t\mathbf{E}^K + Z_K\gamma_\times^T\mathbf{H}^K)\Big]\cdot\overline{\gamma_t\mathbf{E}'^T}.\end{aligned}$$

En développant, nous avons

$$\mathcal{I}_{T,F} = \frac{(Z_K\gamma_t\mathbf{E}^T + Z_T\gamma_t\mathbf{E}^K)}{Z_K+Z_T}\cdot\overline{\gamma_\times^T\mathbf{H}'^T} + \frac{(Z_T\gamma_\times^T\mathbf{H}^T + Z_K\gamma_\times^T\mathbf{H}^K)}{Z_K+Z_T}\cdot\overline{\gamma_t\mathbf{E}'^T} - a_{T,F}^{EE'} - a_{T,F}^{HH'}.$$

Nous remarquons ensuite que

$$\frac{Z_K}{Z_K+Z_T} + \frac{Z_T}{Z_K+Z_T} = 1 \quad \text{et} \quad \frac{Z_{\partial\Omega}}{Z_{\partial\Omega}+Z_T} + \frac{Z_T}{Z_{\partial\Omega}+Z_T} = 1.$$

Cela mène à

$$\mathcal{I}_{T,F} = \gamma_t\mathbf{E}^T\cdot\overline{\gamma_\times^T\mathbf{H}'^T} + \gamma_\times^T\mathbf{H}^T\cdot\overline{\gamma_t\mathbf{E}'^T} - a_{T,F}^{EE'} - a_{T,F}^{EH'} - a_{T,F}^{HE'} - a_{T,F}^{HH'}. \tag{2.49}$$

(*b*) Si $F \in \mathcal{F}_{\rm ext}$, nous avons de la même façon

$$\mathcal{I}_{T,F} = \gamma_t\mathbf{E}^T\cdot\overline{\gamma_\times^T\mathbf{H}'^T} + \gamma_\times^T\mathbf{H}^T\cdot\overline{\gamma_t\mathbf{E}'^T} \;+\; \ell_{T,F}^{E'} \;+\; \ell_{T,F}^{H'} - a_{T,F}^{EE'} - a_{T,F}^{EH'} - a_{T,F}^{HE'} - a_{T,F}^{HH'}. \tag{2.50}$$

Le Problème 10 suit de (2.49), (2.50) et de la formule de réciprocité (2.7). □

2.3.5 Coercivité faible de la formulation upwind

Le caractère bien posé du Problème 9 repose sur la coercivité de la forme sesquilinéaire a définie par (2.47a). Cette propriété est établie vis à vis d'une norme sous-jacente à la méthode de Galerkin Discontinue, que nous nommerons : norme GD. Cette dernière met en jeu les sauts de la composante (*resp*. trace) tangentielle du champ électrique (*resp*. magnétique), définis par (2.25).

Proposition 2.2 (Coercivité faible pour la norme GD)**.** *La forme sesquilinéaire a est positive comme $\Re(a(\mathbb{E},\mathbb{E})) \geq 0$. Nous définissons la norme* GD *par*

$$\|\mathbb{E}\|_{\mathrm{GD}} = \sqrt{\Re(a(\mathbb{E},\mathbb{E}))}, \quad \textit{pour tout} \quad \mathbb{E} \in \mathbb{X}_{\mathcal{T}},$$

pour laquelle nous avons (similairement à (2.23))

$$\|\mathbb{E}\|_{\mathrm{GD}}^2 = \|\mathbb{E}\|_{\mathrm{int}}^2 + \|\mathbb{E}\|_{\mathrm{ext}}^2,$$

avec

$$\begin{aligned}
\|\mathbb{E}\|_{\mathrm{int}}^2 &= \sum_{F\in\mathcal{F}_{\mathrm{int}}} \int_F \Big(\frac{1}{Z_T+Z_K} [\![\gamma_t \mathbf{E}]\!]_{T,F} \cdot \overline{[\![\gamma_t \mathbf{E}]\!]_{T,F}} + \frac{Z_T\, Z_K}{Z_T+Z_K} [\![\gamma_\times \mathbf{H}]\!]_F \cdot \overline{[\![\gamma_\times \mathbf{H}]\!]_F} \Big), \\
\|\mathbb{E}\|_{\mathrm{ext}}^2 &= \sum_{F\in\mathcal{F}_{\mathrm{ext}}} \int_F \Big(\frac{1}{Z_T+Z_{\partial\Omega}} \gamma_t \mathbf{E} \cdot \overline{\gamma_t \mathbf{E}} + \frac{Z_T\, Z_{\partial\Omega}}{Z_T+Z_{\partial\Omega}} \gamma_t \mathbf{H} \cdot \overline{\gamma_t \mathbf{H}} \Big).
\end{aligned}$$

La preuve de la Proposition 2.2 utilise la remarque suivante.

Remarque 2.13 (Point de vue par face)**.** *Nous pouvons décomposer l'intégrale sur ∂T en utilisant un point de vue par face*

$$\sum_{T\in\mathcal{T}} \int_{\partial T} f^T = \sum_{F\in\mathcal{F}_{\mathrm{ext}}} \int_F f^T + \sum_{F\in\mathcal{F}_{\mathrm{int}}} \int_F f^T + f^K,$$

où, dans le membre de droite, T (resp. T et K) est un élément (resp. sont deux éléments) avec une face F.

Démonstration. En utilisant (2.47a), nous pouvons écrire la partie réelle de a comme

$$\Re\left(a(\mathbb{E},\mathbb{E})\right) = \sum_{T\in\mathcal{T}} \sum_{F\in\mathcal{F}_T} \int_F \Re\Big(a_{T,F}^{EE} + a_{T,F}^{EH} + a_{T,F}^{HE} + a_{T,F}^{HH} \Big).$$

(a) Nous montrons dans un premier temps que

$$\mathcal{I}^1 = \sum_{T\in\mathcal{T}} \sum_{F\in\mathcal{F}_T} \int_F \underbrace{\Re(a_{T,F}^{EH} + a_{T,F}^{HE})}_{:=\mathcal{I}_{T,F}^1} = 0. \tag{2.51}$$

Nous évaluons $\mathcal{I}_{T,F}^1$ en distinguant les cas des faces intérieures $F \in \mathcal{F}_{\mathrm{int}}$ et des faces extérieures $F \in \mathcal{F}_{\mathrm{ext}}$.
Pour $F \in \mathcal{F}_{\mathrm{int}}$, en utilisant les définitions des formes (2.48a) et en rappelant que $\Re(x\overline{y}) =$

$\Re(y\overline{x})$, nous avons

$$\begin{aligned}\mathcal{I}^1_{T,F} &= \Re\Big(\frac{Z_T}{Z_T+Z_K}(\gamma_t\mathbf{E}^T-\gamma_t\mathbf{E}^K)\cdot\overline{\gamma_\times^T\mathbf{H}^T}+\frac{Z_K}{Z_T+Z_K}\gamma_t\mathbf{E}^T\cdot\overline{(\gamma_\times^T\mathbf{H}^T-\gamma_\times^T\mathbf{H}^K)}\Big),\\ &= \Re\Big(\gamma_t\mathbf{E}^T\cdot\overline{\gamma_\times^T\mathbf{H}^T}-\frac{Z_T}{Z_T+Z_K}\gamma_t\mathbf{E}^K\cdot\overline{\gamma_\times^T\mathbf{H}^T}-\frac{Z_K}{Z_T+Z_K}\gamma_t\mathbf{E}^T\cdot\overline{\gamma_\times^T\mathbf{H}^K}\Big).\end{aligned}$$

De manière similaire, en utilisant (2.48b) pour $F\in\mathcal{F}_{\rm ext}$, nous obtenons

$$\mathcal{I}^1_{T,F} = \Re\Big(\frac{Z_T}{Z_T+Z_{\partial\Omega}}\gamma_t\mathbf{E}^T\cdot\overline{\gamma_\times^T\mathbf{H}^T}+\frac{Z_{\partial\Omega}}{Z_T+Z_{\partial\Omega}}\gamma_t\mathbf{E}^T\cdot\overline{\gamma_\times^T\mathbf{H}^T}\Big) = \Re(\gamma_t\mathbf{E}^T\cdot\overline{\gamma_\times^T\mathbf{H}^T}).$$

Comme $\Re\left(r_T(\mathbb{E},\mathbb{E})\right)=2\sum_{F\in\mathcal{F}_T}\int_F\Re(\gamma_t\mathbf{E}^T\cdot\overline{\gamma_\times^T\mathbf{H}^T})$, voir la Remarque 2.4, et en sommant sur les faces $F\in\mathcal{F}_T$, nous obtenons

$$\begin{aligned}\sum_{F\in\mathcal{F}_T}\int_F\mathcal{I}^1_{T,F} &= \frac{1}{2}\Re\left(r_T(\mathbb{E},\mathbb{E})\right)\\ &\quad-\sum_{F\in\mathcal{F}_T\cap\mathcal{F}_{\rm int}}\int_F\Re\Big(\frac{Z_T}{Z_T+Z_K}\gamma_t\mathbf{E}^K\cdot\overline{\gamma_\times^T\mathbf{H}^T}+\frac{Z_K}{Z_T+Z_K}\gamma_t\mathbf{E}^T\cdot\overline{\gamma_\times^T\mathbf{H}^K}\Big).\end{aligned}$$

Nous remarquons que $\Re\left(r_T(\mathbb{E},\mathbb{E})\right)=0$, d'après (2.7). En sommant sur les éléments, nous prenons ensuite le point de vue par face de la Remarque 2.13

$$\begin{aligned}\mathcal{I}^1 &= -\sum_{F\in\mathcal{F}_{\rm int}}\int_F\Re\Big(\frac{Z_T}{Z_T+Z_K}(\gamma_t\mathbf{E}^K\cdot\overline{\gamma_\times^T\mathbf{H}^T}+\gamma_t\mathbf{E}^K\cdot\overline{\gamma_\times^K\mathbf{H}^T})\Big)\\ &\quad+\sum_{F\in\mathcal{F}_{\rm int}}\int_F\Re\Big(\frac{Z_K}{Z_T+Z_K}(\gamma_t\mathbf{E}^T\cdot\overline{\gamma_\times^T\mathbf{H}^K}+\gamma_t\mathbf{E}^T\cdot\overline{\gamma_\times^K\mathbf{H}^K})\Big).\end{aligned}$$

Comme $\mathbf{n}_T=-\mathbf{n}_K$, nous remarquons que

$$\gamma_\times^T\mathbf{H}^K+\gamma_\times^K\mathbf{H}^K=0 \quad\text{et}\quad \gamma_\times^K\mathbf{H}^T+\gamma_\times^T\mathbf{H}^T=0.$$

Nous avons finalement (2.51).

(*b*) Il reste à évaluer

$$\mathcal{I}^2=\sum_{T\in\mathcal{T}}\sum_{F\in\mathcal{F}_T}\int_F\Re(a^{EE}_{T,F}) \quad\text{et}\quad \mathcal{I}^3=\sum_{T\in\mathcal{T}}\sum_{F\in\mathcal{F}_T}\int_F\Re(a^{HH}_{T,F}).$$

Nous utilisons à nouveau (2.48) et nous sommons sur les éléments. En différenciant les cas

des faces intérieures et des faces extérieures, nous avons

$$\begin{aligned}\mathcal{J}^2 \;=\; & \sum_{T\in\mathcal{T}}\sum_{F\in\mathcal{F}_T\cap\mathcal{F}_{\text{int}}}\int_F \Re\Big(\frac{1}{Z_T+Z_K}(\gamma_t\mathbf{E}^T-\gamma_t\mathbf{E}^K)\cdot\overline{\gamma_t\mathbf{E}^T}\Big)\\ & +\sum_{T\in\mathcal{T}}\sum_{F\in\mathcal{F}_T\cap\mathcal{F}_{\text{ext}}}\int_F \Re\Big(\frac{1}{Z_T+Z_{\partial\Omega}}\gamma_t\mathbf{E}^T\cdot\overline{\gamma_t\mathbf{E}^T}\Big).\end{aligned}$$

Nous assemblons grâce à un point de vue par face, voir la Remarque 2.13,

$$\begin{aligned}\mathcal{J}^2 \;=\; & \sum_{F\in\mathcal{F}_{\text{int}}}\int_F \Re\Big(\frac{1}{Z_T+Z_K}(\gamma_t\mathbf{E}^T-\gamma_t\mathbf{E}^K)\cdot\overline{\gamma_t\mathbf{E}^T}\Big)\\ & +\sum_{F\in\mathcal{F}_{\text{int}}}\int_F \Re\Big(\frac{1}{Z_K+Z_T}(\gamma_t\mathbf{E}^K-\gamma_t\mathbf{E}^T)\cdot\overline{\gamma_t\mathbf{E}^K}\Big)\\ & +\sum_{F\in\mathcal{F}_{\text{ext}}}\int_F \Re\Big(\frac{1}{Z_T+Z_{\partial\Omega}}\gamma_t\mathbf{E}^T\cdot\overline{\gamma_t\mathbf{E}^T}\Big).\end{aligned}$$

En utilisant la définition du saut du champ électrique (2.25), nous obtenons la formule suivante pour $\mathcal{J}^2$, qui s'annule pour $F\in\mathcal{F}_{\text{int}}$,

$$\mathcal{J}^2=\sum_{F\in\mathcal{F}_{\text{int}}}\int_F\frac{1}{Z_T+Z_K}[\![\gamma_t\mathbf{E}]\!]_{T,F}\cdot\overline{[\![\gamma_t\mathbf{E}]\!]_{T,F}}+\sum_{F\in\mathcal{F}_{\text{ext}}}\int_F\frac{1}{Z_T+Z_{\partial\Omega}}\gamma_t\mathbf{E}^T\cdot\overline{\gamma_t\mathbf{E}^T}.$$

Par un raisonnement similaire et avec la définition du saut du champ magnétique dans (2.25), nous obtenons aussi

$$\mathcal{J}^3=\sum_{F\in\mathcal{F}_{\text{int}}}\int_F\frac{Z_TZ_K}{Z_T+Z_K}[\![\gamma_\times\mathbf{H}]\!]_F\cdot\overline{[\![\gamma_\times\mathbf{H}]\!]_F}+\sum_{F\in\mathcal{F}_{\text{ext}}}\int_F\frac{Z_TZ_{\partial\Omega}}{Z_T+Z_{\partial\Omega}}\gamma_\times^T\mathbf{H}^T\cdot\overline{\gamma_\times^T\mathbf{H}^T}.$$

Nous avons finalement prouvé

$$\|\mathbb{E}\|_{\text{GD}}^2=\Re\left(a(\mathbb{E},\mathbb{E})\right)=\mathcal{J}^2+\mathcal{J}^3.$$

Ceci termine la preuve. □

La coercivité de la formulation upwind assure l'unicité de la solution du problème de Maxwell hétérogène en adaptant le Théorème 2.2. Ainsi, l'unicité de la solution du Problème 9 (et de manière équivalente du Problème 10) est prouvée. D'autre part, l'existence de la solution numérique est assurée par le théorème du rang lorsque celle-ci est utilisée au niveau discret.

2.4 Résultats numériques pour le solveur de Trefftz direct

L'espace de Trefftz discret $\mathbb{X}_{\mathcal{T}}^h$ mène à l'introduction du système matriciel linéaire $\mathbf{A}[\mathbb{E}^h] = \mathbf{F}$ associé aux méthodes de Trefftz. En particulier, la matrice $\mathbf{A}$ est composée de blocs matriciels étant chacun lié à deux éléments T et K. Nous interprétons ces blocs comme des interactions élémentaires entre les éléments de $\mathcal{T}$. Ces liens entre les blocs matriciels et les interactions élémentaires sont mathématiquement représentés par ce que nous avons appelé précédemment : les formes consistantes élémentaires, voir la Section 2.2.1. Elles ont mené à des formulations variationnelles Trefftz coercives. Nous assemblons ici la matrice Trefftz $\mathbf{A}$ au format creux [1]. Puis, nous étudions la convergence de la solution numérique en résolvant le système matriciel à l'aide du solveur MUMPS©. Enfin, nous mettons en lumière le faible coût mémoire de la méthode de Trefftz directe face aux solveurs directs du Chapitre 1.

Le maillage $\mathcal{T}$ étudié est cartésien et est schématisé par la Figure 2.8, où

— $\mathcal{D}_\Omega$ ($\mathcal{D}_\Omega = 5\lambda$ ici) est la longueur du cube, en longueur d'onde λ,

— T est un élément cubique quelconque du maillage de côté h ($h = 1$ ici),

— l'origine du repère est $(0, 0, 0)$.

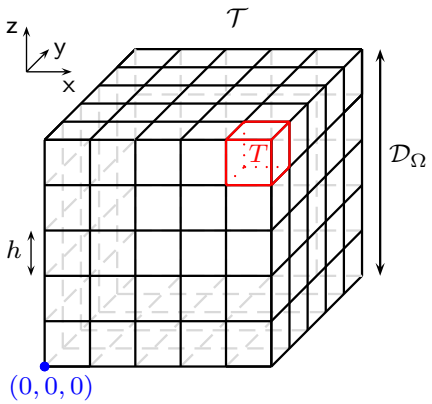

FIGURE 2.8 – Schéma du maillage cartésien $\mathcal{T}$, pour $\mathcal{D}_\Omega = 5\lambda$ et $h = 1$.

2.4.1 Caractéristiques du système linéaire

Nous rappelons que $\mathbb{X}_{\mathcal{T}}^h$ est défini par (2.4), et que l'espace de Trefftz local discret est un espace d'ondes planes. L'espace $\mathbb{X}_{\mathcal{T}}^h$ est de dimension $\#\text{ddl} := N\,\#\text{elem}$, avec $\#\text{elem} :=$

1. Ce format permet de diminuer le coût mémoire de la méthode en ne stockant que les valeurs non nulles de la matrice $\mathbf{A}$.

$\mathsf{card}(\mathcal{T})$, le nombre total d'éléments du maillage. Le système matriciel associé au Problème 9 s'écrit comme suit.

Problème 11 (Problème matriciel). *Trouver* $\mathbb{E}^h \in \mathbb{X}^h_{\mathcal{T}}$ *tel que pour tout* $\mathbb{E}' \in \mathbb{X}^h_{\mathcal{T}}$

$$a(\mathbb{E}^h, \mathbb{E}') = l(\mathbb{E}') \quad \Longleftrightarrow \quad \mathbf{A}[\mathbb{E}^h] = \mathbf{F}, \tag{2.52}$$

avec $\mathbf{A} \in \mathbb{C}^{\#\mathrm{ddl}\times\#\mathrm{ddl}}$ *et* $\mathbf{F} \in \mathbb{C}^{\#\mathrm{ddl}}$ *définis composantes par composantes par*

$$\mathbf{A}_{\mathrm{iglob},\mathrm{jglob}} := a(\mathbf{w}^{\mathrm{jglob}}, \mathbf{w}^{\mathrm{iglob}}) \text{ et } \mathbf{F}_{\mathrm{iglob}} := l(\mathbf{w}^{\mathrm{iglob}}) \quad \textit{pour} \quad \mathrm{iglob}, \mathrm{jglob} = 1, \#\mathrm{ddl}.$$

La matrice globale $\mathbf{A}$ est décomposée en blocs de matrices élémentaires

$$\mathbf{A}_{\mathrm{i_{elem}},\mathrm{j_{elem}}} := \left(\mathbf{A}^{\ell,k}_{\mathrm{i_{elem}},\mathrm{j_{elem}}}\right)_{\ell,k=1}^{N} \in \mathbb{C}^{N\times N}, \quad \text{avec} \quad \mathrm{i_{elem}}, \mathrm{j_{elem}} = 1, \#\mathrm{elem}, \tag{2.53}$$

où

$$\mathbf{A}^{\ell,k}_{\mathrm{i_{elem}},\mathrm{j_{elem}}} = a(\mathbf{w}^{k}_{\mathrm{j_{elem}}}, \mathbf{w}^{\ell}_{\mathrm{i_{elem}}}), \quad \text{avec} \quad \ell, k = 1, N.$$

Une fois assemblés, les blocs élémentaires forment la matrice $\mathbf{A}$. Elle est constituée de lignes et de colonnes de blocs élémentaires correspondant à des éléments, voir la Figure 2.9. Autrement dit, la ligne bloc de numéro $\mathrm{i_{elem}}$ contient des informations relatives à l'élément de numéro $\mathrm{i_{elem}}$. Elle inclut un terme diagonal $\mathbf{A}_{\mathrm{i_{elem}},\mathrm{i_{elem}}}$ qui représente l'interaction de $\mathbb{E}^T$ avec lui-même, voir la Définition 2.1, et aussi avec le bord, voir la Définition 2.4. La matrice contient aussi des termes $\mathbf{A}_{\mathrm{j_{elem}},\mathrm{i_{elem}}}$, correspondant à une interaction entre l'inconnue relative à l'élément $\mathrm{i_{elem}}$, notée $\mathbb{E}^T$, et celle relative à l'élément $\mathrm{j_{elem}}$, notée $\mathbb{E}^K$. Les éléments T et K sont distincts et partagent une face commune F. Ce sont des interactions de l'élément T avec un voisin, voir la Définition 2.1.

Dans cette thèse, nous considérons le cas particulier d'un maillage cartésien. Soit une ligne bloc de numéro $\mathrm{i_{elem}}$. En fonction de la colonne considérée (correspondant à un élément K ou K_{ext}) et de la nature de la face, nous distinguons différentes couleurs d'interactions (voir la Figure 2.9) :

— En noir : interactions de l'élément de numéro $\mathrm{i_{elem}}$ avec lui-même, voir la Figure 2.10,

— En rouge : interactions voisines intervenant sur les faces gauche et droite de l'élément de numéro $\mathrm{i_{elem}}$, voir la Figure 2.10,

— En bleu : interactions voisines intervenant sur les faces avant et arrière de l'élément de numéro $\mathrm{i_{elem}}$, voir la Figure 2.11,

— En vert : interactions voisines intervenant sur les faces basse et haute de l'élément de numéro $\mathrm{i_{elem}}$, voir la Figure 2.11.

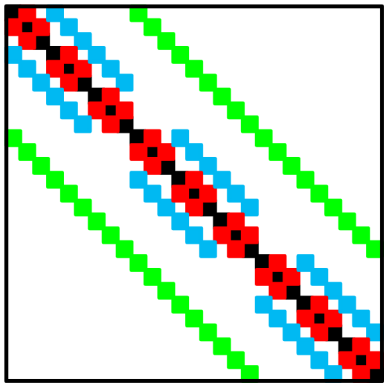

FIGURE 2.9 – Structure de la matrice $\mathbf{A}$ pour un domaine $\mathcal{D}_\Omega = 3\lambda$ (*ie* $\#\text{elem} = 27$).

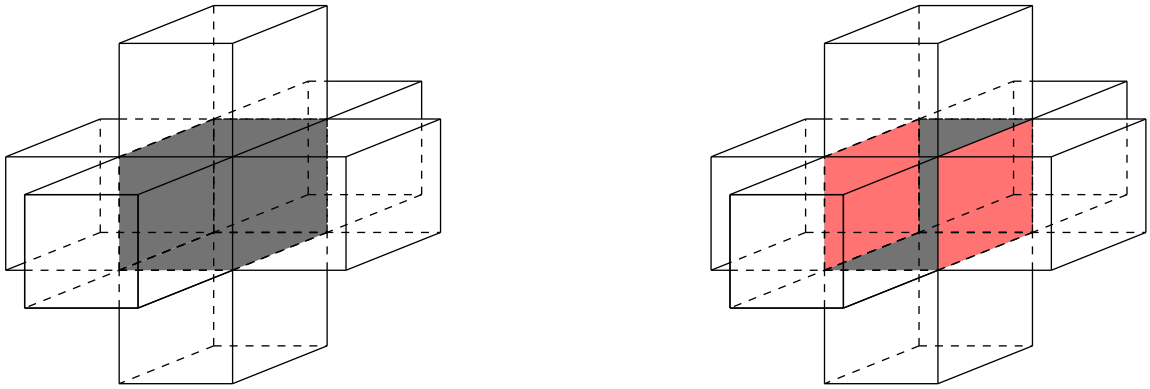

FIGURE 2.10 – Interactions d'un élément avec lui-même (à gauche) et avec ses voisins ou avec le bord via ses faces gauche et droite (à droite).

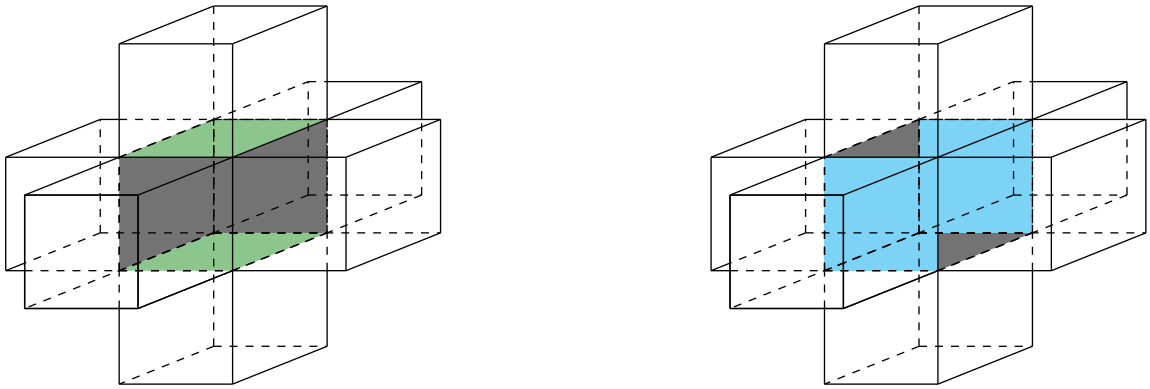

FIGURE 2.11 – Interactions d'un élément avec ses voisins ou avec le bord, via ses faces basse et haute (à gauche) et avant et arrière (à droite).

La discrétisation du problème implique un vecteur colonne complexe $[\mathbb{E}^h]$ de dimension $\#\text{ddl}$, représentant $\mathbf{E}^h \in \mathbb{X}_\mathcal{T}^h$. Ses composantes $[\mathbb{E}^h]_{\text{i}_{\text{glob}}}$ sont les amplitudes des fonctions de base $(\mathbf{w}_{\text{i}_{\text{loc}}}^\ell)_{\text{i}_{\text{loc}}=1,N} \in \mathbb{X}_T^h$, où $\mathbb{X}_T^h$ est la base engendrée par (2.5). Cela conduit à l'expression

suivante

$$\mathbb{E}^h = \sum_{\mathrm{iglob}=1}^{\#\mathrm{ddl}} [\mathbb{E}^h]_{\mathrm{iglob}} \mathbf{w}^{\mathrm{iglob}} = \sum_{\mathrm{i_{elem}}=1}^{\#\mathrm{elem}} \sum_{\mathrm{i_{loc}}=1}^{N} [\mathbb{E}^h]_{\mathrm{i_{elem}}}^{\mathrm{i_{loc}}} \mathbf{w}_{\mathrm{i_{elem}}}^{\mathrm{i_{loc}}}, \tag{2.54}$$

avec $\mathrm{iglob} := (\mathrm{i_{elem}} - 1)\,\#\mathrm{elem} + \mathrm{i_{loc}}$, et

$$\begin{cases} [\mathbb{E}^h]_{\mathrm{iglob}} := [\mathbb{E}^h]_{\mathrm{i_{elem}}}^{\mathrm{i_{loc}}}, \\ \mathbf{w}^{\mathrm{iglob}} := \mathbf{w}_{\mathrm{i_{elem}}}^{\mathrm{i_{loc}}}. \end{cases}$$

Remarque 2.14. *En pratique, le tableau* LocToGlob *assurant un lien entre les numérotations locale et globale des degrés de liberté prend la valeur*

$$\mathrm{i_{glob}} = \mathrm{LocToGlob}(\mathrm{i_{elem}}, \mathrm{i_{loc}}) = (\mathrm{i_{elem}} - 1)\,\#\mathrm{elem} + \mathrm{i_{loc}}.$$

2.4.2 Convergence du solveur de Trefftz direct

Dans le but d'étudier la convergence du schéma numérique en fonction du pas de maillage h, nous considérons un domaine cubique de côté fixé à $\mathcal{D}_\Omega = 1\lambda$. Nous simulons un dipôle assimilé à une source ponctuelle $\mathbf{x}_0 := (x_0, y_0, z_0)$ placée à l'avant du domaine $\mathbf{x}_0 = (0.5, -0.5, 0.5)$. Ce dipôle est défini par son champ électrique et son champ magnétique [69]

$$\begin{aligned} \mathbf{E}^{\mathrm{ex}}(\mathbf{x}) &= \frac{e^{ik_0 r}}{4\pi r}\Big[\big(-\frac{1}{r^2} + \frac{ik_0}{r} + k_0^2\big)(\hat{\mathbf{x}} \times (\mathbf{d} \times \hat{\mathbf{x}})) + 2\big(\frac{1}{r^2} - \frac{ik_0}{r}\big)(\mathbf{d} \cdot \hat{\mathbf{x}})\hat{\mathbf{x}}\Big], \\ \mathbf{H}^{\mathrm{ex}}(\mathbf{x}) &= \frac{e^{ik_0 r}}{4\pi r}\big(\frac{ik_0}{r} + k_0^2\big)(\hat{\mathbf{x}} \times \mathbf{d}), \end{aligned}$$

où $\mathbf{x} := (x, y, z) \in \Omega$, $\hat{\mathbf{x}} \in \Omega$, $r = \sqrt{(x - x_0)^2 + (y - y_0)^2 + (z - z_0)^2}$ et où $\mathbf{d} := (1, 0, 0)^{\mathsf{T}}$ est le moment dipolaire.

Le tableau de coefficients 2.6 définit la formulation variationnelle utilisée. L'erreur absolue ponctuelle entre $\mathbb{E}^h$ et $\mathbb{E}^{\mathrm{ex}}$ est définie par

$$e^h := \big|\mathbb{E}^h(\mathbf{x}_1) - \mathbb{E}^{\mathrm{ex}}(\mathbf{x}_1)\big|, \quad \text{avec } \mathbf{x}_1 := (0.5, 0.5, 0.5), \text{ le centre de } \Omega.$$

Remarque 2.15. *L'utilisation d'une erreur ponctuelle est critiquable comme elle est associée à une semi-norme. Toutefois, nous n'avons pas rencontré de cas, jusqu'à maintenant, où cet indicateur d'erreur converge alors que la solution numérique diverge.*

La convergence du schéma est claire d'après la Figure 2.12. En effet, e^h diminue lorsque la taille h des éléments diminue. Ainsi, nous avons développé une méthode de Trefftz convergente associée au Tableau 2.6 et simulant un dipôle électromagnétique. Un exemple d'un tel dipôle est donné en Figure 2.13 pour une taille de domaine $\mathcal{D}_\Omega = 35\lambda$, $N = 52$, $R_\Omega = 0$ et $\mathbf{x}_0 = (17.5, 17.5, -0.5)$.

Dans GoTEM3, nous pouvons aussi considérer des domaines contenant des obstacles. Nous visualisons en Figure 2.14 l'amplitude du champ électromagnétique généré par un dipôle situé en $\mathbf{x}_0 = (5, -2.5, 5)$. Le domaine $\Omega = [0, 10]^3$, dans lequel un gobelet parfaitement métallique a été placé, est maillé avec $h = \frac{1}{3}$. Nous laissons l'onde se propager à l'infini, *ie* $R_{\partial\Omega} = 0$ et nous utilisons $N = 52$ ondes planes. Pour ce cas numérique, nous avons utilisé au total 420Go de mémoire.

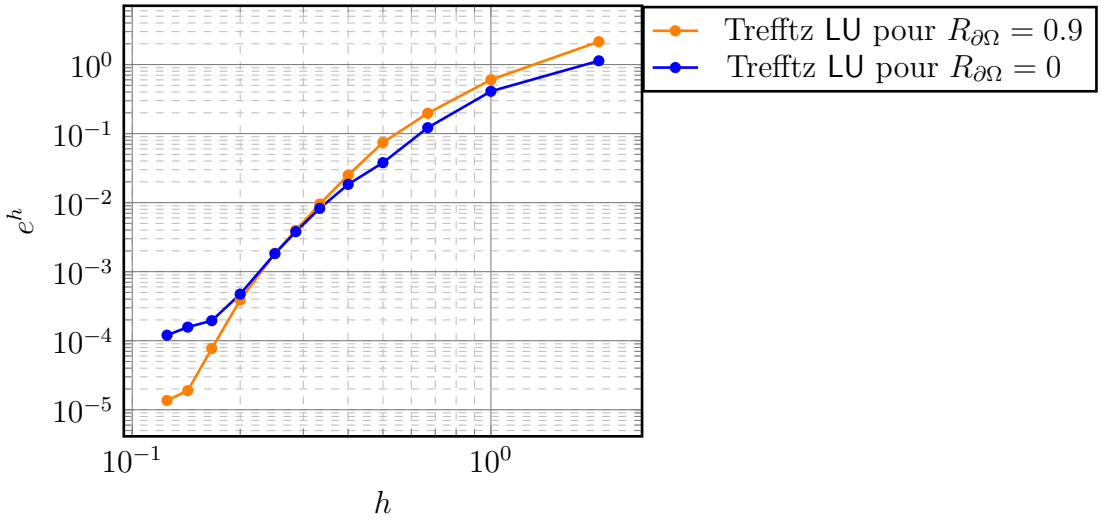

FIGURE 2.12 – Erreur absolue entre la solution numérique $\mathbb{E}^h$ et la solution exacte $\mathbb{E}^{\text{ex}}$ en fonction de la taille des éléments h pour le solveur de Trefftz direct.

Interactions de T	$a^{\mathbf{E},\mathbf{E}}$	$a^{\mathbf{E},\mathbf{H}}$	$a^{\mathbf{H},\mathbf{E}}$	$a^{\mathbf{H},\mathbf{H}}$
Lui-même	$\frac{1}{4}$	$\frac{1}{4}$	$\frac{1}{4}$	$\frac{1}{4}$
Voisin	$-\frac{1}{4}$	$-\frac{1}{4}$	$-\frac{1}{4}$	$-\frac{1}{4}$
Bord	$\frac{1}{4}$	$\frac{1}{4}$	$\frac{1}{4}$	$\frac{1}{4}$
2^{nd} membre	$\frac{1}{4}$	$\frac{1}{4}$	$\frac{1}{4}$	$\frac{1}{4}$

TABLE 2.6 – Tableau de coefficients de la formulation Trefftz utilisée pour les expériences numériques de convergence.

2.4.3 Coût mémoire de la méthode

Pour étudier le coût mémoire du solveur de Trefftz direct, nous simulons une onde électromagnétique sur un domaine Ω cubique. Les éléments du maillage sont des cubes de côté $h = 1$. La factorisation LU de la matrice $\mathbf{A}$ du Problème 11 requiert beaucoup plus de mémoire que la matrice elle-même. En effet, sans permutation, elle nécessite deux matrices

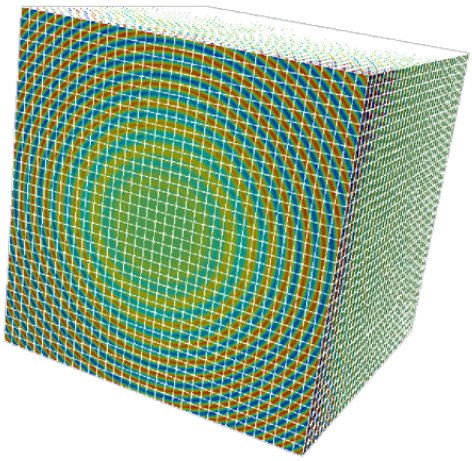

FIGURE 2.13 – Simulation d'un champ électromagnétique généré par un dipôle situé en $\mathbf{x}_0 = (17.5, 17.5, -0.5)$, par la méthode de Trefftz directe.

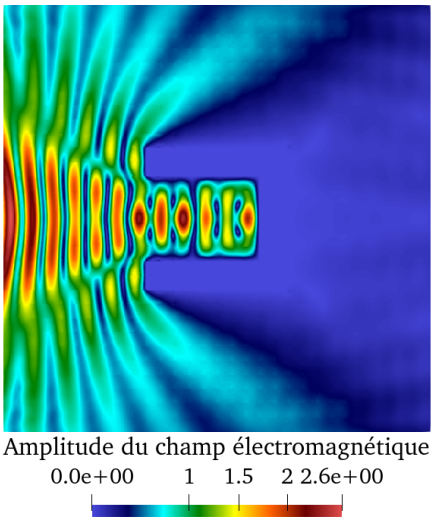

FIGURE 2.14 – Visualisation de l'amplitude du champ électromagnétique généré par un dipôle situé en $\mathbf{x}_0 = (5, -2.5, 5)$, pour $\mathcal{D}_\Omega = 10\lambda$, $h = \frac{1}{3}$, $R_{\partial\Omega} = 0$ et $N = 52$.

bandes, voir la Figure 2.15. De plus, même avec des stratégies de pivot, cette factorisation a un coût mémoire de l'ordre de $(\mathcal{D}_\Omega)^4$, voir [5]. Il s'agit du facteur bloquant de la méthode de Trefftz directe. L'augmentation du coût mémoire de la factorisation de MUMPS© évolue comme pour la méthode d'EF de Nédélec et de GD, voir la Figure 2.16. Néanmoins, la méthode de Trefftz directe permet de simuler des ondes électromagnétiques sur de plus grands domaines, *ie* jusqu'à la taille maximale $\mathcal{D}_\Omega^{\max} = 35\lambda$ (pour 1To), ce qui apparaît clairement sur la Figure 2.17. Les méthodes classiques introduites dans le Chapitre 1 sont limitées aux domaines de taille $\mathcal{D}_\Omega^{\max} = 19\lambda$. En ce sens, nous pouvons dire que nous avons obtenu de

FIGURE 2.15 – Structure de la décomposition LU de $\mathbf{A}$: à gauche, la matrice triangulaire inférieure L ; à droite, la matrice triangulaire supérieure U.

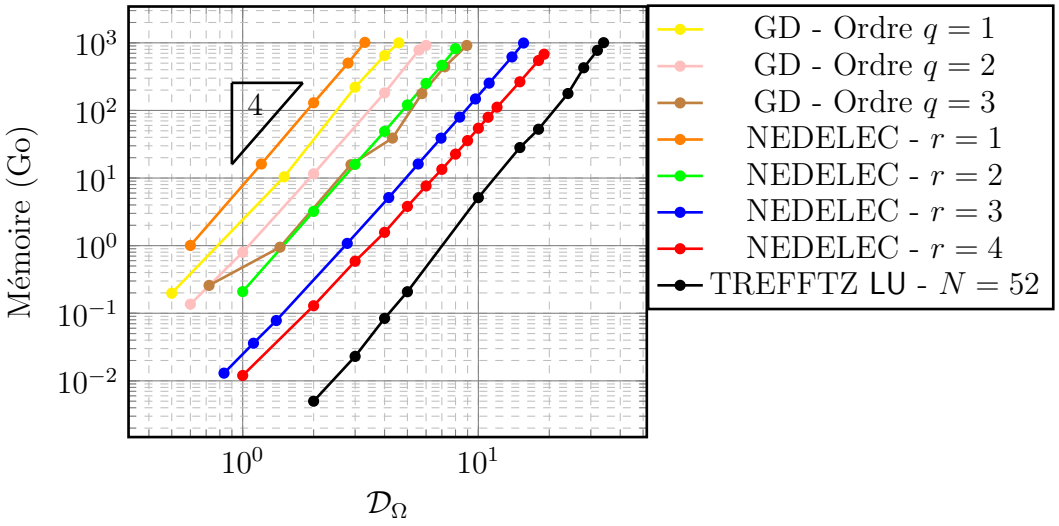

FIGURE 2.16 – Comparaison des coûts mémoire des méthodes de Trefftz directe, de Nédélec et de GD, en fonction de la taille du domaine $\mathcal{D}_\Omega$ (échelle loglog).

meilleurs résultats avec un solveur de Trefftz direct sur un maillage cubique.

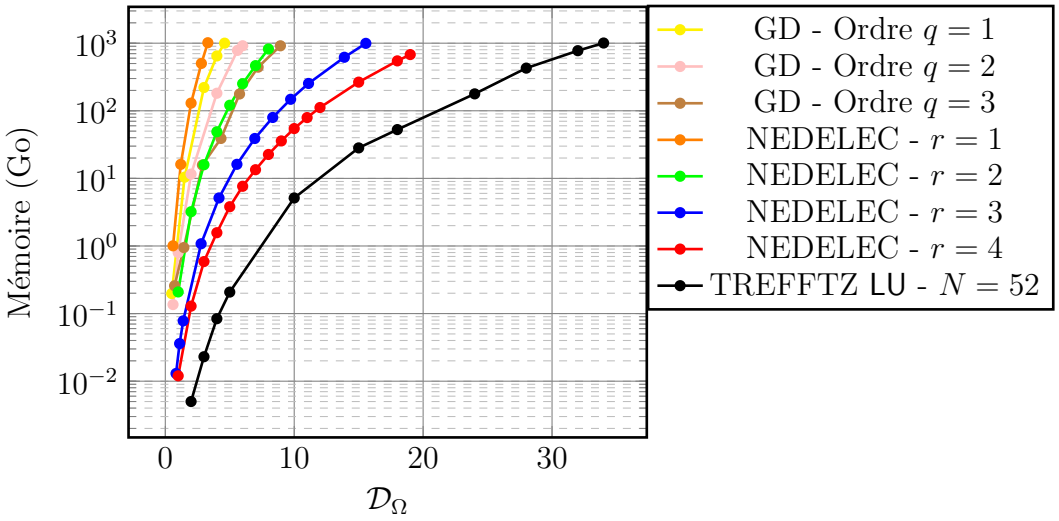

FIGURE 2.17 – Comparaison des coûts mémoire des méthodes de Trefftz directe, de Nédélec et de GD, en fonction de la taille du domaine $\mathcal{D}_\Omega$ (échelle semilog).

2.5 Conclusion

Dans ce chapitre, nous avons construit des formulations variationnelles Trefftz consistantes. Nous avons exhibé des critères assurant la coercivité de la méthode. Puis, nous avons proposé un choix optimal de paramètres grâce aux solveurs de Riemann, au sens où ils n'engendrent pas de réflexions à incidence normale. Cette approche a été généralisée au cas de milieux hétérogènes, grâce aux traces numériques upwind qui se révèlent équivalentes aux traces numériques de Riemann. Les formulations homogènes (*resp.* hétérogènes) sont

Interactions de T	$a^{\mathbf{E},\mathbf{E}}$	$a^{\mathbf{E},\mathbf{H}}$	$a^{\mathbf{H},\mathbf{E}}$	$a^{\mathbf{H},\mathbf{H}}$
Lui-même	1	1	1	1
Voisin	-1	-1	-1	-1
Bord	$1 - R_{\partial\Omega}$	$1 - R_{\partial\Omega}$	$1 + R_{\partial\Omega}$	$1 + R_{\partial\Omega}$
2^{nd} membre	$1 - R_{\partial\Omega}$	$1 - R_{\partial\Omega}$	$1 + R_{\partial\Omega}$	$1 + R_{\partial\Omega}$

TABLE 2.7 – Tableau de coefficients d'une formulation Trefftz basée sur les traces numériques de Riemann pour $F \in \mathcal{F}_T$ avec une condition de bord d'impédance.

associées au Tableau 2.7 (*resp.* Tableau 2.8).

Remarque 2.16. *Le Tableau 2.7 s'obtient à partir des Théorèmes 2.4 et 2.5 en insérant les traces de Riemann dans la formule de réciprocité* (2.7). *Le Tableau 2.8 s'obtient à partir du Problème 10.*

Interactions de T	$a^{\mathbf{E},\mathbf{E}}$	$a^{\mathbf{E},\mathbf{H}}$	$a^{\mathbf{H},\mathbf{E}}$	$a^{\mathbf{H},\mathbf{H}}$
Lui-même	$\frac{1}{Z_T + Z_K}$	$\frac{Z_T}{Z_T + Z_K}$	$\frac{Z_K}{Z_T + Z_K}$	$\frac{Z_T\, Z_K}{Z_T + Z_K}$
Voisin	$-\frac{1}{Z_K + Z_T}$	$-\frac{Z_T}{Z_K + Z_T}$	$-\frac{Z_K}{Z_K + Z_T}$	$-\frac{Z_T\, Z_K}{Z_K + Z_T}$
Bord	$\frac{1}{Z_T + Z_{\partial\Omega}}$	$\frac{Z_T}{Z_T + Z_{\partial\Omega}}$	$\frac{Z_{\partial\Omega}}{Z_T + Z_{\partial\Omega}}$	$\frac{Z_T\, Z_{\partial\Omega}}{Z_T + Z_{\partial\Omega}}$
2^{nd} membre	$\frac{1}{Z_T + Z_{\partial\Omega}}$	$\frac{Z_T}{Z_T + Z_{\partial\Omega}}$	$\frac{Z_{\partial\Omega}}{Z_T + Z_{\partial\Omega}}$	$\frac{Z_T\, Z_{\partial\Omega}}{Z_T + Z_{\partial\Omega}}$

TABLE 2.8 – Tableau de coefficients d'une formulation Trefftz basée sur les traces numériques upwind pour $F \in \mathcal{F}_T$ avec une condition de bord d'impédance. Dans le cas où $F \in \mathcal{F}_{\text{int}}$, nous prenons $F := \partial T \cap \partial K$.

La méthode de Trefftz directe permet de considérer des tailles de domaine environ deux fois plus grandes qu'avec des méthodes classiques du Chapitre 1, voir la Figure 2.18. Malheureusement, la résolution directe du problème de Trefftz est extrêmement coûteuse. Afin de remédier au problème de coût mémoire, l'idée est d'appliquer un solveur itératif pour résoudre le Problème 11 de Trefftz.

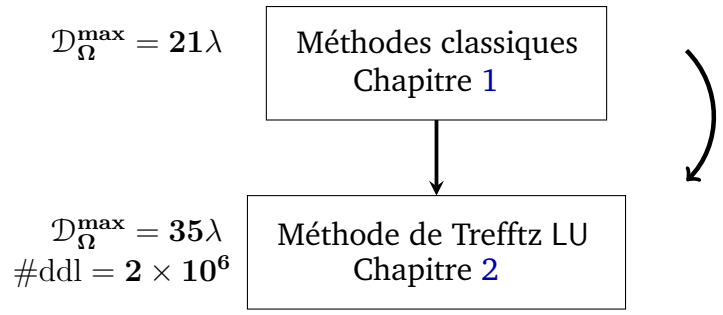

FIGURE 2.18 – Augmentation de la taille $\mathcal{D}_{\Omega}$ qu'il est possible de considérer grâce à la mise en place d'une méthode de Trefftz, où $\mathcal{D}_{\Omega}^{\max}$ est la taille maximale atteinte dans chacun des chapitres pour 1To de mémoire, et où $\#\text{ddl} = N \times \#\text{elem}$ avec $N = 52$ et $h = 1$.

CHAPITRE 3

SOLVEUR ITÉRATIF HÉTÉROGÈNE DE TYPE TREFFTZ

Sommaire

Le Chapitre précédent nous a permis de construire une formulation variationnelle Trefftz hétérogène, grâce aux traces numériques upwind. Sa version homogène a été obtenue à l'aide des traces numériques de Riemann. Elle a mené à un solveur de résolution directe qui améliore les capacités mémoire des méthodes d'EF et de GD du Chapitre 1. Ce progrès est toutefois limité, au sens où nous ne pouvons pas considérer des domaines de taille supérieure à 35λ avec 1To de mémoire. Bien que ce solveur pourrait être davantage optimisé, ses capacités resteraient freinées par l'utilisation d'une factorisation LU. C'est pourquoi nous décidons de mettre en place un solveur de Trefftz itératif.

Premièrement, nous introduisons des nouvelles traces numériques : celles de Cessenat-Després [21] menant à une UWVF. Cette dernière s'avère être équivalente au Problème 10 du Chapitre 2. Néanmoins, l'UWVF Trefftz permet d'introduire naturellement une décomposition dite "régulière-singulière" adaptée à un algorithme itératif de bloc-Jacobi. Malheureusement, nous verrons que ce dernier rencontre des problèmes de convergence dans certains cas.

Les méthodes itératives [23] permettent de considérablement diminuer le coût mémoire d'une méthode en évitant la factorisation LU de la matrice $\mathbf{A}$. Dans le Chapitre 3, nous mettons donc en œuvre deux méthodes itératives basées sur des espaces de type Krylov [91] :

— une méthode de General Minimum RESidual (GMRES), reposant sur la théorie établie dans [91], et pour laquelle nous utilisons l'algorithme de GMRES développé par le CERFACS© [44],

— une méthode de Krylov Galerkin (KG), dont la théorie est établie à partir du point de vue de Galerkin de la méthode de Trefftz.

Dans un premier temps, nous rappelons quelques fondements de la méthode de GMRES : les espaces de Krylov (utilisés par les deux méthodes énoncées ci-dessus), et la stratégie de *restart*, permettant de diminuer le coût mémoire de l'algorithme. L'idée principale est de minimiser un résidu sur l'espace de Krylov. Ce dernier est construit grâce à une orthogonalisation des vecteurs le composant. Pour toutes les matrices considérées, la méthode de GMRES implique le produit auto-adjoint $\mathbf{A}^*\mathbf{A}$ de telle sorte que le spectre associé sera toujours à partie réelle positive. Cela conduit à de bonnes propriétés de convergence. Mais cette méthode ne tire alors pas profit du caractère coercif de la forme sesquilinéaire a de l'UWVF. Ainsi, nous développons la deuxième méthode qui, basée sur la théorie de Galerkin, bénéficie d'un spectre à partie réelle positive dû à la coercivité et non à la construction de la méthode elle-même.

Dans un second temps, nous expliquons la méthode de KG en détaillant la construction des espaces de Krylov par un algorithme de pseudo-orthogonalisation. Ensuite, nous

établissons la convergence de la méthode à partir des propriétés du problème UWVF de Cessenat-Després. De plus, nous écrivons une version préconditionnée de ce problème dans le but d'accélérer la convergence du solveur de KG.

Finalement, nous comparons ces deux méthodes de Krylov. Puis, nous choisissons d'employer la méthode de GMRES pour observer les vitesses de convergence de la méthode UWVF itérative préconditionnée ou non.

3.1 Formulation Trefftz par le schéma UWVF de Cessenat-Després pour les milieux hétérogènes

Cette section a pour but d'introduire les traces de Cessenat-Després et la formulation Trefftz UWVF hétérogène associée. Celle-ci est équivalente à la formulation upwind. De cette façon, nous prouvons aussi l'équivalence entre les traces numériques de Riemann, upwind et de Cessenat-Després. Nous développons ensuite un algorithme itératif de Jacobi basé sur une décomposition de Cessenat-Després. Cette dernière mène au problème itératif UWVF discret dont la matrice associée est strictement contractante. Néanmoins, nous constaterons que cette propriété n'est pas toujours vraie numériquement à cause des erreurs d'arrondis. De plus, nous montrerons aussi que cet algorithme peut diverger dans certaines configurations. Cela nous conduira aux méthodes de Krylov de la Section 3.2.

3.1.1 Dérivation des traces numériques de Cessenat-Després dans le cas hétérogène

Le but de cette partie est d'expliciter les traces numériques de Cessenat-Després. Pour cela nous introduisons, pour chaque $T \in \mathcal{T}$, l'opérateur de Cessenat-Després $\mathcal{U}^T$ hétérogène associant à la trace sortante $\mathbf{x}^T$ d'un champ électromagnétique

$$\mathbf{x}^T := \gamma_{\text{out}}^T \mathbb{E}^T \underset{(2.39)}{=} \gamma_t \mathbf{E}^T - Z_T \gamma_\times^T \mathbf{H}^T, \tag{3.1}$$

sa trace entrante $\mathcal{U}^T \mathbf{x}^T$:

$$\mathcal{U}^T \mathbf{x}^T := \gamma_{\text{in}}^T \mathbb{E}^T \underset{(2.39)}{=} \gamma_t \mathbf{E}^T + Z_T \gamma_\times^T \mathbf{H}^T. \tag{3.2}$$

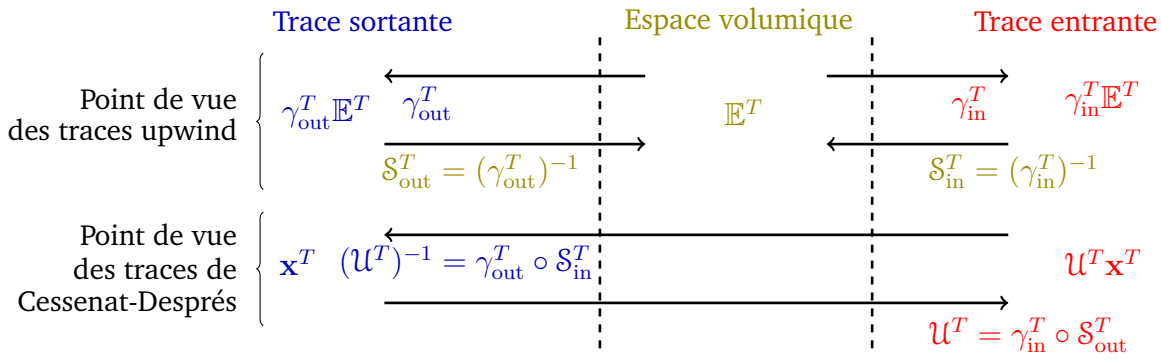

FIGURE 3.1 – Comparaison entre les points de vue des traces numériques upwind et de Cessenat-Després.

Plus précisément, l'opérateur $\mathcal{U}^T : \mathrm{L}^2_t(\partial T) \to \mathrm{L}^2_t(\partial T)$ est défini par $\mathcal{U}^T := \gamma^T_{\text{in}} \circ \mathcal{S}^T_{\text{out}}$, avec l'opérateur solution

$$\mathcal{S}^T_{\text{out}} : \mathrm{L}^2_t(\partial T) \to \mathbb{X}_T, \quad \mathbf{x}^T \mapsto \mathbb{E}^T \text{ vérifiant (2.1)}, \tag{3.3}$$

et $\gamma^T_{\text{out}}\mathbb{E}^T = \mathbf{x}^T$. Ces opérateurs sont représentés et liés aux traces upwind dans la Figure 3.1. En utilisant les définitions ci-dessus, la composante tangentielle électrique et la trace tangentielle magnétique sont déduites

$$\gamma_t\mathbf{E}^T = \frac{1}{2}\left(\mathcal{U}^T\mathbf{x}^T + \mathbf{x}^T\right) \quad \text{et} \quad \gamma^T_\times\mathbf{H}^T = \frac{1}{2Z_T}\left(\mathcal{U}^T\mathbf{x}^T - \mathbf{x}^T\right). \tag{3.4}$$

Le Problème 9 peut alors être reformulé avec les notations de Cessenat-Després. En effet, les traces numériques upwind (2.45) pour les faces intérieures, s'écrivent en fonction de $\mathbf{x}^T$ et de $\mathbf{x}^K$

$$\begin{cases} \widehat{(\gamma_t\mathbf{E})}_{|F} &= \dfrac{Z_K}{Z_K+Z_T}\mathbf{x}^T + \dfrac{Z_T}{Z_K+Z_T}\mathbf{x}^K \\ &= \dfrac{1}{2}\left(\dfrac{Z_K-Z_T}{Z_K+Z_T}\mathbf{x}^T + \dfrac{2Z_T}{Z_K+Z_T}\mathbf{x}^K + \mathbf{x}^T\right), \\ \widehat{(\gamma^T_\times\mathbf{H})}_{|F} &= -\dfrac{1}{Z_K+Z_T}\mathbf{x}^T + \dfrac{1}{Z_K+Z_T}\mathbf{x}^K \\ &= \dfrac{1}{2Z_T}\left(\dfrac{Z_K-Z_T}{Z_K+Z_T}\mathbf{x}^T + \dfrac{2Z_T}{Z_K+Z_T}\mathbf{x}^K - \mathbf{x}^T\right), \end{cases}$$

et les traces upwind de bord (2.46) pour les faces extérieures deviennent aussi

$$\begin{cases} \widehat{(\gamma_t \mathbf{E})}_{|F} &= \dfrac{Z_{\partial\Omega}}{Z_{\partial\Omega}+Z_T}\mathbf{x}^T + \dfrac{Z_T}{Z_{\partial\Omega}+Z_T}\mathbf{g}^T \\ &= \dfrac{1}{2}\left(\dfrac{Z_{\partial\Omega}-Z_T}{Z_{\partial\Omega}+Z_T}\mathbf{x}^T + \dfrac{2Z_T}{Z_{\partial\Omega}+Z_T}\mathbf{g}^T + \mathbf{x}^T\right), \\ \widehat{(\gamma_\times^T \mathbf{H})}_{|F} &= -\dfrac{1}{Z_{\partial\Omega}+Z_T}\mathbf{x}^T + \dfrac{1}{Z_{\partial\Omega}+Z_T}\mathbf{g}^T \\ &= \dfrac{1}{2Z_T}\left(\dfrac{Z_{\partial\Omega}-Z_T}{Z_{\partial\Omega}+Z_T}\mathbf{x}^T + \dfrac{2Z_T}{Z_{\partial\Omega}+Z_T}\mathbf{g}^T - \mathbf{x}^T\right). \end{cases}$$

Nous pouvons directement définir la trace entrante numérique $\widehat{\mathcal{U}^T\mathbf{x}}$ en identifiant aux formules (3.4). Nous avons alors pour $F \in \mathcal{F}_T$,

$$\widehat{(\gamma_t \mathbf{E})}_{|F} = \frac{1}{2}\Big((\widehat{\mathcal{U}^T\mathbf{x}})_{|F} + \mathbf{x}^T\Big) \quad \text{et} \quad \widehat{(\gamma_\times^T \mathbf{H})}_{|F} = \frac{1}{2\,Z_T}\Big((\widehat{\mathcal{U}^T\mathbf{x}})_{|F} - \mathbf{x}^T\Big), \tag{3.5}$$

où $\widehat{\mathcal{U}^T\mathbf{x}}$ est la trace numérique de Cessenat-Després associée à l'élément T définie par

$$(\widehat{\mathcal{U}^T\mathbf{x}})_{|F} := \frac{Z_K - Z_T}{Z_K + Z_T}\mathbf{x}^T + \frac{2Z_T}{Z_K+Z_T}\mathbf{x}^K \quad \text{pour } F \in \mathcal{F}_{\text{int}}, \tag{3.6a}$$

$$(\widehat{\mathcal{U}^T\mathbf{x}})_{|F} := \frac{Z_{\partial\Omega} - Z_T}{Z_{\partial\Omega} + Z_T}\mathbf{x}^T + \frac{2Z_T}{Z_{\partial\Omega}+Z_T}\mathbf{g}^T \quad \text{pour } F \in \mathcal{F}_{\text{ext}}. \tag{3.6b}$$

Remarque 3.1. *Comme pour la Remarque 2.11, la formule (3.6b) peut être vue comme une interaction avec un élément virtuel extérieur à Ω, où nous aurions $Z_{\partial\Omega} = Z_K$ et $\mathbf{g}^T = \mathbf{x}^K$.*

En remplaçant les traces numériques du Problème 9 par (3.5), nous obtenons pour tout $\mathbf{x}' \in \mathrm{L}^2_t(\partial\mathcal{T}) := \prod_{T\in\mathcal{T}} \mathrm{L}^2_t(\partial T)$

$$\sum_{T\in\mathcal{T}} \int_{\partial T} Z_T^{-1}\left(\mathbf{x}^T \cdot \overline{\mathbf{x}'^T} - \widehat{\mathcal{U}^T\mathbf{x}} \cdot \overline{\mathcal{U}^T\mathbf{x}'^T}\right) = 0.$$

Cela mène à une nouvelle façon d'écrire la formulation variationnelle Trefftz. Celle-ci est la Formulation Variationnelle Ultra-Faible (UWVF) hétérogène de Cessenat-Després.

Problème 12 (Problème de Cessenat-Després)**.** *Trouver $\mathbf{x} \in \mathrm{L}^2_t(\partial\mathcal{T})$ tel que pour tout $\mathbf{x}' \in \mathrm{L}^2_t(\partial\mathcal{T})$ nous avons*

$$(\mathbf{x}, \mathbf{x}')_{\mathrm{L}^2_t(\partial\mathcal{T})} - (\widehat{\mathcal{U}\mathbf{x}}, \mathcal{U}\mathbf{x}')_{\mathrm{L}^2_t(\partial\mathcal{T})} = 0, \tag{3.7}$$

où $\left(\widehat{\mathcal{U}\mathbf{x}}\right)^T := \widehat{\mathcal{U}^T\mathbf{x}}$, $(\mathcal{U}\mathbf{x}')^T := \mathcal{U}^T\mathbf{x}'^T$ et où le produit scalaire sur $\mathrm{L}^2_t(\partial\mathcal{T})$ est défini par

$$(\mathbf{x}, \mathbf{x}')_{\mathrm{L}^2_t(\partial\mathcal{T})} := \sum_{T\in\mathcal{T}} \int_{\partial T} Z_T^{-1}\mathbf{x}^T \cdot \overline{\mathbf{x}'^T}. \tag{3.8}$$

Comme illustré sur la Figure 3.1, un lien est établi entre le Problème 12 basé sur les traces hétérogènes de Cessenat-Després, voir (3.1), (3.2) et (3.5), et le Problème 9 basé sur les traces upwind, voir (2.45) et (2.46).

Théorème 3.1 (Équivalence des formulations)**.** *Le Problème 9 est équivalent au Problème 12 au sens où*

— *Si* $\mathbf{x}$ *est solution du Problème 12 alors* $\mathbb{E}$*, défini par* $\mathbb{E}^T = \mathbb{S}^T_{\text{out}}\mathbf{x}^T$*, est solution du Problème 9.*

— *Si* $\mathbb{E}$ *est solution du Problème 9 alors* $\mathbf{x}$*, défini par* $\mathbf{x}^T = \gamma^T_{\text{out}}\mathbb{E}^T$*, est solution du Problème 12.*

3.1.2 Construction de l'algorithme itératif de type Jacobi

En prenant en compte les traces numériques de Cessenat-Després définies par (3.6), nous obtenons la formulation variationnelle du Problème 12.

Problème 13 (Formulation variationnelle de Cessenat-Després)**.** *Trouver* $\mathbf{x} \in \mathrm{L}^2_t(\partial\mathcal{T})$ *tel que pour tout* $\mathbf{x}' \in \mathrm{L}^2_t(\partial\mathcal{T})$

$$\mathbf{a}(\mathbf{x},\mathbf{x}') \;=\; \mathbf{l}(\mathbf{x}'), \quad \textit{avec} \quad \mathbf{a}(\mathbf{x},\mathbf{x}') = (\mathbf{x},\mathbf{x}')_{L^2_t(\partial\mathcal{T})} - \mathbf{k}(\mathbf{x},\mathbf{x}'), \tag{3.9}$$

où la forme sesquilinéaire $\mathbf{k} : \mathrm{L}^2_t(\partial\mathcal{T}) \times \mathrm{L}^2_t(\partial\mathcal{T}) \to \mathbb{C}$ *et la forme antilinéaire* $\mathbf{l} : \mathrm{L}^2_t(\partial\mathcal{T}) \to \mathbb{C}$ *sont données par*

$$\mathbf{k}(\mathbf{x},\mathbf{x}') := (\Pi_{\mathcal{U}}\mathbf{x}, \mathcal{U}\mathbf{x}')_{\mathrm{L}^2_t(\partial\mathcal{T})} \quad \textit{et} \quad \mathbf{l}(\mathbf{x}') := (\mathbf{g}_{\mathcal{U}}, \mathcal{U}\mathbf{x}')_{\mathrm{L}^2_t(\partial\mathcal{T})}, \tag{3.10}$$

avec $\Pi_{\mathcal{U}} : \mathrm{L}^2_t(\partial\mathcal{T}) \to \mathrm{L}^2_t(\partial\mathcal{T})$ *l'opérateur de trace et* $\mathbf{g}_{\mathcal{U}} \in \mathrm{L}^2_t(\partial\mathcal{T})$ *le second membre, tous les deux définis sur chaque* $T \in \mathcal{T}$ *et pour toute face* $F \in \mathcal{F}_T$*, par*

$$(\Pi_{\mathcal{U}}\mathbf{x})^T_{|F} := (\Pi^T_{\mathcal{U}}\mathbf{x})_{|F} := \begin{cases} \left(\widehat{\mathcal{U}^T\mathbf{x}}\right)_{|F} & \textit{si } F \in \mathcal{F}_{\text{int}}, \\ \dfrac{Z_{\partial\Omega} - Z_T}{Z_{\partial\Omega} + Z_T}\mathbf{x}^T & \textit{si } F \in \mathcal{F}_{\text{ext}}, \end{cases} \tag{3.11}$$

$$(\mathbf{g}^T_{\mathcal{U}})_{|F} := \begin{cases} 0 & \textit{si } F \in \mathcal{F}_{\text{int}}, \\ \dfrac{2Z_T}{Z_{\partial\Omega} + Z_T}\mathbf{g}^T & \textit{si } F \in \mathcal{F}_{\text{ext}}. \end{cases}$$

La proposition suivante assure que l'opérateur $\mathbf{k}$ est contractant. Ainsi, (3.9) et (3.10) conduisent à un problème de point fixe.

Proposition 3.1 (Opérateurs contractants)**.** *Si V est un espace de Hilbert, les normes de l'opérateur linéaire $\mathcal{A} : V \to V$ et de la forme sesquilinéaire $\mathsf{a} : V \times V \longrightarrow \mathbb{C}$ sont définies par*

$$|||\mathcal{A}|||_V := \sup_{\mathbf{x}\in V\setminus\{0\}} \sup_{\mathbf{x}'\in V\setminus\{0\}} \frac{|(\mathcal{A}\,\mathbf{x},\mathbf{x}')_V|}{\|\mathbf{x}\|_V\ \|\mathbf{x}'\|_V} \quad \textit{et} \quad |||\mathsf{a}|||_V := \sup_{\mathbf{x}\in V\setminus\{0\}} \sup_{\mathbf{x}'\in V\setminus\{0\}} \frac{|\mathsf{a}(\mathbf{x},\mathbf{x}')_V|}{\|\mathbf{x}\|_V\ \|\mathbf{x}'\|_V}.$$

Nous avons les trois propriétés suivantes :

(i) L'opérateur $\Pi_{\mathfrak{U}} : \mathrm{L}^2_t(\partial\mathcal{T}) \to \mathrm{L}^2_t(\partial\mathcal{T})$ vérifie $|||\Pi_{\mathfrak{U}}|||_{\mathrm{L}^2_t(\partial\mathcal{T})} \leq 1$.

(ii) L'opérateur $\mathfrak{U} : \mathrm{L}^2_t(\partial\mathcal{T}) \to \mathrm{L}^2_t(\partial\mathcal{T})$ vérifie $|||\mathfrak{U}|||_{\mathrm{L}^2_t(\partial\mathcal{T})} = 1$.

(iii) L'opérateur $\mathbf{k} : \mathrm{L}^2_t(\partial\mathcal{T}) \times \mathrm{L}^2_t(\partial\mathcal{T}) \to \mathbb{C}$ est contractant, ie $|||\mathbf{k}|||_{\mathrm{L}^2_t(\partial\mathcal{T})} \leq 1$.

Démonstration. (i) En prenant le point de vue de la sommation par face (voir la Remarque 2.13), nous avons pour tout $\mathbf{x} \in \mathrm{L}^2_t(\partial\mathcal{T})$

$$\sum_{T\in\mathcal{T}} \int_{\partial T} Z_T^{-1} |\Pi^T_{\mathfrak{U}}\mathbf{x}|^2 = \sum_{F\in\mathcal{F}} \mathcal{K}_F \text{ avec } \mathcal{K}_F := \begin{cases} \displaystyle\int_F \underbrace{Z_T^{-1}|\Pi^T_{\mathfrak{U}}\mathbf{x}|^2}_{:=\mathcal{J}_1} + \underbrace{Z_K^{-1}|\Pi^K_{\mathfrak{U}}\mathbf{x}|^2}_{:=\mathcal{J}_2} & \text{pour } F \in \mathcal{F}_{\mathrm{int}}, \\ \displaystyle\int_F Z_T^{-1}|\Pi^T_{\mathfrak{U}}\mathbf{x}|^2 & \text{pour } F \in \mathcal{F}_{\mathrm{ext}}. \end{cases}$$

Pour $F \in \mathcal{F}_{\mathrm{int}}$ séparant deux éléments T et K, nous avons par (3.11)

$$\mathcal{J}_1 = \int_F Z_T^{-1}\Big(\frac{Z_K - Z_T}{Z_K + Z_T}\Big)^2 |\mathbf{x}^T|^2 + \frac{2(Z_T - Z_K)}{(Z_K + Z_T)^2}(\mathbf{x}^T \cdot \overline{\mathbf{x}^K} + \mathbf{x}^K \cdot \overline{\mathbf{x}^T}) + \frac{4Z_T}{(Z_K + Z_T)^2}|\mathbf{x}^K|^2,$$

et

$$\mathcal{J}_2 = \int_F Z_K^{-1}\Big(\frac{Z_T - Z_K}{Z_T + Z_K}\Big)^2 |\mathbf{x}^K|^2 + \frac{2(Z_K - Z_T)}{(Z_T + Z_K)^2}(\mathbf{x}^K \cdot \overline{\mathbf{x}^T} + \mathbf{x}^T \cdot \overline{\mathbf{x}^K}) + \frac{4Z_K}{(Z_T + Z_K)^2}|\mathbf{x}^T|^2.$$

Ainsi, il suit que

$$\mathcal{K}_F = \int_F Z_T^{-1}|\mathbf{x}^T|^2 \ + \ Z_K^{-1}|\mathbf{x}^K|^2.$$

Nous obtenons pour $F \in \mathcal{F}_{\mathrm{ext}}$

$$\mathcal{K}_F = \int_F Z_T^{-1}|\Pi^T_{\mathfrak{U}}\mathbf{x}^T|^2 = \int_F Z_T^{-1}\Big(\frac{Z_{\partial\Omega} - Z_T}{Z_{\partial\Omega} + Z_T}\Big)^2 |\mathbf{x}^T|^2. \tag{3.13}$$

En remarquant que $\dfrac{Z_{\partial\Omega} - Z_T}{Z_{\partial\Omega} + Z_T} \leq 1$ et en utilisant la définition du produit scalaire global (3.8), nous avons

$$\|\Pi_{\mathfrak{U}}\mathbf{x}\|^2_{\mathrm{L}^2_t(\partial\mathcal{T})} \ = \ \sum_{F\in\mathcal{F}} \mathcal{K}_F \text{ avec } \mathcal{K}_F \leq \begin{cases} \displaystyle\int_F Z_T^{-1}|\mathbf{x}^T|^2 + Z_K^{-1}|\mathbf{x}^K|^2 & \text{pour } F \in \mathcal{F}_{\mathrm{int}}, \\ \displaystyle\int_F Z_T^{-1}|\mathbf{x}^T|^2 & \text{pour } F \in \mathcal{F}_{\mathrm{ext}}. \end{cases}$$

En appliquant la Remarque 2.13, nous obtenons un assemblage par éléments menant à

$$\|\Pi_{\mathcal{U}}\mathbf{x}\|^2_{\mathrm{L}^2_t(\partial\mathcal{T})} \leq \sum_{T\in\mathcal{T}}\sum_{F\in\mathcal{F}_T}\int_F Z_T^{-1}|\mathbf{x}^T|^2 = \|\mathbf{x}\|^2_{\mathrm{L}^2_t(\partial\mathcal{T})} \quad\Longrightarrow\quad \|\Pi_{\mathcal{U}}\|_{\mathrm{L}^2_t(\partial\mathcal{T})} \leq 1.$$

□

Démonstration. (ii) Nous rappelons que nous avons la formule de réciprocité (2.7). Nous remarquons alors que

$$r_T(\mathbb{E},\mathbb{E}') = \int_{\partial T}\left(\gamma_\times^T\mathbf{H}^T\cdot\overline{\gamma_t\mathbf{E}'^T} + \gamma_t\mathbf{E}^T\cdot\overline{\gamma_\times^T\mathbf{H}'^T}\right) = \frac{\mathcal{I}_1-\mathcal{I}_2}{2} = 0,$$

avec $\mathcal{I}_1$ et $\mathcal{I}_2$ définis par

$$\mathcal{I}_1 := \int_{\partial T} Z_T^{-1}\gamma_{\mathrm{in}}^T\mathbb{E}^T\cdot\overline{\gamma_{\mathrm{in}}^T\mathbb{E}'^T} = \int_{\partial T} Z_T^{-1}(\gamma_t\mathbf{E}^T + Z_T\gamma_\times^T\mathbf{H}^T)\cdot\left(\overline{\gamma_t\mathbf{E}'^T} + Z_T\overline{\gamma_\times^T\mathbf{H}'^T}\right),$$

et

$$\mathcal{I}_2 := \int_{\partial T} Z_T^{-1}\gamma_{\mathrm{out}}^T\mathbb{E}^T\cdot\overline{\gamma_{\mathrm{out}}^T\mathbb{E}'^T} = \int_{\partial T} Z_T^{-1}(\gamma_t\mathbf{E}^T - Z_T\gamma_\times^T\mathbf{H}^T)\cdot\left(\overline{\gamma_t\mathbf{E}'^T} - Z_T\overline{\gamma_\times^T\mathbf{H}'^T}\right).$$

Par conséquent, pour tout $\mathbb{E}^T\in\mathbb{X}_T$, $\mathbb{E}'^T\in\mathbb{X}_T$ et pour tout $\mathbf{x}^T\in\mathrm{L}^2_t(\partial T)$, $\mathbf{x}'^T\in\mathrm{L}^2_t(\partial T)$, nous avons

$$\int_{\partial T} Z_T^{-1}\gamma_{\mathrm{out}}^T\mathbb{E}^T\cdot\overline{\gamma_{\mathrm{out}}^T\mathbb{E}'^T} = \int_{\partial T} Z_T^{-1}\gamma_{\mathrm{in}}^T\mathbb{E}^T\cdot\overline{\gamma_{\mathrm{in}}^T\mathbb{E}'^T} = (\mathbb{E}^T,\mathbb{E}'^T)_{\mathbb{X}_T} = (\mathcal{S}_{\mathrm{out}}^T\mathbf{x}^T,\mathcal{S}_{\mathrm{out}}^T\mathbf{x}'^T)_{\mathbb{X}_T}.$$

Comme $\mathcal{S}_{\mathrm{out}}^T : \mathrm{L}^2_t(\partial T)\to\mathbb{X}_T$ est bijectif, alors l'espace $\mathbb{X}_T$ peut être paramétrisé par $\mathrm{L}^2_t(\partial T)$. En sommant sur les éléments $T\in\mathcal{T}$, cela mène à $\|\mathbf{x}\|^2_{\mathrm{L}^2_t(\partial\mathcal{T})} = \|\mathcal{U}\mathbf{x}\|^2_{\mathrm{L}^2_t(\partial\mathcal{T})}$.
Ainsi, nous avons $\|\mathcal{U}\|_{\mathrm{L}^2_t(\partial\mathcal{T})} = 1$. □

Démonstration. (iii) En utilisant le point de vue de Cessenat-Després, le Problème 12 et l'inégalité de Cauchy-Schwarz, nous avons

$$\mathbf{k}(\mathbf{x},\mathbf{x}') = (\Pi_{\mathcal{U}}\mathbf{x},\mathcal{U}\mathbf{x}')_{\mathrm{L}^2_t(\partial\mathcal{T})} \leq \|\Pi_{\mathcal{U}}\mathbf{x}\|_{\mathrm{L}^2_t(\partial\mathcal{T})}\,\|\mathcal{U}\mathbf{x}'\|_{\mathrm{L}^2_t(\partial\mathcal{T})} \underset{\text{(i),(ii)}}{\leq} \|\mathbf{x}\|_{\mathrm{L}^2_t(\partial\mathcal{T})}\,\|\mathbf{x}'\|_{\mathrm{L}^2_t(\partial\mathcal{T})}.$$

Ainsi $\|\mathbf{k}\|_{\mathrm{L}^2_t(\partial\mathcal{T})}\leq 1$ et cela conclut la preuve. □

Remarque 3.2. *La propriété* (ii) *de la Proposition 3.1 peut être appliquée au Problème 12. Plus précisément, nous pouvons remplacer* $(\mathbf{x},\mathbf{x}')_{\mathrm{L}^2_t(\partial\mathcal{T})}$ *par* $(\mathcal{U}\mathbf{x},\mathcal{U}\mathbf{x}')_{\mathrm{L}^2_t(\partial\mathcal{T})}$ *dans le membre de gauche de l'équation* (3.7). *Nous obtenons ainsi une autre formulation UWVF hétérogène*

Trouver $\mathbf{x}\in\mathrm{L}^2_t(\partial\mathcal{T})$ *tel que pout tout* $\mathbf{x}'\in\mathrm{L}^2_t(\partial\mathcal{T})$, *nous avons*

$$(\mathcal{U}\mathbf{x} - \widehat{\mathcal{U}\mathbf{x}},\mathcal{U}\mathbf{x}')_{\mathrm{L}^2_t(\partial\mathcal{T})} = 0.$$

Cette formulation variationnelle traduit la continuité de la solution, ie $\mathcal{U}^T\mathbf{x}^T = (\widehat{\mathcal{U}^T\mathbf{x}})$.

La Proposition 3.1 n'assure pas que l'opérateur $\mathbf{k}$ est strictement contractant. Nous choisissons d'introduire un problème itératif de Trefftz seulement à l'échelle discrète, où il est possible d'obtenir le caractère strictement contractant de la matrice associée à $\mathbf{k}$.

Le Problème 12 est discrétisé sur l'espace de dimension finie $\mathbb{Y}^h_{\mathcal{T}} \subset \mathrm{L}^2_t(\partial\mathcal{T})$

$$\mathbb{Y}^h_{\mathcal{T}} := \prod_{T\in\mathcal{T}} \mathbb{Y}^h_T, \quad \text{avec } \mathbb{Y}^h_T := \gamma^T_{\mathrm{out}}\mathbb{X}^h_T.$$

Remarque 3.3. *L'espace discret global* $\mathbb{Y}^h_{\mathcal{T}}$ *de Cessenat-Després, où nous considérons des traces sortantes d'ondes planes, est de la même nature que l'espace discret global* $\mathbb{X}^h_{\mathcal{T}}$ *upwind ou de Riemann.*

Plus précisément, pour chaque $T \in \mathcal{T}$, l'espace discret local de dimension finie $\mathbb{Y}^h_T$ est composé des traces sortantes des ondes planes de $\mathbb{X}^h_T$, définies dans (2.5). Comme γ^T_{out} est un opérateur bijectif, l'espace $\mathbb{Y}^h_T$ est de dimension $N := \mathsf{dim}(\mathbb{Y}^h_T) = \mathsf{dim}(\mathbb{X}^h_T) = N_T$. Par conséquent, la base $\mathbb{Y}^h_T$ est donnée par

$$\mathbb{Y}^h_T := \mathrm{span}\Big(\{\mathbf{w}^\ell_{\mathrm{i_{elem}}} := \gamma^T_{\mathrm{out}}\mathbf{v}^\ell_T \text{ tel que } \mathbf{v}^\ell_T \in \mathbb{X}^h_T \text{ pour } \ell = 1, N \text{ et } \mathrm{i_{elem}} \text{ le n° de } T\}\Big).$$

L'espace $\mathbb{Y}^h_{\mathcal{T}}$ a la dimension $\#\mathrm{ddl} := N\,\#\mathrm{elem}$, avec $\#\mathrm{elem} := \mathsf{card}(\mathcal{T})$ le nombre total d'éléments dans le maillage. Chaque solution numérique $\mathbf{x}^h \in \mathbb{Y}^h_{\mathcal{T}}$ est représentée par un vecteur complexe $[\mathbf{x}^h]$ de dimension $\#\mathrm{ddl}$. Ainsi, la solution numérique $\mathbf{x}^h$ s'écrit

$$\begin{aligned} \mathbf{x}^h \;&=\; \sum_{\mathrm{i_{glob}}=1}^{\#\mathrm{ddl}} [\mathbf{x}^h]_{\mathrm{i_{glob}}}\mathbf{w}^{\mathrm{i_{glob}}} = \sum_{\mathrm{i_{elem}}=1}^{\#\mathrm{elem}}\sum_{\ell=1}^{N}[\mathbf{x}^h]^\ell_{\mathrm{i_{elem}}}\mathbf{w}^\ell_{\mathrm{i_{elem}}}, \\ \text{avec}\quad \mathrm{i_{glob}} &:= (\mathrm{i_{elem}}-1)\,\#\mathrm{elem} + \ell, \quad \text{et} \quad \begin{cases} [\mathbf{x}^h]_{\mathrm{i_{glob}}} := [\mathbf{x}^h]^\ell_{\mathrm{i_{elem}}}, \\ \mathbf{w}^{\mathrm{i_{glob}}} := \mathbf{w}^\ell_{\mathrm{i_{elem}}}. \end{cases} \end{aligned} \tag{3.14}$$

La formulation UWVF hétérogène discrète de Cessenat-Després est donnée.

Problème 14 (Problème UWVF discret)**.** *Trouver* $\mathbf{x}^h \in \mathbb{Y}^h_{\mathcal{T}}$ *tel que pour tout* $\mathbf{x}' \in \mathbb{Y}^h_{\mathcal{T}}$*, nous avons*

$$\mathbf{a}(\mathbf{x}^h, \mathbf{x}') = \mathbf{l}(\mathbf{x}') \quad \Longleftrightarrow \quad \mathbf{A}[\mathbf{x}] = \mathbf{F}, \tag{3.15}$$

avec $\mathbf{A} \in \mathbb{C}^{\#\mathrm{ddl}\times\#\mathrm{ddl}}$ *et* $\mathbf{F} \in \mathbb{C}^{\#\mathrm{ddl}}$ *définis composante par composante par*

$$\mathbf{A}_{\mathrm{i_{glob}},\mathrm{j_{glob}}} := \mathbf{a}(\mathbf{w}^{\mathrm{j_{glob}}}, \mathbf{w}^{\mathrm{i_{glob}}}) \text{ et } \mathbf{F}_{\mathrm{i_{glob}}} := \mathbf{l}(\mathbf{w}^{\mathrm{i_{glob}}}) \quad \text{pour } \mathrm{i_{glob}}, \mathrm{j_{glob}} = 1, \#\mathrm{ddl}.$$

Remarque 3.4. *Comme dans le Chapitre 2, la matrice globale* $\mathbf{A}$ *est composée des blocs matriciels* $\mathbf{A}_{\mathrm{i_{elem}},\mathrm{j_{elem}}}$ *pour* $\mathrm{i_{elem}}, \mathrm{j_{elem}} = 1, \#\mathrm{elem}$ *définis par* (2.53).

La Proposition 3.1 mène à une décomposition régulière-singulière de la matrice $\mathbf{A}$. Dans le cas homogène, cela correspond à une décomposition bloc-Jacobi [23]. Cela nous conduit, même dans notre configuration hétérogène, à considérer la possibilité d'appliquer un algorithme de Jacobi. Nous introduisons alors un problème itératif UWVF discret.

Problème 15 (Problème itératif UWVF discret)**.** *À partir de la récurrence suivante, déterminer la suite* $\mathbf{x}_n^{\mathrm{jac}} \in \mathbb{Y}_{\mathcal{T}}^h$, $n \in \mathbb{N}^*$, *avec* $\mathbf{x}_0^{\mathrm{jac}} = 0$,

$$(\mathbf{x}_{n+1}^{\mathrm{jac}}, \mathbf{x}')_{\mathrm{L}_t^2(\partial\mathcal{T})} - \mathbf{k}(\mathbf{x}_n^{\mathrm{jac}}, \mathbf{x}') = \mathbf{l}(\mathbf{x}'), \ \forall \mathbf{x}' \in \mathbb{Y}_{\mathcal{T}}^h,$$

ou de manière équivalente, en utilisant une décomposition régulière-singulière de la matrice $\mathbf{A} := \mathbf{M} - \mathbf{N}$ *: déterminer la suite* $[\mathbf{x}_n^{\mathrm{jac}}] \in \mathbb{C}^{\#\mathrm{ddl}}$, $n \in \mathbb{N}^*$, *avec* $[\mathbf{x}_0^{\mathrm{jac}}] = 0$

$$\mathbf{M}[\mathbf{x}_{n+1}^{\mathrm{jac}}] = \mathbf{N}[\mathbf{x}_n^{\mathrm{jac}}] + \mathbf{F}, \tag{3.16}$$

avec $\mathbf{M} \in \mathbb{C}^{\#\mathrm{ddl}\times\#\mathrm{ddl}}$ *et* $\mathbf{N} \in \mathbb{C}^{\#\mathrm{ddl}\times\#\mathrm{ddl}}$ *définies par*

$$\mathbf{M}_{\mathrm{i_{glob}},\mathrm{j_{glob}}} := (\mathbf{w}^{\mathrm{j_{glob}}}, \mathbf{w}^{\mathrm{i_{glob}}})_{\mathrm{L}_t^2(\partial\mathcal{T})} \quad \textit{et} \quad \mathbf{N}_{\mathrm{i_{glob}},\mathrm{j_{glob}}} := \mathbf{k}(\mathbf{w}^{\mathrm{j_{glob}}}, \mathbf{w}^{\mathrm{i_{glob}}}), \tag{3.17}$$

pour $\mathrm{i_{glob}}, \mathrm{j_{glob}} = 1, \#\mathrm{ddl}$.

Remarque 3.5. *Grâce aux propriétés des supports des fonctions de base, la matrice* $\mathbf{M}$ *prend la forme d'une matrice hermitienne diagonale par blocs, voir la Figure 3.2,*

$$\mathbf{M}^{\ell,k}_{\mathrm{i_{elem}},\mathrm{j_{elem}}} = (\mathbf{w}^k_{\mathrm{i_{elem}}}, \mathbf{w}^\ell_{\mathrm{i_{elem}}})_{\mathrm{L}_t^2(\partial\mathcal{T})}\, \delta_{\mathrm{i_{elem}},\mathrm{j_{elem}}} \quad \textit{pour } \mathrm{i_{elem}}, \mathrm{j_{elem}} = 1, \#\mathrm{elem} \textit{ et } \ell, k = 1, N,$$

ce qui permet une inversion rapide du système linéaire (3.16)*. La structure de la matrice* $\mathbf{N}$ *est aussi illustrée en Figure 3.2.*

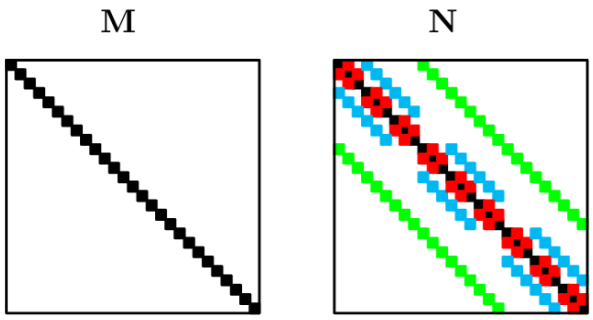

FIGURE 3.2 – Structures des matrices $\mathbf{M}$ et $\mathbf{N}$ de la décomposition de Cessenat-Després de la matrice $\mathbf{A}$.

Afin d'assurer la convergence du Problème itératif 15, la matrice $\mathbf{M}^{-1}\mathbf{N}$ associée à la forme $\mathbf{k}$ doit être strictement contractante. En d'autres termes, son rayon spectral, noté $\rho(\mathbf{M}^{-1}\mathbf{N})$, doit vérifier $\rho(\mathbf{M}^{-1}\mathbf{N}) < 1$. La discrétisation est naturellement conforme, *ie* $\mathbb{Y}^h_{\mathcal{T}} \subset \mathrm{L}^2_t(\partial\mathcal{T})$, et nous mène au résultat suivant

$$\rho(\mathbf{M}^{-1}\mathbf{N}) \leq |\!|\!|\mathbf{k}|\!|\!|_{\mathbb{Y}^h_{\mathcal{T}}} = \sup_{\mathbf{x}\in\mathbb{Y}^h_{\mathcal{T}}\setminus\{0\}} \sup_{\mathbf{x}'\in\mathbb{Y}^h_{\mathcal{T}}\setminus\{0\}} \frac{|\mathbf{k}(\mathbf{x},\mathbf{x}')|}{\|\mathbf{x}\|_{\mathbb{Y}^h_{\mathcal{T}}}\,\|\mathbf{x}'\|_{\mathbb{Y}^h_{\mathcal{T}}}} \leq |\!|\!|\mathbf{k}|\!|\!|_{\mathrm{L}^2_t(\partial\mathcal{T})} \leq 1,$$

avec $\|\mathbf{x}\|_{\mathbb{Y}^h_{\mathcal{T}}} = \|\mathbf{x}\|_{\mathrm{L}^2_t(\partial\mathcal{T})}$. Il nous reste donc à exclure toutes les valeurs propres présentes sur le cercle unité.

Proposition 3.2 (Matrice strictement contractante)**.** *La matrice* $\mathbf{M}^{-1}\mathbf{N}$ *est strictement contractante, ie* $\rho(\mathbf{M}^{-1}\mathbf{N}) < 1$.

Démonstration. Nous agissons par l'absurde. Nous supposons qu'il existe $\mathbf{x}^h \in \mathbb{Y}^h_{\mathcal{T}}$ et $\lambda \in \mathbb{C}$, tels que $\mathbf{x}^h \neq 0$ et $|\lambda| = 1$. Nous avons

$$\mathbf{k}(\mathbf{x}^h,\mathbf{x}') = \lambda\,(\mathbf{x}^h,\mathbf{x}')_{\mathrm{L}^2_t(\partial\mathcal{T})}, \quad \forall \mathbf{x}' \in \mathbb{Y}^h_{\mathcal{T}} \quad \Longleftrightarrow \quad \mathbf{N}[\mathbf{x}^h] = \lambda\,\mathbf{M}[\mathbf{x}^h],$$

avec $\mathbf{x}^h$ représenté par $[\mathbf{x}^h]$ à travers (3.14). Compte tenu de la définition de $\mathbf{k}$, en (3.10), cela conduit à

$$\begin{aligned}\|\Pi_{\mathcal{U}}\mathbf{x}^h - \lambda\mathcal{U}\mathbf{x}^h\|^2_{\mathrm{L}^2_t(\partial\mathcal{T})} &= \|\Pi_{\mathcal{U}}\mathbf{x}^h\|^2_{\mathrm{L}^2_t(\partial\mathcal{T})} - \lambda(\mathcal{U}\mathbf{x}^h,\Pi_{\mathcal{U}}\mathbf{x}^h)_{\mathrm{L}^2_t(\partial\mathcal{T})} \\ &\quad -\overline{\lambda}(\Pi_{\mathcal{U}}\mathbf{x}^h,\mathcal{U}\mathbf{x}^h)_{\mathrm{L}^2_t(\partial\mathcal{T})} + |\lambda|^2\|\mathcal{U}\mathbf{x}^h\|_{\mathrm{L}^2_t(\partial\mathcal{T})}.\end{aligned}$$

Grâce à la Proposition 3.1, nous obtenons

$$\|\Pi_{\mathcal{U}}\mathbf{x}^h - \lambda\mathcal{U}\mathbf{x}^h\|_{\mathrm{L}^2_t(\partial\mathcal{T})} \leq \|\mathbf{x}^h\|^2_{\mathrm{L}^2_t(\partial\mathcal{T})} - \lambda\overline{\lambda}\,\|\mathbf{x}^h\|^2_{\mathrm{L}^2_t(\partial\mathcal{T})} - \overline{\lambda}\lambda\,\|\mathbf{x}^h\|^2_{\mathrm{L}^2_t(\partial\mathcal{T})} + |\lambda|^2\,\|\mathbf{x}^h\|^2_{\mathrm{L}^2_t(\partial\mathcal{T})}.$$

Ainsi, nous avons

$$\|\Pi_{\mathcal{U}}\mathbf{x}^h - \lambda\mathcal{U}\mathbf{x}^h\|_{\mathrm{L}^2_t(\partial\mathcal{T})} \underset{|\lambda|=1}{\leq} 0 \quad \Longrightarrow \quad \Pi_{\mathcal{U}}\mathbf{x}^h = \lambda\,\mathcal{U}\mathbf{x}^h. \tag{3.18}$$

Nous montrons maintenant que $\mathbf{x}^h = 0$. Ceci nous permettra de conclure.
Nous introduisons l'espace

$$\widetilde{\mathcal{T}} := \left\{T \in \mathcal{T} \;\middle|\; \mathbf{x}^T \equiv 0\right\} \subset \mathcal{T}.$$

Il reste à prouver que $\widetilde{\mathcal{T}} = \mathcal{T}$ pour obtenir la Proposition 3.2.
(a) Nous montrons que $\widetilde{\mathcal{T}}$ est non-vide. D'après (3.18) et comme $\mathcal{U}$ est unitaire d'après la

Proposition 3.1, nous avons

$$\|\Pi_{\mathcal{U}}\mathbf{x}^h\|_{\mathrm{L}^2_t(\partial\mathcal{T})} = \|\mathcal{U}\mathbf{x}^h\|_{\mathrm{L}^2_t(\partial\mathcal{T})} = \|\mathbf{x}^h\|_{\mathrm{L}^2_t(\partial\mathcal{T})}.$$

D'après (3.13) qui peut se reformuler comme

$$\int_F Z_T^{-1}|\Pi_{\mathcal{U}}^T\mathbf{x}^T|^2 = \int_F Z_T^{-1}|\mathbf{x}^T|^2 - \int_F \frac{4Z_{\partial\Omega}}{(Z_{\partial\Omega}+Z_T)^2}|\mathbf{x}^T|^2,$$

il suit que

$$0 \;=\; \|\mathbf{x}^h\|^2_{\mathrm{L}^2_t(\partial\mathcal{T})} - \|\Pi_{\mathcal{U}}\mathbf{x}^h\|^2_{\mathrm{L}^2_t(\partial\mathcal{T})} \;=\; \sum_{T\in\mathcal{T}}\sum_{F\in\mathcal{F}_T\cap\mathcal{F}_{\mathrm{ext}}}\int_F \underbrace{\frac{4Z_{\partial\Omega}}{(Z_{\partial\Omega}+Z_T)^2}}_{>0}|\mathbf{x}^T|^2,$$

comme $Z_{\partial\Omega}$ est une fonction à partie réelle strictement positive, voir la page 1 et où nous avons écrit, pour simplifier les notations, $\mathbf{x}^T$ comme la restriction de $\mathbf{x}^h$ à T. Nous avons alors $\mathbf{x}^T = 0$ sur $F \in \mathcal{F}_T \cap \mathcal{F}_{\mathrm{ext}}$.
De plus, avec la Définition (3.11) de l'opérateur de trace restreint à un élément $\Pi_{\mathcal{U}}^T$ et en utilisant (3.18), nous avons aussi

$$\mathcal{U}^T\mathbf{x}^T = \frac{1}{\lambda}\frac{Z_{\partial\Omega}-Z_T}{Z_{\partial\Omega}+Z_T}\mathbf{x}^T = 0 \quad \text{sur } F \in \mathcal{F}_T \cap \mathcal{F}_{\mathrm{ext}}. \tag{3.19}$$

En combinant que $\mathbf{x}^T = 0$ sur $F \in \mathcal{F}_T \cap \mathcal{F}_{\mathrm{ext}}$ et (3.19), et en utilisant (3.4), nous obtenons par le théorème du prolongement unique

$$\gamma_t\mathbf{E}^T = \gamma_\times\mathbf{H}^T = 0 \text{ sur } F \quad\Longrightarrow\quad \mathbb{E}^T \equiv 0 \text{ et } \mathbf{x}^T \equiv 0 \text{ dans } T.$$

(*b*) Soit $T \in \mathcal{T}$ et $K \in \widetilde{\mathcal{T}}$ avec une face commune $F \in \mathcal{F}_T \cap \mathcal{F}_K$. Nous devons montrer que $T \in \widetilde{\mathcal{T}}$.
Comme $K \in \widetilde{\mathcal{T}}$, il suit que $\mathbf{x}^K = 0$ et $\mathcal{U}^K\mathbf{x}^K = 0$. D'après (3.18), nous avons sur F

$$\begin{cases} \Pi_{\mathcal{U}}^T\mathbf{x}^h &= \lambda\,\mathcal{U}^T\mathbf{x}^T, \\ \Pi_{\mathcal{U}}^K\mathbf{x}^h &= \lambda\,\mathcal{U}^K\mathbf{x}^K. \end{cases}$$

Ensuite, comme les traces numériques intérieures de Cessenat-Després (3.6a) définissent $(\Pi_{\mathcal{U}}^K)_{|F}$ et $(\Pi_{\mathcal{U}}^T)_{|F}$ d'après (3.11), nous obtenons

$$\frac{Z_T-Z_K}{Z_K+Z_T}\mathbf{x}^K + \frac{2Z_K}{Z_K+Z_T}\mathbf{x}^T = \lambda\,\mathcal{U}^K\mathbf{x}^K \quad\text{et}\quad \frac{Z_K-Z_T}{Z_K+Z_T}\mathbf{x}^T + \frac{2Z_T}{Z_K+Z_T}\mathbf{x}^K = \lambda\,\mathcal{U}^T\mathbf{x}^T,$$

menant à $\mathbf{x}^T = 0$ et $\mathcal{U}^T\mathbf{x}^T = 0$ sur F, comme $Z_T > 0$, $Z_K > 0$ et $\lambda \neq 0$. Ainsi, $T \in \widetilde{\mathcal{T}}$. D'après

la définition des traces en (3.4), nous avons aussi $\gamma_t \mathbf{E}^T = \gamma_\times \mathbf{H}^T = 0$ sur F, conduisant à $\mathbb{E}^T \equiv 0$ et $\mathbf{x}^T \equiv 0$ dans T d'après le théorème du prolongement unique.
(*c*) En utilisant le caractère complet et connexe du graphe de voisinage, nous concluons finalement que $\mathcal{T} = \widetilde{\mathcal{T}}$. □

Remarque 3.6. *Cette preuve assure l'unicité de la solution de la formulation variationnelle* (3.15). *Combinée avec le théorème du rang, ceci mène au caractère bien posé du Problème 15 de dimension finie.*

Remarque 3.7. *Un résultat plus faible a été prouvé dans le contexte d'une méthode itérative de sous-relaxation par Cessenat-Després [22]. Ce résultat prend la forme de* $\rho\left((1-\beta)\mathbf{I}_{\#\text{ddl}\times\#\text{ddl}} + \beta\,\mathbf{M}^{-1}\mathbf{N}\right) < 1$ *pour tout* $\beta \in]0,1[$. *En particulier, la Proposition 3.2 affine ce résultat en traitant le cas limite* $\beta = 1$. *Il sera ainsi par la suite inutile d'avoir recours à des algorithmes de relaxation pour faire converger les algorithmes de point fixe [22].*

3.1.3 Problèmes d'erreurs d'arrondis dans l'algorithme itératif de Cessenat-Després

Nous avons prouvé théoriquement le caractère strictement contractant de la matrice $\mathbf{M}^{-1}\mathbf{N}$. Cependant, des erreurs d'arrondis peuvent ralentir ou même faire diverger l'algorithme itératif en pratique. Comme montré dans le Tableau 3.1, la Proposition 3.2 n'est plus vraie numériquement. En effet, l'algorithme itératif diverge numériquement et aléatoirement selon les cas. En particulier, il diverge ici pour $\mathcal{D}_\Omega = 80\lambda$ (Tableau 3.1).

$\mathcal{D}_\Omega\,(\lambda)$	10	30	50	80	100
$\rho(\mathbf{M}^{-1}\mathbf{N})$	$1-0.0097$	$1-0.0033$	$1-0.0024$	$1+0.00064$	$1-0.00043$

TABLE 3.1 – Comparaisons du rayon spectral $\rho(\mathbf{M}^{-1}\mathbf{N})$ de la matrice $\mathbf{M}^{-1}\mathbf{N}$ grâce à la méthode de la puissance itérée, sur différentes tailles de domaine $\mathcal{D}_\Omega$.

Un autre aspect problématique d'une méthode de Jacobi est qu'elle peine à résoudre tous les problèmes. Nous étudions ce phénomène sur un exemple très simple. Nous considérons le cas où $\mathbf{A}$ est une matrice diagonale de dimension 1000. Cette matrice est de même taille qu'une matrice Trefftz pour un domaine de côté 2.7λ. Elle prend la forme

$$\mathbf{A} := \begin{pmatrix} \lambda_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \lambda_{1000} \end{pmatrix},$$

où $(\lambda_{\mathrm{i}})_{\mathrm{i}=1,1000}$ sont ses valeurs propres. Nous désignons par λ_{min} (*resp.* λ_{max}) la plus petite (*resp.* grande) valeur propre de $\mathbf{A}$. La décomposition

$$\mathbf{A} = \mathbf{M} - \mathbf{N} = \mathbf{M}(\mathbf{I}_{\#\mathrm{ddl}\times\#\mathrm{ddl}} - \mathbf{M}^{-1}\mathbf{N}),$$

peut s'écrire matriciellement comme

$$\begin{pmatrix} \lambda_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \lambda_{\#\mathrm{ddl}} \end{pmatrix} = \begin{pmatrix} \lambda_0 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \lambda_0 \end{pmatrix} \left(\mathbf{I}_{\#\mathrm{ddl}\times\#\mathrm{ddl}} + \begin{pmatrix} \frac{\lambda_1-\lambda_0}{\lambda_0} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \frac{\lambda_{\#\mathrm{ddl}}-\lambda_0}{\lambda_0} \end{pmatrix} \right),$$

où nous posons $\lambda_0 = \dfrac{\lambda_{min}+\lambda_{max}}{2}$, dans le but d'assurer $\rho(\mathbf{M^{-1}N}) < 1$. Nous choisissons comme solution de référence la solution numérique obtenue par l'inversion du problème $A[\mathbf{x}] = \mathbf{F}$, avec $\mathbf{F}$ un vecteur aléatoire. L'erreur entre $[\mathbf{x}]$ et le vecteur solution obtenu par une méthode de Jacobi à l'itération n, noté $[\mathbf{x}_n^{\mathrm{jac}}] \in \mathbb{C}^{1000}$, est

$$e_n^2 = \|[\mathbf{x}_n^{\mathrm{jac}}] - [\mathbf{x}]\|_2.$$

Nous souhaitons comparer cette erreur à celle obtenue avec une autre méthode itérative, par exemple la méthode de GMRES. En effet, le but de cette étude est à la fois de prouver les limites d'un algorithme de Jacobi mais aussi d'en trouver une alternative. De la même façon, l'erreur entre $[\mathbf{x}] \in \mathbb{C}^{1000}$ et la solution obtenue par une méthode de GMRES à l'itération n, représentée par $[\mathbf{x}_n] \in \mathbb{C}^{1000}$, est

$$e_n^2 = \|[\mathbf{x}_n] - [\mathbf{x}]\|_2.$$

Dans cette étude, nous observons l'influence des valeurs propres sur la convergence des solveurs de Jacobi et de GMRES. Dans un premier temps, nous considérons 1000 valeurs propres choisies entre 0.25 et 0.75. Elles sont calculées aléatoirement, grâce à γ_n un réel aléatoire,

$$\lambda_n = 0.25 + \gamma_n\, 0.5, \quad \text{pour } n = 1, 1000 \text{ et } \gamma_n \in [0, 1].$$

Comme illustré sur la Figure 3.3, les deux méthodes convergent vers la solution numérique de référence. Dans un second temps, nous remplaçons la valeur propre maximale par $\lambda_{max} = 100$. Dans ce cas, le message est clair : la méthode de Jacobi ne converge plus tandis que le solveur de GMRES est quasiment insensible à cette modification, voir la Figure 3.4. Ainsi, il paraît préférable de privilégier une méthode de GMRES plutôt qu'une méthode de Jacobi pour traiter un spectre plus étalé de matrices.

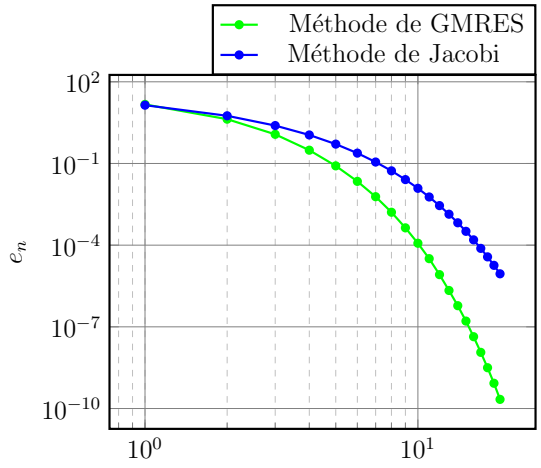

FIGURE 3.3 – Comparaison des erreurs obtenues avec la méthode de GMRES et la méthode de Jacobi en fonction du nombre d'itérations, pour $\lambda_{min} = 0.25$ et $\lambda_{max} = 0.75$.

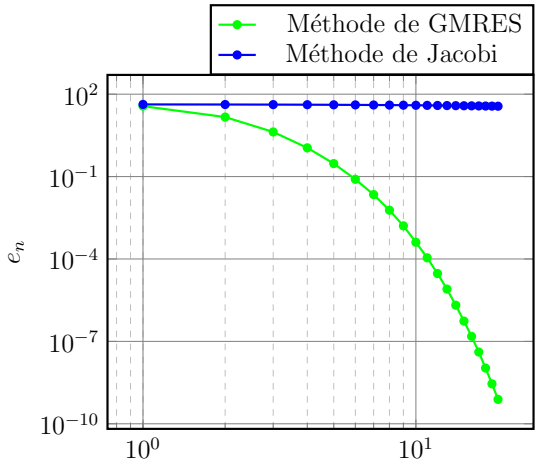

FIGURE 3.4 – Comparaison des erreurs obtenues avec la méthode de GMRES et la méthode de Jacobi en fonction du nombre d'itérations, pour $\lambda_{min} = 0.25$ et $\lambda_{max} = 100$.

3.2 Solveur GMRES basé sur le code du CERFACS©

Dans la section précédente, nous avons exhibé des cas pour lesquels les algorithmes itératifs classiques ne convergent pas. Cela nous pousse à utiliser la méthode de GMRES, dont nous rappelons la théorie de convergence dans cette section. L'algorithme de GMRES est connu pour sa stabilité. Cette propriété est particulièrement utile pour des scènes de calcul géométriquement complexes, hétérogènes ou encore remplies d'obstacles. De plus, son principe est, comme son nom l'indique, de mettre en jeu un algorithme de minimisation de la norme du résidu

$$r([\mathbf{x}]) = \frac{\|\mathbf{F} - \mathbf{A}[\mathbf{x}]\|_2}{\|\mathbf{F}\|_2},$$

sur un sous-espace de Krylov de dimension finie N_{kry} défini par :

$$[\mathbb{K}_{N_{\text{kry}}}] := \text{span}\Big(\{\mathbf{A}^k\mathbf{F},\ \text{pour } 0 \leq k \leq N_{\text{kry}} - 1\}\Big), \tag{3.20}$$

où N_{kry} est choisi petit face au nombre d'inconnues $\#\text{ddl}$. Lorsqu'elle est associée à une stratégie de *restart*, cette méthode a un coût mémoire bien moins important qu'une méthode de factorisation LU et est appropriée pour simuler des ondes sur de grands domaines. Nous introduisons la technique de *restart* dans cette section.

3.2.1 Théorie générale de convergence de la méthode de GMRES appliquée au problème UWVF

Une méthode de GMRES est un algorithme itératif ayant recours à des espaces de Krylov qui permettent de réduire la dimension de l'espace d'approximation [91, 92]. L'algorithme de GMRES a pour but de résoudre le problème de minimisation suivant.

Problème 16 (Problème de GMRES)**.** *Trouver la solution* $[\mathbf{x}_{N_{\text{kry}}}] \in [\mathbb{K}_{N_{\text{kry}}}]$ *qui minimise* r *défini par*

$$r([\mathbf{x}]) := \frac{\|\mathbf{F} - \mathbf{A}[\mathbf{x}]\|_2}{\|\mathbf{F}\|_2}, \tag{3.21}$$

avec l'espace de Krylov $[\mathbb{K}_{N_{\text{kry}}}]$ *défini par*

$$\begin{aligned} [\mathbb{K}_{N_{\text{kry}}}] &:= \text{span}\Big(\{\mathbf{A}^k\mathbf{F},\ \textit{pour}\ 0 \leq k \leq N_{\text{kry}} - 1\}\Big) \\ &= \text{span}\Big(\Big\{\mathbf{F},\ \mathbf{AF},\ \mathbf{A}^2\mathbf{F},\ \ldots,\ \mathbf{A}^{N_{\text{kry}}-1}\mathbf{F}\Big\}\Big). \end{aligned} \tag{3.22}$$

Remarque 3.8. *Le résidu GMRES, voir* (3.21)*, utilisé dans le Problème 16 est le résidu standard utilisé dans le code du CERFACS*© *[44].*

À chaque itération du solveur GMRES, le résidu défini par (3.21) est estimé et employé comme un critère d'arrêt pour l'algorithme. Le Problème de minimisation 16 s'écrit aussi comme la résolution d'une équation normale.

Problème 17 (Problème de GMRES matriciel)**.** *Trouver* $[\mathbf{x}_{N_{\text{kry}}}] \in [\mathbb{K}_{N_{\text{kry}}}]$ *tel que pour tout* $[\mathbf{x}'] \in [\mathbb{K}_{N_{\text{kry}}}]$

$$[\mathbf{x}']^*\mathbf{A}^*\mathbf{A}[\mathbf{x}_{N_{\text{kry}}}] = [\mathbf{x}']^*\mathbf{A}^*\mathbf{F}.$$

Cela définit une fonction $[\mathbf{x}_{N_{\text{kry}}}] = \mathsf{KRYLOV}(\mathbf{A}, \mathbf{F}, N_{\text{kry}})$. En écrivant le problème de minimisation sous la forme d'équations normales, nous mettons en avant les risques d'accumulation d'erreurs d'arrondis dues au produit $\mathbf{A}^*\mathbf{A}$ dont le nombre de conditionnement est beaucoup plus grand que celui de $\mathbf{A}$.

Dans la littérature, beaucoup de bornes de convergence existent pour majorer la norme du résidu GMRES [40, 70]. Mais ces bornes sont généralement pessimistes. En pratique, la méthode de GMRES a de bien meilleurs comportements. Bien que non optimal, nous rappelons le résultat de convergence classique des méthodes de GMRES [92].

Proposition 3.3. *[Résultat de convergence GMRES - [92]] Supposons que la matrice* $\mathbf{A}$ *est diagonalisable, ie* $\mathbf{A} = \mathbf{XDX}^{-1}$*, et posons*

$$\varepsilon^m = \min_{\substack{p \in \mathcal{P}^m \\ p(0)=1}} \max_{\lambda_i \in \rho(\mathbf{A})} |p(\lambda_i)|,$$

où $\rho(\mathbf{A})$ est le rayon spectral de $\mathbf{A}$, $\mathcal{P}^m$ l'ensemble des polynômes de degré inférieur ou égal à m. Alors la norme du résidu à l'étape m du GMRES satisfait

$$\|[r^{m+1}]\| \leq \varepsilon^m \|\mathbf{X}\|\|\mathbf{X}^{-1}\| \|[r^0]\| \quad \textit{avec} \quad \|[r^{m+1}]\| := \|[\mathbf{x}_{m+1}] - [\mathbf{x}_m]\|.$$

Remarque 3.9. *Il est théoriquement discutable que $\mathbf{A}$ soit diagonalisable, bien que nous ayons observé cette propriété en pratique.*

La convergence de la méthode dépend du spectre de la matrice $\mathbf{A}$. Nous prenons de nouveau l'exemple simple de la Sous-section 3.1.3. Nous plaçons artificiellement une valeur propre proche de 0 : $\lambda_{min} = 10^{-2}$. Cette fois-ci, la méthode de GMRES peine à converger. Un effet plateau persiste jusqu'à la sixième itération pour ce cas pourtant de taille très petite, voir la Figure 3.5. La méthode de Jacobi, quant à elle, ne converge toujours pas (en comparaison avec la Figure 3.4). Par cet exemple, nous illustrons que la méthode de GMRES peut avoir des difficultés à converger lors de la présence de petites valeurs propres. Nous

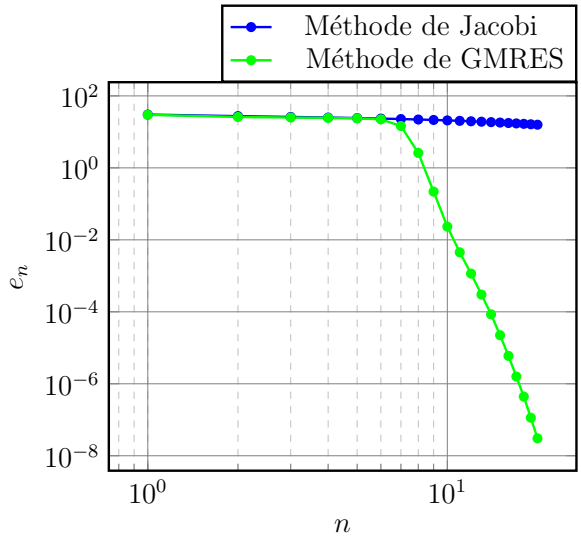

FIGURE 3.5 – Comparaison des erreurs obtenues avec la méthode de GMRES et la méthode de Jacobi en fonction du nombre d'itérations, pour $\lambda_{min} = 10^{-2}$ et $\lambda_{max} = 0.75$.

montrons en Figure 3.6 le spectre de $\mathbf{A}$ pour une taille de domaine $\mathcal{D}_\Omega = 6\lambda$. Il contient de nombreuses valeurs propres proches de 0, en rouge sur la Figure 3.6. Cela pénalise la vitesse de convergence de la solution numérique.

Nous avons pour but d'écarter le spectre de la matrice $\mathbf{A}$ de la valeur 0. C'est pourquoi nous introduirons un préconditionneur modifiant ce spectre dans la Sous-section 3.3.3.

De plus, la méthode de GMRES implique la matrice $\mathbf{A}^*\mathbf{A}$ dont la partie réelle est positive, sans pour autant se servir de la propriété de coercivité de la forme sesquilinéaire a. Dans le

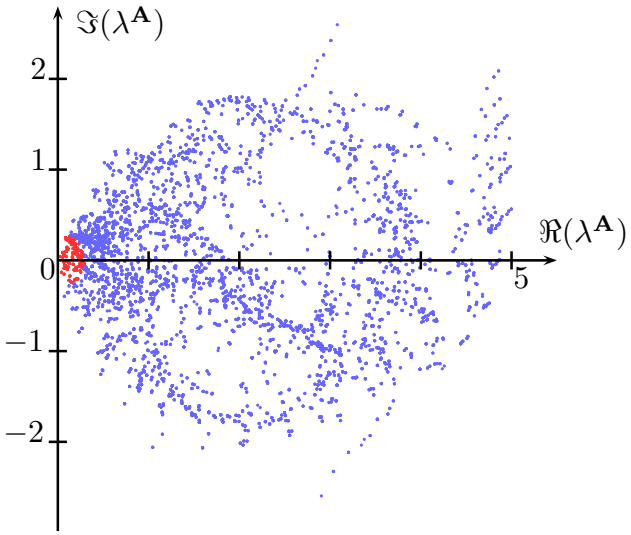

FIGURE 3.6 – Parties réelle et complexe, *resp.* $\Re(\lambda^{\mathbf{A}})$ et $\Im(\lambda^{\mathbf{A}})$, du spectre $\lambda^{\mathbf{A}}$ de la matrice $\mathbf{A}$ pour $\mathcal{D}_\Omega = 6\lambda$ et $N = 52$.

but d'exploiter cette particularité, *ie* $\Re\left([\mathbf{x}]^* \mathbf{A}[\mathbf{x}]\right) > 0$, nous choisissons de mettre en place une méthode de Krylov Galerkin dans la Section 3.3.

3.2.2 Stratégie de *restart*

Une méthode de GMRES converge en théorie en $\#\text{ddl}$ itérations. Cette convergence théorique induit un coût mémoire trop important pour stocker l'espace de Krylov en double précision. En effet, en utilisant 1To de mémoire, nous ne pouvons pas considérer des cas au delà de $\mathcal{D}_\Omega \approx 15\lambda$ (*ie* $\#\text{ddl} \approx 175500$), ce qui n'est même pas la taille maximale que peut traiter la méthode de Trefftz directe (factorisation LU), voir la Figure 2.16. C'est pourquoi nous utilisons en pratique un espace de Krylov de taille petite, où $N_{\text{kry}} << \#\text{ddl}$. Malheureusement, ce nombre devient trop petit pour espérer avoir une approximation précise de la solution numérique du système. Ainsi, il est nécessaire de mettre en oeuvre une stratégie de *restart* qui consiste à effectuer des résolutions GMRES successives. Nous choisissons η comme seuil de convergence. L'algorithme prend alors la forme suivante

function GMRES − RESTART ($\mathbf{A}$, $[\mathbf{x}_0^{\text{init}}]$)

$[\mathbf{r}_{N_{\text{kry}}}] = \mathbf{F} - \mathbf{A}[\mathbf{x}_0^{\text{init}}]$; $[\mathbf{x}_{N_{\text{kry}}}] = [\mathbf{x}_0^{\text{init}}]$;

do while $(r_{N_{\text{kry}}} > \eta)$

$[\boldsymbol{\delta}\mathbf{x}_{N_{\text{kry}}}] = \text{KRYLOV}(\mathbf{A}, [\mathbf{r}_{N_{\text{kry}}}], N_{\text{kry}})$

$[\mathbf{x}_{N_{\text{kry}}}] = [\mathbf{x}_{N_{\text{kry}}}] + [\boldsymbol{\delta}\mathbf{x}_{N_{\text{kry}}}]$

$r_{N_{\text{kry}}} = \dfrac{\|\mathbf{F} - \mathbf{A}[\mathbf{x}_{N_{\text{kry}}}]\|_2}{\|\mathbf{F}\|_2}$! norme du residu GMRES

```
    end do
end function
```

où $\boldsymbol{\delta}\mathbf{x}_{N_{\text{kry}}}$ est une approximation de l'erreur $\mathbf{x} - \mathbf{x}_{N_{\text{kry}}}$ commise par la solution numérique $\mathbf{x}_{N_{\text{kry}}}$ à l'itération N_{kry}.

Cette stratégie est implémentée dans l'algorithme GMRES du CERFACS© [44]. Ce dernier est optimisé et adapté pour traiter de grandes scènes de calcul. Dans ce solveur, la méthode d'Arnoldi, incluant un algorithme de Gram-Schmidt Modifié (GSM), est employée pour le calcul de la base de Krylov. Elle est explicitée dans [92] par exemple.

Remarque 3.10. *À partir de maintenant, nous ne précisons plus systématiquement qu'il s'agit d'une méthode "avec restart", bien que cette stratégie est toujours utilisée afin de considérer de grandes scènes de calcul.*

3.3 Solveur de Krylov Galerkin

Nous développons une nouvelle méthode de Krylov nommée : Krylov Galerkin (KG). Elle est similaire à celle de GMRES au sens où elle utilise aussi des espaces de Krylov. Plus précisément, il ne s'agit plus de minimiser un résidu, mais de résoudre le problème de Galerkin : Trouver $[\mathbf{x}_{N_{\text{kry}}}] \in [\mathbb{K}_{N_{\text{kry}}}]$ tel que

$$[\mathbf{x}']^* \mathbf{A} [\mathbf{x}_{N_{\text{kry}}}] = [\mathbf{x}']^* \mathbf{F}, \quad \text{pour tout } \mathbf{x}' \in [\mathbb{K}_{N_{\text{kry}}}].$$

Toutefois, nous ne bénéficions plus du caractère auto-adjoint de la matrice $\mathbf{A}^*\mathbf{A}$. Cela nous conduit alors à employer un algorithme de pseudo-orthogonalisation plutôt que d'orthogonalisation (comme dans le solveur GMRES). Cela a pour conséquence négative d'avoir une matrice de dimension réduite triangulaire et non plus diagonale.

Dans un premier temps, nous allons détailler la construction de la base de l'espace de Krylov $[\mathbb{K}_{N_{\text{kry}}}]$ en utilisant un algorithme de Gram-Schmidt. Celle-ci est adaptée à la résolution du problème et rend la matrice associée triangulaire inférieure. Dans un second temps, nous introduisons le problème de Krylov UWVF préconditionné. Grâce à la théorie de Galerkin, nous établissons un résultat de convergence. En particulier, nous prouverons que, si l'algorithme itératif de Cessenat-Després décrit par le Problème 15 converge, alors la solution obtenue par une méthode de KG appliquée au problème UWVF préconditionné converge elle aussi.

3.3.1 Construction des espaces de Krylov associés au problème UWVF de Galerkin

Dans cette section, nous proposons un nouvel algorithme de résolution itérative du Problème 14 UWVF qui respecte le cadre Galerkin du schéma numérique étudié.

Ce nouveau problème est nommé Problème de KG UWVF et est défini de la façon suivante.

Problème 18 (Problème de KG UWVF). *Pour $N_{\text{kry}} \in \mathbb{N}^*$, déterminer $\mathbf{x}_{N_{\text{kry}}} \in \mathbb{K}_{N_{\text{kry}}}$ solution de*

$$\textit{Trouver } \mathbf{x}_{N_{\text{kry}}} \in \mathbb{K}_{N_{\text{kry}}},\ \mathbf{a}(\mathbf{x}_{N_{\text{kry}}}, \mathbf{x}') = \mathbf{l}(\mathbf{x}'),\ \forall \mathbf{x}' \in \mathbb{K}_{N_{\text{kry}}},$$

ou de manière équivalente

$$\textit{Trouver } [\mathbf{x}_{N_{\text{kry}}}] \in [\mathbb{K}_{N_{\text{kry}}}],\ [\mathbf{x}']^* \mathbf{A} [\mathbf{x}_{N_{\text{kry}}}] = [\mathbf{x}']^* \mathbf{F},\ \forall [\mathbf{x}'] \in [\mathbb{K}_{N_{\text{kry}}}],$$

avec $\mathbb{K}_{N_{\text{kry}}}$ le sous-espace vectoriel de $\mathbb{Y}_{\mathcal{T}}^h$ défini par l'ensemble des fonctions $\mathbf{x}_{N_{\text{kry}}}$ représentées par $[\mathbf{x}_{N_{\text{kry}}}] \in [\mathbb{K}_{N_{\text{kry}}}]$ à travers la bijection définie par (3.14), où $[\mathbb{K}_{N_{\text{kry}}}]$ est l'espace de Krylov associé à $\mathbf{A}$ et à $\mathbf{F}$ défini par (3.22).

La résolution du Problème 18 nécessite l'utilisation d'une base de l'espace de Krylov $[\mathbb{K}_{N_{\text{kry}}}]$ défini par (3.22). Il est connu que les vecteurs générateurs $\mathbf{A}^k\mathbf{F}$ ne sont pas des bons choix. Ainsi, comme pour l'approche de GMRES, un processus d'orthogonalisation relativement à la matrice associée au problème doit être mis en place pour la construction d'une base adaptée. Néanmoins, contrairement au GMRES qui est basé sur un système hermitien, nous ne pouvons pas ici assurer une stricte orthogonalité des vecteurs. Effectivement, bien que $\mathbf{A}$ soit positive (au sens où $\Re(\mathbf{x}^*\mathbf{A}\mathbf{x}) > 0, \forall \mathbf{x} \neq 0$), elle est aussi non symétrique, et seule une pseudo-orthogonalité de la base peut être assurée.

Un vecteur $[\mathbf{y}] \in \mathbb{C}^{\#\text{ddl}}$ est pseudo-orthogonal à $[\mathbf{z}] \in \mathbb{C}^{\#\text{ddl}}$, relativement à la matrice $\mathbf{A}$, s'il vérifie

$$\mathbf{a}(\mathbf{y}, \mathbf{z}) = 0 \quad \Longrightarrow \quad [\mathbf{z}]^* \mathbf{A} [\mathbf{y}] = 0.$$

Remarque 3.11. *La spécificité de la pseudo-orthogonalité par rapport à l'orthogonalité est que $\mathbf{y}$ peut être orthogonal à $\mathbf{z}$ sans que $\mathbf{z}$ ne soit orthogonal à $\mathbf{y}$.*

L'idée est de construire une base $(\mathbf{z}_i)_{0 \leq i \leq N_{\text{kry}}-1} \in \mathbb{K}_{N_{\text{kry}}}$ qui vérifie

$$|\mathbf{a}(\mathbf{z}_j, \mathbf{z}_j)| = 1 \quad \text{et} \quad \mathbf{a}(\mathbf{z}_j, \mathbf{z}_i) = 0 \quad \text{pour } i < j. \tag{3.23}$$

Ainsi, si la seconde condition de (3.23) est vérifiée, nous avons aussi $[\mathbf{z}_i]^* \mathbf{A} [\mathbf{z}_j] = 0$ pour $i < j$. Autrement dit, la pseudo-orthogonalisation rend la matrice associée au Problème 18

triangulaire inférieure. Ce point n'est pas fondamental comme N_{kry} est choisi petit dans notre cas. La réelle utilité de ce procédé de pseudo-orthogonalisation est de lutter contre les erreurs d'arrondis en reconditionnant la base de Krylov constituée des vecteurs $\mathbf{A}^k\,\mathbf{F}$. De plus, le fait que la matrice $\mathbf{A}$ soit de grande dimension rend la construction de la base de l'espace de Krylov $[\mathbb{K}_{N_{\text{kry}}}]$ compliquée compte tenu du nombre de produits matriciels à effectuer. Nous expliquons maintenant sa construction.

L'espace de Krylov utilisé pour la résolution du Problème 18 peut s'écrire

$$[\mathbb{K}_{N_{\text{kry}}}] = \text{span}\Big(\{[\mathbf{z}_k],\ \text{pour } 0 \leq k \leq N_{\text{kry}} - 1\}\Big).$$

La base $\{[\mathbf{z}_k],\ \text{pour } 0 \leq k \leq N_{\text{kry}} - 1\}$ est construite à partir d'un processus itératif. En particulier, le $(k+1)^{ème}$ vecteur, pseudo-orthogonal à tous les vecteurs $[\mathbf{z}_{\text{i}}] \in \mathbb{C}^{\#\text{ddl}}$, pour $\text{i} = 0, k$, est de la forme

$$[\mathbf{z}_{k+1}] = \beta_{k+1}\big(\mathbf{A}[\mathbf{z}_k] - \sum_{i=0}^{k} \alpha_{k+1}^{\text{i}}[\mathbf{z}_{\text{i}}]\big),$$

avec $\beta_{k+1} \in \mathbb{R}^+$ et $\alpha_{k+1}^{\text{i}} \in \mathbb{C}$. Il est nécessaire de déterminer les valeurs des coefficients β_{k+1} et α_{k+1}^{i} pour tout $0 \leq \text{i} \leq k$. Tout d'abord, nous cherchons à calculer les α_{k+1}^{i}. D'après (3.23), nous avons que

$$[\mathbf{z}_\ell]^*\mathbf{A}\big(\mathbf{A}[\mathbf{z}_k] - \sum_{\text{i}=0}^{k} \alpha_{k+1}^{\text{i}}[\mathbf{z}_{\text{i}}]\big) = 0, \quad \text{pour } \ell \leq k.$$

De plus, comme $\mathbf{a}(\mathbf{z}_i, \mathbf{z}_\ell) = 0$ pour $\ell < i$, nous obtenons

$$[\mathbf{z}_\ell]^*\mathbf{A}\big(\mathbf{A}[\mathbf{z}_k] - \sum_{\text{i}=0}^{\ell} \alpha_{k+1}^{\text{i}}[\mathbf{z}_{\text{i}}]\big) = 0, \quad \text{pour } \ell \leq k. \tag{3.24}$$

Nous posons $[\mathbf{z}_{k+1}^{\ell}] := \mathbf{A}[\mathbf{z}_k] - \sum_{\text{i}=0}^{\ell-1} \alpha_{k+1}^{\text{i}}[\mathbf{z}_{\text{i}}]$. En pratique, cette suite sera calculée à l'aide de la récurrence qui évitera de recalculer la somme à chaque itération

$$\left\{\begin{array}{rcl} [\mathbf{z}_{k+1}^{0}] & = & \mathbf{A}[\mathbf{z}_k], \\ [\mathbf{z}_{k+1}^{\ell+1}] & = & [\mathbf{z}_{k+1}^{\ell}] - \alpha_{k+1}^{\ell}[\mathbf{z}_\ell]. \end{array}\right.$$

Ainsi, l'équation (3.24) devient

$$[\mathbf{z}_\ell]^*\mathbf{A}[\mathbf{z}_{k+1}^{\ell}] - \alpha_{k+1}^{\ell}[\mathbf{z}_\ell]^*\mathbf{A}[\mathbf{z}_\ell] = 0, \quad \text{pour } \ell \leq k.$$

Finalement, nous obtenons la valeur du coefficient α_{k+1}^{ℓ}

$$\alpha_{k+1}^{\ell} = \frac{[\mathbf{z}_\ell]^*\mathbf{A}[\mathbf{z}_{k+1}^{\ell}]}{[\mathbf{z}_\ell]^*\mathbf{A}[\mathbf{z}_\ell]}, \quad \text{pour } \ell \leq k.$$

Le deuxième coefficient β_{k+1} permet de normaliser le vecteur obtenu et a pour valeur

$$\beta_{k+1} = \frac{1}{\sqrt{\left|[\mathbf{z}_{k+1}^{k+1}]^* \mathbf{A} [\mathbf{z}_{k+1}^{k+1}]\right|}}.$$

La solution $\mathbf{x}_{N_{\text{kry}}} \in \mathbb{K}_{N_{\text{kry}}}$ du problème de KG est représentée par le vecteur $[\mathbf{x}_{N_{\text{kry}}}] \in \mathbb{C}^{\#\text{ddl}}$, qui est exprimé dans la base des $[\mathbf{z}_k] \in [\mathbb{K}_{N_{\text{kry}}}]$ comme

$$[\mathbf{x}_{N_{\text{kry}}}] = \sum_{k=0}^{N_{\text{kry}}-1} \mathbf{u}_{N_{\text{kry}}}^k [\mathbf{z}_k].$$

Elle est solution du système triangulaire suivant

$$\mathbf{L}\, \mathbf{u}_{N_{\text{kry}}} = \mathbf{B}, \quad \text{avec} \quad \mathbf{L}_{\text{i,j}} := \mathbf{a}(\mathbf{z}_{\text{j}}, \mathbf{z}_{\text{i}}) \text{ et } \mathbf{B}_{\text{i}} := \ell(\mathbf{z}_{\text{i}}), \tag{3.26}$$

ou de manière équivalente

$$\mathbf{L}_{\text{i,j}} := [\mathbf{z}_{\text{i}}]^* \mathbf{A} [\mathbf{z}_{\text{j}}] \quad \text{et} \quad \mathbf{B}_{\text{i}} := [\mathbf{z}_{\text{i}}]^* \mathbf{F},$$

avec $\mathbf{u}_{N_{\text{kry}}} = [\mathbf{u}_{N_{\text{kry}}}^0, \mathbf{u}_{N_{\text{kry}}}^1, \cdots, \mathbf{u}_{N_{\text{kry}}}^{N_{\text{kry}}-1}]^{\mathsf{T}}$.

Remarque 3.12. *Dans le système* (3.26)*, nous avons* $\mathbf{L} \in \mathbb{C}^{N_{\text{kry}} \times N_{\text{kry}}}$ *et* $\mathbf{B} \in \mathbb{C}^{N_{\text{kry}}}$*. Ainsi, nous réduisons considérablement le coût mémoire de la méthode comme* N_{kry} *est en pratique choisi petit (voir la stratégie de restart Section 3.2.2).*

D'une part, la pseudo-orthonormalisation se réalise grâce à une méthode de Gram-Schmidt par exemple. D'autre part, la résolution du problème matriciel (3.26) s'effectue rapidement grâce à un algorithme de Descente. L'algorithme de KG pour le Problème 18 prend alors la forme

```
function KrylovGalerkin(A, F) result ([u_{N_kry}])
    [z_0] = F
    do k = 0, N_kry − 2
        ! algorithme de pseudo-orthonormalisation
        [z_{k+1}] = GramSchmidt(A, [z_0], ..., [z_k])
        do ℓ = 0, k + 1 ! construction matrice
            L_{k+1,ℓ} = [z_{k+1}]^* A [z_ℓ]
        end do
    end do
    do i = 0, N_kry − 1
        B_i = [z_i]^* F_i
```

```
    end do
    [u_{N_kry}] = Descente(L, B)
end function
```

En pratique, le procédé de pseudo-orthonormalisation peut se faire de différentes façons. En premier, nous donnons l'exemple d'un algorithme de Gram-Schmidt classique qui prend la forme suivante

```
function GramSchmidt(A, [z_0], ..., [z_k]) result([z_{k+1}])
    α = 0
    [z̃^0_{k+1}] = A[z_k]
    ! pseudo-orthogonalisation avec les autres
    ! vecteurs de la base
    do i=0,k
        α^i_{k+1} = ([z_i]* A [z̃^i_{k+1}]) / ([z_i]* A [z_i])
        [z̃^{i+1}_{k+1}] = [z̃^i_{k+1}] − α^i_{k+1} [z_i]
    end do
    ! pseudo-normalisation
    [z_{k+1}] = [z̃^{k+1}_{k+1}] / sqrt(|[z̃^{k+1}_{k+1}]* A [z̃^{k+1}_{k+1}]|)
end function
```

Ce dernier peut se simplifier en évitant le stockage des informations :

```
function GramSchmidt(A, [z_0], ..., [z_k]) result([z_{k+1}])
    α = 0
    [z_{k+1}] = A[z_k]
    ! pseudo-orthogonalisation avec les autres
    ! vecteurs de la base
    do i=0,k
        α = ([z_i]* A [z_{k+1}]) / ([z_i]* A [z_i])
        [z_{k+1}] = [z_{k+1}] − α [z_i]
    end do
    ! pseudo-normalisation
    [z_{k+1}] = [z_{k+1}] / sqrt(|[z_{k+1}]* A [z_{k+1}]|)
end function
```

Dans ce processus de pseudo-orthonormalisation, les produits par $\mathbf{A}$ peuvent devenir très coûteux en termes de temps de calcul lorsque $\#\mathrm{ddl}$ devient grand. Il faut donc éviter le plus possible de calculer $\mathbf{A}[\mathbf{z}_k] \in \mathbb{C}^{\#\mathrm{ddl}}$. Il est possible de construire d'une autre façon la base de Krylov, en préférant un stockage restreint des informations plutôt que leur calcul à chaque itération. L'idée est donc de stocker en mémoire la matrice triangulaire inférieure $\mathbf{L}$. De manière équivalente, nous stockons les PMV $\mathbf{A}[\mathbf{z}_k]$. Ainsi, un seul PMV $\mathbf{A}[\mathbf{z}_{k+1}]$ est réalisé à chaque itération et nous utilisons les données précédemment calculées pour réaliser la pseudo-orthonormalisation. La fonction associée prend la forme suivante

```
function GramSchmidt1Prod(A, [z_0], ..., [z_k], [z^A]) result([z_{k+1}], [z^A_{k+1}])
    α. = 0
    [z^AA] = A[z^A] ! PMV de l'algorithme
    do i = 1, k ! coefficients pour la pseudo-orthogonalisation
        α_i = [z_i] · [z^AA]
        do ℓ = 1, i − 1
            α_i = α_i − α_ℓ L_{i,ℓ}
        end do
        α_i = α_i / L_{i,i}
    end do
    [z_{k+1}] = [z^A_k] ! nouveaux vecteurs de la base
    [z^A_{k+1}] = [z^AA]
    do i = 1, k ! pseudo-orthogonalisation des vecteurs
        [z_{k+1}] = [z_{k+1}] − α_i [z_i]
        [z^A_{k+1}] = [z^A_{k+1}] − α_i [z^A_i]
    end do
    ! coefficient de normalisation
    β_{k+1} = 1 / sqrt(|[z_{k+1}] · [z^A_{k+1}]|)
    ! normalisation des nouveaux vecteurs de la base
    [z_{k+1}] = β_{k+1} [z_{k+1}]
    [z^A_{k+1}] = β_{k+1} [z^A_{k+1}]
end function
```

Remarque 3.13. *Bien que le stockage des informations engendre davantage d'erreurs d'arrondis au fur et à mesure des itérations, cet algorithme diminue les temps de calcul en pratique.*

Remarque 3.14. *L'algorithme de pseudo-orthonormalisation à privilégier (entre Gram-Schmidt classique ou modifié par exemple) dépend finalement de la valeur de N_{kry}. Pour les cas les plus petits, l'algorithme de Gram-Schmidt classique reste un bon compromis.*

La méthode de KG a mis en valeur l'intérêt de construire $[\mathbb{K}_{N_{\text{kry}}}]$. Elle permet aussi de lutter contre les erreurs d'arrondis, comme elle ne requiert pas l'inversion d'une matrice à chaque itération contrairement à la méthode itérative de Jacobi par exemple. De plus, la dimension réduite N_{kry} permet de diminuer le coût mémoire de la méthode.

Dans la partie suivante, nous cherchons à montrer que cette méthode de KG UWVF est bien posée et convergente.

3.3.2 Méthode de Krylov UWVF bien posée et convergente

Le Problème de KG UWVF 18 est bien posé comme il est une approximation conforme du Problème UWVF 12, lui-même équivalent au Problème upwind 9, dont la forme sesquilinéaire (Problème 10) est coercive (Proposition 2.2). De plus, la convergence de la solution du Problème 18 repose sur la théorie de Galerkin établie pour le Problème UWVF 12.

Proposition 3.4 (Convergence méthode de KG UWVF)**.** *Soient $\mathbf{x}^h \in \mathbb{Y}^h_{\mathfrak{T}}$ et $\mathbf{x}_{N_{\text{kry}}} \in \mathbb{K}_{N_{\text{kry}}}$, des solutions respectives des Problèmes 12 et 18, où $\mathbb{K}_{N_{\text{kry}}}$ est le sous-espace vectoriel de $\mathbb{Y}^h_{\mathfrak{T}}$ défini par : $\mathbf{x}_{N_{\text{kry}}} \in \mathbb{K}_{N_{\text{kry}}}$ représenté par $[\mathbf{x}_{N_{\text{kry}}}] \in [\mathbb{K}_{N_{\text{kry}}}]$ à travers la bijection (3.14), où $[\mathbb{K}_{N_{\text{kry}}}]$ est l'espace de Krylov défini par (3.22). La convergence de la méthode de KG UWVF est assurée par*

$$\|\mathbb{E}^h - \mathbb{E}_{N_{\text{kry}}}\|_{\text{GD}} \leq \sqrt{2} \min_{\mathbf{y} \in \mathbb{K}_{N_{\text{kry}}}} \|\mathbf{x}^h - \mathbf{y}\|_{\text{L}^2_t(\partial\mathfrak{T})},$$

où $\mathbb{E}_{N_{\text{kry}}} = \mathcal{S}_{\text{out}} \mathbf{x}_{N_{\text{kry}}}$ et $\mathbb{E}^h = \mathcal{S}_{\text{out}} \mathbf{x}^h$, voir la définition de cette bijection en (3.3) et dans la Figure 3.1.

Démonstration. Nous posons $\mathcal{E} := \|\mathbb{E}^h - \mathbb{E}_{N_{\text{kry}}}\|^2_{\text{GD}}$. Nous rappelons que

$$\mathcal{E} = \Re(\mathbf{a}(\mathbf{x}^h - \mathbf{x}_{N_{\text{kry}}}, \mathbf{x}^h - \mathbf{x}_{N_{\text{kry}}})).$$

Comme $\mathbf{a}(\mathbf{x}^h - \mathbf{x}_{N_{\text{kry}}}, \mathbf{x}_{N_{\text{kry}}} - \mathbf{y}) = 0$ pour tout $\mathbf{y} \in \mathbb{K}_{N_{\text{kry}}}$, nous avons

$$\mathcal{E} = \Re(\mathbf{a}(\mathbf{x}^h - \mathbf{x}_{N_{\text{kry}}}, \mathbf{x}^h - \mathbf{y})) = \Re([\mathbf{x}^h - \mathbf{y}]^* \mathbf{A} [\mathbf{x}^h - \mathbf{x}_{N_{\text{kry}}}]).$$

Comme $\mathbf{M}$ est symétrique définie positive, nous avons

$$\begin{aligned} [\mathbf{x}^h - \mathbf{y}]^* \mathbf{A} [\mathbf{x}^h - \mathbf{x}_{N_{\text{kry}}}] &= [\mathbf{x}^h - \mathbf{y}]^* \mathbf{M}^{\frac{1}{2}} \mathbf{M}^{-\frac{1}{2}} \mathbf{A} [\mathbf{x}^h - \mathbf{x}_{N_{\text{kry}}}] \\ &= [\mathbf{M}^{\frac{1}{2}} [\mathbf{x}^h - \mathbf{y}]]^* \mathbf{M}^{-\frac{1}{2}} \mathbf{A} [\mathbf{x}^h - \mathbf{x}_{N_{\text{kry}}}]. \end{aligned}$$

Nous appliquons ensuite l'inégalité de Cauchy-Schwarz, et nous obtenons

$$\begin{cases} \mathcal{E} \leq \sqrt{[\mathbf{M}^{\frac{1}{2}} [\mathbf{x}^h - \mathbf{y}]]^* \mathbf{M}^{\frac{1}{2}} [\mathbf{x}^h - \mathbf{y}]} \sqrt{[\mathbf{M}^{-\frac{1}{2}} \mathbf{A} [\mathbf{x}^h - \mathbf{x}_{N_{\text{kry}}}]]^* \mathbf{M}^{-\frac{1}{2}} \mathbf{A} [\mathbf{x}^h - \mathbf{x}_{N_{\text{kry}}}]}, \\ \leq \sqrt{[\mathbf{x}^h - \mathbf{y}]^* \mathbf{M} [\mathbf{x}^h - \mathbf{y}]} \sqrt{[\mathbf{A} [\mathbf{x}^h - \mathbf{x}_{N_{\text{kry}}}]]^* \mathbf{M}^{-1} \mathbf{A} [\mathbf{x}^h - \mathbf{x}_{N_{\text{kry}}}]}. \end{cases}$$

Nous montrons maintenant que, pour tout $[\mathbf{x}] \in \mathbb{C}^{\#\mathrm{ddl}}$, nous avons

$$[\mathbf{A}[\mathbf{x}]]^* \mathbf{M}^{-1} [\mathbf{A}[\mathbf{x}]] \;\leq\; 2\,\Re([\mathbf{x}]^* \mathbf{A}[\mathbf{x}]).$$

Nous utilisons la décomposition de Cessenat-Després $\mathbf{A} = \mathbf{M} - \mathbf{N}$. Nous obtenons

$$[\mathbf{A}[\mathbf{x}]]^* \mathbf{M}^{-1} [\mathbf{A}[\mathbf{x}]] = [\mathbf{x}]^* \mathbf{M}[\mathbf{x}] - [\mathbf{x}]^* \mathbf{N}[\mathbf{x}] - [\mathbf{x}]^* \mathbf{N}^*[\mathbf{x}] + [\mathbf{x}]^* \mathbf{N}^* \mathbf{M}^{-1} \mathbf{N}[\mathbf{x}].$$

Soit $[\mathbf{z}] \in \mathbb{C}^{\#\mathrm{ddl}}$, ayant pour valeur $[\mathbf{z}] = \mathbf{M}^{-1}\mathbf{N}[\mathbf{x}]$. Il suit que

$$[\mathbf{x}]^* \mathbf{N}^* \mathbf{M}^{-1} \mathbf{N}[\mathbf{x}] = [\mathbf{z}]^* \mathbf{N}[\mathbf{x}] = \mathbf{k}(\mathbf{x}, \mathbf{z}).$$

En utilisant (iii) de la Proposition 3.1, nous avons

$$[\mathbf{x}]^* \mathbf{N}^* \mathbf{M}^{-1} \mathbf{N}[\mathbf{x}] \;\leq\; \|\mathbf{x}\|_{\mathrm{L}^2_t(\partial\mathcal{T})} \|\mathbf{z}\|_{\mathrm{L}^2_t(\partial\mathcal{T})} = \sqrt{[\mathbf{z}]^* \mathbf{M}[\mathbf{z}]} \sqrt{[\mathbf{x}]^* \mathbf{M}[\mathbf{x}]}.$$

Cela conduit à

$$\sqrt{[\mathbf{z}]^* \mathbf{M}[\mathbf{z}]} \sqrt{[\mathbf{x}]^* \mathbf{M}[\mathbf{x}]} \;\leq\; \sqrt{[\mathbf{x}]^* \mathbf{N}^* \mathbf{M}^{-1} \mathbf{N}[\mathbf{x}]} \sqrt{[\mathbf{x}]^* \mathbf{M}[\mathbf{x}]}.$$

Ainsi, nous obtenons $[\mathbf{x}]^* \mathbf{N}^* \mathbf{M}^{-1} \mathbf{N}[\mathbf{x}] \;\leq [\mathbf{x}]^* \mathbf{M}[\mathbf{x}]$, menant à

$$[\mathbf{A}[\mathbf{x}]]^* \mathbf{M}^{-1} [\mathbf{A}[\mathbf{x}]] \;\leq\; 2\,[\mathbf{x}]^* \mathbf{M}[\mathbf{x}] - 2\,\Re([\mathbf{x}]^* \mathbf{N}[\mathbf{x}]) = 2\,\Re([\mathbf{x}]^* \mathbf{A}[\mathbf{x}]) = 2\,\|\mathbb{E}\|^2_{\mathrm{GD}}.$$

où $\mathbb{E} = \mathcal{S}_{\mathrm{out}}\mathbf{x}$. Le résultat suit de $\|\mathbb{E}^h - \mathbb{E}_{N_{\mathrm{kry}}}\|_{\mathrm{GD}} \;\leq\; \sqrt{2}\,\|\mathbf{x}^h - \mathbf{y}\|_{\mathrm{L}^2_t(\partial\mathcal{T})}$. □

3.3.3 Préconditionneur de Cessenat-Després

Il est crucial d'utiliser un préconditionneur afin d'améliorer la convergence des méthodes itératives de type Krylov que nous développons dans ce chapitre. En particulier, nous portons notre attention sur le préconditionneur de Cessenat-Després basé sur la matrice $\mathbf{M}$. Nous proposons ici une version adaptée à l'approche de KG UWVF et nous en étudions la convergence.

Problème 19 (Problème de KG UWVF préconditionné)**.** *Pour $N_{\mathrm{kry}} \in \mathbb{N}^*$, trouver $\mathbf{x}^{\mathrm{prec}}_{N_{\mathrm{kry}}} \in \mathbb{K}^{\mathrm{prec}}_{N_{\mathrm{kry}}}$, ou de manière équivalente $[\mathbf{x}^{\mathrm{prec}}_{N_{\mathrm{kry}}}] \in [\mathbb{K}^{\mathrm{prec}}_{N_{\mathrm{kry}}}]$, tel que nous avons*

$$\mathbf{a}(\mathbf{x}^{\mathrm{prec}}_{N_{\mathrm{kry}}}, \mathbf{x}') = \mathbf{l}(\mathbf{x}'),\ \forall \mathbf{x}' \in \mathbb{K}^{\mathrm{prec}}_{N_{\mathrm{kry}}} \iff [\mathbf{x}']^* \mathbf{A}[\mathbf{x}^{\mathrm{prec}}_{N_{\mathrm{kry}}}] = [\mathbf{x}']^* \mathbf{F},\ \forall [\mathbf{x}'] \in [\mathbb{K}^{\mathrm{prec}}_{N_{\mathrm{kry}}}],$$

où l'espace de dimension finie $\mathbb{K}^{\mathrm{prec}}_{N_{\mathrm{kry}}}$ est l'image de $[\mathbb{K}^{\mathrm{prec}}_{N_{\mathrm{kry}}}]$ à travers (3.14). L'espace $[\mathbb{K}^{\mathrm{prec}}_{N_{\mathrm{kry}}}] \subset$

$\mathbb{C}^{\#\text{ddl}}$ *est défini par*

$$[\mathbb{K}^{\text{prec}}_{N_{\text{kry}}}] := \mathbf{M}^{-\frac{1}{2}}\{\mathbb{K}^{\text{prec}}_{N_{\text{kry}}}\},$$

avec $\{\mathbb{K}^{\text{prec}}_{N_{\text{kry}}}\}$ *un sous-espace vectoriel de* $\mathbb{C}^{\#\text{ddl}}$

$$\begin{aligned}\{\mathbb{K}^{\text{prec}}_{N_{\text{kry}}}\} &:= \text{span}\Big(\{\widetilde{\mathbf{A}}^k\widetilde{\mathbf{F}},\ \textit{pour}\ 0 \le k \le N_{\text{kry}}-1\}\Big)\\ &= \text{span}\Big(\{\widetilde{\mathbf{F}}, \widetilde{\mathbf{A}}\widetilde{\mathbf{F}}, \ldots, \widetilde{\mathbf{A}}^{N_{\text{kry}}-1}\widetilde{\mathbf{F}}\}\Big),\end{aligned}$$

où

$$\widetilde{\mathbf{A}} = \mathbf{M}^{-\frac{1}{2}}\mathbf{A}\mathbf{M}^{-\frac{1}{2}} \quad \textit{et} \quad \widetilde{\mathbf{F}} = \mathbf{M}^{-\frac{1}{2}}\mathbf{F},$$

menant à

$$[\mathbb{K}^{\text{prec}}_{N_{\text{kry}}}] = \text{span}\Big(\{\mathbf{M}^{-1}\mathbf{F}, \ldots, (\mathbf{M}^{-1}\mathbf{N})^{N_{\text{kry}}-1}\mathbf{M}^{-1}\mathbf{F}\}\Big).$$

Remarque 3.15. *Nous remarquons que la matrice* $\widetilde{\mathbf{A}}$ *devient*

$$\widetilde{\mathbf{A}} = \mathbf{I}_{\#\text{ddl}\times\#\text{ddl}} - \widetilde{\mathbf{N}} \quad \textit{avec} \quad \widetilde{\mathbf{N}} := \mathbf{M}^{-\frac{1}{2}}\mathbf{N}\mathbf{M}^{-\frac{1}{2}},$$

car $\mathbf{M}^{-\frac{1}{2}}\mathbf{M}\mathbf{M}^{-\frac{1}{2}} = \mathbf{I}_{\#\text{ddl}\times\#\text{ddl}}$ *comme* $\mathbf{M}$ *est la matrice représentant le produit scalaire symétrique dans* $\mathbb{Y}^h_{\mathcal{T}}$*. Cela engendre un gain en mémoire et en temps de calcul considérable, en particulier pour les grands domaines, ie* $\mathcal{D}_\Omega > 100\lambda$*, comme l'algorithme itératif ne nécessite plus l'inversion de* $\mathbf{M}$*. De plus, nous avons alors*

$$\{\mathbb{K}^{\text{prec}}_{N_{\text{kry}}}\} = \text{span}\Big(\{\widetilde{\mathbf{F}}, \widetilde{\mathbf{A}}\widetilde{\mathbf{F}}, \ldots, \widetilde{\mathbf{A}}^{N_{\text{kry}}-1}\widetilde{\mathbf{F}}\}\Big) = \text{span}\Big(\{\widetilde{\mathbf{F}}, \widetilde{\mathbf{N}}\widetilde{\mathbf{F}}, \ldots, \widetilde{\mathbf{N}}^{N_{\text{kry}}-1}\widetilde{\mathbf{F}}\}\Big).$$

Remarque 3.16. *Il existe d'autres préconditionneurs associés à la résolution du système*

$$\mathbf{M}_1^{-1}\mathbf{A}\mathbf{M}_2^{-1}\{\mathbf{x}^{\text{prec}}_{N_{\text{kry}}}\} = \mathbf{M}_1^{-1}\mathbf{F}, \quad \textit{avec}\ [\mathbf{x}_{N_{\text{kry}}}] = \mathbf{M}_2^{-1}\{\mathbf{x}^{\text{prec}}_{N_{\text{kry}}}\}.$$

Ils généralisent l'approche précédente où $\mathbf{M}_1 = \mathbf{M}_2 = \mathbf{M}^{\frac{1}{2}}$*. Les préconditionneurs à gauche (*$\mathbf{M}_1 = \mathbf{M}$ *et* $\mathbf{M}_2 = \mathbf{I}_{\#\text{ddl}\times\#\text{ddl}}$*) et à droite (*$\mathbf{M}_1 = \mathbf{I}_{\#\text{ddl}\times\#\text{ddl}}$ *et* $\mathbf{M}_2 = \mathbf{M}$*) donnent des résultats similaires à la méthode employée. Pour rappel, le résidu GMRES s'écrit dans le cas général*

$$e^{\text{prec}}_{N_{\text{kry}}} := \frac{\|\mathbf{M}_1^{-1}\mathbf{A}\mathbf{M}_2^{-1}\{\mathbf{x}_{N_{\text{kry}}}\} - \mathbf{M}_1^{-1}\mathbf{F}\|_2}{\alpha^{\text{prec}}\|\{\mathbf{x}_{N_{\text{kry}}}\}\|_2 + \beta^{\text{prec}}}, \tag{3.27}$$

où $\alpha^{\text{prec}} \approx \|\mathbf{M}_1^{-1}\mathbf{A}\|_2$*,* $\beta^{\text{prec}} \approx \|\mathbf{M}_1^{-1}\mathbf{F}\|_2$*, voir [44].*

Nous remarquons que $[\mathbf{x}^{\text{jac}}_{N_{\text{kry}}}] \in \mathbb{C}^{\#\text{ddl}}$, la solution du Problème itératif UWVF discret 15, appartient à $[\mathbb{K}^{\text{prec}}_{N_{\text{kry}}}]$, *ie* $\mathbf{x}^{\text{jac}}_{N_{\text{kry}}} \in \mathbb{K}^{\text{prec}}_{N_{\text{kry}}}$. Ainsi, nous établissons le théorème de convergence suivant pour le problème de KG UWVF préconditionné.

Théorème 3.2 (Convergence de la méthode de KG UWVF préconditionnée)**.** *La méthode de KG UWVF préconditionnée converge au moins au même taux de convergence que la méthode itérative UWVF 15. Nous avons*

$$\|\mathbb{E}^h - \mathbb{E}^{\mathrm{prec}}_{N_{\mathrm{kry}}}\|_{\mathrm{GD}} \;\leq\; \sqrt{2}\|\mathbf{x}^h - \mathbf{x}^{\mathrm{jac}}_{N_{\mathrm{kry}}}\|_{\mathrm{L}^2_t(\partial\mathcal{T})},$$

où $\mathbb{E}^{\mathrm{prec}}_{N_{\mathrm{kry}}} := \mathcal{S}_{\mathrm{out}}\mathbf{x}^{\mathrm{prec}}_{N_{\mathrm{kry}}}$ *et* $\mathbb{E}^h = \mathcal{S}_{\mathrm{out}}\mathbf{x}^h$*, voir la définition de cette bijection en* (3.3) *et dans la Figure 3.1.*

Remarque 3.17. *Le Théorème 3.2 affirme que l'algorithme de KG converge dès que l'algorithme de Cessenat-Després converge. D'autre part, la Proposition 3.2 énonce que* $\mathbf{x}^{\mathrm{jac}}_{N_{\mathrm{kry}}} \in \mathbb{Y}^h_{\mathcal{T}}$ *converge vers* $\mathbf{x}^h \in \mathbb{Y}^h_{\mathcal{T}}$ *avec un taux de convergence* $\rho(\mathbf{M}^{-\frac{1}{2}}\mathbf{N}\mathbf{M}^{-\frac{1}{2}}) < 1$*. Ainsi, nous avons*

$$\|\mathbf{x}^h - \mathbf{x}^{\mathrm{jac}}_{N_{\mathrm{kry}}}\|_{\mathrm{L}^2_t(\partial\mathcal{T})} \underset{N_{\mathrm{kry}}\to+\infty}{\longrightarrow} 0\,.$$

Comme nous l'avons vu grâce à la Proposition 3.3, la convergence de la solution numérique obtenue par une méthode de Krylov est liée aux spectres des matrices étudiées. Le taux de convergence du solveur itératif UWVF (*resp.* de KG préconditionné) dépend du spectre de $\mathbf{A}$ (*resp.* $\widetilde{\mathbf{A}}$), voir la Figure 3.6 (*resp.* Figure 3.7). Dans cet exemple, nous vérifions bien la propriété de contraction décrite par la Proposition 3.2 (*ie* valeurs propres de $\widetilde{\mathbf{A}}$ dans le cercle unité). Nous verrons que la présence de moins de valeurs propres proches de 0 (en rouge Figure 3.7) améliorera la convergence des solveurs de type Krylov. De plus, $\widetilde{\mathbf{A}}$ a un nombre de conditionnement plus petit que $\mathbf{A}$, voir le Tableau 3.2. Cela témoigne des meilleures propriétés pour la convergence.

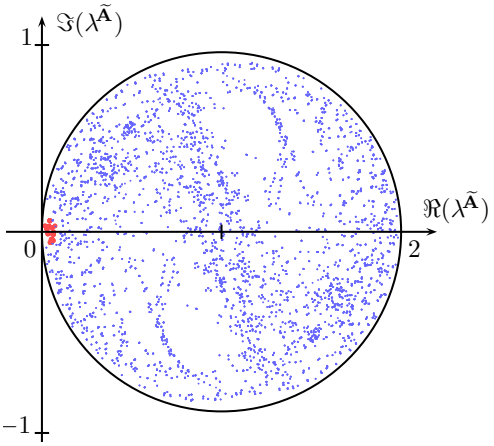

FIGURE 3.7 – Parties réelle et complexe, *resp.* $\Re(\lambda^{\widetilde{\mathbf{A}}})$ et $\Im(\lambda^{\widetilde{\mathbf{A}}})$, du spectre $\lambda^{\widetilde{\mathbf{A}}}$ de la matrice $\widetilde{\mathbf{A}}$ pour $\mathcal{D}_\Omega = 6\lambda$ et $N = 52$.

$\mathcal{D}_\Omega\,(\lambda)$	1	2	3	4	5	6
$\kappa(\mathbf{A})$	18.5	23.4	45.1	60.7	88.8	113
$\kappa\left(\widetilde{\mathbf{A}}\right)$	4.82	10.9	25.6	29.9	59.8	60.9

TABLE 3.2 – Comparaison des nombres de conditionnement de $\mathbf{A}$ et de $\widetilde{\mathbf{A}}$ en fonction de $\mathcal{D}_\Omega$.

3.4 Résultats numériques pour les méthodes itératives de Krylov

Dans un premier temps, nous mettons en avant les bénéfices en termes de coût mémoire de l'utilisation d'un algorithme itératif. Nous donnons en particulier celui associé au solveur GMRES qui se révèle être du même ordre pour l'algorithme de KG. Dans un second temps, nous souhaitons comparer les performances des deux méthodes de Krylov développées dans les Sections 3.2 et 3.3. Nous avons implémenté dans GoTEM3 l'algorithme de KG associé au Problème 19. Il utilise la stratégie de *restart* et la méthode de Gram-Schmidt à un PMV par $\mathbf{A}$ (voir page 128).

Remarque 3.18. *Les solutions $\mathbf{x}_{N_{\text{kry}}} \in \mathbb{Y}_\mathcal{T}^h$ et les fonctions test $\mathbf{x}' \in \mathbb{Y}_\mathcal{T}^h$ employées dans les Problèmes 16 (ou de manière équivalente 17), 18 et 19 sont en bijection avec $\mathbb{E} \in \mathbb{X}_\mathcal{T}^h$ et $\mathbb{E}' \in \mathbb{X}_\mathcal{T}^h$ du Problème variationnel 10, voir (3.3) et la Figure 3.1. Comme nous l'avions brièvement suggéré grâce à la Remarque 3.3, il existe bien une bijection entre les espaces de dimension finie $\mathbb{Y}_\mathcal{T}^h$ et $\mathbb{X}_\mathcal{T}^h$. Nous utiliserons la solution numérique volumique, ie $\mathbb{E} \in \mathbb{X}_\mathcal{T}^h$, pour interpréter nos résultats numériques dans la section suivante.*

3.4.1 Gains mémoire face aux solveurs directs

Nous étudions dans cette partie le coût mémoire d'une méthode de GMRES en fonction de la taille de l'espace de Krylov N_{kry} et face aux solveurs utilisant une factorisation LU. La méthode de GMRES est moins coûteuse qu'une méthode directe pour la résolution du problème de Maxwell, voir la Figure 3.8. Cependant, le solveur GMRES reste relativement limité, au sens où il n'est pas possible de considérer une taille de domaine supérieure à $\mathcal{D}_\Omega = 150^{\max}\lambda$ avec 1To de mémoire. Ce phénomène est dû au coût de stockage de l'ensemble de la base de Krylov. Plus précisément, son coût mémoire [44] en complexes double précision est environ donné par

$$\text{MEM}^{\text{GMRES}} := \left(N_{\text{kry}} \times N_{\text{kry}} + N_{\text{kry}} \times \#\text{ddl} + 5 \times \#\text{ddl} + \text{MEM}^{\mathbf{A}}\right) \times 16, \tag{3.28}$$

où nous avons

— $\times 16$ pour le stockage en double précision,

— $N_{\text{kry}} \times N_{\text{kry}}$ pour le stockage de la matrice $\mathbf{L}$, définie par (3.26),

— $N_{\text{kry}} \times \#\text{ddl}$ pour le stockage de l'espace Krylov $[\mathbb{K}_{N_{\text{kry}}}]$, défini par (3.20),

— $5 \times \#\text{ddl}$ pour le stockage des vecteurs de travail, comme $\mathbf{F}$ par exemple,

— $\text{MEM}^{\mathbf{A}}$ pour le coût de stockage de la matrice $\mathbf{A}$, qui est approché par

$$\text{MEM}^{\mathbf{A}} := 7 \times N \times \#\text{ddl} = 7 \times N^2 \times \#\text{elem},$$

où nous rappelons que $\#\text{elem}$ est le nombre d'éléments du maillage, et où cette estimation provient de la structure de la matrice $\mathbf{A}$ en Figure 2.9. En effet, nous devons considérer au maximum 7 blocs matriciels (pour les 7 interactions possibles : 1 sur l'élément lui-même et 6 avec ses voisins, voir Sous-section 2.4.1) ayant N colonnes par bloc ligne (*ie* par élément) comme nous avons N fonctions de base par élément.

Remarque 3.19. *Dans l'estimation* (3.28)*, nous négligeons notamment le coût mémoire de stockage des structures associées au maillage.*

Le facteur N_{kry} est déterminant dans le coût mémoire de la méthode. En effet, à coût mémoire équivalent (environ 1To), la méthode de GMRES avec $N_{\text{kry}} = 10 \times 10^3$ peut traiter un domaine de taille maximale $\mathcal{D}_\Omega^{\max} = 50\lambda$ contre $\mathcal{D}_\Omega^{\max} = 150\lambda$ avec $N_{\text{kry}} = 20$, voir la Figure 3.8. Nous rappelons ici l'intérêt, en ce qui concerne la mémoire, d'utiliser une stratégie de *restart* avec N_{kry} petit.

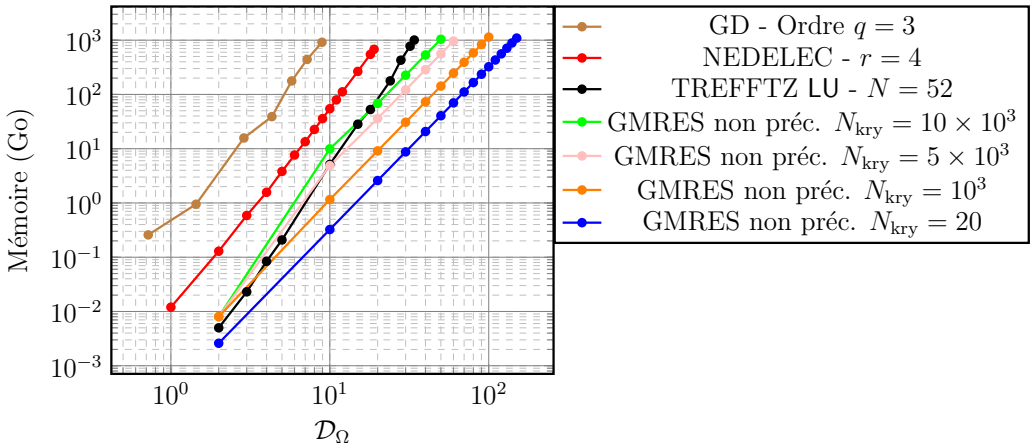

FIGURE 3.8 – Coût mémoire des méthodes directes face à la méthode de GMRES non préconditionnée pour $N = 52$, pour différentes tailles de l'espace de Krylov N_{kry}.

3.4.2 Études de convergence

Nous proposons d'étudier un domaine $\Omega = [0, 10]^3$ en longueur d'onde, où $h = 0.5$ et $R_{\partial\Omega} = 0$. Nous simulons un dipôle électromagnétique, dont nous rappelons la définition du champ électromagnétique, voir [69],

$$\begin{aligned} \mathbf{E}^{\text{ex}}(\mathbf{x}) &= \frac{e^{ik_0 r}}{4\pi r}\Big[(-\frac{1}{r^2} + \frac{ik_0}{r} + k_0^2)(\hat{\mathbf{x}} \times (\mathbf{d} \times \hat{\mathbf{x}})) + 2(\frac{1}{r^2} - \frac{ik_0}{r})(\mathbf{d} \cdot \hat{\mathbf{x}})\hat{\mathbf{x}}\Big], \\ \mathbf{H}^{\text{ex}}(\mathbf{x}) &= \frac{e^{ik_0 r}}{4\pi r}(\frac{ik_0}{r} + k_0^2)(\hat{\mathbf{x}} \times \mathbf{d}), \end{aligned} \tag{3.29}$$

où $r = \sqrt{(x-x_0)^2 + (y-y_0)^2 + (z-z_0)^2}$, $\mathbf{x} := (x, y, z) \in \Omega$, $\hat{\mathbf{x}} \in \widehat{T} := [0, 1]^3$, où $\mathbf{d} := (1, 0, 0)^{\mathsf{T}}$ est le moment dipolaire et $\mathbf{x}_0 = (5, -2.5, 5)$ dans cet exemple.

Nous calculons la solution obtenue grâce à une factorisation LU, notée $\mathbb{E}^{\text{ref}}$ et choisie comme solution de référence. Nous analysons l'erreur entre la solution de référence et les solutions numériques des algorithmes préconditionnés de GMRES UWVF, et de KG UWVF, respectivement notées $\mathbb{E}^{\text{GMRES}}$ et $\mathbb{E}^{\text{KG}}$. Les erreurs relatives associées sont définies par le produit scalaire hermitien de l'espace $\mathbb{X}_T$, voir (2.2). Nous nommons cette erreur : l'erreur relative Trefftz. Nous avons respectivement pour les deux solutions :

$$\begin{cases} e_{\text{tz}}^2 := \dfrac{([\mathbb{E}^{\text{ref}}] - [\mathbb{E}^{\text{GMRES}}])^* \mathbf{M} ([\mathbb{E}^{\text{ref}}] - [\mathbb{E}^{\text{GMRES}}])}{[\mathbb{E}^{\text{ref}}]^* \mathbf{M} [\mathbb{E}^{\text{ref}}]}, \\ e_{\text{tz}}^2 := \dfrac{([\mathbb{E}^{\text{ref}}] - [\mathbb{E}^{\text{KG}}])^* \mathbf{M} ([\mathbb{E}^{\text{ref}}] - [\mathbb{E}^{\text{KG}}])}{[\mathbb{E}^{\text{ref}}]^* \mathbf{M} [\mathbb{E}^{\text{ref}}]}, \end{cases}$$

où $\mathbf{M}$ est définie par (3.17).

Comme nous pouvons le voir sur la Figure 3.9, la diminution de e_{tz} au fur et à mesure des itérations semble légèrement plus rapide pour la méthode de KG. Cela confirme le meilleur contrôle attendu de la norme Trefftz par l'approche KG UWVF. Néanmoins, des investigations numériques plus poussées seront nécessaires afin d'évaluer si nous pouvons tirer profit de cet avantage. En ce qui concerne le coût de ce nouvel algorithme, des études sont encore à mener car il est probable que notre implémentation soit moins optimisée que l'algorithme de GMRES du CERFACS©. En effet, il s'avère qu'en termes de temps d'exécution et de mémoire, cette dernière méthode a de meilleures performances d'environ un tiers. Nous conseillons donc de l'utiliser, comme ses interfaces d'entrées et de sorties et ses nombreuses options sont pratiques à manipuler.

Maintenant, nous souhaitons comparer la convergence des solveurs GMRES UWVF préconditionnés ou non. Nous employons la méthode de GMRES du CERFACS©. Pour comparer les convergences selon un même indicateur d'erreur, nous utilisons la norme Trefftz plutôt que le résidu GMRES. Elle nous paraît plus adaptée pour étudier, au sens physique, la

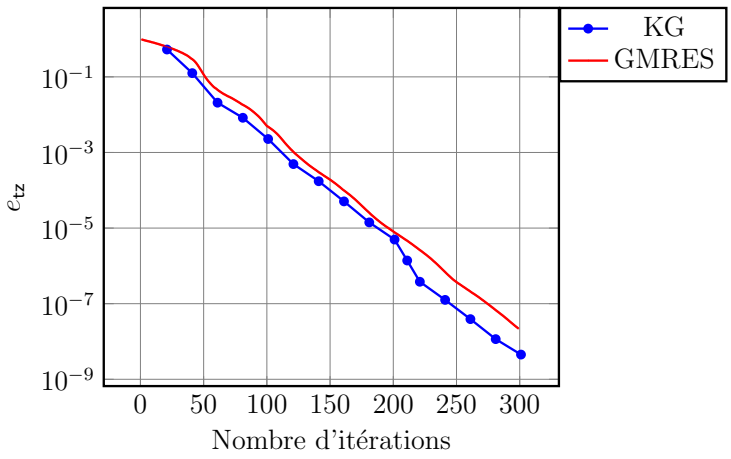

FIGURE 3.9 – Erreur relative Trefftz en fonction du nombre d'itérations pour la méthode de GMRES et pour la méthode de KG.

convergence de la solution numérique vers la solution exacte du problème.

Nous choisissons comme solution de référence $[\mathbb{E}^{\mathrm{ref}}_{N_{kry}}] \in [\mathbb{K}_{N_{\mathrm{kry}}}]$ la solution numérique du solveur GMRES UWVF non préconditionné convergée à $e^{\mathrm{prec}}_{N_{\mathrm{kry}}} = 10^{-12}$ (résidu défini par (3.27)). Ensuite nous étudions l'erreur relative Trefftz entre cette solution de référence et la solution numérique $[\mathbb{E}_{N_{kry}}] \in [\mathbb{K}_{N_{\mathrm{kry}}}]$ (*resp.* $[\mathbb{E}^{\mathrm{prec}}_{N_{kry}}] \in [\mathbb{K}^{\mathrm{prec}}_{N_{\mathrm{kry}}}]$) du Problème 18 non préconditionné (*resp.* 19 préconditionné) à chaque itération du GMRES jusqu'à la convergence $e^{\mathrm{prec}}_{N_{\mathrm{kry}}} = 10^{-8}$. Autrement dit nous calculons :

$$e^2_{\mathrm{tz}} := \frac{([\mathbb{E}^{\mathrm{ref}}_{N_{\mathrm{kry}}}] - [\mathbb{E}_{N_{\mathrm{kry}}}])^* \mathbf{M} ([\mathbb{E}^{\mathrm{ref}}_{N_{\mathrm{kry}}}] - [\mathbb{E}_{N_{\mathrm{kry}}}])}{[\mathbb{E}^{\mathrm{ref}}_{N_{\mathrm{kry}}}]^* \mathbf{M} [\mathbb{E}^{\mathrm{ref}}_{N_{\mathrm{kry}}}]}. \tag{3.30}$$

Nous analysons les convergences des solutions des solveurs GMRES UWVF non préconditionné et de GMRES UWVF préconditionné. Nous étudions un domaine sans obstacle $\Omega = [1, 40]^3$ en longueur d'onde, avec une taille $h = 1$ pour les éléments du maillage. Nous simulons un dipôle assimilé à une source ponctuelle placée à l'avant du domaine $\mathbf{x}_0 = (20, -5, 20)$. Dans le cas où $R_{\partial\Omega} = 0$, voir Figure 3.10, (*resp.* $R_{\partial\Omega} = 0.9$, voir Figure 3.11), le solveur GMRES préconditionné utilise 25% (*resp.* 20%) moins d'itérations que le solveur GMRES non préconditionné. Ainsi, la méthode avec préconditionneur de Cessenat-Després (en orange et en bleu) a besoin de moins d'itérations que la méthode non préconditionnée.

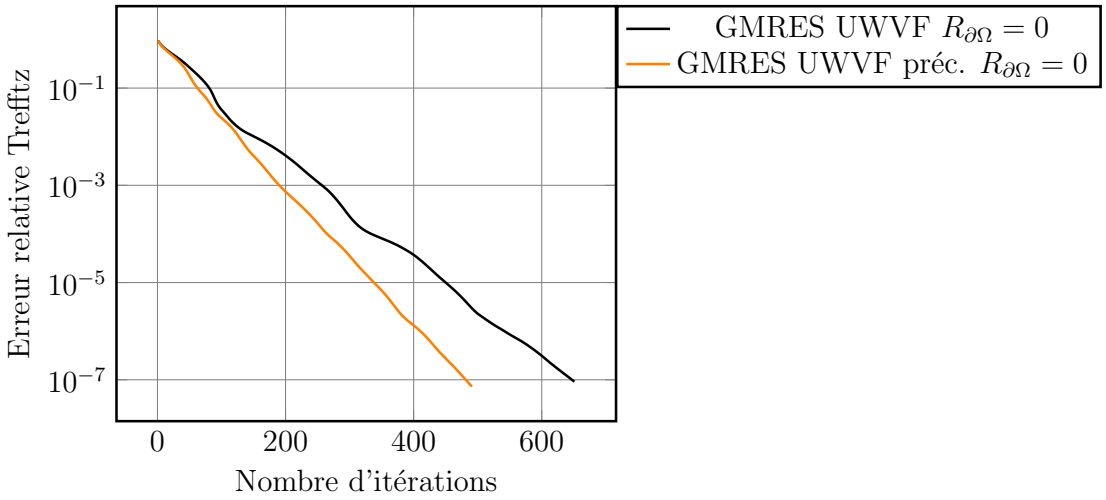

FIGURE 3.10 – Courbes de convergence du solveur GMRES UWVF non préconditionné et preconditionné, pour $\mathcal{D}_\Omega = 40\lambda$, $h = 1$, $N_{\text{kry}} = 500$ et $R_{\partial\Omega} = 0$.

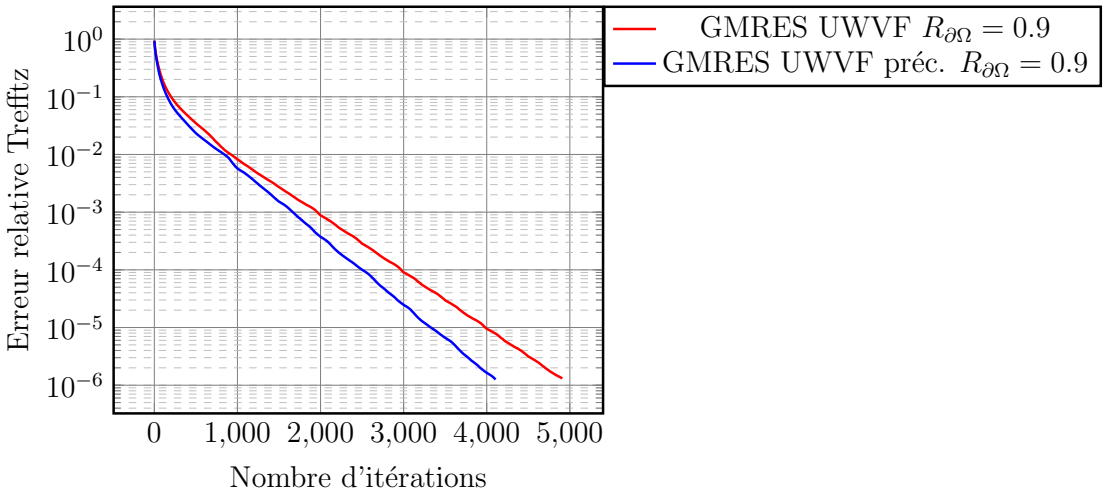

FIGURE 3.11 – Courbes de convergence du solveur GMRES UWVF non préconditionné et preconditionné, pour $\mathcal{D}_\Omega = 40\lambda$, $h = 1$, $N_{\text{kry}} = 500$ et $R_{\partial\Omega} = 0.9$.

3.5 Conclusion

Dans ce chapitre, nous avons explicité une UWVF basée sur les traces numériques de Cessenat-Després. Il s'avère que celle-ci est équivalente aux formulations dérivées dans le Chapitre 2. En effet, le Problème 12 de Cessenat-Després est équivalent au Problème 9 obtenu à partir des traces numériques upwind. Plus particulièrement, les traces numériques de Riemann, de type upwind ou de Cessenat-Després mènent à des formulations équivalentes. Le processus de construction des formulations Trefftz et leur caractère bien posé est résumé dans la Figure 3.12.

Ces différents points de vue peuvent être utilisés au sein de deux méthodes de Krylov : une méthode de GMRES ou une méthode de Krylov Galerkin (KG). Elles sont toutes deux définies sur un espace de Krylov, voir le Tableau 3.3 où nous détaillons aussi d'autres similitudes et différences. Dans les deux cas, nous tirons profit de la stratégie de *restart*, qui permet de réduire le coût mémoire de la méthode. Dans la Section 3.3, nous avons décrit le procédé de pseudo-orthogonalisation mis en jeu par la méthode de KG (Gram-Schmidt à un PMV). De plus, la théorie de Galerkin nous a permis d'établir un résultat de convergence pour le problème UWVF de Cessenat-Després. Cette convergence a enfin pu être accélérée grâce à la mise en place d'un préconditionneur.

L'emploi des méthodes de Krylov permet de diminuer drastiquement le coût mémoire de la méthode de Trefftz, qui est estimé par (3.28) dans le cas du solveur GMRES par exemple. Comme observé dans les résultats numériques de ce chapitre, il est intéressant d'utiliser une petite dimension pour la taille de la base de l'espace de Krylov (N_{kry} petit). Cela rend nécessaire l'utilisation de *restart* dans les algorithmes de Krylov mais permet la considération de grands domaines de calcul, jusqu'à $\mathcal{D}_\Omega^{\max} = 150\lambda$ longueurs d'onde, voir la Figure 3.13, pour $N_{\text{kry}} = 20$ et pour un coût $\text{MEM}^{\text{GMRES}} = 1092$Go.

De plus, en comparant les résultats numériques associés à ces méthodes de Krylov, nous nous sommes aperçus qu'un bon compromis est l'utilisation de la méthode de GMRES du CERFACS©. Grâce à son implémentation optimisée, l'algorithme de GMRES s'est révélé être plus efficace en termes de temps de calcul et de coût mémoire. Toutefois, Le nombre d'itérations de la méthode de KG étant légèrement inférieur, de futures optimisations pourraient mener à une meilleure efficacité.

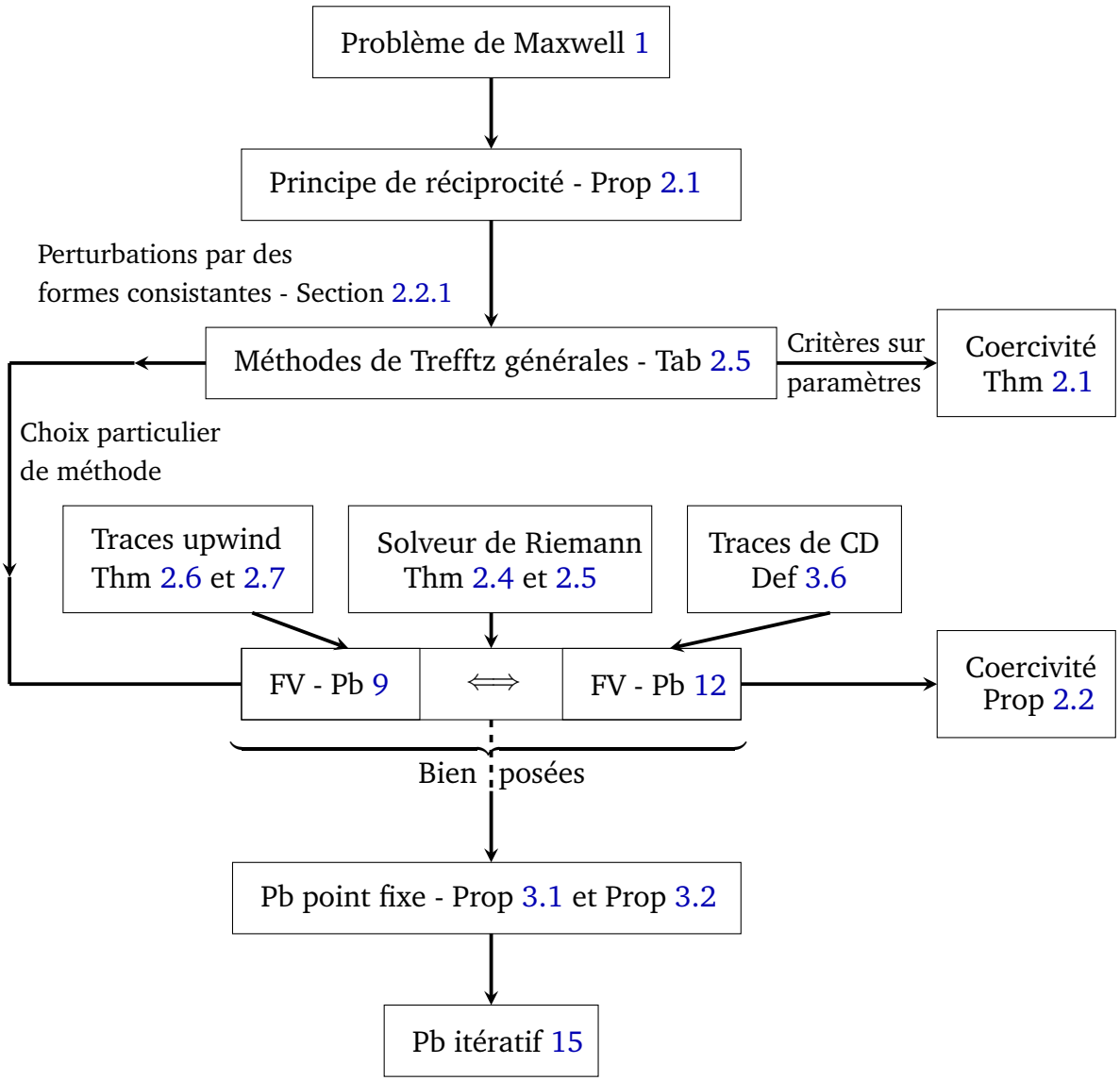

FIGURE 3.12 – Processus de construction des différentes formulations variationnelles de Trefftz bien posées.

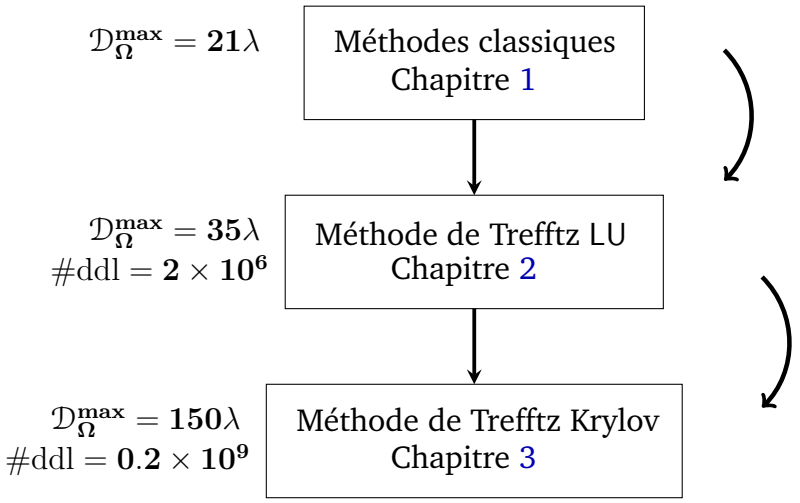

FIGURE 3.13 – Augmentation de la taille $\mathcal{D}_\Omega$ qu'il est possible de considérer grâce à la mise en place d'une méthode de Trefftz directe puis de GMRES, où $\mathcal{D}_\Omega^{\max}$ est la taille maximale atteinte dans chacun des chapitres pour 1To de mémoire, et où $\#\text{ddl} = N \times \#\text{elem}$ avec $N = 52$ et $h = 1$.

	Méthodes de Krylov	
	Méthode de GMRES (Section 3.2)	Méthode de Krylov Galerkin (Section 3.3)
Problème variationnel	Trouver $\mathbf{x} \in \mathbb{Y}_\mathcal{T}^h$ tel que $\forall \mathbf{x}' \in \mathbb{Y}_\mathcal{T}^h, \quad a(\mathbf{x}, \mathbf{x}') = \ell(\mathbf{x}')$	
Vu sous la forme	Trouver $[\mathbf{x}_{N_{\text{kry}}}^{\text{prec}}] \in \left[\mathbb{K}_{N_{\text{kry}}}^{\text{prec}}\right]$ minimisant $\|[\widetilde{\mathbf{A}}[\mathbf{x}_{N_{\text{kry}}}^{\text{prec}}] - \widetilde{\mathbf{F}}]\|_2$	Trouver $[\mathbf{x}_{N_{\text{kry}}}^{\text{prec}}] \in \left[\mathbb{K}_{N_{\text{kry}}}^{\text{prec}}\right]$ tel que $\forall [\mathbf{x}'] \in \left[\mathbb{K}_{N_{\text{kry}}}^{\text{prec}}\right]$ $[\mathbf{x}']^* \widetilde{\mathbf{A}}[\mathbf{x}_{N_{\text{kry}}}^{\text{prec}}] = [\mathbf{x}']^* \widetilde{\mathbf{F}}$
Équation résolue	$[\mathbf{x}']^* \widetilde{\mathbf{A}}^* \widetilde{\mathbf{A}}[\mathbf{x}_{N_{\text{kry}}}^{\text{prec}}] = [\mathbf{x}']^* \widetilde{\mathbf{F}}$	$[\mathbf{x}']^* \widetilde{\mathbf{A}}[\mathbf{x}_{N_{\text{kry}}}^{\text{prec}}] = [\mathbf{x}']^* \widetilde{\mathbf{F}}$
Espace de Krylov	$\left[\mathbb{K}_{N_{\text{kry}}}^{\text{prec}}\right] = \text{span}\left(\left\{ \widetilde{\mathbf{F}}, \widetilde{\mathbf{A}}\widetilde{\mathbf{F}}, \widetilde{\mathbf{A}}^2\widetilde{\mathbf{F}}, \ldots, \widetilde{\mathbf{A}}^{N_{\text{kry}}-1}\widetilde{\mathbf{F}} \right\} \right)$	
Procédé d'ortho.	GSM (Arnoldi) [92]	Gram-Schmidt 1 PMV (page 128)

TABLE 3.3 – Différences entre les deux méthodes de Krylov développées : la méthode de GMRES et la méthode de Krylov Galerkin (préconditionnée en bleu).

CHAPITRE 4

STRATÉGIES DE STABILISATION ET D'ACCÉLÉRATION DE LA RÉSOLUTION ITÉRATIVE ADAPTÉES AU CALCUL HPC

Sommaire

Dans le Chapitre 3, nous avons dérivé des méthodes itératives de type Trefftz : les méthodes de Krylov. En particulier, nous utilisons en pratique la méthode de GMRES dans laquelle la stratégie de *restart* permet de fortement augmenter la taille des domaines que GoTEM3 peut considérer. Mais le stockage de la matrice $\mathbf{A}$ reste important lorsque $\mathcal{D}_\Omega$ augmente. Par exemple, pour $\mathcal{D}_\Omega = 150\lambda$ le coût de stockage approché de $\mathbf{A}$ est $\mathrm{MEM}^{\mathbf{A}} =$

1022Go. Pour $N = 52$, $N_{\text{kry}} = 20$ et pour l'algorithme de GMRES, nous avons alors

$$\frac{\text{MEM}^{\mathbf{A}}}{\text{MEM}^{\text{GMRES}}} \approx 94\%.$$

Ainsi, c'est le stockage de $\mathbf{A}$ qui nous empêche d'étudier des tailles de scènes de calcul $\mathcal{D}_\Omega > 150\lambda$. Avec ce constat, nous remarquons qu'une stratégie doit nécessairement être développée pour réduire $\text{MEM}^{\mathbf{A}}$. En particulier, l'idée est de ne plus stocker la matrice $\mathbf{A}$ mais seulement les informations nécessaires au calcul du PMV par $\mathbf{A}$. Ceci constitue la première section de ce chapitre.

Dans la même optique, nous dérivons d'autres stratégies pour répondre aux deux objectifs de cette partie :

1. diminuer le coût mémoire de la méthode de GMRES de type Trefftz,
2. accélérer la convergence de la solution numérique grâce à des améliorations du conditionnement de la matrice $\mathbf{A}$.

Plus précisément, nous dérivons une stratégie de réduction de la base d'ondes planes $\mathbb{Y}_{\mathcal{T}}^h$. Cette démarche permet d'agir sur les deux objectifs de ce chapitre, à savoir la réduction du coût mémoire de GoTEM3 et l'amélioration du conditionnement de la matrice. Par la suite, nous présentons un préconditionneur global, qui accélère la convergence.

4.1 Stratégie de désassemblage

L'assemblage de la matrice $\mathbf{A}$ engendre un coût mémoire conséquent qui devient de plus en plus grand lorsque la taille du domaine augmente, voir le Tableau 4.1.

$\mathcal{D}_\Omega\,(\lambda)$	5	10	20	100	200
$\#\text{elem}$	512	1000	8000	10^6	8×10^6
Coût avec assemblage $\mathbf{A}$	344 Mo	2.75 Go	22.04 Go	2.75 To	2204 To

TABLE 4.1 – Coûts mémoire pour le stockage de la matrice $\mathbf{A}$ en fonction de la taille du domaine $\mathcal{D}_\Omega$ et du nombre d'éléments $\#\text{elem}$, pour $N = 52$ et $h = 1$.

Tout d'abord, nous mettons en place une technique de désassemblage, qui consiste à segmenter $\mathbf{A}$ en blocs matriciels. Puis, nous analysons les performances de cette technique en termes :

— de gain mémoire, grâce au caractère structuré de la matrice,

— de temps de calcul, grâce à l'utilisation de librairies OpenMP.

4.1.1 Principe de la stratégie

La stratégie de désassemblage est basée sur la structure cartésienne du maillage et sur l'utilisation des blocs de matrice élémentaires, définis par (2.53), qui composent la matrice $\mathbf{A}$. Nous rappelons ces blocs matriciels

$$\mathbf{A}_{\mathrm{i_{elem}},\mathrm{j_{elem}}} = \left(\mathbf{A}^{\ell,k}_{\mathrm{i_{elem}},\mathrm{j_{elem}}}\right)_{\ell,k=1}^{N} \in \mathbb{C}^{N\times N}, \quad \text{où} \quad \mathbf{A}^{\ell,k}_{\mathrm{i_{elem}},\mathrm{j_{elem}}} = \mathbf{a}(\mathbf{w}^{k}_{\mathrm{j_{elem}}}, \mathbf{w}^{\ell}_{\mathrm{i_{elem}}}), \tag{4.1}$$

avec $\mathrm{i_{elem}}, \mathrm{j_{elem}} = 1, \#\mathrm{elem}$ et $\ell, k = 1, N$.

Nous expliquons l'implémentation informatique de ces matrices dans GoTEM3. Nous considérons un domaine cubique Ω maillé de cubes $T \in \mathcal{T}$. Nous rappelons que 3 types d'interactions sont possibles dans chaque élément T :

— élément T avec lui-même,

— élément T avec son voisin K,

— élément T avec le bord, *ie* avec un élément "virtuel" K_{ext}, voir la Remarque 3.1.

Chacune d'entre elle peut intervenir sur les 6 faces du cube T. Ainsi, les matrices élémentaires sont stockées informatiquement en construisant les tableaux suivants :

$$\mathrm{MatInteraction}(k, f, \mathrm{i_{loc}}, \mathrm{j_{loc}}) \quad \text{avec } \mathrm{i_{loc}}, \mathrm{j_{loc}} = 1, N\,, \tag{4.2}$$

où $\mathrm{i_{loc}}, \mathrm{j_{loc}}$ sont les numéros des fonctions de base de $\mathbb{Y}^h_{\mathcal{T}}$, et avec k le type d'interaction de l'élément T et f le type de face :

$$\begin{cases} k &= \text{lui}-\text{même, voisin, bord,} \\ f &= \text{gauche, droite, bas, haut, avant, arrière.} \end{cases}$$

Les matrices $\mathbf{A}$, $\mathbf{M}$ et $\mathbf{N}$ sont constituées des blocs élémentaires, voir les Figures 2.9 et 3.2. L'idée est de ne pas les assembler mais de générer seulement le minimum d'information nécessaire aux PMV bloc par bloc. En effet, de par la structure cartésienne du maillage, les matrices élémentaires se répètent et sont les mêmes pour tous les élements T (couleurs dans les Figures 2.9 et 3.2). Plus précisément, un PMV par $\mathbf{A}$ désassemblé prend la forme suivante.

```
function produitParA(u) result(prod)
    allocate(v(N))
    allocate(prod(#ddl))
    do elem = 1,#elem ! pour tous les elements
       do f = 1,6 ! pour toutes les faces
        ! partie du vecteur correspondant a l'element
          u_ielem = u((elem - 1) × N + 1 : elem × N)
```

```
        j_elem = voisin(i_elem, f)  ! numero du voisin
      ! si face sur le bord du domaine
        if (type(j_elem) ==bord) then
          A^{i_elem,j_elem} = MatInteraction(bord, f, :, :)
          v = v + A^{i_elem,j_elem} u_{i_elem}
        else ! si face commune avec le voisin
          A^{i_elem,j_elem} = MatInteraction(voisin, f, :, :)
          v = v + A^{i_elem,j_elem} u_{i_elem}
          A^{i_elem,j_elem} = MatInteraction(luiMeme, f, :, :)
          v = v + A^{i_elem,j_elem} u_{i_elem}
        end if
      end do
      prod((elem - 1) × N + 1 : elem × N) = v
    end do
end function
```

Le tableau voisin$(\mathrm{i}_{\mathrm{elem}}, f)$ donne l'élément voisin, numéroté $\mathrm{j}_{\mathrm{elem}}$, de $\mathrm{i}_{\mathrm{elem}}$ via sa face f (gauche, droite, bas, haut, avant ou arrière).

Le coût mémoire de GoTEM3 avec désassemblage, noté $\mathrm{MEM}^{\mathrm{GMRES}}_{\mathrm{des}}$, est estimé par (3.28) où $\mathrm{MEM}^{\mathbf{A}}$ est remplacé par le coût de stockage des tableaux élémentaires (4.2). Plus précisément, nous avons

$$\mathrm{MEM}^{\mathrm{GMRES}}_{\mathrm{des}} := (N_{\mathrm{kry}} \times N_{\mathrm{kry}} + N_{\mathrm{kry}} \times \#\mathrm{ddl} + 5 \times \#\mathrm{ddl} + \mathrm{MEM}^{\mathbf{A}}_{\mathrm{des}}) \times 16, \tag{4.3}$$

avec

$$\mathrm{MEM}^{\mathbf{A}}_{\mathrm{des}} := 3 \times 6 \times N^2, \tag{4.4}$$

où le facteur 3 (*resp.* 6) fait référence aux types d'interactions (*resp.* aux six faces d'un cube), et enfin N^2 à la taille des blocs élémentaires (4.1). Par exemple, nous avons $\mathrm{MEM}^{\mathbf{A}}_{\mathrm{des}} = 0.7$Mo avec $N = 52$ et $\mathrm{MEM}^{\mathbf{A}}_{\mathrm{des}} = 11$Mo avec $N = 196$. Le coût mémoire de stockage de la matrice $\mathbf{A}$ (*ie* des interactions entre les éléments) devient alors finalement négligeable devant le coût total du solveur GMRES.

4.1.2 Gain mémoire face à des systèmes assemblés

Nous souhaitons observer le gain mémoire obtenu avec la stratégie de désassemblage. Dans cette partie, nous ne distinguons pas la méthode de GMRES non préconditionnée de celle préconditionnée comme leurs coûts mémoire $\mathrm{MEM}^{\mathrm{GMRES}}$ sont tous les deux définis par (3.28).

Nous comparons les coûts mémoire $\mathrm{MEM}^{\mathrm{GMRES}}$ (*resp.* $\mathrm{MEM}^{\mathrm{GMRES}}_{\mathrm{des}}$) d'une méthode de GMRES assemblée (*resp.* désassemblée), pour différentes tailles N_{kry} de la base de l'espace de Krylov. Il s'avère que les méthodes de GMRES avec la matrice $\mathbf{A}$ assemblée sont peu, et même de moins en moins, impactées par une diminution du nombre de vecteurs de Krylov (N_{kry}) lorsque la taille du domaine augmente, voir la Figure 4.1. De plus, si nous comparons à des versions avec désassemblage, le fait de ne plus stocker la matrice $\mathbf{A}$ n'impacte pas le coût mémoire du solveur GMRES lorsque $N_{\mathrm{kry}} \geq 10^3$, voir la Figure 4.2. En revanche, le désassemblage engendre une diminution importante de $\mathrm{MEM}^{\mathrm{GMRES}}_{\mathrm{des}}$ lorsque $N_{\mathrm{kry}} < 10^3$. Cette différence de coût devient de plus en plus grande lorsque N_{kry} diminue.

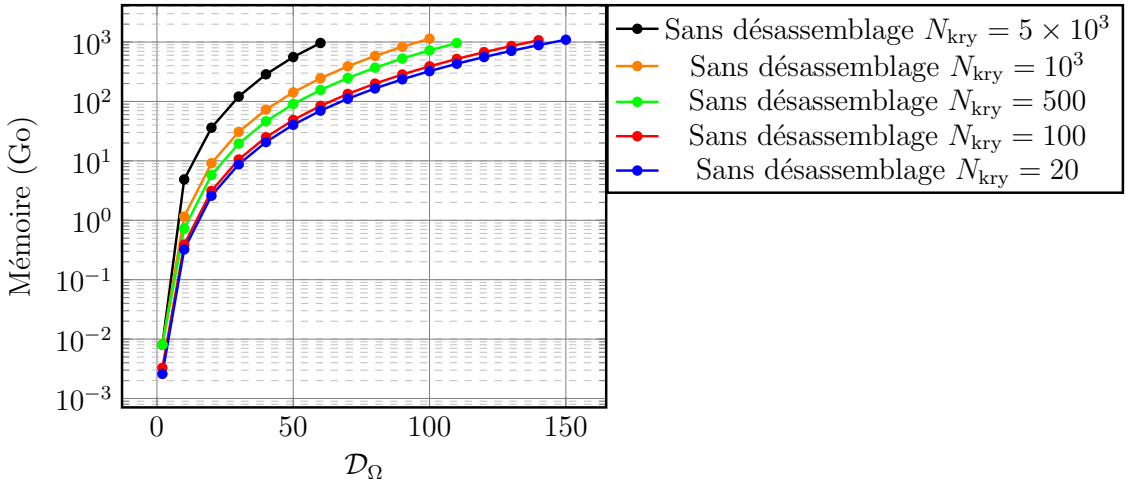

FIGURE 4.1 – Coûts mémoire de la méthode de GMRES sans désassemblage de $\mathbf{A}$ en fonction de la taille du domaine $\mathcal{D}_\Omega$ et pour différents N_{kry}.

En particulier, le gain mémoire avec une technique de désassemblage est représenté dans le Tableau 4.2. Par exemple, pour $N_{\mathrm{kry}} = 20$ et $\mathcal{D}_\Omega = 50\lambda$, nous utilisons, avec le désassemblage, seulement 6% de la mémoire utilisée avec une méthode de GMRES sans désassemblage.

N_{kry}	5×10^3	10^3	500	100	20
Gain (%)	6.7	27	42	78	94

TABLE 4.2 – Gain mémoire avec un solveur GMRES désassemblé pour une taille de domaine $\mathcal{D}_\Omega = 50\lambda$.

Finalement, nous analysons les coûts mémoire d'une méthode de GMRES désassemblée $\mathrm{MEM}^{\mathrm{GMRES}}_{\mathrm{des}}$ face à ceux des méthodes directes ou de GMRES assemblée sur la Figure 4.3. Nous retrouvons le facteur $(\mathcal{D}_\Omega)^4$ pour les méthodes avec factorisation LU, tandis que pour la méthode de GMRES nous avons $(\mathcal{D}_\Omega)^3$. Le gain mémoire engendré par le désassemblage,

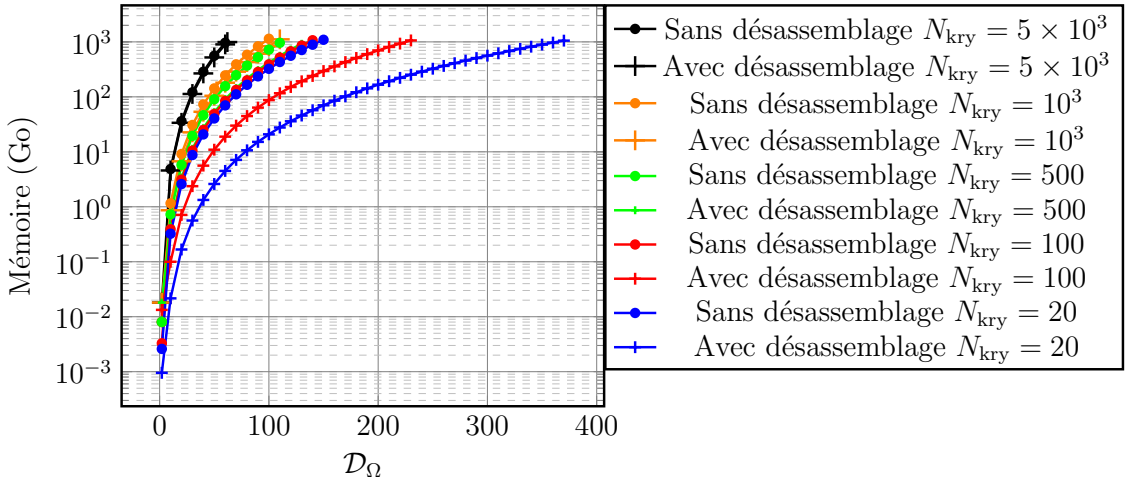

FIGURE 4.2 – Comparaison des coûts mémoire de la méthode de GMRES avec la stratégie de désassemblage ou non, pour différents N_{kry}.

rendu possible grâce à la structure cartésienne de la géométrie, est flagrant, voir la Figure 4.4. En effet, pour $N_{\text{kry}} = 20$, contrairement à un solveur avec assemblage où nous atteignions $\mathcal{D}_\Omega^{\max} = 150\lambda$, nous pouvons désormais traiter des tailles de domaine allant jusqu'à $\mathcal{D}_\Omega^{\max} = 370\lambda$.

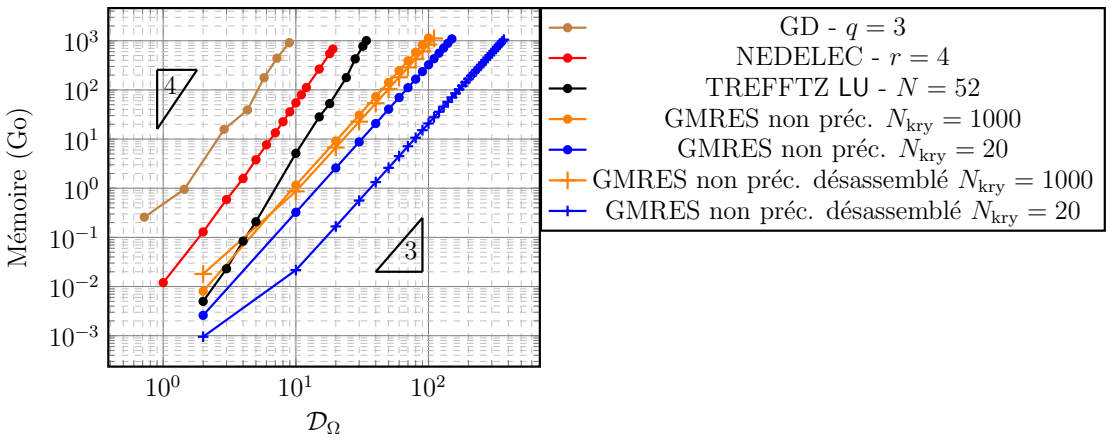

FIGURE 4.3 – Comparaison des coûts mémoire des méthodes directes et de GMRES avec ou sans la stratégie de désassemblage, en échelle loglog, pour différents N_{kry}.

Remarque 4.1. *En pratique, avec* 1*To de mémoire, il n'est pas possible de traiter exactement un cas* $\mathcal{D}_\Omega^{\max} = 370\lambda$*, mais un domaine du même ordre de grandeur :* $\mathcal{D}_\Omega^{\max} = 315\lambda$*. Ceci est dû au coût mémoire causé par le stockage des structures (géométriques notamment) et des tableaux de travail du code GoTEM3. En effet, ce coût est négligé dans l'estimation de* $\text{MEM}_{\text{des}}^{\text{GMRES}}$ *en* (4.3).

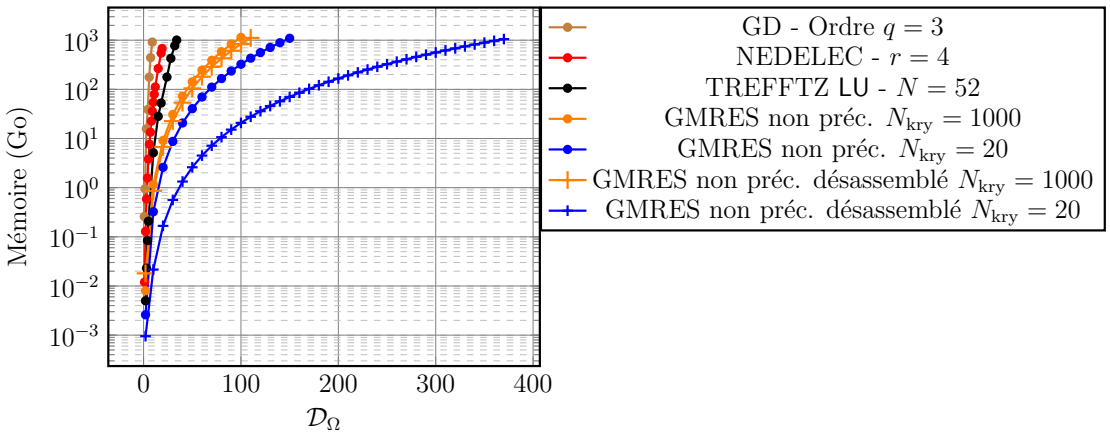

FIGURE 4.4 – Comparaison des coûts mémoire des méthodes directes et de GMRES avec ou sans la stratégie de désassemblage, en échelle semilog, pour différents N_{kry}.

4.1.3 Stratégie adaptée au calcul HPC

Nous étudions maintenant le temps de calcul associé au solveur GMRES. Nous regardons, ici et tout au long du manuscrit, les temps d'exécution obtenus pour des résidus GMRES égaux, définis par (3.21) ou (3.27) selon la méthode employée. De par la structure de la fonction produitParA, voir la Sous-section 4.1.1, nous remarquons qu'elle est particulièrement adaptée à une parallélisation OpenMP. Ceci permet de considérablement réduire les temps de calcul, voir le Tableau 4.3, où les temps d'exécution sont associés aux courbes des Figures 3.10 et 3.11. Avec une stratégie de désassemblage, l'effet de l'utilisation des librairies OpenMP est accentué. Plus précisément, la parallélisation est réalisée sur 24 threads : 2 sockets de 12 cœurs, de type *Broadwell Intel©Xeon©CPU* E5 − 2650v4 à 2.20GHz et à 256Go de mémoire. Les PMV par $\mathbf{A}$ prennent 68% du temps total d'exécution pour une version sans OpenMP, contre 25% du temps total d'exécution pour une version avec OpenMP. Ainsi, l'usage de ces librairies réduit drastiquement les temps de calcul. De plus, pour une taille de domaine $\mathcal{D}_\Omega = 100\lambda$ par exemple, la proportion de temps des PMV par $\mathbf{A}$ est de l'ordre de 20% environ pour la méthode préconditionnée ou non, voir le Tableau 4.4. L'impact de l'OpenMP est alors accentué lorsque nous augmentons la taille du domaine. Concernant les temps totaux d'exécution dans le cas d'une plus grande taille $\mathcal{D}_\Omega$, la méthode préconditionnée prend un temps du même ordre que celle non préconditionnée (voir la deuxième ligne du Tableau 4.3).

Remarque 4.2. *Dans le cas du Tableau 4.3 et des Figures 3.10 et 3.11, où* $\mathcal{D}_\Omega = 40\lambda$, $h = 1$ *et* $N_{\text{kry}} = 100$, *77Go de mémoire sont utilisés avec une méthode de GMRES non préconditionnée assemblée, contre* 51*Go lorsqu'elle est désassemblée.*

Finalement, nous pensons qu'il est préférable de privilégier un préconditionnement de

Cessenat-Després, car il converge plus rapidement et consomme moins de mémoire (même pour une taille de domaine raisonnable).

Temps (s)	GMRES UWVF non préc.	GMRES UWVF préc.
Sans OpenMP	3905.620	3427.433
Avec OpenMP	504.7141	473.3549

TABLE 4.3 – Temps d'exécution de la méthode de GMRES UWVF non préconditionnée et préconditionnée, pour $\mathcal{D}_\Omega = 40\lambda$, $h = 1$ et $R_{\partial\Omega} = 0$, pour des résidus GMRES égaux (10^{-8}).

Temps (s)	GMRES UWVF non préc.	GMRES UWVF préc.
Avec OpenMP	21800.97	21135.12
PMV par $\mathbf{A}$	20.5%	19.5%

TABLE 4.4 – Temps d'exécution de la méthode de GMRES UWVF non préconditionnée et préconditionnée, pour $\mathcal{D}_\Omega = 100\lambda$, $h = 1$ et $R_{\partial\Omega} = 0$, pour des résidus GMRES égaux (10^{-8}).

Remarque 4.3. *Ces temps ont été obtenus sur un même nœud de calcul et avec les mêmes options de compilation. Bien entendu, les temps en eux-mêmes restent contestables dans le sens où ils peuvent changer selon les optimisations (de compilation ou d'implémentation). Toutefois, les comparaisons et les proportions des Tableaux 4.3 et 4.4 sont un bon indicateur.*

4.2 Stratégie de réduction de l'espace de fonctions de base

Dans cette section, nous décrivons une stratégie de sélection des fonctions de base. C'est une manière de limiter le coût mémoire de la méthode de Trefftz GMRES et aussi d'accélérer le calcul de sa solution numérique. Ce dernier aspect est lié à une amélioration du conditionnement de la matrice associée au système matriciel. Des résultats numériques témoigneront de l'intérêt d'appliquer cette stratégie de réduction de base.

4.2.1 Construction d'une base réduite

Le préconditionnement de la matrice $\mathbf{A}$ accélère la convergence du solveur GMRES (Figures 3.10 et 3.11). Mais des travaux précédents [9, 30, 72, 87] ont montré que les ondes

planes de la base $\mathbb{Y}_{\mathcal{T}}^h$ peuvent être linéairement dépendantes numériquement. Cela se traduit par une matrice $\mathbf{M}$ ayant un très grand nombre de conditionnement, qui est le rapport entre la plus grande et la plus petite valeur propre. De plus, des erreurs d'arrondis sur une petite valeur propre de $\mathbf{M}$ donnent naissance à de grandes erreurs d'arrondis pour $\mathbf{M}^{-\frac{1}{2}}$ (ou $\mathbf{M}^{-1}$ dans le cas où nous utilisons une approche classique). Cela impacte alors la convergence du solveur GMRES et peut parfois le faire diverger. Ainsi, le but de la stratégie de réduction est de déterminer un sous espace de $\mathbb{Y}_{\mathcal{T}}^h$ en ne gardant que les vecteurs propres associés aux plus grandes valeurs propres. Cela assure une représentation de $\mathbf{x} \in \mathbb{Y}_{\mathcal{T}}^h$ à travers (3.14), sans le bruit numérique pouvant être causé par le caractère numériquement lié de la base d'ondes planes. La matrice hermitienne $\mathbf{M}$ représente le produit scalaire $\mathrm{L}_t^2(\partial\mathcal{T})$, voir sa définition en (3.17). De par la structure cartésienne du maillage, la matrice $\mathbf{M}$ est en fait une répétition du bloc matriciel élémentaire $\mathbf{M}^{\mathrm{elem}} \in \mathbb{C}^{N\times N}$, voir la Figure 3.2. Cette dernière est définie par

$$\mathbf{M}^{\mathrm{elem}} := \left(\mathbf{M}^{\ell,k}_{\mathrm{i_{elem}},\mathrm{i_{elem}}}\right)_{\ell,k=1}^{N} \in \mathbb{C}^{N\times N}, \quad \text{où} \quad \mathbf{M}^{\ell,k}_{\mathrm{i_{elem}},\mathrm{i_{elem}}} = \left(\mathbf{w}^{\mathrm{i_{glob}}}, \mathbf{w}^{\mathrm{j_{glob}}}\right)_{\mathrm{L}_t^2(\partial\mathcal{T})},$$

où $\mathrm{i_{glob}} = (\mathrm{i_{elem}} - 1)N + \ell$ et $\mathrm{i_{glob}} = (\mathrm{i_{elem}} - 1)N + k$. Sa diagonalisation s'écrit

$$\mathbf{M}^{\mathrm{elem}} = \mathbf{T}^{\mathrm{elem}}\mathbf{\Lambda}^{\mathrm{elem}}(\mathbf{T}^{\mathrm{elem}})^* \quad \text{avec } \mathbf{T}^{\mathrm{elem}} \in \mathbb{R}^{N\times N} \text{ et } \mathbf{\Lambda}^{\mathrm{elem}} \in \mathbb{R}^{N\times N},$$

où $\mathbf{T}^{\mathrm{elem}}$ est la matrice orthogonale de vecteurs propres et $\mathbf{\Lambda}^{\mathrm{elem}}$ est la matrice diagonale des valeurs propres satisfaisant

$$\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_N \quad \text{avec } \mathbf{\Lambda}^{\mathrm{elem}}_{\ell,\ell} = \lambda_\ell \quad \text{pour } \ell = 1, N.$$

La réduction de base consiste dans un premier temps à sélectionner les N_{red} vecteurs propres associés aux plus grandes valeurs propres de $\mathbf{M}$ à partir du critère suivant :

$$\lambda_{N_{\mathrm{red}}} \geq \lambda_1\varepsilon > \lambda_{N_{\mathrm{red}}+1},$$

où la valeur seuil $\varepsilon > 0$ contrôle la précision relative de cette approximation. Ces vecteurs sont alors collectés dans la matrice rectangulaire $\mathbf{T}^{\mathrm{elem}}_{\mathrm{red}}$ définie par

$$(\mathbf{T}^{\mathrm{elem}}_{\mathrm{red}})_{\ell,k} := (\mathbf{T}^{\mathrm{elem}})_{\ell,k} \quad \text{pour } \ell = 1, N \text{ et } k = 1, N_{\mathrm{red}}.$$

La matrice des valeurs propres réduite est donnée par

$$\mathbf{\Lambda}^{\mathrm{elem}}_{\mathrm{red}} := (\mathbf{T}^{\mathrm{elem}}_{\mathrm{red}})^*\mathbf{M}^{\mathrm{elem}}\mathbf{T}^{\mathrm{elem}}_{\mathrm{red}} \in \mathbb{R}^{N_{\mathrm{red}}\times N_{\mathrm{red}}}.$$

Nous associons à chaque vecteur propre une fonction normalisée notée $\mathbf{w}^n_{\mathrm{i_{elem}},\mathrm{red}}$ de l'espace $\mathbb{Y}^h_T$

$$\mathbf{w}^n_{\mathrm{i_{elem}},\mathrm{red}} := \sum_{\ell=1}^{N} (\mathbf{T}^{\mathrm{elem}}_{\mathrm{red}})_{\ell,n} \lambda_n^{-\frac{1}{2}} \mathbf{w}^{\ell}_{\mathrm{i_{elem}}}, \text{ où } n = 1, N_{\mathrm{red}}.$$

Nous définissons enfin l'espace réduit $\mathbb{Y}^h_{T,\mathrm{red}}$ dans lequel la solution réduite sera recherchée

$$\mathbb{Y}^h_{T,\mathrm{red}} := \mathrm{span}(\{\mathbf{w}^n_{\mathrm{i_{elem}},\mathrm{red}} \,:\, n = 1, \ldots, N_{\mathrm{red}}\}) \subset \mathbb{Y}^h_T,$$

où T est l'élement de numéro $\mathrm{i_{elem}}$ du maillage $\mathcal{T}$. La solution numérique réduite s'exprime alors dans la base de départ et dans la base réduite

$$\mathbf{x}^T_{\mathrm{red}} = \sum_{\ell=1}^{N} [\mathbf{x}^T_{\mathrm{red}}]_\ell \, \mathbf{w}^{\ell}_{\mathrm{i_{elem}}} = \sum_{\ell=1}^{N_{\mathrm{red}}} \{\mathbf{x}^T_{\mathrm{red}}\}_\ell \, \mathbf{w}^{\ell}_{\mathrm{i_{elem}},\mathrm{red}},$$

avec $[\mathbf{x}_{\mathrm{red}}]_\ell$ (*resp.* $\{\mathbf{x}_{\mathrm{red}}\}_\ell$) les coordonnées dans la base de départ (*resp.* dans la base orthogonale réduite) liées par

$$[\mathbf{x}^T_{\mathrm{red}}] := \mathbf{T}^{\mathrm{elem}}_{\mathrm{red}} (\mathbf{\Lambda}^{\mathrm{elem}})^{-\frac{1}{2}}_{\mathrm{red}} \{\mathbf{x}^T_{\mathrm{red}}\}.$$

Remarque 4.4. *Nous notons* $\#\mathrm{ddl}_{\mathrm{red}} := N_{\mathrm{red}} \times \#\mathrm{elem}$*, le nombre global réduit de degrés de liberté. Les matrices globales réduites*

$$\mathbf{T}_{\mathrm{red}} \in \mathbb{C}^{\#\mathrm{ddl} \times \#\mathrm{ddl}_{\mathrm{red}}} \quad \textit{et} \quad \mathbf{\Lambda}_{\mathrm{red}} \in \mathbb{C}^{\#\mathrm{ddl}_{\mathrm{red}} \times \#\mathrm{ddl}_{\mathrm{red}}},$$

s'obtiennent facilement par un assemblage diagonal des blocs matriciels élémentaires $\mathbf{T}^{\mathrm{elem}}_{\mathrm{red}} \in \mathbb{C}^{N \times N_{\mathrm{red}}}$ *et* $\mathbf{\Lambda}^{\mathrm{elem}}_{\mathrm{red}} \in \mathbb{C}^{N_{\mathrm{red}} \times N_{\mathrm{red}}}$.

Cela mène à un problème UWVF réduit utilisant moins d'inconnues et défini sur $\mathbb{Y}^h_{\mathcal{T},\mathrm{red}} \subset \mathbb{Y}^h_{\mathcal{T}}$. Il est par conséquent moins coûteux à résoudre numériquement.

Problème 20 (Problème UWVF réduit)**.** *Le problème UWVF réduit s'écrit*
Trouver $\mathbf{x}_{\mathrm{red}} \in \mathbb{Y}^h_{\mathcal{T},\mathrm{red}}$*, tel que nous avons*

$$\mathbf{a}(\mathbf{x}_{\mathrm{red}}, \mathbf{x}'_{\mathrm{red}}) = \mathbf{l}(\mathbf{x}'_{\mathrm{red}}), \quad \forall \mathbf{x}'_{\mathrm{red}} \in \mathbb{Y}^h_{\mathcal{T},\mathrm{red}}, \tag{4.5}$$

ou de manière équivalente :
Trouver $\{\mathbf{x}_{\mathrm{red}}\} \in \mathbb{C}^{\#\mathrm{ddl}_{\mathrm{red}}}$*, tel que nous avons*

$$\{\mathbf{x}'_{\mathrm{red}}\}^* \mathbf{A}_{\mathrm{red}} \{\mathbf{x}_{\mathrm{red}}\} = \{\mathbf{x}'_{\mathrm{red}}\}^* \mathbf{F}_{\mathrm{red}}, \quad \forall \{\mathbf{x}'_{\mathrm{red}}\} \in \mathbb{C}^{\#\mathrm{ddl}_{\mathrm{red}}}, \quad \textit{où } \#\mathrm{ddl}_{\mathrm{red}} := N_{\mathrm{red}} \times \#\mathrm{elem},$$

avec

$$\mathbf{A}_{\mathrm{red}} := \mathbf{\Lambda}^{-\frac{1}{2}}_{\mathrm{red}} \, \mathbf{T}^*_{\mathrm{red}} \, \mathbf{A} \, \mathbf{T}_{\mathrm{red}} \, \mathbf{\Lambda}^{-\frac{1}{2}}_{\mathrm{red}} \quad \textit{et} \quad \mathbf{F}_{\mathrm{red}} := \mathbf{\Lambda}^{-\frac{1}{2}}_{\mathrm{red}} \, \mathbf{T}^*_{\mathrm{red}} \, \mathbf{F}, \tag{4.6}$$

où $\mathbf{T}_{\mathrm{red}}$ *et* $\mathbf{\Lambda}_{\mathrm{red}}$ *sont définies par la Remarque 4.4.*

Remarque 4.5. *La réduction de base associée à l'espace* $\mathbb{Y}^h_{\mathfrak{T},\mathrm{red}}$ *peut être interprétée comme un préconditionneur symétrique, voir (4.6), et est donc en adéquation avec l'approche KG UWVF, voir le Problème 19.*

Comme $\mathbb{Y}^h_{\mathfrak{T},\mathrm{red}} \subset \mathbb{Y}^h_{\mathfrak{T}}$, la théorie de convergence établie dans la Proposition 3.2 reste vraie pour le système réduit (4.5). Par conséquent, le Théorème 3.2 donne une méthode itérative UWVF GMRES réduite convergente.

Par ailleurs, nous remarquons que $\mathbf{M}$ devient la matrice identité, mais de dimension $\#\mathrm{ddl}_{\mathrm{red}}$ cette fois-ci (voir la Remarque 3.15),

$$\mathbf{I}_{\#\mathrm{ddl}_{\mathrm{red}}\times\#\mathrm{ddl}_{\mathrm{red}}} := \mathbf{\Lambda}_{\mathrm{red}}^{-\frac{1}{2}}\mathbf{T}^*_{\mathrm{red}}\mathbf{M}\mathbf{T}_{\mathrm{red}}\mathbf{\Lambda}_{\mathrm{red}}^{-\frac{1}{2}}.$$

Remarque 4.6. *Le préconditionneur de Cessenat-Després est alors automatiquement contenu dans la formulation réduite. Le Problème de KG UWVF 19 peut aussi s'écrire en version réduite : Trouver* $\mathbf{x}^{\mathrm{prec,red}}_{N_{\mathrm{kry}}} \in \mathbb{K}^{\mathrm{prec,red}}_{N_{\mathrm{kry}}}$*, tel que nous avons*

$$\mathbf{a}(\mathbf{x}^{\mathrm{prec,red}}_{N_{\mathrm{kry}}}, \mathbf{x}') = \mathbf{l}(\mathbf{x}'), \quad \forall \mathbf{x}' \in \mathbb{K}^{\mathrm{prec,red}}_{N_{\mathrm{kry}}},$$

ou de manière équivalente $\{\mathbf{x}^{\mathrm{prec,red}}_{N_{\mathrm{kry}}}\} \in \{\mathbb{K}^{\mathrm{prec,red}}_{N_{\mathrm{kry}}}\} \subset \mathbb{C}^{\#\mathrm{ddl}_{\mathrm{red}}}$ *tel que nous avons*

$$\{\mathbf{x}'\}^*\mathbf{A}_{\mathrm{red}}\{\mathbf{x}^{\mathrm{prec,red}}_{N_{\mathrm{kry}}}\} = \{\mathbf{x}'\}^*\mathbf{F}_{\mathrm{red}}, \quad \forall\{\mathbf{x}'\} \in \{\mathbb{K}^{\mathrm{prec,red}}_{N_{\mathrm{kry}}}\},$$

où l'espace de dimension finie $\mathbb{K}^{\mathrm{prec,red}}_{N_{\mathrm{kry}}}$ *est l'image de* $\{\mathbb{K}^{\mathrm{prec,red}}_{N_{\mathrm{kry}}}\}$ *à travers (3.14). L'espace* $\{\mathbb{K}^{\mathrm{prec,red}}_{N_{\mathrm{kry}}}\}$ *est défini par*

$$\begin{aligned}\{\mathbb{K}^{\mathrm{prec,red}}_{N_{\mathrm{kry}}}\} &:= \mathrm{span}\Big(\{\mathbf{F}_{\mathrm{red}}, \mathbf{A}_{\mathrm{red}}\mathbf{F}_{\mathrm{red}}, \ldots, \mathbf{A}_{\mathrm{red}}^{N_{\mathrm{kry}}-1}\mathbf{F}_{\mathrm{red}}\}\Big)\\ &= \mathrm{span}\Big(\{\mathbf{F}_{\mathrm{red}}, \mathbf{N}_{\mathrm{red}}\mathbf{F}_{\mathrm{red}}, \ldots, \mathbf{N}_{\mathrm{red}}^{N_{\mathrm{kry}}-1}\mathbf{F}_{\mathrm{red}}\}\Big).\end{aligned}$$

En utilisant le solveur GMRES pour résoudre le Problème 20, nous avons le coût mémoire suivant

$$\mathrm{MEM}^{\mathrm{GMRES}}_{\mathrm{des,red}} := \big(N_{\mathrm{kry}}\times N_{\mathrm{kry}} + N_{\mathrm{kry}}\times\#\mathrm{ddl}_{\mathrm{red}} + 5\times\#\mathrm{ddl}_{\mathrm{red}} + \mathrm{MEM}^{\mathbf{A}}_{\mathrm{des,red}}\big)\times 16, \tag{4.7}$$

où, similairement à (4.4), nous avons

$$\mathrm{MEM}^{\mathbf{A}}_{\mathrm{des,red}} := 3\times 6\times N^2_{\mathrm{red}}.$$

Le terme dominant est maintenant $N_{\mathrm{kry}} \times \#\mathrm{ddl}_{\mathrm{red}}$ au lieu de $N_{\mathrm{kry}} \times \#\mathrm{ddl}$. La valeur de

$\mathrm{MEM}_{\mathrm{des,red}}^{\mathrm{GMRES}}$ vis à vis de $\mathrm{MEM}_{\mathrm{des}}^{\mathrm{GMRES}}$ dépend clairement de la valeur seuil ε, qui détermine le nombre réduit de fonctions de base N_{red}. Cependant, ε doit être choisi de telle sorte à diminuer le coût mémoire de la méthode sans pour autant dégrader la solution numérique obtenue. C'est pourquoi nous ne présentons pas les courbes mémoire de $\mathrm{MEM}_{\mathrm{des,red}}^{\mathrm{GMRES}}$ lorsque la taille $\mathcal{D}_\Omega$ du domaine augmente. Il s'agit surtout de trouver un bon compromis pour améliorer le conditionnement du système tout en maintenant une précision suffisante pour la solution numérique.

4.2.2 Impact du nombre de fonctions de base sur la solution numérique

En filtrant la base d'ondes planes par l'intermédiaire du produit hermitien $\mathbf{M}$, nous utilisons localement un nombre réduit de fonctions de base, noté N_{red}. Cette quantité dépend de la valeur de ε et de la taille de chaque élément h, voir les Tableaux 4.5 et 4.6.

h \ ε	10^{-6}	10^{-5}	10^{-4}	10^{-3}	10^{-2}	10^{-1}	1
0.25	52	46	36	30	16	6	0
0.5	52	52	52	46	42	30	6
1	52	52	52	52	52	52	42

TABLE 4.5 – Valeurs de N_{red} en fonction de ε et h, lors de la réduction d'une base contenant $N = 52$ ondes planes.

h \ ε	10^{-16}	10^{-15}	10^{-13}	10^{-11}	10^{-9}	10^{-7}	10^{-5}	10^{-4}	10^{-3}	10^{-2}
0.25	180	175	154	126	96	70	48	36	30	16
0.5	196	196	190	186	174	132	96	84	70	48
1	196	196	196	196	196	190	180	174	148	114

TABLE 4.6 – Valeurs de N_{red} en fonction de ε et h, lors de la réduction d'une base contenant $N = 196$ ondes planes.

Les valeurs propres sont représentées dans les Figures 4.5 et 4.6 pour $N = 52$ et $N = 196$ respectivement. Comme nous pouvons le voir elles dépendent de la taille h de chaque élément du maillage. Grâce à ces figures nous voyons rapidement qu'une troncature à un seuil ε donné induira une diminution du nombre de fonctions de base.

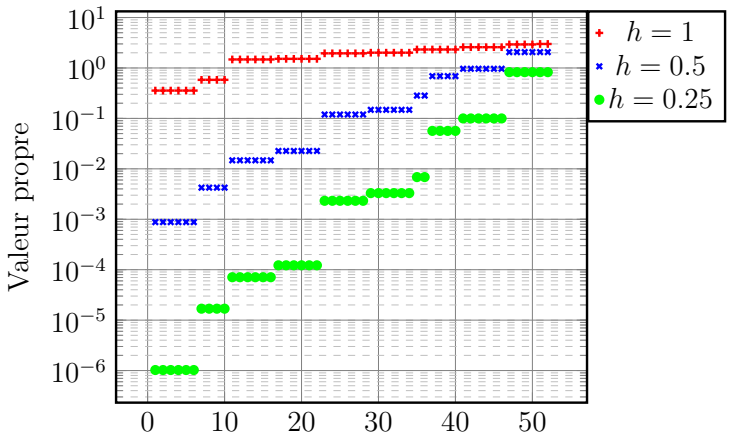

FIGURE 4.5 – Valeurs propres classées par ordre croissant en fonction du pas de maillage h utilisé pour N=52.

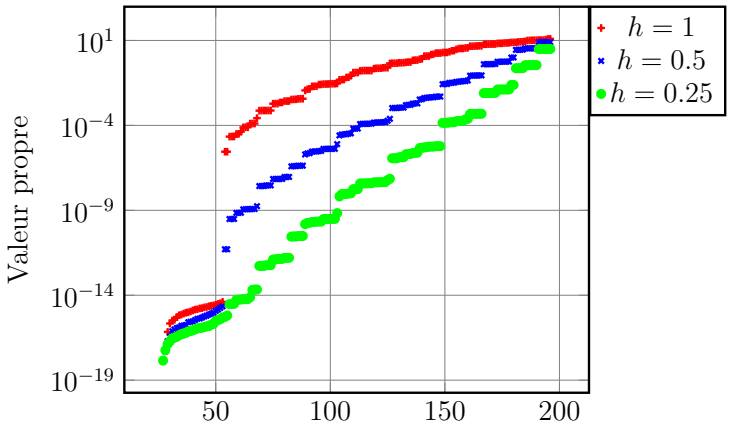

FIGURE 4.6 – Valeurs propres classées par ordre croissant en fonction du pas de maillage h utilisé pour N=196.

Avec la stratégie de réduction, le coût mémoire de la méthode de GMRES désassemblée réduite, voir (4.7), dépend majoritairement de la valeur de $N_{\text{kry}} \times \#\text{ddl}_{\text{red}}$. En effet, dans l'espace de Krylov, nous avons N_{kry} vecteurs de la même taille que $\mathbf{F}_{\text{red}} \in \mathbb{C}^{\#\text{ddl}_{\text{red}}}$. Pour une grande scène de calcul, $\mathcal{D}_\Omega = 200\lambda$ par exemple, les coûts de stockage (en Go) du second membre sont donnés à titre indicatif en Tableau 4.7 en fonction de N_{red}.

$N = 196$	$N = 52$	N_{red}	175	96	48	46	36	30	24	16	6
25.1	6.66	$\mathbf{F}_{\text{red}}$	22.4	12.3	6.14	5.89	4.61	3.84	3.07	2.05	0.77

TABLE 4.7 – Coût mémoire (en Go) pour le stockage du vecteur $\mathbf{F}$ et le vecteur réduit $\mathbf{F}_{\text{red}}$ en fonction de N_{red}, dans le cas où $\mathcal{D}_\Omega = 200\lambda$ (*ie* $\#\text{elem} = 8 \times 10^6$ si $h = 1$).

Nous nous proposons désormais de regarder l'impact du nombre de fonctions de base

sur la solution numérique. Pour cela, nous simulons un dipôle électromagnétique défini par (3.29), placé à l'avant du domaine, *ie* $\mathbf{x}_0 = (9.5, -20, 9.5)$. Celui-ci se propage dans un domaine $\Omega = [0, 20]^3$ où nous plaçons trois tiges pour lesquelles nous imposons une condition de métal parfait sur leur bord, *ie* $(\gamma_t \mathbf{E})_{|F} = 0$. Lors d'une simulation où la taille des éléments est $h = 1$ et où $N = 52$, voir la Figure 4.7, nous observons clairement des discontinuités aux interfaces entre les éléments. De plus, en comparant avec les Figures 4.8 et 4.9, nous remarquons que nous perdons beaucoup d'information en maillant si grossièrement. En effet, un maillage avec $h = 0.25$ (Figure 4.9) donne une visualisation nettement meilleure au sens où la solution numérique ne présente pas de discontinuités apparentes et où le comportement de l'onde électromagnétique paraît plausible compte tenu de la configuration étudiée.

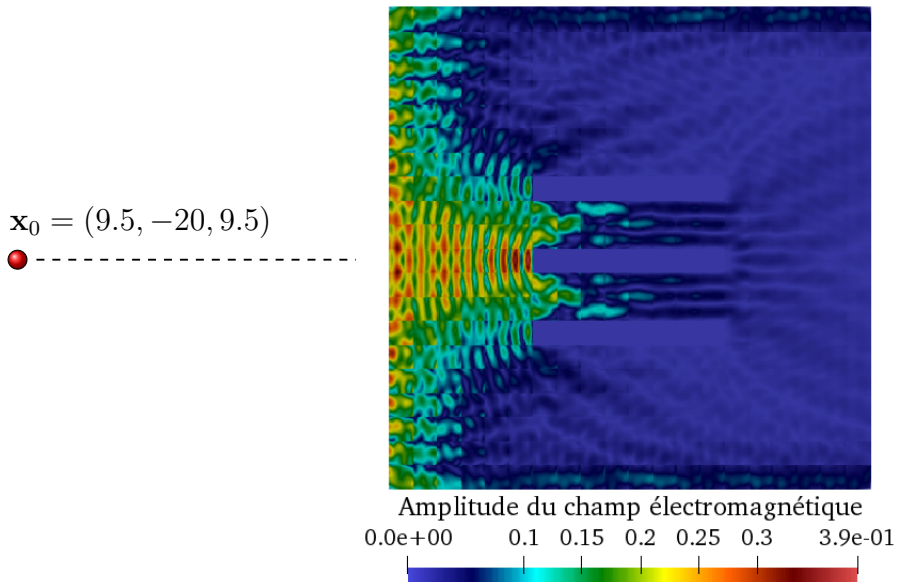

FIGURE 4.7 – Coupe perpendiculaire à l'axe x où l'amplitude du champ électromagnétique associé à un dipôle électromagnétique est représentée, dans le cas où $h = 1$ et où les trois tiges sont des objets parfaitement conducteurs.

Par conséquent, nous étudions, dans le cas où $h = 0.25$, la convergence de la solution numérique du Problème 20 vis à vis du paramètre ε. Afin de comparer la solution numérique avec la solution exacte $\mathbb{E}^{\text{ex}}$, nous simulons un dipôle électromagnétique, toujours défini par (3.29). Nous prenons un domaine $\Omega = [0, 5]^3$ sans obstacle à l'intérieur, maillé avec $h = 0.25$, et où $R_{\partial\Omega} = 0$. Ce domaine est de taille petite face à ce que nous pouvons considérer avec le solveur Trefftz. Néanmoins, cela va nous permettre de comparer, à ε fixé, les précisions de

— la solution $\mathbb{E}^{\varepsilon,\text{LU}}_{\text{red}}$, obtenue à partir d'une factorisation LU,

— la solution obtenue par une méthode de GMRES, notée $\mathbb{E}^{\varepsilon,\text{GMRES}}_{\text{red}}$, convergée à $e^{\text{prec}}_{N_{\text{kry}}} = 10^{-12}$ (voir la définition de $e^{\text{prec}}_{N_{\text{kry}}}$ en (3.27)) et pour $N_{\text{kry}} = 100$.

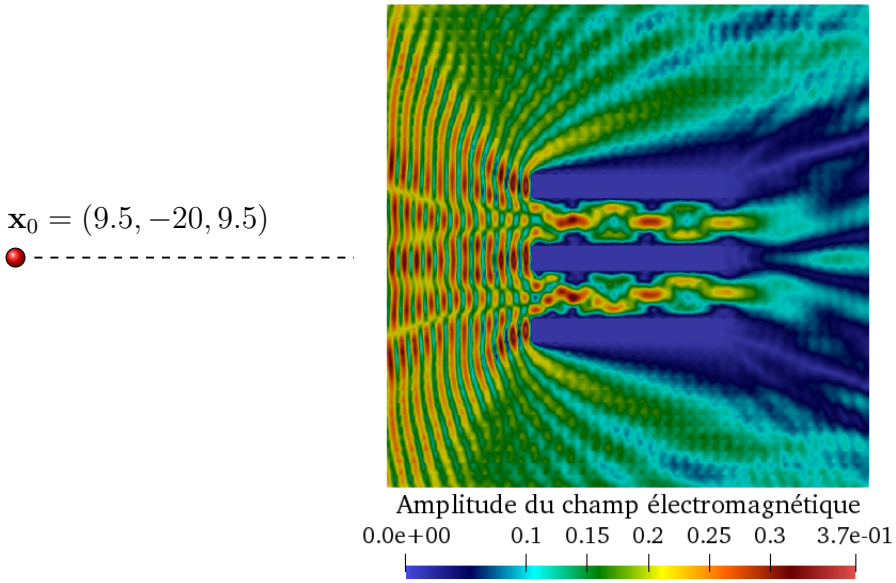

FIGURE 4.8 – Coupe perpendiculaire à l'axe x où l'amplitude du champ électromagnétique associé à un dipôle électromagnétique est représentée, dans le cas où $h = 0.5$ et où les trois tiges sont des objets parfaitement conducteurs.

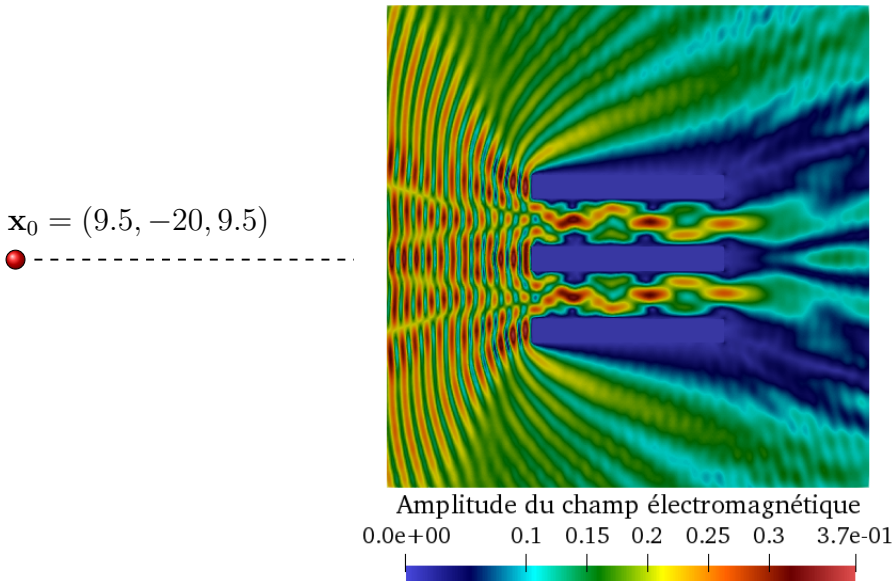

FIGURE 4.9 – Coupe perpendiculaire à l'axe x où l'amplitude du champ électromagnétique associé à un dipôle électromagnétique est représentée, dans le cas où $h = 0.25$ et où les trois tiges sont des objets parfaitement conducteurs.

Les erreurs relatives entre la solution analytique $\mathbb{E}^{\text{ex}}$ et $\mathbb{E}^{\varepsilon,\text{GMRES}}_{\text{red}}$ (ou de manière équivalente

$\mathbb{E}_{\text{red}}^{\varepsilon,\mathsf{LU}}$), sont définies par

$$e_{\infty}^{\varepsilon} := \frac{\|\mathbb{E}_{\text{red}}^{\varepsilon,\text{ref}} - \mathbb{E}^{\text{ex}}\|_{\infty}}{\|\mathbb{E}^{\text{ex}}\|_{\infty}}, \quad \text{où } \|\mathbb{E}^{h}\|_{\infty} = \max_{i=1,27} \mathbb{E}^{h}(\mathbf{x}_i),$$

où $\mathbb{E}_{\text{red}}^{\varepsilon,\text{ref}} = \mathbb{E}_{\text{red}}^{\varepsilon,\text{GMRES}}$ ou $\mathbb{E}_{\text{red}}^{\varepsilon,\mathsf{LU}}$ et avec 27 points $(\mathbf{x}_i)_{i=1,27} \in \Omega$. Ces coordonnées sont astucieusement choisies de telle sorte à calculer la norme dans différentes zones du domaine Ω.

Les valeurs de e_{∞}^{ε} sont de l'ordre de 2% lorsque $\varepsilon \leq 10^{-4}$, voir les Figures 4.10 et 4.11. Cette erreur d'approximation est la même pour $\mathbb{E}_{\text{red}}^{\varepsilon,\mathsf{LU}}$ et pour $\mathbb{E}_{\text{red}}^{\varepsilon,\text{GMRES}}$. Ainsi, nous pouvons dire que nos solveurs, direct et itératif, limitent les erreurs d'arrondis.

Par ailleurs, plus ε devient grand, plus N_{red} devient petit. Cela mène à la fois à une réduction du temps de calcul et du coût mémoire, voir les Tableaux 4.8 et 4.9. Dans ces tableaux, nous élaborons un récapitulatif des résultats numériques obtenus pour calculer la solution LU $\mathbb{E}_{\text{red}}^{\varepsilon,\mathsf{LU}}$ et la solution GMRES $\mathbb{E}_{\text{red}}^{\varepsilon,\text{GMRES}}$ en fonction de ε.

Remarque 4.7. *Le Tableau 4.8 met à nouveau en avant les limites mémoire des solveurs directs. En effet, pour un problème non réduit (ie $\varepsilon = 10^{-6}$), la factorisation* LU *coûte en mémoire plus de 10^3 fois plus cher qu'avec une méthode de GMRES (voir le Tableau 4.9).*

ε	10^{-6}	10^{-5}	10^{-4}	10^{-3}	10^{-2}
N_{red}	52	46	36	30	16
Temps pour $\mathbb{E}_{\text{red}}^{\varepsilon,\mathsf{LU}}$ (s)	5.7624	4.5768	3.2300	1.9434	0.6205
Mémoire (Go)	108.711	86.996	51.772	35.016	9.855

TABLE 4.8 – Résultats numériques pour obtenir la solution de référence $\mathbb{E}_{\text{red}}^{\varepsilon,\mathsf{LU}}$ en fonction du seuil de réduction de base ε pour un domaine $\mathcal{D}_{\Omega} = 5\lambda$, $h = 0.25$ et $N = 52$.

Enfin, nous étudions pour ε fixé l'erreur relative Trefftz (définie par (3.30)) entre la solution GMRES convergée à $e_{N_{\text{kry}}}^{\text{prec}} = 10^{-12}$, $\mathbb{E}_{\text{red}}^{\varepsilon,\text{GMRES}}$, et la solution GMRES convergée à $e_{N_{\text{kry}}}^{\text{prec}} = 10^{-8}$, notée $\mathbb{E}_{\text{red}}^{\varepsilon}$, voir la Figure 4.12. Il s'avère que la réduction n'accélère que très légèrement la convergence du solveur GMRES, en termes d'itérations, sauf pour le cas $\varepsilon = 10^{-2}$ où nous utilisons moins d'itérations. Dans la suite, nous affinons notre étude dans le but de savoir si ce seuil permet d'obtenir une solution numérique précise.

Les gains mentionnés précédemment incitent à considérer des valeurs seuil ε grandes. Toutefois, nous devons trouver un bon compromis afin de réduire les coûts tout en préservant la précision de la solution numérique. Compte tenu de la courbe d'erreur relative avec la solution exacte du dipôle, en Figure 4.11, nous pensons qu'un choix raisonnable est $\varepsilon = 10^{-4}$,

ε	10^{-6}	10^{-5}	10^{-4}	10^{-3}	10^{-2}
N_{red}	52	46	36	30	16
$\#\text{ddl}_{\text{red}}$	6500	5750	4500	3750	2000
Temps pour $\mathbb{E}_{\text{red}}^{\varepsilon,\text{GMRES}}$ (s)	7189	5594	3420	2364	614
Mémoire (Go)	0.023	0.019	0.013	0.010	0.004
e_{∞}^{ε}	1.6×10^{-2}	1.6×10^{-2}	2.2×10^{-2}	4.1×10^{-2}	0.23

TABLE 4.9 – Résultats numériques pour la solution de référence $\mathbb{E}_{\text{red}}^{\varepsilon,\text{GMRES}}$ (convergence à $e_{N_{\text{kry}}}^{\text{prec}} = 10^{-12}$, avec $N_{\text{kry}} = 100$), en fonction du seuil de réduction de base ε pour un domaine $\mathcal{D}_\Omega = 5\lambda$, $h = 0.25$ et $N = 52$.

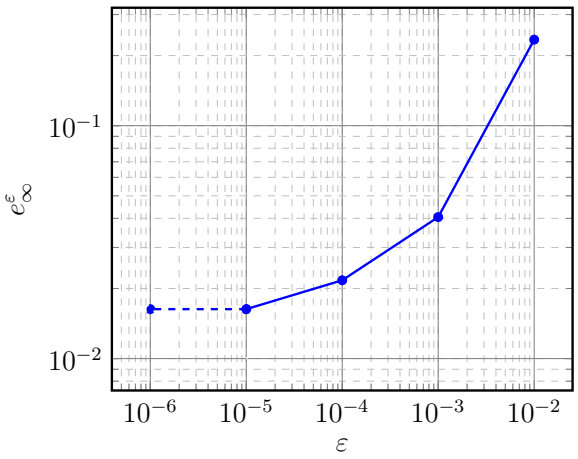

FIGURE 4.10 – Erreur relative infinie entre la solution obtenue par une factorisation LU, notée $\mathbb{E}_{\text{red}}^{\varepsilon,\text{LU}}$, et la solution exacte du champ généré par un dipôle électromagnétique localisé en $\mathbf{x}_0 = (2.5, -0.5, 2.5)$, $\mathcal{D}_\Omega = 5\lambda$, $h = 0.25$, $R_{\partial\Omega} = 0$.

pour lequel nous obtenons $e_{\infty}^{\varepsilon} \approx 2\%$ d'erreur, voir le Tableau 4.9. Nous confirmons aussi ce choix visuellement. Les Figures 4.17, 4.18, 4.19 et 4.20 représentent, pour différentes valeurs de ε, l'amplitude du champ du dipôle électromagnétique et la composante x du champ électrique $\mathbf{E}_{\text{red}}^{\varepsilon,\text{GMRES}}$. L'onde se propage vers la droite dans Ω sans obstacle. Nous comparons ces représentations à celle de la solution non réduite (prise comme référence) en Figure 4.21. Les visualisations conformes à cette dernière sont celles associées à $\varepsilon \leq 10^{-4}$ (Figures 4.19 et 4.20). Ainsi, bien que le solveur GMRES converge pour $\varepsilon = 10^{-2}$, voir la Figure 4.12, nous écartons cette possibilité visuellement (Figure 4.17) et aussi grâce à son erreur relative avec la solution analytique ($e_{\infty}^{\varepsilon} = 23\%$ dans le Tableau 4.9 ou dans la Figure 4.11).

Maintenant, nous étudions le cas d'une stratégie de réduction appliquée à un problème UWVF préconditionné résolu par GMRES où $N >> 52$. Par exemple, nous choisissons $N =$

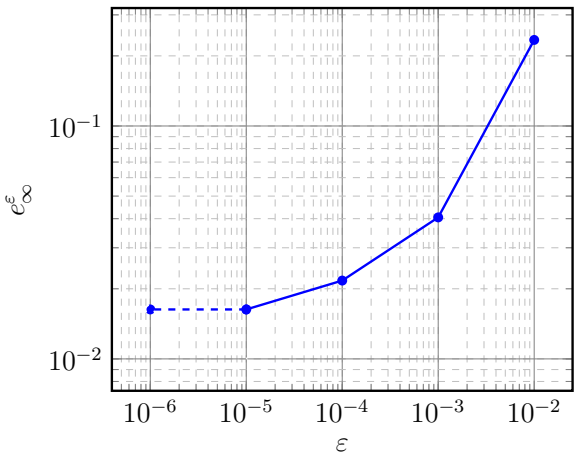

FIGURE 4.11 – Erreur relative infinie entre la solution obtenue par le solveur GMRES, notée $\mathbb{E}_{\text{red}}^{\varepsilon,\text{GMRES}}$, et la solution exacte du champ généré par un dipôle électromagnétique localisé en $\mathbf{x}_0 = (2.5, -0.5, 2.5)$, pour $N = 52$, $\mathcal{D}_\Omega = 5\lambda$, $h = 0.25$, $R_{\partial\Omega} = 0$.

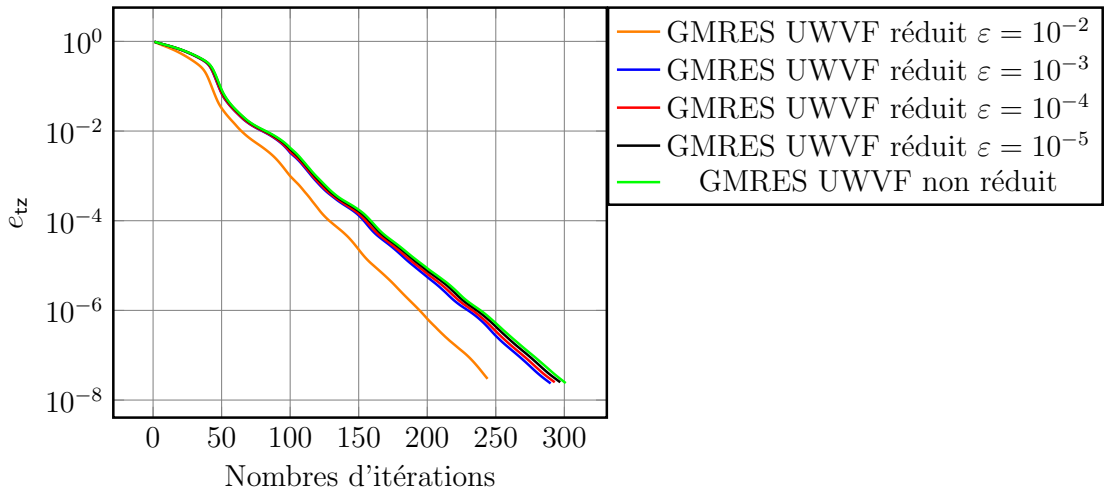

FIGURE 4.12 – Erreur relative Trefftz entre la solution GMRES réduite $\mathbb{E}_{\text{red}}^{\varepsilon}$ et la solution de référence $\mathbb{E}_{\text{red}}^{\varepsilon,\text{GMRES}}$, convergée à $e_{N_{\text{kry}}}^{\text{prec}} = 10^{-12}$ pour $\mathcal{D}_\Omega = 5\lambda$, $h = 0.25$, $R_{\partial\Omega} = 0$ et $N = 52$.

196 ondes planes. Dans ce cas, l'algorithme de GMRES préconditionné non réduit ne converge pas, voir la Figure 4.13. Cette courbe met en avant les problèmes de conditionnement de la méthode lorsque trop d'ondes planes sont considérées face à la taille h de chaque élément. Il est alors nécessaire d'appliquer une stratégie de réduction de base.

De la même façon que précédemment nous analysons l'effet des valeurs de ε sur :

— la précision de la solution GMRES $\mathbb{E}_{\text{red}}^{\varepsilon,\text{GMRES}}$, convergée à $e_{N_{\text{kry}}}^{\text{prec}} = 10^{-12}$ dans le cas $N = 196$ et $N_{\text{kry}} = 100$, par rapport à la solution exacte $\mathbb{E}^{\text{ex}}$,

— la vitesse de convergence du GMRES (en termes d'itérations et de temps d'exécution),

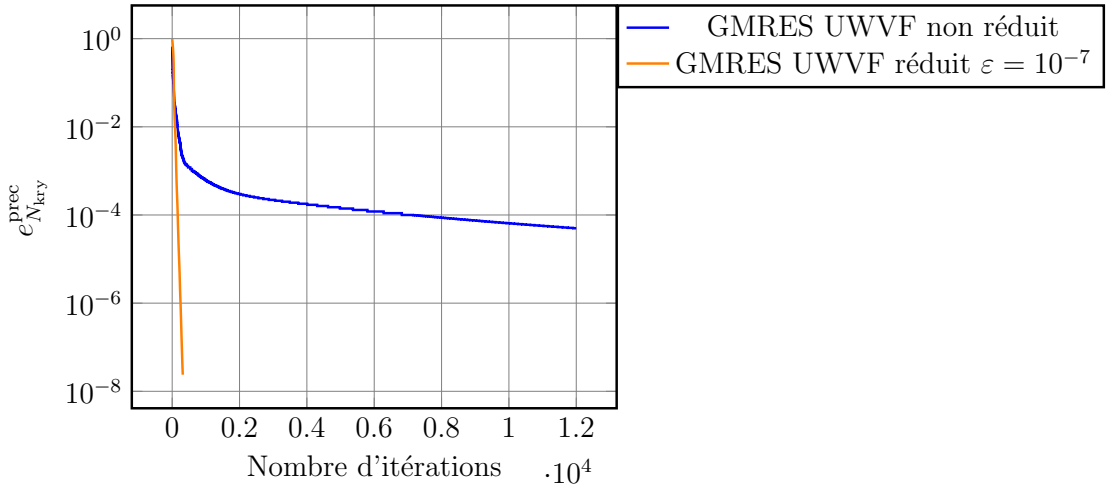

FIGURE 4.13 – Convergence des solutions numériques GMRES UWVF préconditionnées : non réduite et pour $\varepsilon = 10^{-7}$, pour $N = 196$, $\mathbf{x}_0 = (2.5, -0.5, 2.5)$, $\mathcal{D}_\Omega = 5\lambda$, $h = 0.25$, $R_{\partial\Omega} = 0$.

— la mémoire utilisée pour obtenir la solution numérique $\mathbb{E}^{\varepsilon}_{\text{red}}$ convergée à $e^{\text{prec}}_{N_{\text{kry}}} = 10^{-8}$ (ou de manière équivalente $\mathbb{E}^{\varepsilon,\text{GMRES}}_{\text{red}}$).

Tout d'abord, la solution GMRES converge vers la solution exacte, voir la Figure 4.14. Nous observons que l'erreur e^{ε}_{∞} associée au ε le plus petit est nettement plus faible (de l'ordre de 10^{-7}) que lorsque nous avons considéré $N = 52$ (de l'ordre de 10^{-2}), voir la Figure 4.11. En ce sens, nous pouvons dire qu'il semble préférable d'utiliser une taille plus grande pour la base d'ondes planes, en sélectionnant ensuite les meilleurs modes grâce à la réduction. Ainsi, nous ferons face à moins d'erreurs d'arrondis.

Par ailleurs, la vitesse de convergence de la solution $\mathbb{E}^{\varepsilon}_{\text{red}}$, en termes d'itérations, est accélérée de l'ordre de 15% entre la solution à $\varepsilon = 10^{-3}$ et celle à $\varepsilon = 10^{-16}$, voir la Figure 4.15, où nous avons calculé l'erreur Trefftz entre $\mathbb{E}^{\varepsilon}_{\text{red}}$ et $\mathbb{E}^{\varepsilon,\text{GMRES}}_{\text{red}}$ (respectivement convergées à $e^{\text{prec}}_{N_{\text{kry}}} = 10^{-8}$ et $e^{\text{prec}}_{N_{\text{kry}}} = 10^{-12}$).

Maintenant, nous allons observer les gains en termes de temps de calcul et de coût mémoire, lorsque la valeur de ε devient grande. Nous choisissons comme valeur de référence $\varepsilon = 10^{-16}$, voir Figure 4.30. Ce choix est discutable vis à vis de l'erreur d'arrondi mais la Figure 4.14 montre la fiabilité de la solution même pour ce seuil de troncature. Nous comparons sa visualisation à celles obtenues pour toutes les autres valeurs de ε considérées, voir les Figures 4.22, 4.23, 4.24, 4.25, 4.26, 4.27, 4.28, 4.29. Dans cette configuration, nous choisissons $\varepsilon = 10^{-5}$ comme compromis judicieux à adopter pour le seuil de réduction. En effet, nous observons visuellement une différence avec la solution de référence dès que $\varepsilon < 10^{-5}$. Un bilan des résultats numériques pour chacune des solutions $\mathbb{E}^{\varepsilon,\text{GMRES}}_{\text{red}}$ est effectué dans les Tableaux 4.10 et 4.11. Nous remarquons que nous avons une erreur $e^{\varepsilon}_{\infty} > 2\%$ dès que

$\varepsilon \geq 10^{-4}$. En confrontant les résultats pour $\varepsilon = 10^{-16}$ et pour $\varepsilon = 10^{-5}$, nous avons réduit de 94% le temps de calcul et de 75% la mémoire utilisée pour calculer la solution numérique à $\varepsilon = 10^{-5}$.

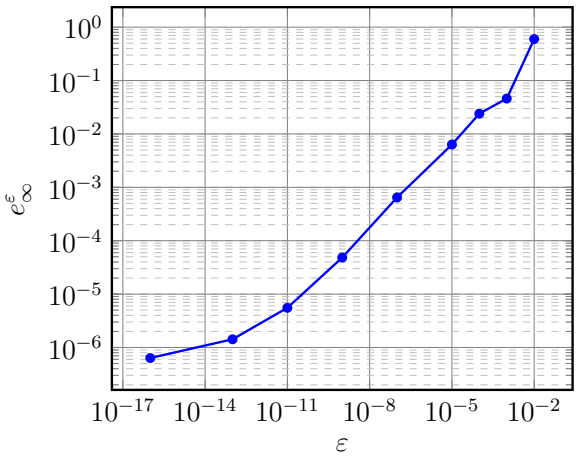

FIGURE 4.14 – Erreur relative infinie entre la solution de référence obtenue par le solveur GMRES UWVF préconditionné réduit, notée $\mathbb{E}^{\varepsilon,\mathrm{GMRES}}_{\mathrm{red}}$, et la solution exacte du dipôle électromagnétique, pour différentes valeurs de ε, pour $N = 196$, $\mathbf{x}_0 = (2.5, -0.5, 2.5)$, $\mathcal{D}_\Omega = 5\lambda$, $h = 0.25$, $R_{\partial\Omega} = 0$.

ε	10^{-16}	10^{-13}	10^{-11}	10^{-9}
N_{red}	180	154	126	96
$\#\mathrm{ddl}_{\mathrm{red}}$ ($\times 10^3$)	1440	1232	1008	768
Temps pour $\mathbb{E}^{\varepsilon}_{\mathrm{red}}$ (s)	4121	2936	1888	1085
Mémoire (Go)	0.151	0.129	0.105	0.080
e^{ε}_{∞}	6.35×10^{-7}	1.42×10^{-6}	5.53×10^{-6}	4.83×10^{-5}

TABLE 4.10 – Résultats numériques pour la solution GMRES UWVF réduite $\mathbb{E}^{\varepsilon,\mathrm{GMRES}}_{\mathrm{red}}$ (convergence à $e^{\mathrm{prec}}_{N_{\mathrm{kry}}} = 10^{-12}$, avec $N_{\mathrm{kry}} = 100$), en fonction du seuil de réduction de base ε pour un domaine $\mathcal{D}_\Omega = 5\lambda$, $h = 0.25$ et $N = 196$.

Nous avons aussi testé l'effet de la réduction sur des cas avec obstacles et pour $N = 196$. Nous doublons la taille du domaine étudié, *ie* $\Omega = [0, 10]^3$, dans lequel nous plaçons un gobelet parfaitement métallique. Nous choisissons toujours $h = 0.25$ et $R_{\partial\Omega} = 0$. Cette fois-ci nous prenons comme solution de référence celle obtenue avec $\varepsilon = 10^{-7}$. Effectivement, comme nous avons pu le constater, ce seuil de troncature donne une solution numérique précise (voir la Figure 4.14). Nous obtenons des résultats semblables à la configuration sans obstacle, au sens où un bon compromis paraît être ici $\varepsilon = 10^{-4}$ et entraîne une diminution du temps de calcul et de la mémoire utilisée, voir le Tableau 4.12 et la Figure 4.16.

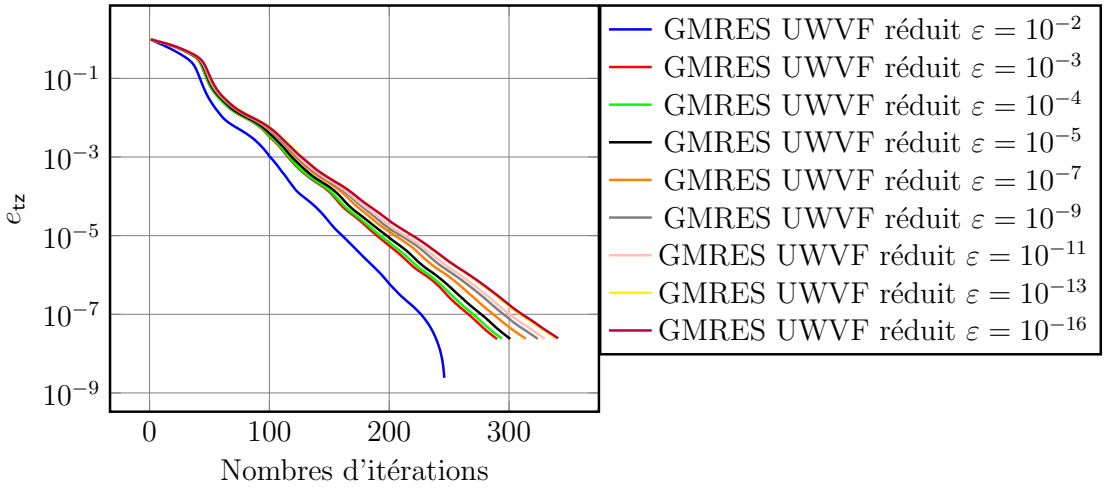

FIGURE 4.15 – Erreur relative Trefftz entre la solution GMRES réduite $\mathbb{E}_{\text{red}}^{\varepsilon}$ et la solution de référence $\mathbb{E}_{\text{red}}^{\varepsilon,\text{GMRES}}$ obtenue par un solveur GMRES convergé à $e_{N_{\text{kry}}}^{\text{prec}} = 10^{-12}$, pour $\mathcal{D}_\Omega = 5\lambda$, $h = 0.25$, $R_{\partial\Omega} = 0$ et $N = 196$.

ε	10^{-7}	10^{-5}	10^{-4}	10^{-3}	10^{-2}
N_{red}	70	48	36	30	16
$\#\text{ddl}_{\text{red}}$ ($\times 10^3$)	560	384	288	240	128
Temps pour $\mathbb{E}_{\text{red}}^{\varepsilon}$ (s)	496.1	254.9	141.8	99.68	25.61
Mémoire (Go)	0.058	0.04	0.03	0.02	0.01
e_{∞}^{ε}	6.46×10^{-4}	6.33×10^{-3}	2.39×10^{-2}	4.59×10^{-2}	0.59

TABLE 4.11 – Résultats numériques pour la solution GMRES UWVF réduite $\mathbb{E}_{\text{red}}^{\varepsilon,\text{GMRES}}$ (convergence à $e_{N_{\text{kry}}}^{\text{prec}} = 10^{-12}$, avec $N_{\text{kry}} = 100$), en fonction du seuil de réduction de base ε pour un domaine $\mathcal{D}_\Omega = 5\lambda$, $h = 0.25$ et $N = 196$.

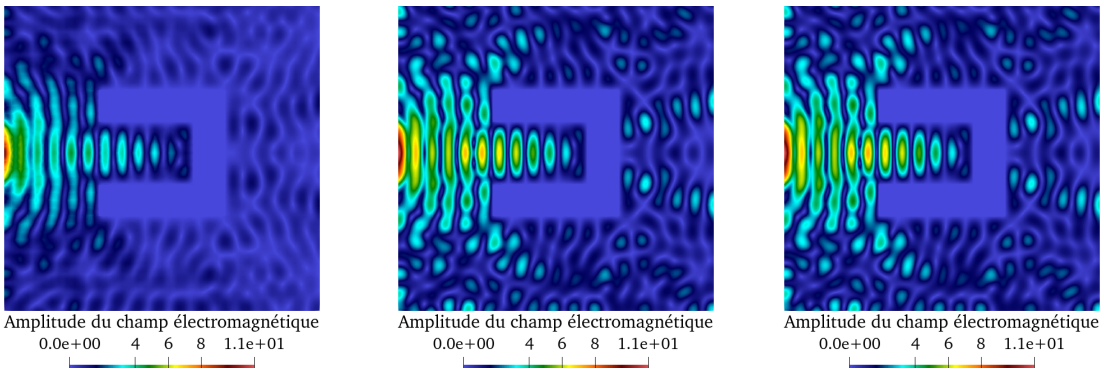

FIGURE 4.16 – Amplitude du champ électromagnétique se propageant dans un gobelet parfaitement métallique, pour $N = 196$ et pour (de gauche à droite) : $\varepsilon = 10^{-2}$, $\varepsilon = 10^{-4}$ et $\varepsilon = 10^{-7}$.

ε	10^{-7}	10^{-4}	10^{-2}
Mémoire (Go)	0.56	0.28	0.13
Temps (h)	55	11.3	0.58

TABLE 4.12 – Résultats numériques du cas avec gobelet associés à la Figure 4.16, pour $N = 196$.

Ces constats témoignent très clairement des gains obtenus en utilisant une stratégie de réduction de base. D'une part, dans le cas où $N = 196$, la stratégie est primordiale au sens où elle permet à la solution de converger. D'autre part, elle permet d'améliorer le conditionnement du problème de Trefftz en sélectionnant des fonctions de base appropriées pour décrire la solution numérique. Enfin, elle apporte aussi des gains en termes de temps de calcul et de mémoire du solveur GoTEM3.

Pour perfectionner cette stratégie, une idée serait de choisir des valeurs grandes de ε au début des itérations GMRES. Cela privilégierait l'aspect propagatif de l'onde, qui a besoin de moins d'ondes planes pour être décrit par la solution numérique. Au bout d'un certain nombre d'itérations (selon un critère de précision de la solution par exemple), nous sélectionnerions une valeur plus petite pour ε afin de traduire correctement les continuités dans la description de la solution numérique. Toutefois, nous n'avons pas investigué cette possibilité dans cette thèse et cela reste une perspective à exploiter.

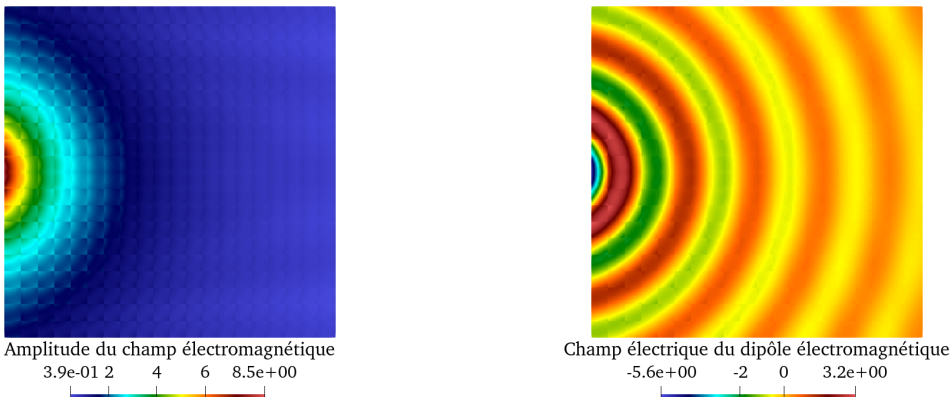

FIGURE 4.17 – Visualisation du champ électromagnétique généré par un dipôle, situé en $\mathbf{x}_0 = (2.5, -0.5, 2.5)$, associé à la solution numérique $\mathbb{E}_{\mathrm{red}}^{\varepsilon,\mathrm{GMRES}}$ pour $N = 52$ et $\varepsilon = 10^{-2}$; sa magnitude (à gauche) et la composante x de son champ électrique $\mathbf{E}_{\mathrm{red}}^{\varepsilon,\mathrm{GMRES}}$ (à droite) pour $\mathcal{D}_\Omega = 5\lambda$, $h = 0.25$ et $R_{\partial\Omega} = 0$.

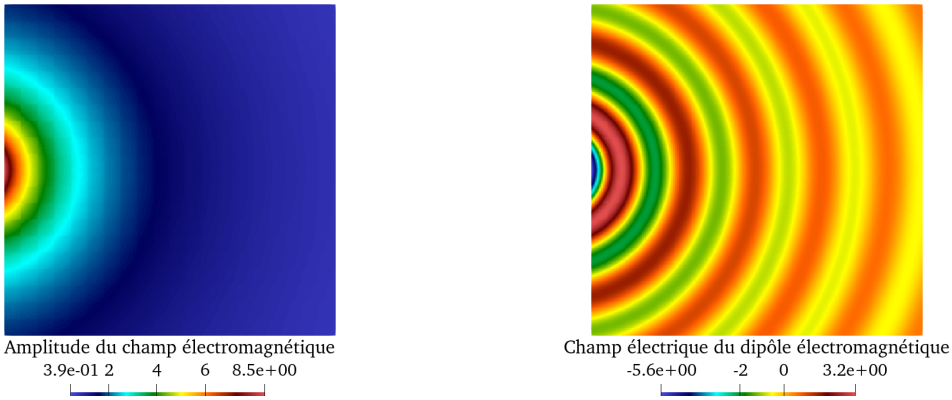

FIGURE 4.18 – Visualisation du champ électromagnétique généré par un dipôle, situé en $\mathbf{x}_0 = (2.5, -0.5, 2.5)$, associé à la solution numérique $\mathbb{E}_{\text{red}}^{\varepsilon,\text{GMRES}}$ pour $N = 52$ et $\varepsilon = 10^{-3}$; sa magnitude (à gauche) et la composante x de son champ électrique $\mathbf{E}_{\text{red}}^{\varepsilon,\text{GMRES}}$ (à droite) pour $\mathcal{D}_\Omega = 5\lambda$, $h = 0.25$ et $R_{\partial\Omega} = 0$.

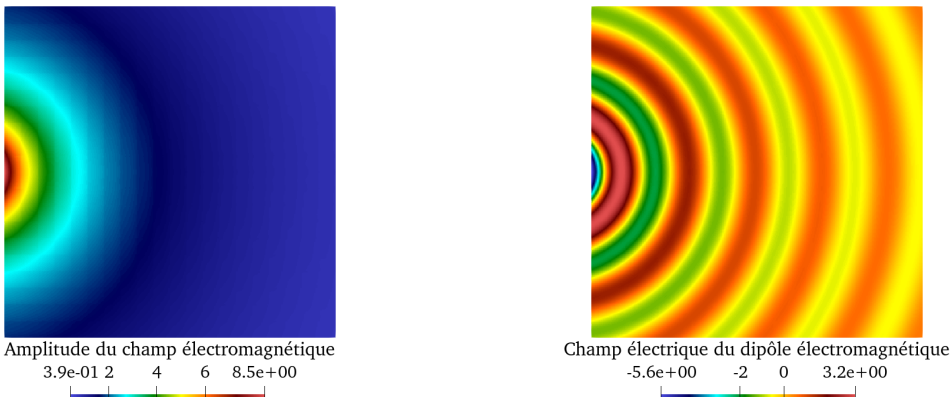

FIGURE 4.19 – Visualisation du champ électromagnétique généré par un dipôle, situé en $\mathbf{x}_0 = (2.5, -0.5, 2.5)$, associé à la solution numérique $\mathbb{E}_{\text{red}}^{\varepsilon,\text{GMRES}}$ pour $N = 52$ et $\varepsilon = 10^{-4}$; sa magnitude (à gauche) et la composante x de son champ électrique $\mathbf{E}_{\text{red}}^{\varepsilon,\text{GMRES}}$ (à droite) pour $\mathcal{D}_\Omega = 5\lambda$, $h = 0.25$ et $R_{\partial\Omega} = 0$.

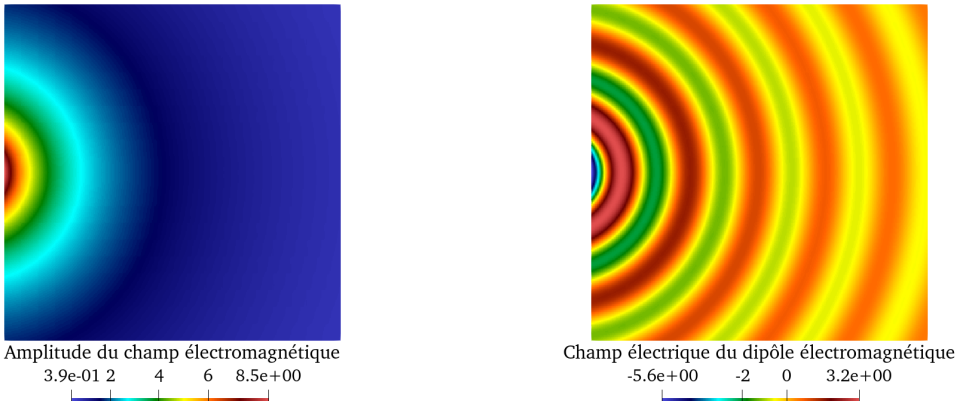

FIGURE 4.20 – Visualisation du champ électromagnétique généré par un dipôle, situé en $\mathbf{x}_0 = (2.5, -0.5, 2.5)$, associé à la solution numérique $\mathbb{E}_{\text{red}}^{\varepsilon,\text{GMRES}}$ pour $N = 52$ et $\varepsilon = 10^{-5}$; sa magnitude (à gauche) et la composante x de son champ électrique $\mathbf{E}_{\text{red}}^{\varepsilon,\text{GMRES}}$ (à droite) pour $\mathcal{D}_\Omega = 5\lambda$, $h = 0.25$ et $R_{\partial\Omega} = 0$.

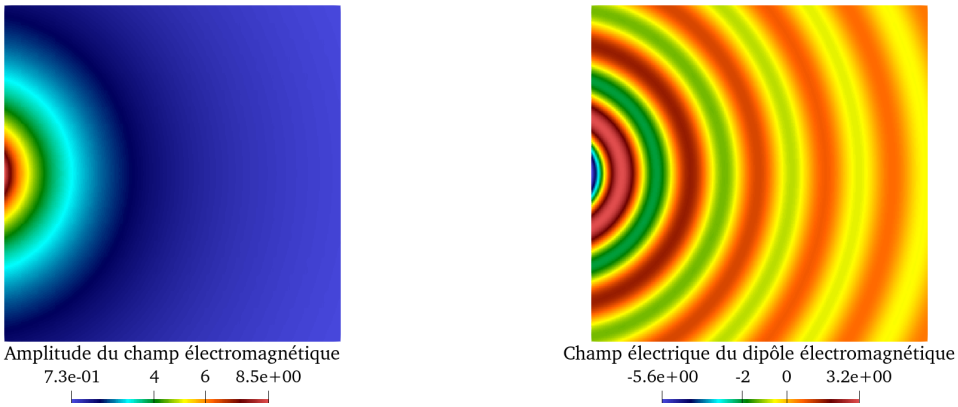

FIGURE 4.21 – Visualisation du champ électromagnétique généré par un dipôle, situé en $\mathbf{x}_0 = (2.5, -0.5, 2.5)$, associé à la solution numérique $\mathbb{E}_{\text{red}}^{\varepsilon,\text{GMRES}}$ pour $N = 52$ et $\varepsilon = 10^{-6}$; sa magnitude (à gauche) et la composante x de son champ électrique $\mathbf{E}_{\text{red}}^{\varepsilon,\text{GMRES}}$ (à droite) pour $\mathcal{D}_\Omega = 5\lambda$, $h = 0.25$ et $R_{\partial\Omega} = 0$.

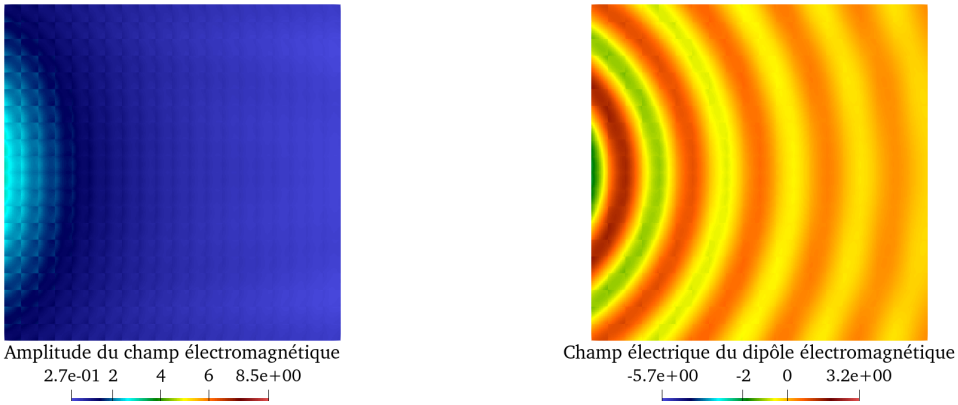

FIGURE 4.22 – Visualisation du champ électromagnétique généré par un dipôle, situé en $\mathbf{x}_0 = (2.5, -0.5, 2.5)$, associé à la solution numérique $\mathbb{E}_{\text{red}}^{\varepsilon,\text{GMRES}}$ pour $N = 196$ et $\varepsilon = 10^{-2}$; sa magnitude (à gauche) et la composante x de son champ électrique $\mathbf{E}_{\text{red}}^{\varepsilon,\text{GMRES}}$ (à droite) pour $\mathcal{D}_\Omega = 5\lambda$, $h = 0.25$ et $R_{\partial\Omega} = 0$.

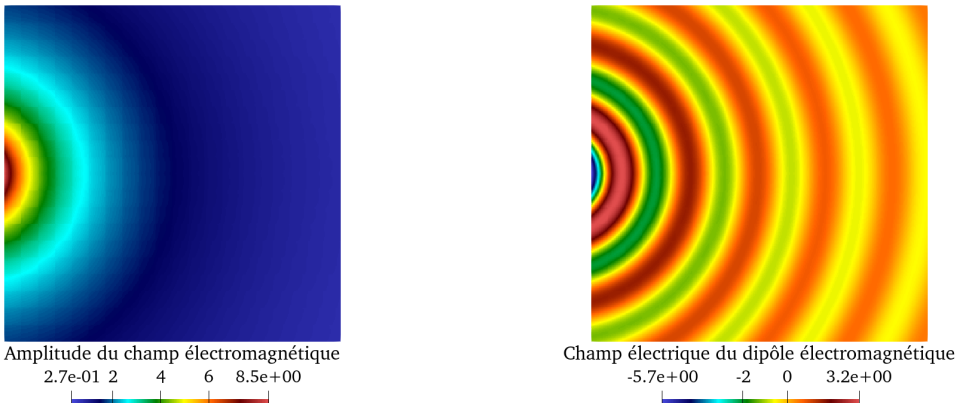

FIGURE 4.23 – Visualisation du champ électromagnétique généré par un dipôle, situé en $\mathbf{x}_0 = (2.5, -0.5, 2.5)$, associé à la solution numérique $\mathbb{E}_{\text{red}}^{\varepsilon,\text{GMRES}}$ pour $N = 196$ et $\varepsilon = 10^{-3}$; sa magnitude (à gauche) et la composante x de son champ électrique $\mathbf{E}_{\text{red}}^{\varepsilon,\text{GMRES}}$ (à droite) pour $\mathcal{D}_\Omega = 5\lambda$, $h = 0.25$ et $R_{\partial\Omega} = 0$.

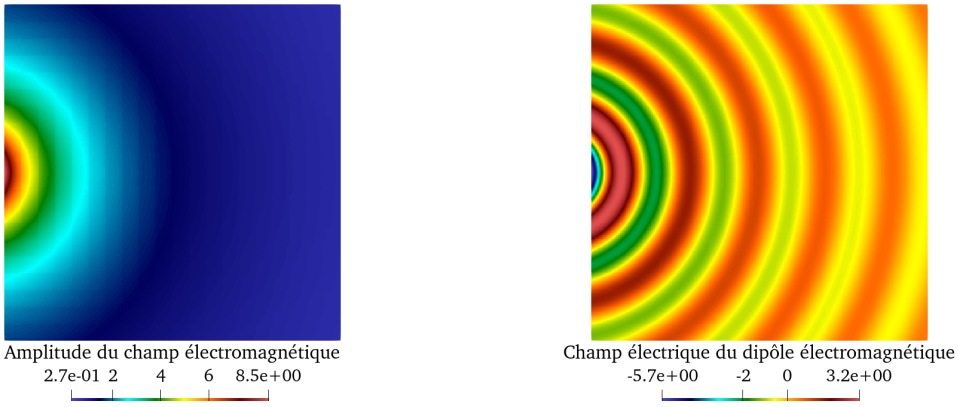

FIGURE 4.24 – Visualisation du champ électromagnétique généré par un dipôle, situé en $\mathbf{x}_0 = (2.5, -0.5, 2.5)$, associé à la solution numérique $\mathbb{E}_{\text{red}}^{\varepsilon,\text{GMRES}}$ pour $N = 196$ et $\varepsilon = 10^{-4}$; sa magnitude (à gauche) et la composante x de son champ électrique $\mathbf{E}_{\text{red}}^{\varepsilon,\text{GMRES}}$ (à droite) pour $\mathcal{D}_\Omega = 5\lambda$, $h = 0.25$ et $R_{\partial\Omega} = 0$.

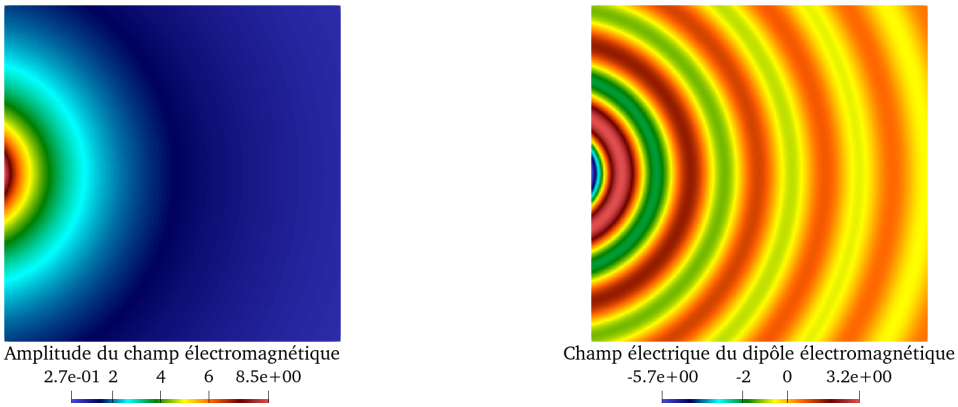

FIGURE 4.25 – Visualisation du champ électromagnétique généré par un dipôle, situé en $\mathbf{x}_0 = (2.5, -0.5, 2.5)$, associé à la solution numérique $\mathbb{E}_{\text{red}}^{\varepsilon,\text{GMRES}}$ pour $N = 196$ et $\varepsilon = 10^{-5}$; sa magnitude (à gauche) et la composante x de son champ électrique $\mathbf{E}_{\text{red}}^{\varepsilon,\text{GMRES}}$ (à droite) pour $\mathcal{D}_\Omega = 5\lambda$, $h = 0.25$ et $R_{\partial\Omega} = 0$.

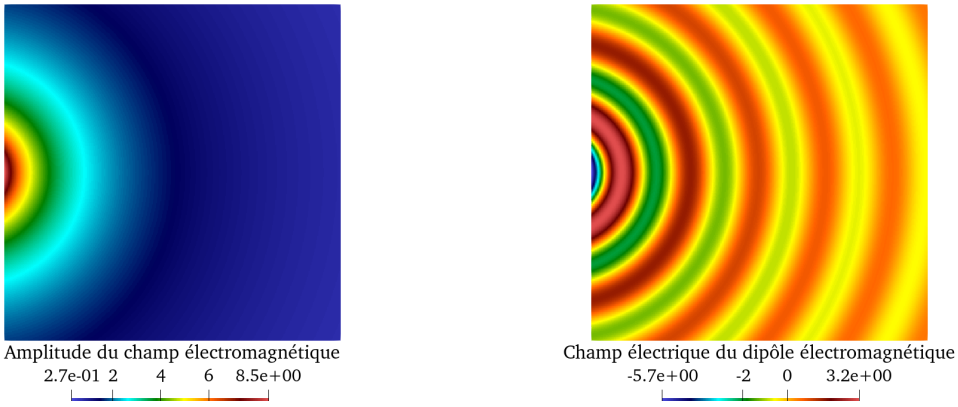

FIGURE 4.26 – Visualisation du champ électromagnétique généré par un dipôle, situé en $\mathbf{x}_0 = (2.5, -0.5, 2.5)$, associé à la solution numérique $\mathbb{E}_{\text{red}}^{\varepsilon,\text{GMRES}}$ pour $N = 196$ et $\varepsilon = 10^{-7}$; sa magnitude (à gauche) et la composante x de son champ électrique $\mathbf{E}_{\text{red}}^{\varepsilon,\text{GMRES}}$ (à droite) pour $\mathcal{D}_\Omega = 5\lambda$, $h = 0.25$ et $R_{\partial\Omega} = 0$.

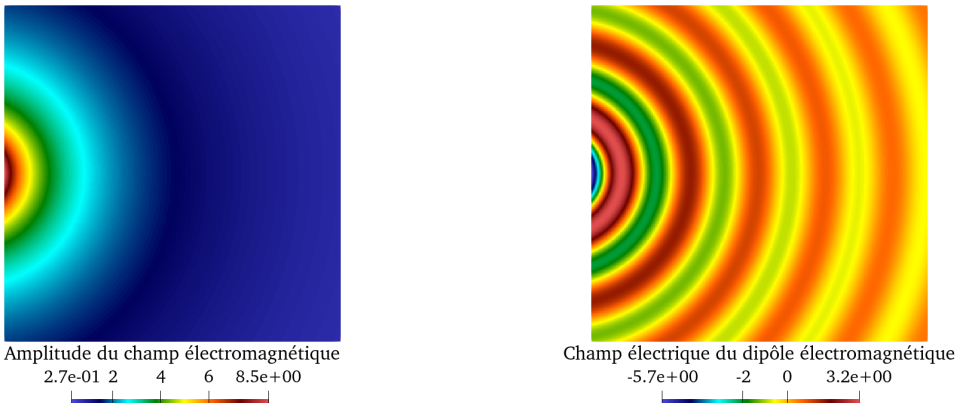

FIGURE 4.27 – Visualisation du champ électromagnétique généré par un dipôle, situé en $\mathbf{x}_0 = (2.5, -0.5, 2.5)$, associé à la solution numérique $\mathbb{E}_{\text{red}}^{\varepsilon,\text{GMRES}}$ pour $N = 196$ et $\varepsilon = 10^{-9}$; sa magnitude (à gauche) et la composante x de son champ électrique $\mathbf{E}_{\text{red}}^{\varepsilon,\text{GMRES}}$ (à droite) pour $\mathcal{D}_\Omega = 5\lambda$, $h = 0.25$ et $R_{\partial\Omega} = 0$.

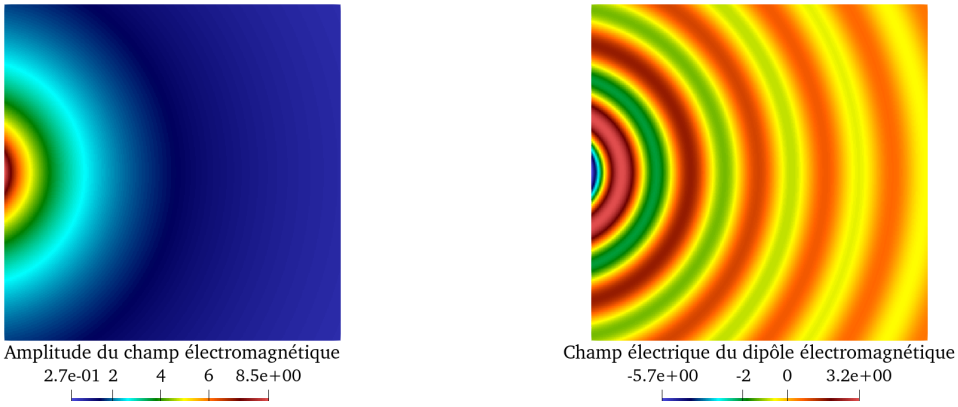

FIGURE 4.28 – Visualisation du champ électromagnétique généré par un dipôle, situé en $\mathbf{x}_0 = (2.5, -0.5, 2.5)$, associé à la solution numérique $\mathbb{E}_{\text{red}}^{\varepsilon,\text{GMRES}}$ pour $N = 196$ et $\varepsilon = 10^{-11}$; sa magnitude (à gauche) et la composante x de son champ électrique $\mathbf{E}_{\text{red}}^{\varepsilon,\text{GMRES}}$ (à droite) pour $\mathcal{D}_\Omega = 5\lambda$, $h = 0.25$ et $R_{\partial\Omega} = 0$.

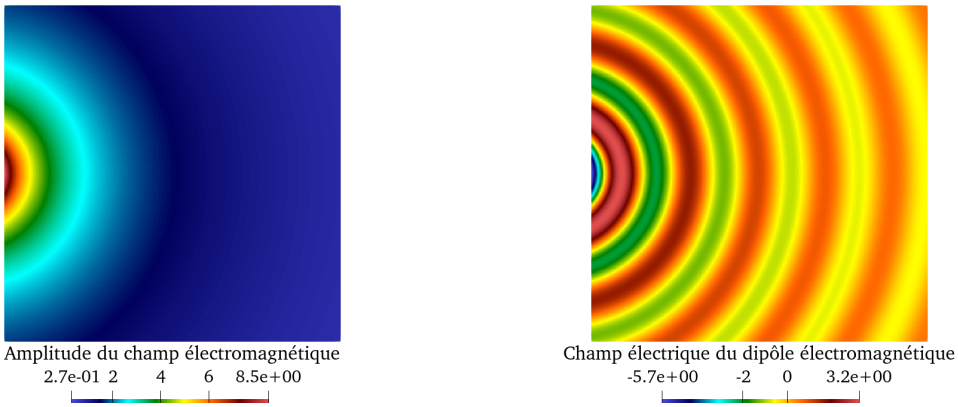

FIGURE 4.29 – Visualisation du champ électromagnétique généré par un dipôle, situé en $\mathbf{x}_0 = (2.5, -0.5, 2.5)$, associé à la solution numérique $\mathbb{E}_{\text{red}}^{\varepsilon,\text{GMRES}}$ pour $N = 196$ et $\varepsilon = 10^{-13}$; sa magnitude (à gauche) et la composante x de son champ électrique $\mathbf{E}_{\text{red}}^{\varepsilon,\text{GMRES}}$ (à droite) pour $\mathcal{D}_\Omega = 5\lambda$, $h = 0.25$ et $R_{\partial\Omega} = 0$.

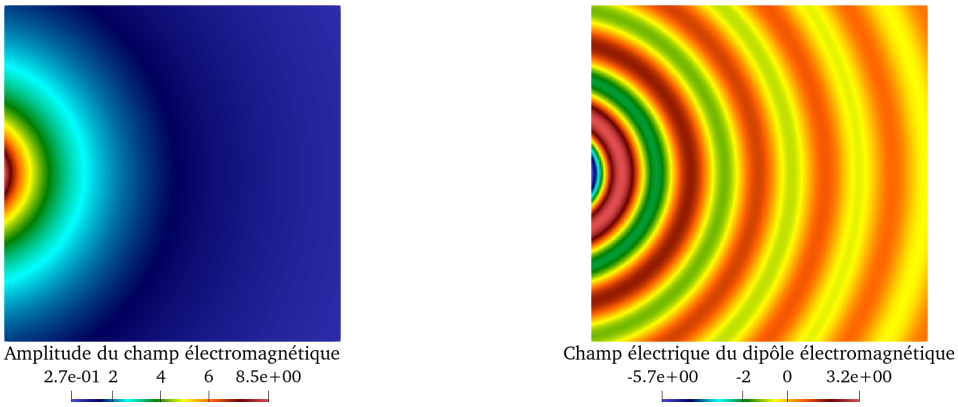

FIGURE 4.30 – Visualisation du champ électromagnétique généré par un dipôle, situé en $\mathbf{x}_0 = (2.5, -0.5, 2.5)$, associé à la solution numérique $\mathbb{E}_{\text{red}}^{\varepsilon,\text{GMRES}}$ pour $N = 196$ et $\varepsilon = 10^{-16}$; sa magnitude (à gauche) et la composante x de son champ électrique $\mathbf{E}_{\text{red}}^{\varepsilon,\text{GMRES}}$ (à droite) pour $\mathcal{D}_\Omega = 5\lambda$, $h = 0.25$ et $R_{\partial\Omega} = 0$.

4.3 Stratégie d'un préconditionneur global

Dans cette section nous décrivons un préconditionneur qui a pour but d'accélérer la convergence de la méthode. Pour mettre en lumière cette approche, nous étudions le spectre de la matrice préconditionnée et les gains en temps et en itérations de cette mise en place.

4.3.1 Définition du préconditionneur

Les ondes électromagnétiques étudiées se propagent dans tout le domaine. Cet aspect encourage à la mise en place d'un préconditionneur "global", au sens où il n'implique pas seulement des interactions locales. Une approche similaire a été introduite dans [97]. Nous en proposons une alternative basée sur différents sous-ensembles de faces (gauche/droite, avant/arrière, bas/haut) des éléments cubiques du maillage. Cela mène à trois décompositions régulières-singulières [23] $\mathbf{M} - \mathbf{N}$ de $\mathbf{A}$. Ces dernières sont associées aux trois formes sesquilinéaires suivantes, définies à partir du Problème 13 de Cessenat-Després et de l'espace d'ondes planes $\mathbb{Y}_{\mathcal{T}}^h$.

Définition 4.1. *Pour tout* $\mathbf{x} \in \mathbb{Y}_{\mathcal{T}}^h$ *et pour tout* $\mathbf{x}' \in \mathbb{Y}_{\mathcal{T}}^h$*, nous définissons les trois formes sesquilinéaires* $\mathbf{k}^{x/y/z}$ *(*$\mathbf{k}^x$*,* $\mathbf{k}^y$*,* $\mathbf{k}^z$*) et leurs matrices associées* $\mathbf{N}^{x/y/z} \in \mathbb{C}^{\#\mathrm{ddl}\times\#\mathrm{ddl}}$ *par*

$$\mathbf{k}^{x/y/z}(\mathbf{x},\mathbf{x}') := \sum_{T\in\mathcal{T}} \sum_{F\in\mathcal{F}_{x/y/z}^T\cap\mathcal{F}_{\mathrm{int}}} \left(\Pi_{\mathcal{U}}\mathbf{x}, \mathcal{U}^T\mathbf{x}'\right)_{\mathrm{L}_t^2(F)} = [\mathbf{x}']^*\mathbf{N}^{x/y/z}[\mathbf{x}],$$

où nous définissons par

- $\mathcal{F}_x^T$ *l'ensemble des faces gauche et droite de* T*,*
- $\mathcal{F}_y^T$ *l'ensemble des faces avant et arrière de* T*,*
- $\mathcal{F}_z^T$ *l'ensemble des faces basse et haute de* T*.*

De plus, les trois matrices régulières associées sont définies comme $\mathbf{M}^{x/y/z} := \mathbf{A} + \mathbf{N}^{x/y/z}$.

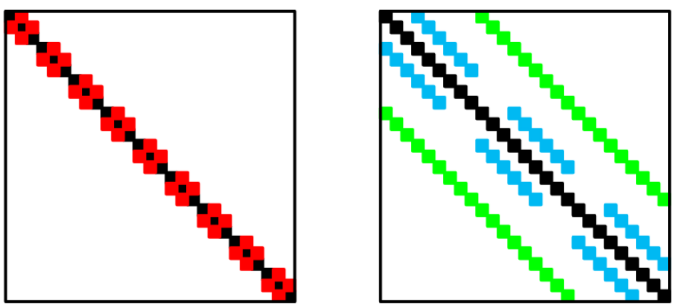

FIGURE 4.31 – Structures des matrices de la décomposition dans la direction x pour $\mathcal{D}_\Omega = 3\lambda$: la matrice $\mathbf{M}^x$ à gauche, et la matrice $\mathbf{N}^x$ à droite.

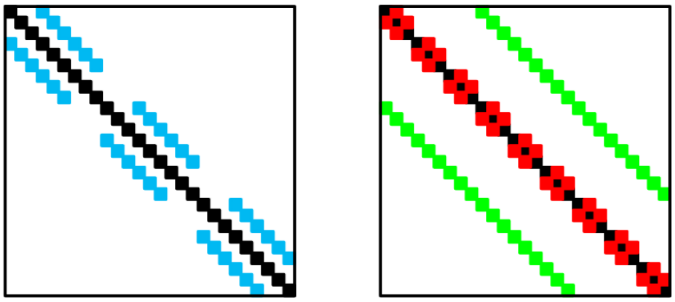

FIGURE 4.32 – Structures des matrices de la décomposition dans la direction y pour $\mathcal{D}_\Omega = 3\lambda$: la matrice $\mathbf{M}^{\mathsf{y}}$ à gauche, et la matrice $\mathbf{N}^{\mathsf{y}}$ à droite.

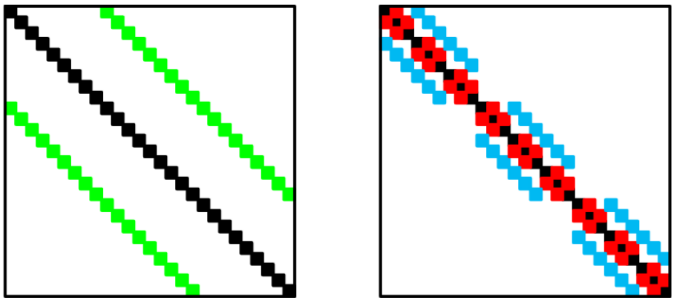

FIGURE 4.33 – Structures des matrices de la décomposition dans la direction z pour $\mathcal{D}_\Omega = 3\lambda$: la matrice $\mathbf{M}^{\mathsf{z}}$ à gauche, et la matrice $\mathbf{N}^{\mathsf{z}}$ à droite.

Ainsi, nous associons à ces trois décompositions régulières-singulières, trois préconditionneurs $[\mathbf{y}] = \mathbf{P}^{\mathsf{x/y/z}}\mathbf{F}$. Soient $[\mathbf{y}] \in \mathbb{C}^{\#\text{ddl}}$ et $[\mathbf{x}] \in \mathbb{C}^{\#\text{ddl}}$, nous introduisons les trois schémas itératif suivants

$$\mathbf{M}^{\mathsf{x/y/z}}[\mathbf{y}] = \mathbf{F}.$$

Mais en appliquant seulement l'un de ces préconditionneurs, nous ne transmettons pas l'information dans tout le domaine à chaque itération du solveur GMRES. En pratique, nous proposons alors de les combiner, de telle sorte à obtenir un préconditionneur global impliquant les trois directions x, y et z :

$$\begin{cases} \mathbf{M}^{\mathsf{x}}[\mathbf{x}^0] &= \mathbf{F}, \\ \mathbf{M}^{\mathsf{y}}[\mathbf{x}^1] &= \mathbf{F} + \mathbf{N}^{\mathsf{y}}[\mathbf{x}^0], \\ \mathbf{M}^{\mathsf{z}}[\mathbf{y}] &= \mathbf{F} + \mathbf{N}^{\mathsf{z}}[\mathbf{x}^1]. \end{cases}$$

Ce nouveau préconditionneur associe à $\mathbf{F}$ le vecteur $[\mathbf{y}] \in \mathbb{C}^{\#\text{ddl}}$. Ainsi, nous obtenons une méthode itérative de la forme $[\mathbf{y}] = \mathbf{P}^{\mathsf{xyz}}\mathbf{F}$. Cela nous introduit un inverse approché du système linéaire $\mathbf{A}[\mathbf{x}] = \mathbf{F}$. C'est pourquoi cette méthode itérative définie un préconditionneur

global appliqué au Problème non préconditionné de KG UWVF 18,

$$\mathbf{P}^{\mathsf{xyz}}\mathbf{A}[\mathbf{x}] = \mathbf{P}^{\mathsf{xyz}}\mathbf{F},$$

où nous utilisons un préconditionnement à gauche et où

$$\mathbf{P}^{\mathsf{xyz}} := (\mathbf{M}^{\mathsf{z}})^{-1} + (\mathbf{M}^{\mathsf{z}})^{-1}\mathbf{N}^{\mathsf{z}}(\mathbf{M}^{\mathsf{y}})^{-1} + (\mathbf{M}^{\mathsf{z}})^{-1}\mathbf{N}^{\mathsf{z}}(\mathbf{M}^{\mathsf{y}})^{-1}\mathbf{N}^{\mathsf{y}}(\mathbf{M}^{\mathsf{x}})^{-1}.$$

La version réduite de ce préconditionnement existe aussi. Le préconditionneur global réduit est

$$\mathbf{P}^{\mathsf{xyz}}_{\mathrm{red}} := (\mathbf{M}^{\mathsf{z}}{}_{\mathrm{red}})^{-1} + (\mathbf{M}^{\mathsf{z}}{}_{\mathrm{red}})^{-1}\mathbf{N}^{\mathsf{z}}_{\mathrm{red}}(\mathbf{M}^{\mathsf{y}}{}_{\mathrm{red}})^{-1} + (\mathbf{M}^{\mathsf{z}}{}_{\mathrm{red}})^{-1}\mathbf{N}^{\mathsf{z}}_{\mathrm{red}}(\mathbf{M}^{\mathsf{y}}{}_{\mathrm{red}})^{-1}\mathbf{N}^{\mathsf{y}}_{\mathrm{red}}(\mathbf{M}^{\mathsf{x}}_{\mathrm{red}})^{-1}.$$

Nous ne donnons pas ici de théorie assurant la convergence du solveur GMRES préconditionné à gauche associé à $\mathbf{P}^{\mathsf{xyz}}_{\mathrm{red}}$. Cependant, deux caractéristiques de ce nouveau préconditionneur peuvent être mises en avant. D'une part, l'idée de définir un préconditionneur "global" provient, comme nous l'avons mentionné précédemment, de l'aspect propagatif des ondes électromagnétiques. Plus précisément, avec les sous-ensembles de faces et les applications successives des préconditionneurs $\mathbf{P}^{\mathsf{x/y/z}}$, nous espérons communiquer l'information d'un bout à l'autre du domaine (dans les trois directions x, y et z), et cela plus rapidement qu'avec un préconditionneur de Cessenat-Després. D'autre part, dans les nombreux cas que nous avons testés, le spectre de $\mathbf{P}^{\mathsf{xyz}}_{\mathrm{red}}\mathbf{A}_{\mathrm{red}}$ est concentré autour de 1 (voir la Figure 4.35), et enlève les petites valeurs propres présentes pour $\mathbf{A}_{\mathrm{red}}$ (en rouge dans la Figure 4.34) qui ralentissent la convergence. Dans la Figure 4.36, nous étudions l'erreur relative de la norme du résidu définie par

$$e_r := \frac{\|\mathbf{A}[\mathbf{x}] - \mathbf{F}\|}{\|\mathbf{F}\|}.$$

Le domaine étudié est $\Omega = [0, 200] \times [0, 40] \times [0, 40]$ et a des éléments de taille $h = 1$. Nous optons pour un seuil de troncature $\varepsilon = 10^{-9}$ et $N = 196$, de telle sorte qu'il n'y a pas de "réelle" réduction, voir le Tableau 4.6. L'utilisation de la matrice préconditionnée $\mathbf{P}^{\mathsf{xyz}}_{\mathrm{red}}\mathbf{A}_{\mathrm{red}}$ permet d'avoir nettement moins d'itérations qu'avec la matrice $\mathbf{A}_{\mathrm{red}}$, voir la Figure 4.36. En effet, à précision égale, de l'ordre de 1% pour le résidu GMRES, nous utilisons 343 itérations avec le préconditionneur global contre 1441 avec le préconditionneur de Cessenat-Després, soit 4 fois moins d'itérations. Le préconditionneur global accélère aussi les temps de calcul. En effet, l'utilisation de ce nouveau préconditionneur à la place de celui de Cessenat-Després diminue d'un facteur 2 environ le temps de calcul.

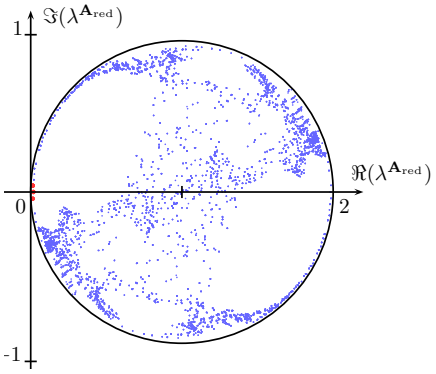

FIGURE 4.34 – Parties réelle et complexe, *resp.* $\Re(\lambda^{\mathbf{A}_{\text{red}}})$ et $\Im(\lambda^{\mathbf{A}_{\text{red}}})$, du spectre $\lambda^{\mathbf{A}_{\text{red}}}$ de $\mathbf{A}_{\text{red}}$, pour $\mathcal{D}_\Omega = 6\lambda$ et $N = 52$.

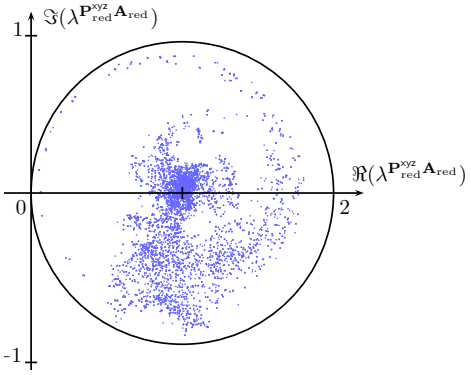

FIGURE 4.35 – Parties réelle et complexe, *resp.* $\Re(\lambda^{\mathbf{P}^{\text{xyz}}_{\text{red}}\mathbf{A}_{\text{red}}})$ et $\Im(\lambda^{\mathbf{P}^{\text{xyz}}_{\text{red}}\mathbf{A}_{\text{red}}})$, du spectre $\lambda^{\mathbf{P}^{\text{xyz}}_{\text{red}}\mathbf{A}_{\text{red}}}$ de $\mathbf{P}^{\text{xyz}}_{\text{red}}\mathbf{A}_{\text{red}}$, pour $\mathcal{D}_\Omega = 6\lambda$ et $N = 52$.

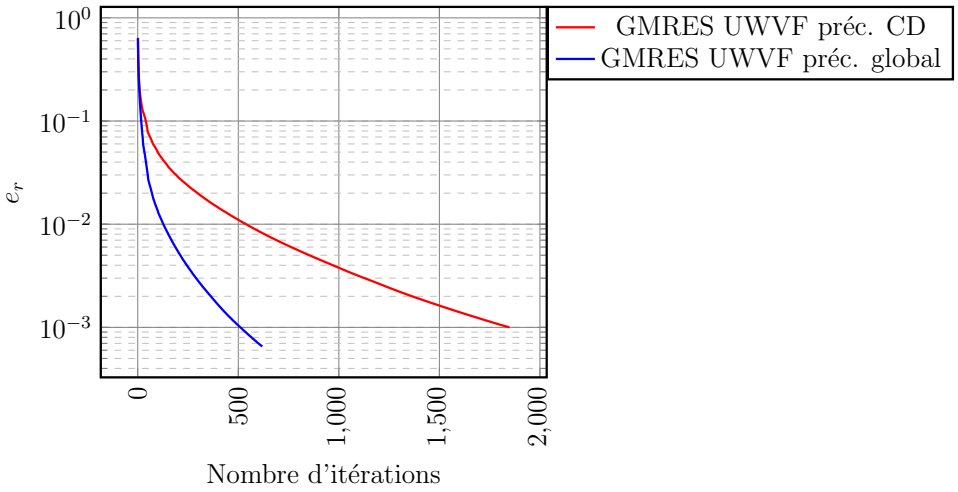

FIGURE 4.36 – Erreur relative e_r de la norme du résidu pour la solution GMRES UWVF préconditionnée par Cessenat-Després et pour la solution GMRES UWVF préconditionnée par $\mathbf{P}^{\text{xyz}}_{\text{red}}$, où $N_{\text{kry}} = 25$.

4.3.2 Stratégies d'implémentation

Le préconditionneur global implique les inverses $(\mathbf{M}^{\text{x}}_{\text{red}})^{-1}$, $(\mathbf{M}^{\text{y}}_{\text{red}})^{-1}$ et $(\mathbf{M}^{\text{z}}_{\text{red}})^{-1}$. Afin de les calculer, nous mettons en place des permutations, transformant les matrices $\mathbf{M}^{\text{y}}_{\text{red}}$ et $\mathbf{M}^{\text{z}}_{\text{red}}$ sous la même forme que $\mathbf{M}^{\text{x}}_{\text{red}}$. Ainsi, nous obtenons des matrices tridiagonales, ou bandes, et leurs factorisations LU sont moins coûteuses en mémoire et en temps de calcul. Pour les obtenir, nous avons recours en pratique à la fonction zgbtrf des librairies Lapack©. De plus, une factorisation LU de l'intégralité des matrices n'est pas nécessaire. En effet, nous utilisons de nouveau l'aspect cartésien du maillage et nous remarquons que seules les inversions des blocs matriciels en Figure 4.37 sont nécessaires. Ces plus petites matrices correspondent aux

barres 1D dans chaque direction du domaine, voir la Figure 4.38.

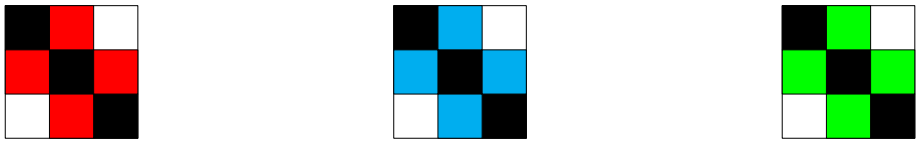

FIGURE 4.37 – Matrices à inverser pour obtenir, (de gauche à droite sur la figure) : $(\mathbf{M}_{\text{red}}^{\mathsf{x}})^{-1}$ (pour les interactions gauche/droite), $(\mathbf{M}_{\text{red}}^{\mathsf{y}})^{-1}$ (pour les interactions avant/arrière) et $(\mathbf{M}_{\text{red}}^{\mathsf{z}})^{-1}$ (pour les interactions bas/haut).

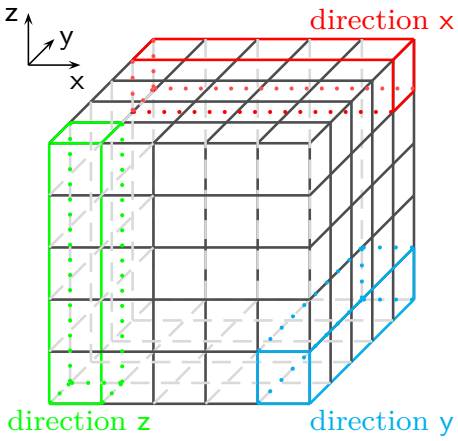

FIGURE 4.38 – Sous-domaines globaux ou "barres 1D" dans un cube, pour $\mathcal{D}_\Omega = 5\lambda$.

Plus précisément, leur stockage a un coût mémoire estimé à

$$\text{MEM}^{\mathsf{xyz}} := (\#\text{elem}^{\mathsf{x}} + \#\text{elem}^{\mathsf{y}} + \#\text{elem}^{\mathsf{z}}) \times 3 \times N_{\text{red}}^2 \times 16,$$

où

— $\#\text{elem}^{\mathsf{x}}$ (*resp.* $\#\text{elem}^{\mathsf{y}}$ et $\#\text{elem}^{\mathsf{z}}$) est le nombre d'éléments dans la direction x (*resp.* y et z),

— le facteur $3 \times N_{\text{red}}^2$ fait référence aux blocs des matrices permutées tridiagonales, voir la Figure 4.37,

— le facteur 16 fait référence au stockage en double précision.

Par ailleurs, pour le calcul des PMV, l'astuce présentée dans la stratégie de désassemblage est employée à nouveau. Les produits $(\mathbf{M}_{\text{red}}^{\mathsf{x/y/z}})^{-1}\mathbf{u}$, pour $\mathbf{u} \in \mathbb{C}^{\#\text{ddl}_{\text{red}}}$, sont réalisés grâce à la fonction zgbtrs des librairies LAPACK©. Cette dernière résout de façon optimisée les systèmes associés à la factorisation $\mathsf{L}_{\mathsf{x/y/z}}\mathsf{U}_{\mathsf{x/y/z}}$ de chacune des matrices $(\mathbf{M}_{\text{red}}^{\mathsf{x/y/z}})^{-1}$.

4.4 Conclusion

Dans ce chapitre, nous avons établi plusieurs stratégies afin de réduire le coût mémoire de la méthode et d'améliorer son conditionnement.

D'une part, la structure cartésienne du maillage a rendu possible la mise en place de la stratégie de désassemblage de la matrice de Trefftz, engendrant un gain mémoire considérable pour la méthode itérative. En effet, nous pouvons traiter un domaine de taille deux fois plus grande : de $\mathcal{D}_\Omega^{\max} = 150\lambda$ à $\mathcal{D}_\Omega^{\max} = 370\lambda$ (qui s'avère être $\mathcal{D}_\Omega^{\max} = 315\lambda$ en pratique), voir la Figure 4.39. De plus, le désassemblage permet de réduire les temps de calcul grâce à sa structure adaptée aux librairies OpenMP.

D'autre part, la stratégie de réduction de base conduit à la fois à l'amélioration du conditionnement de la matrice et à la réduction du coût mémoire de la méthode. Cette réduction des fonctions de base donne des matrices de plus petites tailles tout en ne perdant pas d'informations pour décrire la solution numérique en choisissant un seuil de troncature ε adapté. En particulier, nous avons vu que pour $N = 196$ ondes planes, $\varepsilon = 10^{-5}$ ou $\varepsilon = 10^{-4}$ sont des bons compromis dans l'exemple de la simulation d'un dipôle dans Ω.

Enfin, la stratégie de préconditionnement global implique trois décompositions régulières-singulières associées aux différentes directions du domaine (x, y et z). Sur un exemple de domaine parallélépipédique, nous avons montré que ce préconditionneur global utilise, à précision fixée, quatre fois moins d'itérations que le préconditionneur de Cessenat-Després et presque deux fois moins de temps pour atteindre cette convergence. Il accélère donc le solveur GMRES en termes de temps de calcul et du nombre d'itérations bien qu'il s'avère avoir un coût mémoire très légèrement plus important.

Pour conclure, ces stratégies ont apporté de nettes améliorations aux méthodes de type Trefftz Krylov : diminution du coût mémoire et accélération de la convergence de l'algorithme sans dégradation de la solution numérique. Elles doivent toutefois être utilisées à bon escient et en fonction du cas d'application.

Jusqu'ici nous n'avons pas investigué un aspect important de la méthode de Trefftz : le choix des fonctions de base, qui sont des solutions locales du problème de Maxwell. Les ondes planes utilisées peuvent avoir des difficultés à approcher avec précision les effets de pointe ou encore les ondes piégées dans des géométries complexes. De plus, ce choix de fonctions de base induit de nombreux problèmes de conditionnement, provoquant à leur tour de nombreuses erreurs d'arrondis. Nous avons proposé des stratégies pour remédier à ce phénomène (en jouant sur le nombre de fonctions de base par exemple, Section 4.2) mais aucune d'entre elles ne consistait à employer un autre type de fonctions de base. Par conséquent, nous mettons en oeuvre dans le chapitre suivant une nouvelle version de la méthode de Trefftz, utilisant des fonctions de base qui sont des solutions numériques (donc

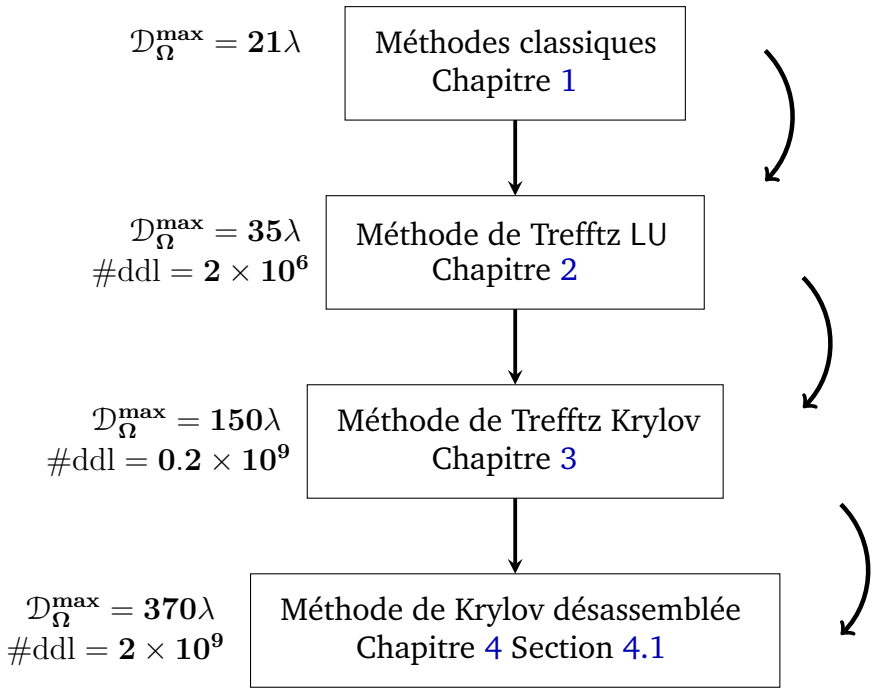

FIGURE 4.39 – Augmentation de la taille $\mathcal{D}_\Omega$ qu'il est possible de considérer grâce à la mise en place d'une méthode de Trefftz directe puis de GMRES et enfin de GMRES désassemblée, où $\mathcal{D}_\Omega^{\max}$ est la taille maximale atteinte dans chacun des chapitres pour 1To de mémoire, et où $\#\text{ddl} = N \times \#\text{elem}$ avec $N = 52$ et $h = 1$.

approchées) locales du problème. Cette méthode est connue sous le nom de Quasi-Trefftz.

CHAPITRE 5

MÉTHODE QUASI-TREFFTZ

Sommaire

Une attention particulière a été apportée au coût mémoire de la méthode de Trefftz (et donc de GoTEM3) dans les chapitres précédents. Cela a permis d'augmenter la taille des scènes de calcul, voir la Figure 4.39. En parallèle, nous avons aussi étudié le conditionnement des matrices, dont l'amélioration accélère la convergence du solveur GMRES. Nous choisissons désormais d'orienter notre recherche scientifique vers une méthode employant d'autres fonctions de base. En effet, les ondes planes employées pour dériver le problème de

Trefftz dans les Chapitres 2 et 3 ne prennent pas en compte les modes évanescents ou les effets de coin, où des singularités peuvent apparaître.

Le présent chapitre est une transcription de l'article [45]. Il s'agit de l'étude d'une méthode de type quasi-Trefftz pour la simulation d'ondes électromagnétiques en deux dimensions. Les fonctions de base d'une méthode quasi-Trefftz sont des solutions locales approchées du problème étudié. Elles peuvent être construites à partir d'approximations de Taylor par exemple [65], ou encore à partir de la résolution local du problème, grâce à un solveur classique, tel que celui d'EF de Nédélec du Chapitre 1 par exemple. Ces fonctions de base sont ensuite utilisées pour créer la formulation variationnelle Trefftz.

Par manque de temps durant la thèse, nous n'avons pas appliqué ce type de méthode au cas tri-dimensionnel. Cependant, bien que traitant uniquement des cas en dimension 2, cet article révèle le potentiel des méthodes quasi-Trefftz pour simuler les ondes électromagnétiques avec précision. En particulier, des expériences numériques montrent que les phénomènes de propagation d'ondes aussi bien à basses qu'à hautes fréquences sont pris en considération. Ces méthodes sont alors une bonne alternative aux méthodes de Trefftz utilisant des ondes planes. Elles sont une perspective d'amélioration qu'il serait possible de mettre en place dans GoTEM3 dans un futur proche.

Introduction

Numerical methods like the Finite Element Method (FEM), see [3, 66, 78], and the Finite Difference Method (FDM), see [96] for example, are widely used to solve time-harmonic electromagnetic wave equations. One limitation they all face is called the pollution effect. When considering the numerical solution of a propagation phenomenon with wavenumbers k posed on domains with length L, the numerical accuracy deteriorates when kL becomes large. This has been highlighted in the following articles [62, 63]. This phenomenon is related to a numerical dispersion and is called numerical pollution. A detailed analysis with error estimates has been proposed in [74]. This issue is of particular importance at high frequency or on large domains where the number of degrees of freedom per wavelength should be chosen large to achieve a given accuracy.

Classically, numericians resort to one or more of the following remedies, which can be combined. The first one consists in considering high-order FEM, see for example [2]. A second answer to the numerical pollution issue is Discontinuous Galerkin (DG) finite elements that are less dispersive [1]. This latter approach can be of particular interest in the context of inverse problems.

The German mathematician Erich Trefftz proposed a paradigm in which basis functions are taken to be local solutions to the partial differential equations system of interest. The

intuition behind this idea is that such basis functions can better approximate physical phenomena, and be especially less dispersive than classical polynomial basis functions. Trefftz methods gather simultaneously the possibility of being flexible like high-order method and of being less dispersive at low-order as pointed out in [95].

Another important feature of Trefftz methods relies on the difficulty to solve wave problems on huge domains due to memory limitation. This fact is less relevant in homogeneous media where accelerated Boundary Element Method (BEM) [33, 93, 100] or integral equation collocation method [14] can be used. A natural idea is also to resort to a domain decomposition method to successively deal with different subdomains. Trefftz formalism, especially in the context of the Ultra-Weak Variational Formulation (UWVF) [17, 21, 22, 35, 98], seems to be one of the most relevant solutions.

Moreover, Trefftz methods have also a variety of formulations which offer a high flexibility. In addition to the UWVF method and to the approach based on the reciprocity principle considered in this paper, the least square method [47] is another Trefftz method which is of great interest. On the other hand, the partition of unity method [6] shares also some similarities with Trefftz methods even if local basis functions are not exact solutions of Maxwell equations. This is particularly true for problems involving corner singularities but this subject is out of scope for our paper.

In the context of long range propagation, most Trefftz methods are associated to plane wave basis [46, 48, 49, 55, 77] or other analytical solution like Bessel functions. Hiptmair et al. have contributed significantly in this paradigm on the theoretical level. In particular, a good starting point would be [54] which is a survey of Trefftz type methods. The main drawback of these methods is that they are associated to ill-conditioned linear system which leads in double precision to a lack of accuracy. So much so that in [30, 72], methods improving linear system properties have been presented. Moreover, these types of basis approximate accurately only a few kind of wave functions. For example, plane wave basis functions have difficulties to approximate evanescent modes, corner singularities or transitions between light and shadows which are located behind obstacles. In some sense these methods have the same limitation than geometrical optic methods and have difficulties to compute near field accurately.

An important advance was proposed in [56, 57] where the local solutions were not analytical but constructed thanks to a local solver, which was in this case a BEM. This has led to a method where all kinds of waves are correctly computed. A similar idea was later developed in [10] where a symmetric Trefftz method was associated to a local BEM to compute solutions of the Helmholtz equation in two dimensions. In the latter paper, it was remarked that an accurate resolution of the local solver impressively increases the global accuracy. This is rather surprising since this improvement does not cause any extra cost. Our approach is

comparable to static condensation methods [99] and to hybridizable discontinuous Galerkin methods [84] where interior degrees of freedom are eliminated from the final system. However, a local BEM solver is not always appropriate. It cannot easily handle heterogeneous or anisotropic domains. BEM is also less popular in the numerical simulation community. In this paper, the local basis will be computed thanks to a local FEM solver. A wide class of wave equations will then be studied. To highlight this last aspect, we have considered electromagnetic wave propagation in heterogeneous media.

In the present paper, we propose to parameterize the local Trefftz basis functions by an impedance condition (also called Fourier-Robin condition). This is a second improvement of [10] where the elements should be included in a strip of width of half a wavelength to avoid numerical resonance frequencies. This enables to consider a wide class of elements with no size restriction.

The objective of this paper is to develop a Trefftz electromagnetic method and to investigate its performance in particular with respect to the pollution effect. It is structured as follows. Section 5.1 reviews Maxwell equations and introduce the second-order Maxwell problem and its hypotheses. Next, our Trefftz scheme is elaborated in Section 5.2. The associated continuous formulation is provided on a general mesh. Section 5.3 is devoted to the discretization of this latter formulation, which leads to approximations of both Trefftz spaces and local Trefftz basis functions. An example of FEM local solver is presented and uses Nédélec FEM. The final section relates results from our Trefftz method implementation. Numerical error analysis will illustrate improvements on the pollution effect. Examples of wave propagation on different domains will show the method's robustness and accuracy.

5.1 Model problem : the simplified Maxwell equations

The Maxwell system models the propagation of an electromagnetic wave. It reads, in absence of charges and currents and for an isotropic linear medium [1] :

$$\begin{cases} \nabla \cdot \mathbf{d} = 0, & \nabla \times \mathbf{e} = -\dfrac{\partial \mathbf{b}}{\partial t}, & \mathbf{d} = \varepsilon_0 \varepsilon_r \mathbf{e}, \\ \nabla \cdot \mathbf{b} = 0, & \nabla \times \mathbf{h} = \dfrac{\partial \mathbf{d}}{\partial t}, & \mathbf{b} = \mu_0 \mu_r \mathbf{h}, \end{cases}$$

where ε_0 (*resp.* ε_r) and μ_0 (*resp.* μ_r) are the permittivity (*resp.* relative permittivity) and the permeability (*resp.* relative permeability) of the vacuum (*resp.* of the medium).

This system involves the electric and magnetic field intensities $\mathbf{e}$ and $\mathbf{h}$ and the electric displacement and the magnetic induction $\mathbf{d}$ and $\mathbf{b}$. In this article we suppose that all

1. All along this paper, bold terms refer to either vectors or vectorial functions.

these fields are time-harmonic. They can therefore be represented by four complex valued functions :

$$\begin{cases} \mathbf{e}(\mathbf{x},t) = & \Re\left(exp(-i\omega t)\mathbf{E}(\mathbf{x})\right), \qquad \mathbf{h}(\mathbf{x},t) = & \Re\left(exp(-i\omega t)\mathbf{H}(\mathbf{x})\right), \\ \mathbf{d}(\mathbf{x},t) = & \Re\left(exp(-i\omega t)\mathbf{D}(\mathbf{x})\right), \qquad \mathbf{b}(\mathbf{x},t) = & \Re\left(exp(-i\omega t)\mathbf{B}(\mathbf{x})\right), \end{cases}$$

where ω is the angular frequency that accounts for time-harmonic dependency. It leads to the following time-harmonic Maxwell system :

$$\begin{cases} \nabla\cdot\mathbf{D} = 0, \qquad \nabla\times\mathbf{E} = -i\omega\mathbf{B}, \quad \mathbf{D}=\varepsilon_0\varepsilon_r\mathbf{E}, \\ \nabla\cdot\mathbf{B} = 0, \qquad \nabla\times\mathbf{H} = i\omega\mathbf{D}, \quad \mathbf{B}=\mu_0\mu_r\mathbf{H}. \end{cases}$$

We suppose, moreover, that the propagation domain $\Omega\times\mathbb{R}$, with $\Omega\subset\mathbb{R}^2$, is invariant in the z-direction. We consider only the transverse electric polarization :

$$\begin{cases} \mathbf{E}(\mathbf{x}) = E_x(x,y)\mathsf{e}_x + E_y(x,y)\mathsf{e}_y, \qquad \mathbf{D}(\mathbf{x}) = D_x(x,y)\mathsf{e}_x + D_y(x,y)\mathsf{e}_y, \\ \mathbf{B}(\mathbf{x}) = B_z(x,y)\mathsf{e}_z, \qquad\qquad\qquad\qquad \mathbf{H}(\mathbf{x}) = H_z(x,y)\mathsf{e}_z. \end{cases}$$

In this paper, we assume that the computational domain $\Omega\subset\mathbb{R}^2$ is a connected domain with polygonal boundary, see Fig. 5.1. This is representative of most applications, see Fig 5.11 for an example. Moreover, we assume that Ω can be decomposed into P subdomains denoted Ω_p *i.e* $\overline{\Omega} = \bigcup_{p=1,\dots,P}\overline{\Omega}_p$ and $\Omega_p\cap\Omega_q=\emptyset$, if $p\neq q$, such that ε_r and μ_r are constant on each subdomain.

Finally, we consider the following second-order Maxwell equation subjected to an impedance boundary condition :

$$\nabla\times\left(\frac{1}{\mu_r}\nabla\times\mathbf{E}\right) - k_0^2\varepsilon_r\mathbf{E} = 0 \qquad \text{in } \Omega,$$

$$ik_0\,Y\,\mathbf{n}\times(\mathbf{E}\times\mathbf{n}) \;-\; \mathbf{n}\times\left(\frac{1}{\mu_r}\nabla\times\mathbf{E}\right) = \mathbf{h} \quad \text{on } \partial\Omega, \tag{5.1a}$$

where

— $c_0 = \dfrac{1}{\sqrt{\varepsilon_0\mu_0}}$ and $k_0 = \dfrac{\omega}{c_0}$ are the wave speed of light and the wavenumber in vacuum respectively,

— the function $Y\in L^\infty(\partial\Omega)$ is an admittance coefficient with strictly non-positive real part,

— the normal $\mathbf{n}\in\mathbb{R}^2$ is the outward unit normal to $\partial\Omega$,

— the boundary term $\mathbf{h} : \partial\Omega\longrightarrow\mathbb{C}^2$ is taken to be a purely tangential function in the

functional space

$$L_t^2(\partial\Omega) := \left\{ \mathbf{u} \in \left(L^2(\partial\Omega)\right)^2, \quad \mathbf{u}\cdot\mathbf{n} = 0 \right\},$$

— the solution of this problem is in the following classical Sobolev space $H(\mathrm{rot},\Omega)$

$$H(\mathrm{rot},\Omega) = \left\{ \boldsymbol{\omega} : \Omega \to \mathbb{C}^2, \int_\Omega |\boldsymbol{\omega}|^2 dx < \infty, \int_\Omega |\nabla\times\boldsymbol{\omega}|^2 dx < \infty \right\}.$$

Remarque 5.1. *In this paper, we consider a two dimensional problem. However, this problem can be seen as a three dimensional problem on $\Omega\times\mathbb{R}$ whose solution is independent of the z-component. The two dimensional vector field $\mathbf{E}$ is composed of the two first components of a three dimensional electromagnetic field problem which is independent of z and with no z-component. Equivalently, $\mathbb{C}^2$-vectors (E_x, E_y) are identified to $\mathbb{C}^3$-vectors $(E_x, E_y, 0)$ and $H_z \in \mathbb{C}$ to $\mathbb{C}^3$-vectors $(0,0,H_z)$. In this context, the $\nabla\times$ operators acting on vector fields or on scalar fields are defined as*

$$\nabla\times\mathbf{E} \;=\; \partial_x E_y - \partial_y E_x \quad \textit{and} \quad \nabla\times H_z \;=\; \Big(\partial_y H_z, -\partial_x H_z\Big).$$

In the same way, $\mathbf{n}\times$ operators are given by

$$\mathbf{n}\times\mathbf{E} \;=\; n_x E_y - n_y E_x \quad \textit{and} \quad \mathbf{n}\times H_z \;=\; \Big(n_y H_z, -n_x H_z\Big).$$

These hypotheses being considered, (5.1) is well-posed in the Hadamard sense. For proof in the context of homogeneous media, we refer to [20, 78]. Physical interpretation of the impedance boundary condition can also be found in [94].

This model is simple enough for numerical implementation, and sufficiently rich to perform a comparison with different numerical methods.

5.2 Construction of a Trefftz scheme

5.2.1 Notations and definitions

The domain Ω is meshed by non-overlapping polygonal elements T. These open sets are called macro-elements, see Fig. 5.2. The mesh does not need to include only triangles or quadrangles (see the colored pentagon in Fig. 5.2) and allows to consider a large variety of geometrical configurations. The set of all macro-elements T is denoted by $\mathcal{T}$ and respects the partition of Ω. In other words, one element $T \in \mathcal{T}$ cannot intersect two different subdomains Ω_p, $p = 1, ..., P$, at the same time. In particular, since there exists a unique $p_0 = 1, ..., P$ such that $T \subset \Omega_{p_0}$, ε_r and μ_r are constant on each macro-element T.

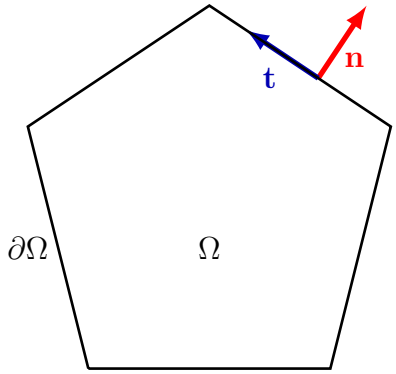

FIGURE 5.1 – Schematic view of the computational domain Ω. The domain boundary is denoted by $\partial\Omega$. The unit normal and tangent vectors are represented in red and blue respectively at one point of the boundary.

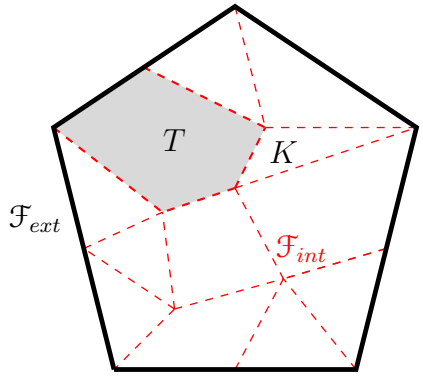

FIGURE 5.2 – An example of Trefftz mesh. Interior (*resp.* exterior) faces are denoted by $\mathcal{F}_{int}$ (*resp.* $\mathcal{F}_{ext}$), see dashed red segments (*resp.* bold segments). Two neighboring macro-elements are for example K and T.

All along this paper, we call edges of an element T as faces and we define the following sets :

— the set $\mathcal{F}_{int}$ of interior macro-faces :

$$\mathcal{F}_{int} := \{\partial T \cap \partial K \ : \ T, K \in \mathcal{T} \text{ with } T \neq K \text{ and } \textit{length}(\partial T \cap \partial K) \neq 0\}, \tag{5.2}$$

where $\textit{length}(I)$ refers to the length of the segment I (that is zero for isolated points),

— the set $\mathcal{F}_{ext}$ of exterior macro-faces :

$$\mathcal{F}_{ext} := \{\partial T \cap \partial\Omega \ : \ T \in \mathcal{T} \text{ and } \textit{length}(\partial T \cap \partial\Omega) \neq 0\},$$

— the set $\mathcal{F}_T$ of macro-faces associated to a macro-element T :

$$\mathcal{F}_T := \{F \in \mathcal{F}_{int} \cup \mathcal{F}_{ext} \ : \ \textit{length}(F \cap \partial T) \neq 0\}. \tag{5.3}$$

We construct an infinite dimensional Trefftz space X_T, which will be instrumental in the construction of the proposed method. It is defined for $T \in \mathcal{T}$ as the set of functions $\boldsymbol{\omega}_T$ satisfying

$$\begin{aligned}
&\boldsymbol{\omega}_T \in H(\mathrm{rot}, T),\\
&\nabla \times \left(\frac{1}{\mu_T}\nabla \times \boldsymbol{\omega}_T\right) - k_0^2 \varepsilon_T \boldsymbol{\omega}_T = 0,\\
&\boldsymbol{\omega}_T \times \mathbf{n}_T \in \mathrm{L}_t^2(\partial T), \qquad (5.4a)\\
&\mathbf{n}_T \times (\nabla \times \boldsymbol{\omega}_T) \in \mathrm{L}_t^2(\partial T), \qquad (5.4b)
\end{aligned}$$

where $\mu_T = \mu_{r|T}$, $\varepsilon_T = \varepsilon_{r|T}$ and $\mathbf{n}_T$ is the outward unit normal to ∂T.

Remarque 5.2. *If* (2.1a) *and* (2.1c) *are both satisfied, then either* (5.4a) *implies* (5.4b) *or* (5.4b) *implies* (5.4a). *In two dimensions, this result is a consequence of the regularity theory of Helmholtz equation which can be found in [73].*

It also leads to the global variational space $X_{\mathcal{T}}$ defined element by element,

$$X_{\mathcal{T}} := \prod_{T\in\mathcal{T}} X_T. \qquad (5.5)$$

Remarque 5.3. *Any element $\boldsymbol{\omega} = (\boldsymbol{\omega}_T)_{T\in\mathcal{T}} \in X_{\mathcal{T}}$ is a vector with components $\boldsymbol{\omega}_T$ in $L^2(T)$. However, $\boldsymbol{\omega}$ can also be identified to a complex valued function of Ω whose restriction to T is equal to $\boldsymbol{\omega}_T$. As a function of Ω, $\boldsymbol{\omega} \in X_{\mathcal{T}}$ is discontinuous across faces and does not satisfy Maxwell equation in the whole domain Ω but only in every element $T \in \mathcal{T}$.*

5.2.2 The Trefftz continuous formulation

The Trefftz variational formulation of (5.1) is deduced from the following duality formula

$$\int_{\partial T} \left(\mathbf{E}_T \cdot \mathbf{n}_T \times \left(\frac{1}{\mu_T}\nabla \times \overline{\boldsymbol{\omega}_T}\right) - \mathbf{n}_T \times \left(\frac{1}{\mu_T}\nabla \times \mathbf{E}_T\right) \cdot \overline{\boldsymbol{\omega}_T}\right) ds = 0, \text{ for } \boldsymbol{\omega}_T \in X_T, \qquad (5.6)$$

where $\mathbf{E}_T$ is the restriction to T of the analytical solution $\mathbf{E}$ to the Maxwell equations (5.1).

Indeed, since $\mathbf{E}_T$ satisfies the Maxwell equation in every element we apply the following Green formula :

$$\begin{aligned}
&\int_T \left(\left(\frac{1}{\mu_T}\nabla \times \mathbf{E}_T\right) \cdot (\nabla \times \overline{\boldsymbol{\omega}_T}) - \underbrace{\nabla \times \left(\frac{1}{\mu_T}\nabla \times \mathbf{E}_T\right)}_{k_0^2 \varepsilon_T \mathbf{E}_T} \cdot \overline{\boldsymbol{\omega}_T} \right) dx \qquad (5.7a)\\
= &\int_{\partial T} \left(\mathbf{n}_T \times \left(\frac{1}{\mu_T}\nabla \times \mathbf{E}_T\right) \cdot \overline{\boldsymbol{\omega}_T}\right) ds.
\end{aligned}$$

The test function $\boldsymbol{\omega}$ also satisfies the Maxwell equation in every element. Therefore, by in-

terchanging the roles of $\mathbf{E}$ and $\boldsymbol{\omega}$, we get

$$\begin{aligned} & \sum_{T\in\mathcal{T}} \int_T \left(\left(\frac{1}{\mu_T} \nabla \times \mathbf{E}_T \right) \cdot (\nabla \times \overline{\boldsymbol{\omega}_T}) - k_0^2 \varepsilon_T \mathbf{E}_T \cdot \overline{\boldsymbol{\omega}_T} \right) dx \\ = & \sum_{T\in\mathcal{T}} \int_{\partial T} \left(\mathbf{E}_T \cdot \mathbf{n}_T \times \left(\frac{1}{\mu_T} \nabla \times \overline{\boldsymbol{\omega}_T} \right) \right) ds. \end{aligned} \tag{5.7b}$$

It remains to subtract (5.7b) to (5.7a) to get (5.6). The expression on the left hand side of (5.6) defines a sesquilinear form. The integral over ∂T will be decomposed following the set of faces $\mathcal{F}_T$, see (5.3), of the element T

$$\partial T = \bigcup_{F\in\mathcal{F}_T} F.$$

It leads to : $\forall \boldsymbol{\omega} \in X_{\mathcal{T}}$

$$\widetilde{a}(\mathbf{E}, \boldsymbol{\omega}) \; = \; \sum_{T\in\mathcal{T}} \sum_{F\in\mathcal{F}_T} \widetilde{a}_{T,F}(\mathbf{E}, \boldsymbol{\omega}) \; = \; 0,$$

with

$$\widetilde{a}_{T,F}(\mathbf{E}, \boldsymbol{\omega}) \; := \; \int_F \left(\mathbf{E}_T \cdot \mathbf{n}_T \times \left(\frac{1}{\mu_T} \nabla \times \overline{\boldsymbol{\omega}_T} \right) - \mathbf{n}_T \times \left(\frac{1}{\mu_T} \nabla \times \mathbf{E}_T \right) \cdot \overline{\boldsymbol{\omega}_T} \right) ds,$$

where the solution $\mathbf{E}$ is assumed to belong to the space $X_{\mathcal{T}}$ and $\boldsymbol{\omega}$ to $H(\mathrm{rot}, T)$, see (2.1a).

We should distinguish the case of an interior face belonging to the interior skeleton $\mathcal{F}_{int}$ to the case of an exterior face belonging to $\mathcal{F}_{ext}$.

— If $F \in \mathcal{F}_{int}$, this face is shared with another element K, see Figure 5.2. Taking into account the continuity of the solution across the interface F

$$\mathbf{n}_T \times (\mathbf{n}_T \times \mathbf{E}_T) \; = \; \mathbf{n}_K \times (\mathbf{n}_K \times \mathbf{E}_K),$$

and

$$\mathbf{n}_T \times \left(\frac{1}{\mu_T} \nabla \times \mathbf{E}_T \right) \; + \; \mathbf{n}_K \times \left(\frac{1}{\mu_K} \nabla \times \mathbf{E}_K \right) \; = \; 0,$$

we have, for $F \in \mathcal{F}_{int}$,

$$\widetilde{a}_{T,F}(\mathbf{E}, \boldsymbol{\omega}) \; = \; \widehat{a}_{T,F}(\mathbf{E}, \boldsymbol{\omega})$$

with $\widehat{a}_{T,F} : X_{\mathcal{T}} \times X_{\mathcal{T}} \longrightarrow \mathbb{C}$ defined by

$$\widehat{a}_{T,F}(\mathbf{E}, \boldsymbol{\omega}) \; := \; \int_F \left(\mathbf{E}_K \cdot \mathbf{n}_T \times \left(\frac{1}{\mu_T} \nabla \times \overline{\boldsymbol{\omega}_T} \right) + \mathbf{n}_K \times \left(\frac{1}{\mu_K} \nabla \times \mathbf{E}_K \right) \cdot \overline{\boldsymbol{\omega}_T} \right) ds. \tag{5.8}$$

— If $F \in \mathcal{F}_{ext}$, this face is subject to the impedance condition (5.1a). The form $\widetilde{a}_{T,F}$ is

rewritten in a way that the symmetry of the final sesquilinear form is respected

$$\widetilde{a}_{T,F}(\mathbf{E},\boldsymbol{\omega}) \;=\; \int_F \left(\mathbf{E}_T\cdot\mathbf{n}_T\times\left(\frac{1}{\mu_T}\nabla\times\overline{\boldsymbol{\omega}_T}\right) + \mathbf{n}_T\times\left(\frac{1}{\mu_T}\nabla\times\mathbf{E}_T\right)\cdot\overline{\boldsymbol{\omega}_T}\right)ds$$
$$-2\int_F \Bigg(\underbrace{\mathbf{n}_T\times\left(\frac{1}{\mu_T}\nabla\times\mathbf{E}_T\right)}_{ik_0Y\,\mathbf{n}_T\times(\mathbf{E}_T\times\mathbf{n}_T)-\mathbf{h}}\cdot\overline{\boldsymbol{\omega}_T}\Bigg)ds.$$

We have, for $F\in\mathcal{F}_{ext}$,

$$\widetilde{a}_{T,F}(\mathbf{E},\boldsymbol{\omega}) \;=\; \widehat{a}_{T,F}(\mathbf{E},\boldsymbol{\omega}) \;-\; \widehat{\ell}_{T,F}(\boldsymbol{\omega})$$

with $\widehat{a}_{T,F} : X_{\mathcal{T}}\times X_{\mathcal{T}}\longrightarrow\mathbb{C}$ and $\widehat{\ell}_{T,F} : X_{\mathcal{T}}\longrightarrow\mathbb{C}$ defined by

$$\widehat{a}_{T,F}(\mathbf{E},\boldsymbol{\omega}) := \int_F \left(\mathbf{E}_T\cdot\mathbf{n}_T\times\left(\frac{1}{\mu_T}\nabla\times\overline{\boldsymbol{\omega}_T}\right) + \overline{\boldsymbol{\omega}_T}\cdot\mathbf{n}_T\times\left(\frac{1}{\mu_T}\nabla\times\mathbf{E}_T\right)\right)ds$$
$$-2ik_0\int_F Y\,(\mathbf{E}_T\times\mathbf{n}_T)\cdot(\overline{\boldsymbol{\omega}_T}\times\mathbf{n}_T)ds \tag{5.9}$$

and

$$\widehat{\ell}_{T,F}(\boldsymbol{\omega}) \;=\; -2\int_F \mathbf{h}\cdot\overline{\boldsymbol{\omega}_T}\,ds. \tag{5.10}$$

Finally, with $\widehat{a}_{T,F}$ defined by (5.8) and (5.9), $\widehat{\ell}_{T,F}$ defined by (5.10) for boundary faces and by taking the convention : $\widehat{\ell}_{T,F}(\boldsymbol{\omega}) \;=\; 0, \forall F\in\mathcal{F}_{int}$, the Trefftz formulation takes the form : find $\mathbf{E}\in X_{\mathcal{T}}$ such that for all $\boldsymbol{\omega}\in X_{\mathcal{T}}$,

$$\sum_{T\in\mathcal{T}}\sum_{F\in\mathcal{F}_T}\widehat{a}_{T,F}(\mathbf{E},\boldsymbol{\omega}) \;=\; \sum_{T\in\mathcal{T}}\sum_{F\in\mathcal{F}_T}\widehat{\ell}_{T,F}(\boldsymbol{\omega}). \tag{5.11}$$

Remarque 5.4. *It is rather natural to add penalization terms to this formulation like in interior penalty DG method (see for example [61]). It is surprising that these additional terms do not lead to any accuracy improvement and deteriorate the linear system conditioning. We refer to [54] for an elaborate theoretical discussion of such penalization in Trefftz formulations.*

5.3 Discretization of the Trefftz formulation

We have now established the variational formulation. The next step consists in approximating the space $X_{\mathcal{T}}$. We are first pointing out the existence of an isomorphism between $X_{\mathcal{T}}$ and $\mathrm{L}^2_t(\partial\mathcal{T})$, a trace space defined on the skeleton of the mesh. A Galerkin approximation then leads to a discretization of $\mathrm{L}^2_t(\partial\mathcal{T})$. Many methods exist to choose associated basis func-

tions. For example plane waves or Bessel functions are analytical local solutions on Trefftz elements and can be employed as basis functions. A comparative discussion of plane waves and Bessel functions discretizations can be found in [47].

In this article, a different approach is proposed. Local basis functions are defined as solutions of a boundary value problem (5.12). An impedance boundary condition is considered to design a well-posed problem. Perfect conductor condition could have been as well considered but resonances phenomena could occur. Once those Trefftz basis functions are defined, (5.11) can not be computed analytically, since the solution operator associated to the isomorphism between $\mathrm{L}_t^2(\partial\mathcal{T})$ and $X_{\mathcal{T}}$ is not explicit. To overcome this difficulty, we propose to use a second approximation based on a FEM. In [10, 56, 57] BEM is proposed to compute the local basis functions for Helmholtz equations. However, FEM are, to our opinion, more flexible and can deal with a wider variety of physical models.

5.3.1 Identification of spaces X_T and $X_{\mathcal{T}}$ with $\mathrm{L}_t^2(\partial T)$ and $\mathrm{L}_t^2(\partial\mathcal{T})$

We start with the identification of the local space X_T with $\mathrm{L}_t^2(\partial T)$ for each $T \in \mathcal{T}$. It is based on the following result : let $\mathbf{g}_T \in \mathrm{L}_t^2(\partial T)$ and $Y_T \in L^\infty(\partial T)$ with a strictly non-positive real part, then there exists a unique $\boldsymbol{\omega}_T \in H(\mathrm{rot}, T)$ (see [78]) such that

$$\begin{aligned} \nabla \times \left(\frac{1}{\mu_T}\nabla \times \boldsymbol{\omega}_T\right) - k_0^2 \varepsilon_T \boldsymbol{\omega}_T = 0 \qquad & \text{in} \quad T, \\ ik_0\, Y_T\, \mathbf{n}_T \times (\boldsymbol{\omega}_T \times \mathbf{n}_T) \;-\; \mathbf{n}_T \times \left(\frac{1}{\mu_T}\nabla \times \boldsymbol{\omega}_T\right) = \mathbf{g}_T \qquad & \text{on} \quad \partial T, \end{aligned} \tag{5.12a}$$

where we recall that physical parameters are constant in the element T. The solution of (5.12) satisfies the regularity statements (5.4a) and (5.4b). Consequently, $\boldsymbol{\omega}_T$ belongs to the space X_T. Reciprocally all functions of X_T are obviously solutions of the problem (5.12). In particular, the local solution operator $\mathbf{S}_T : \mathrm{L}_t^2(\partial T) \longrightarrow X_T$ defined by

$$\mathbf{S}_T \mathbf{g}_T \;=\; \boldsymbol{\omega}_T, \tag{5.13}$$

where $\boldsymbol{\omega}_T$ is the solution of (5.12), induces an isomorphism between the spaces $\mathrm{L}_t^2(\partial T)$ and X_T. This leads to the following characterization of the space X_T :

$$X_T \;=\; \Big\{ \mathbf{S}_T \mathbf{g}_T \text{ such that } \mathbf{g}_T \in \mathrm{L}_t^2(\partial T) \Big\}.$$

Remarque 5.5. *At most frequencies, it would have been also possible to consider Perfect Electric Condition (PEC) or alternatively Perfect Magnetic Condition (PMC) to parametrize X_T. In this*

case (5.12a) *would have been replaced by one of the following boundary conditions*

$$\begin{aligned} \mathbf{n}_T \times (\boldsymbol{\omega}_T \times \mathbf{n}_T) &= \mathbf{g}_T \quad \text{on} \quad \partial T, \quad (5.12a') \\ \mathbf{n}_T \times \left(\frac{1}{\mu_T} \nabla \times \boldsymbol{\omega}_T \right) &= \mathbf{g}_T \quad \text{on} \quad \partial T. \quad (5.12a'') \end{aligned}$$

However, the PEC and PMC are associated to spurious modes predicted by the Maxwell spectral theory in bounded domains. These resonances phenomena do not occur with an impedance boundary condition.

The global Trefftz space $X_{\mathcal{T}}$, see (5.5), is therefore in bijection with the space

$$\mathrm{L}_t^2(\partial\mathcal{T}) := \prod_{T\in\mathcal{T}} \mathrm{L}_t^2(\partial T),$$

through the global solution operator $\mathbf{S} : \mathrm{L}_t^2(\partial\mathcal{T}) \longrightarrow X_{\mathcal{T}}$. We associate to the solution $\mathbf{E}$ (*resp.* the test functions $\boldsymbol{\omega}$) an element of the trace space $\mathbf{f}$ (*resp.* $\mathbf{g}$) by $\mathbf{Sf} = \mathbf{E}$ (*resp.* $\mathbf{Sg} = \boldsymbol{\omega}$) which is defined by (5.13) on every element by $\mathbf{E}_T = \mathbf{S}_T\mathbf{f}_T$ (resp. $\boldsymbol{\omega}_T = \mathbf{S}_T\mathbf{g}_T$), see Fig. 5.3.

The Trefftz formulation (5.11) can be written in the following form : find $\mathbf{f} \in \mathrm{L}_t^2(\partial\mathcal{T})$ such that for all $\mathbf{g} \in \mathrm{L}_t^2(\partial\mathcal{T})$,

$$\sum_{T\in\mathcal{T}} \sum_{F\in\mathcal{F}_T} \widehat{a}_{T,F}(\mathbf{Sf},\mathbf{Sg}) = \sum_{T\in\mathcal{T}} \sum_{F\in\mathcal{F}_T} \widehat{\ell}_{T,F}(\mathbf{Sg}). \tag{5.14}$$

The variational problem (5.14) involves the tangential trace of the rotational. Most of numerical methods are not adapted to the evaluation of this quantity. An important feature of the presented numerical method consists in computing the rotational thanks to (5.12a) that is satisfied by $\boldsymbol{\omega} = \mathbf{Sg}$. Similarly, $\mathbf{E} = \mathbf{Sf}$ satisfies the following equation :

$$\mathbf{n}_T \times \left(\frac{1}{\mu_T} \nabla \times \mathbf{S}_T\mathbf{f}_T \right) = ik_0\, Y_T\, \mathbf{n}_T \times (\mathbf{S}_T\mathbf{f}_T \times \mathbf{n}_T) - \mathbf{f}_T.$$

Finally, we have the variational problem : find $\mathbf{f} \in \mathrm{L}^2(\partial\mathcal{T})$ such that for all $\mathbf{g} \in \mathrm{L}^2(\partial\mathcal{T})$,

$$\sum_{T\in\mathcal{T}} \sum_{F\in\mathcal{F}_T} a_{T,F}(\mathbf{f},\mathbf{g}) = \sum_{T\in\mathcal{T}} \sum_{F\in\mathcal{F}_T} \ell_{T,F}(\mathbf{g}),$$

where $a_{T,F} : \mathrm{L}_t^2(\partial\mathcal{T}) \times \mathrm{L}_t^2(\partial\mathcal{T}) \to \mathbb{C}$ and $\ell_{T,F} : \mathrm{L}_t^2(\partial\mathcal{T}) \to \mathbb{C}$, defined by

$$\begin{aligned} a_{T,F}(\mathbf{f},\mathbf{g}) &:= \widehat{a}_{T,F}(\mathbf{Sf},\mathbf{Sg}) \\ &= ik_0 \int_F (\overline{Y_T} - Y_K)(\mathbf{n}_K \times \mathbf{S}_K\mathbf{f}_K) \cdot (\mathbf{n}_T \times \overline{\mathbf{S}_T\mathbf{g}_T}) ds \\ &\quad - \int_F \left(\overline{\mathbf{g}_T} \cdot \mathbf{S}_K\mathbf{f}_K + \mathbf{f}_K \cdot \overline{\mathbf{S}_T\mathbf{g}_T} \right) ds \quad \text{for } F \in \mathcal{F}_{int}, \end{aligned} \tag{5.15a}$$

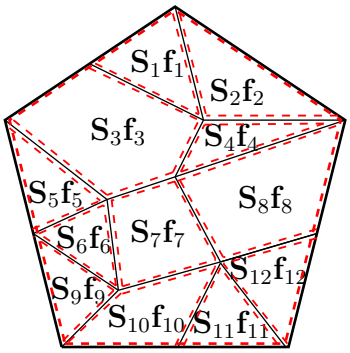

FIGURE 5.3 – Decomposition of the solution with the global solution operator S element by element.

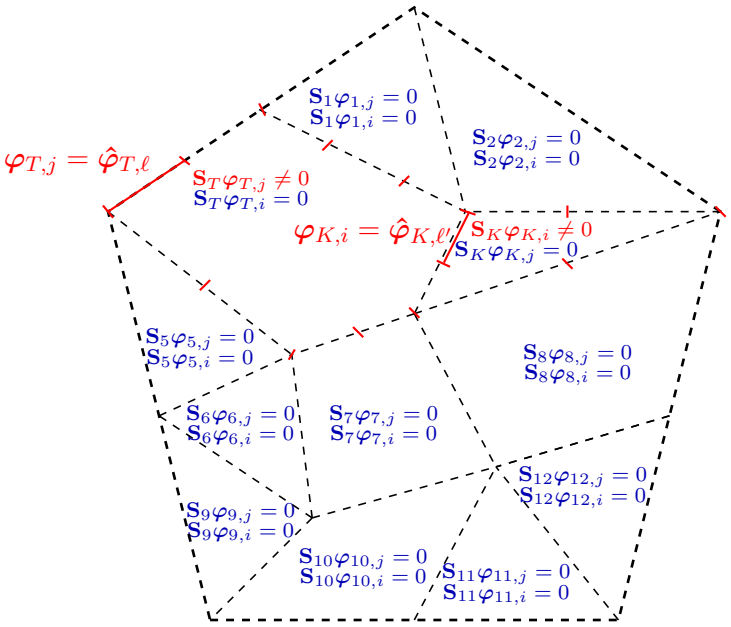

FIGURE 5.4 – Interaction between two local basis functions $\boldsymbol{\varphi}_{T,i}$ and $\boldsymbol{\varphi}_{K,j}$, between two neighboring elements, where $j = loc2glob(T,\ell)$ and $i = loc2glob(K,\ell')$.

$$
\begin{aligned}
a_{T,F}(\mathbf{f},\mathbf{g}) &:= \widehat{a}_{T,F}(\mathbf{Sf},\mathbf{Sg}) \\
&= ik_0 \int_F (Y_T + \overline{Y_T} - 2Y)(\mathbf{n}_T \times \mathbf{S}_T\mathbf{f}_T)\cdot(\mathbf{n}_T \times \overline{\mathbf{S}_T\mathbf{g}_T})\,ds \\
&\quad - \int_F \left(\overline{\mathbf{g}_T}\cdot\mathbf{S}_T\mathbf{f}_T + \mathbf{f}_T\cdot\overline{\mathbf{S}_T\mathbf{g}_T}\right)ds \quad \text{for } F \in \mathcal{F}_{ext},
\end{aligned}
\tag{5.15b}
$$

$$
\ell_{T,F}(\mathbf{g}) := \widehat{\ell}_{T,F}(\mathbf{Sg}) = 0 \quad \text{for} \quad F \in \mathcal{F}_{int},
$$

$$
\ell_{T,F}(\mathbf{g}) := \widehat{\ell}_{T,F}(\mathbf{Sg}) = -2\int_F \mathbf{h}\cdot\overline{\mathbf{S}_T\mathbf{g}_T}\,ds \quad \text{for} \quad F \in \mathcal{F}_{ext}.
$$

Remarque 5.6. *Trefftz methods lead to variational formulations based on trace spaces. The construction of* (5.15) *supports this particularity.*

Remarque 5.7. *In case $Y_T = Y_K = Y \in \mathbb{R}$, the formulation is drastically simplified for both cases $F \in F_{int}$ and $F \in F_{ext}$.*

Remarque 5.8. *Considering a complex bilinear form instead of a sesquilinear form will lead to an alternative interesting simpler formulation. Indeed, the problem* (5.15) *will be simplified by erasing the complex conjugate symbol. Both formulations should be numerically tested to determine advantages and disadvantages of the two formulations. Both formulations are equivalent when Y_T is real. In the case $Y_T = Y_K = Y \in \mathbb{C}$, this second formulation can also be drastically simplified.*

5.3.2 Galerkin approximation of the Trefftz formulation

The discretization of (5.15) goes through the choice of a finite dimensional space to approximate $L^2_t(\partial\mathcal{T})$. This space is constructed from a partition $\mathcal{F}_{F,\mu}$ in segments of each macro-face $F \in \mathcal{F}_{int} \cup \mathcal{F}_{ext}$. They will be called micro-faces.

From these partitions, we then define for each $T \in \mathcal{T}$, the set $\mathcal{F}_{T,\mu} := \bigcup_{F\in\mathcal{F}_T} \mathcal{F}_{F,\mu}$ of micro-faces associated to T and an approximation space of $\mathrm{L}^2_t(\partial T)$:

$$V^q_{T,\mu} := \{\mathbf{v} : \partial T \to \mathbb{C}^2 \ : \ \forall F_\mu \in \mathcal{F}_{T,\mu}, \mathbf{v}_{|F_\mu} \in [\mathcal{P}^q(F_\mu)]^2 \text{ such that } \mathbf{v}_{|F_\mu} \cdot \mathbf{n}_T = 0\},$$

where $\mathcal{P}^q(F_\mu) := \{p : F_\mu \to \mathbb{C} \ : \ p \circ \Phi_{F_\mu} \in \mathcal{P}^q([0,1])\}$ with Φ_{F_μ} being the affine mapping from $[0,1]$ to F_μ and $\mathcal{P}^q([0,1])$ is the space of polynomials of degree $q \in \mathbb{N}$. An alternative definition can be found in [78], that is : $\mathcal{P}^q(F_\mu)$ is the space of polynomials of maximum degree $q \in \mathbb{N}$ in arc length on F_μ.

Finally a conforming approximation space of $\mathrm{L}^2_t(\partial\mathcal{T})$ is defined by

$$V^q_\mu := \prod_{T\in\mathcal{T}} V^q_{T,\mu}.$$

We propose a first discrete Trefftz formulation of (5.15) as follows : find $\mathbf{f}^q_\mu \in V^q_\mu$ such that for all $\mathbf{g}^q_\mu \in V^q_\mu$,

$$a(\mathbf{f}^q_\mu, \mathbf{g}^q_\mu) = \ell(\mathbf{g}^q_\mu), \tag{5.16}$$

where

$$a(\mathbf{f}^q_\mu, \mathbf{g}^q_\mu) := \sum_{T\in\mathcal{T}} \sum_{F\in\mathcal{F}_T} a_{T,F}(\mathbf{f}^q_\mu, \mathbf{g}^q_\mu) \text{ and } \ell(\mathbf{g}^q_\mu) := \sum_{T\in\mathcal{T}} \sum_{F\in\mathcal{F}_T} \ell_{T,F}(\mathbf{g}^q_\mu).$$

A linear system is then associated to this variational formulation by introducing a basis of V^q_μ. As for any discontinuous finite element, we refer to local and global basis functions.

— Local basis functions are defined on ∂T and are denoted by $\hat{\boldsymbol{\varphi}}_{T,\ell} \in V^q_{T,\mu}$ with $1 \le \ell \le (q+1)\ N_{T,\mu}$, where $N_{T,\mu}$ is the number of micro-faces of T and $(q+1)$ is the dimension of the local polynomial approximation space.

— Global basis vectors can be seen as global "functions". They are defined on $\partial\mathcal{T}$ by $\boldsymbol{\varphi}_i \in V^q_\mu$ with $1 \le i \le \mathrm{card}(V^q_\mu)$. The dimension of the space V^q_μ is given by

$$\mathrm{card}(V^q_\mu) = \sum_{T\in\mathcal{T}} (q+1)\ N_{T,\mu}.$$

Local (*resp.* global) basis functions are defined on the boundary ∂T (*resp.* $\partial\mathcal{T}$) and lead to local functions $\boldsymbol{\omega}_T = \mathbf{S}_T\boldsymbol{\varphi}_T$ (*resp.* global functions $\boldsymbol{\omega} = \mathbf{S}\boldsymbol{\varphi}$) defined on the whole element

T (*resp.* the whole set $\mathcal{T}$).

To construct matrices associated to (5.16), we must define a link between $\hat{\boldsymbol{\varphi}}_{T,\ell}$ and $\boldsymbol{\varphi}_i$. That is why we introduce a local numbering and a global numbering linked by the following bijective operator $\mathrm{loc2glob}$:

$$i = \mathrm{loc2glob}(T, \ell),$$

where i is the global number associated to the local number $\ell = 1, ..., (q+1)N_{T,\mu}$ of the macro-element $T \in \mathcal{T}$.

The global basis functions $(\boldsymbol{\varphi}_i)_{i=1,...,\mathrm{card}(V_\mu^q)}$ are constructed as follows :
if $i = \mathrm{loc2glob}(T, \ell)$ then $\boldsymbol{\varphi}_i := (\boldsymbol{\varphi}_{K,i})_{K\in\mathcal{T}}$ are defined by

$$\begin{cases} \boldsymbol{\varphi}_{K,i} = 0, & \text{if } K \neq T, \\ \boldsymbol{\varphi}_{K,i} = \hat{\boldsymbol{\varphi}}_{T,\ell}, & \text{if } K = T. \end{cases} \tag{5.17}$$

The basis defined above leads to a characterization of the space V_μ^q. Indeed, we can describe $\mathbf{f}_\mu^q = (\mathbf{f}_{T,\mu}^q)_{T\in\mathcal{T}} \in V_\mu^q$ using $\boldsymbol{\varphi}_i$:

$$\mathbf{f}_\mu^q = \sum_{j=1}^{\mathrm{card}(V_\mu^q)} f_j \boldsymbol{\varphi}_j, \text{ with } f_j \in \mathbb{C}.$$

Let us consider the restriction $\mathbf{f}_{T,\mu}^q$ of $\mathbf{f}_\mu^q$ on $T \in \mathcal{T}$,

$$\mathbf{f}_{T,\mu}^q = \sum_{j=1}^{\mathrm{card}(V_\mu^q)} f_j \boldsymbol{\varphi}_{T,j},$$

where $\boldsymbol{\varphi}_{T,j}$ is the restriction of $\boldsymbol{\varphi}_j$ on T. According to (5.17), only components from the element T are non-zero, see Fig. 5.4. This sum can thus be formulated for local basis functions of T only :

$$\mathbf{f}_{T,\mu}^q = \sum_{\ell=1}^{(q+1)N_{T,\mu}} f_{\mathrm{loc2glob}(T,\ell)} \hat{\boldsymbol{\varphi}}_{T,\ell}.$$

Finally, the formulation (5.16) can be rewritten as follows : find $(f_j)_{j=1,...,\mathrm{card}(V_\mu)}$ such that $\forall i \in 1, ..., \mathrm{card}(V_\mu)$,

$$\sum_{j=1}^{\mathrm{card}(V_\mu)} f_j a(\boldsymbol{\varphi}_j, \boldsymbol{\varphi}_i) = \ell(\boldsymbol{\varphi}_i). \tag{5.18}$$

If $j = \mathrm{loc2glob}(T, \ell)$ and $i = \mathrm{loc2glob}(K, \ell')$ then $a(\boldsymbol{\varphi}_j, \boldsymbol{\varphi}_i) = 0$ if, and only if,

$$\Big(T \neq K \quad \text{and } \textit{length}(\partial T \cap \partial K) \neq 0\Big) \text{ or } \Big(T = K \text{ s.t } \textit{length}(\partial\Omega \cap \partial T) = 0\Big),$$

where *length* is defined in (5.2).

Therefore the matrix associated to the sesquilinear form a is sparse.

The linear system (5.18) can then be used to find $\mathbf{f}_\mu^q \in V_\mu^q$. To solve this system we must compute $\mathbf{S}\boldsymbol{\varphi}_i$ (see (5.15a) and (5.15b)). However, $\mathbf{S}$ is usually not known, such that $\boldsymbol{\omega} = \mathbf{S}\boldsymbol{\varphi}_i \in X_T$, $i = 1, ..., (q+1)N_{T,\mu}$, cannot be known explicitly. We have to determine an approximation $\mathbf{S}_\nu$ of the linear operator $\mathbf{S}$. The latter can be described by using several methods as a FDM, an integral equations method, or a FEM. In this paper, Nédélec finite elements [78] are taken as an example and are described in the next section.

Remarque 5.9. *For this reason our Trefftz method should be called quasi-Trefftz. Indeed, effective basis functions are not exact solutions of our Maxwell problem.*

5.3.3 Example of finite element approximation of S

The construction of the Trefftz linear system requires to compute an approximation of $\mathbf{S}\boldsymbol{\varphi}_j = (\mathbf{S}_K\boldsymbol{\varphi}_{K,j})_{K\in\mathcal{T}}$. The function $\mathbf{S}_K\boldsymbol{\varphi}_{K,j}$ is equal to zero except on the macro-element $K = T$ such that $j = \mathrm{loc2glob}(T, \ell)$. In this element, an approximation of $\mathbf{S}_T\boldsymbol{\varphi}_j$ is obtained by discretizing the local problem (5.12) for $\mathbf{g}_T = \hat{\boldsymbol{\varphi}}_{T,\ell}$. This problem reads : find $\hat{\boldsymbol{\omega}}_{T,\ell} = \mathbf{S}_T\boldsymbol{\varphi}_{T,j} \in H(\mathrm{rot}, T)$, such that

$$\begin{aligned} \nabla\times\left(\frac{1}{\mu_T}\nabla\times\hat{\boldsymbol{\omega}}_{T,\ell}\right) - k_0^2\varepsilon_T\hat{\boldsymbol{\omega}}_{T,\ell} &= 0 \qquad \text{in} \quad T, \\ ik_0\,Y_T\,\mathbf{n}_T\times(\hat{\boldsymbol{\omega}}_{T,\ell}\times\mathbf{n}_T) \;-\; \mathbf{n}_T\times\left(\frac{1}{\mu_T}\nabla\times\hat{\boldsymbol{\omega}}_{T,\ell}\right) &= \hat{\boldsymbol{\varphi}}_{T,\ell} \qquad \text{on} \quad \partial T. \end{aligned}$$

It will be approximated with a high-order Nédélec FEM, see for example [81, 82]. The associated variational formulation of (5.19) takes the form : find $\hat{\boldsymbol{\omega}}_{T,\ell} = \mathbf{S}_T\boldsymbol{\varphi}_{T,j} \in H(\mathrm{rot}, T)$ such that $\forall\boldsymbol{\psi} \in H(\mathrm{rot}, T)$

$$\begin{aligned} \int_{\partial T}(\hat{\boldsymbol{\varphi}}_{T,\ell}\cdot\overline{\boldsymbol{\psi}})ds &= -k_0^2\int_T\varepsilon_T(\hat{\boldsymbol{\omega}}_{T,\ell}\cdot\overline{\boldsymbol{\psi}})dx + \int_T\frac{1}{\mu_T}(\nabla\times\hat{\boldsymbol{\omega}}_{T,\ell})\cdot(\nabla\times\overline{\boldsymbol{\psi}})dx \\ &\quad + ik_0\int_{\partial T}Y_T(\mathbf{n}_T\times\hat{\boldsymbol{\omega}}_{T,\ell})\cdot(\mathbf{n}_T\times\overline{\boldsymbol{\psi}})ds. \end{aligned}$$

Now, let $\mathcal{T}_\nu(T)$ be a triangular mesh of the macro-element $T \in \mathcal{T}$, see Fig 5.5.

The Nédélec finite element space $V_{T,\nu}^p$ of order $p \in \mathbb{N}$, see [78], is defined by :

$$V_{T,\nu}^p := \{\mathbf{v} \in H(\mathrm{rot}, T) : \forall T_\nu \in \mathcal{T}_\nu(T), \mathbf{v}_{|T_\nu} \in \mathcal{N}^p(T_\nu)\},$$

where

$$\mathcal{N}^p(T_\nu) = \mathcal{P}^p(T_\nu)^2 + \mathcal{S}^{p+1}(T_\nu)$$

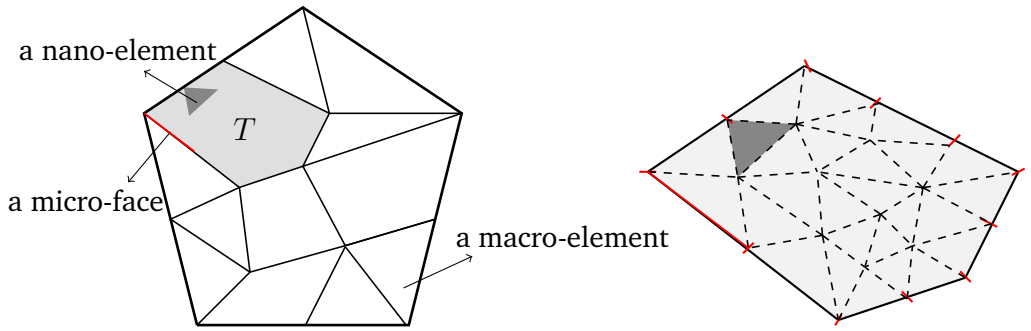

FIGURE 5.5 – Discretization of a macro-element T.

with $\mathcal{S}^p(T_\nu) = \{\mathbf{p} \in (\tilde{\mathcal{P}}^p(T_\nu))^2 : \begin{pmatrix} x \\ y \end{pmatrix} \cdot \mathbf{p} = 0\}$ and $\tilde{\mathcal{P}}^p(T_\nu)$ the space of homogeneous polynomial functions of degree p.

Finally, a finite element approximation $\mathbf{S}^p_{T,\nu}\boldsymbol{\varphi}_{T,j}$ of $\mathbf{S}_T\boldsymbol{\varphi}_{T,j}$ is computed by solving the following discrete problem : find $\hat{\boldsymbol{\omega}}^{\nu,p}_{T,\ell} = \mathbf{S}^p_{T,\nu}\hat{\boldsymbol{\varphi}}_{T,\ell} \in V^p_{T,\nu}$, such that $\forall \boldsymbol{\psi} \in V^p_{T,\nu}$

$$-k_0^2 \int_T \varepsilon_T (\hat{\boldsymbol{\omega}}^{\nu,p}_{T,\ell} \cdot \overline{\boldsymbol{\psi}}) dx + \int_T \frac{1}{\mu_T} \left(\nabla \times \hat{\boldsymbol{\omega}}^{\nu,p}_{T,\ell}\right) \cdot (\nabla \times \overline{\boldsymbol{\psi}}) dx + ik_0 \int_{\partial T} Y_T (\mathbf{n}_T \times \hat{\boldsymbol{\omega}}^{\nu,p}_{T,\ell}) \cdot (\mathbf{n}_T \times \overline{\boldsymbol{\psi}}) ds = \int_{\partial T} (\hat{\boldsymbol{\varphi}}_{T,\ell} \cdot \overline{\boldsymbol{\psi}}) ds.$$

The Nédélec FEM is implemented by using basis functions deduced from the ones proposed in [42] for a high-order Raviart-Thomas approximation, by applying a simple $\pi/2$ rotation.

5.4 Numerical investigation of the proposed Trefftz method

This section presents some numerical results from the implementation of the Trefftz method. The first part focuses on a numerical analysis of the convergence and the pollution error. The second part is qualitative and deals with classical examples.

Our method is natively adapted to small or large macro-elements since the parametrization of the space X_T by an impedance condition avoids spurious modes, see Remark 5.5. In this paper, we subdivide the computation domain into macro-elements whose sizes are at most a few wavelengths. This size is a good compromise in terms of computational effort. Indeed, it induces an interesting sparsity of the Trefftz linear system, and it avoids increases of both the system conditioning and the memory required for the local solver.

Our implementation is based on a particular choice of local basis functions. Indeed, they have distinct local numberings, ranging from 1 to $(q+1)\,N_{T,\mu}$ (see 5.3.2), and have Lagrange interpolation polynomial values. They are defined as follows :

— the subscript ℓ is defined by $\ell = (i_\mu - 1)(q+1) + l_\mu$, where i_μ is the micro-face number of $F_\mu \in \mathcal{F}_{T,\mu}$ such that $1 \leq i_\mu \leq N_{T,\mu}$, and ℓ_μ is the local number of the basis function such that $1 \leq \ell_\mu \leq q+1$,

— for all $F_\mu \in \mathcal{F}_{T,\mu}$,

$$\hat{\boldsymbol{\varphi}}_{T,\ell|F_\mu} \circ \Phi_{F_\mu} = \begin{cases} (0,0)^T \text{ if } F_\mu \neq F_{i_\mu} \\ L^q_{\ell_\mu} \mathbf{t}_{\partial T} \text{ if } F_\mu = F_{i_\mu} \end{cases},$$

where the unit tangent vector function is defined by $\mathbf{t}_{\partial T} = \mathbf{n}_T \times (0,0,1)^T$ and $L^q_{\ell_\mu}$ is the ${\ell_\mu}^{th}$ Lagrange interpolation polynomial constructed from the $(q+1)$ Gauss points in $[0,1]$.

In our implementation, the mesh $\mathcal{T}_\nu(T)$ associated to each macro-element T verifies the following property :

$$\forall T_\nu \in \mathcal{T}_\nu(T), \quad \text{either} \quad \textit{length}(\partial T_\nu \cap \partial T) = 0 \quad \text{or} \quad \exists F_\mu \in \mathcal{F}_{T,\mu}$$
$$\text{such that} \quad \partial T_\nu \cap \partial T = F_\mu \text{ (see Fig. 5.5)}.$$

In other words, we don't refine the mesh $\mathcal{T}_\nu(T)$ in order to improve the quality of the approximation $S^p_{T,\nu}$. Consequently, only p-convergence of Nédélec FEM has yet been implemented. Nevertheless, in presence of singularities, a more efficient approach would be an hp-version of the local solver. A DG method using geometrically graded meshes is especially adapted.

Finally, the local admittance Y_T used in the definition of $\mathbf{S}$ (see (5.12a)) is chosen to be equal to $\sqrt{\varepsilon_T/\mu_T}$.

5.4.1 Numerical error analysis

Relationship between $V^q_{T,\mu}$ and $V^p_{T,\nu}$

We study the link between approximation orders q and p by considering the following problem :

$$\nabla \times (\nabla \times \mathbf{E}) - k_0^2 \mathbf{E} = 0 \qquad \text{in } \Omega,$$

$$ik_0\, \mathbf{n} \times (\mathbf{E} \times \mathbf{n}) \; - \; \mathbf{n} \times (\nabla \times \mathbf{E}) = \mathbf{h}^{inc} \quad \text{on } \partial\Omega,$$

where $\mathbf{h}^{inc} = ik_0\, \mathbf{n} \times (\mathbf{E}^{inc} \times \mathbf{n}) \; - \; \mathbf{n} \times (\nabla \times \mathbf{E}^{inc})$, with $\mathbf{E}^{inc}$ an incident plane wave whose direction of propagation is $(1,1)^T$, and the computational domain Ω is represented in Fig. 5.6.

The analytical solution of this problem is obviously $\mathbf{E} = \mathbf{E}^{inc}$ due to its well-posedness character.

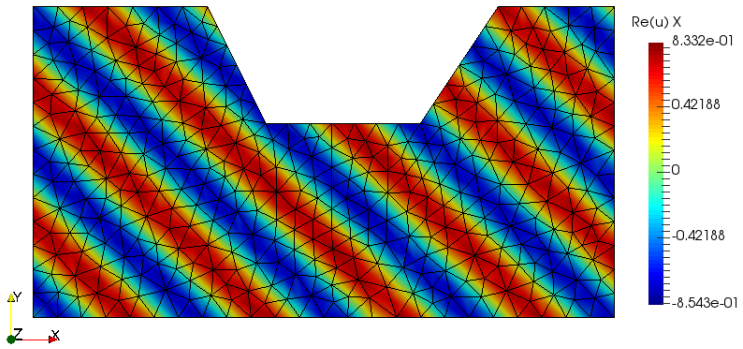

FIGURE 5.6 – The macro-element Ω, an example of triangular mesh $\mathcal{T}_\nu(\Omega)$ and the representation of the analytical solution.

For the simulation, only one macro-element is used *i.e* $\mathcal{T} = \{\Omega\}$. As the example displayed in Fig. 5.6, $\mathcal{T}_\nu(\Omega)$ is a triangular mesh. In order to study their relationship, approximation orders q and p are respectively varying from 0 to 2 and from 0 to 4.

Let us introduce the relative error in H-curl norm :

$$\mathbf{e} := \frac{\sqrt{\|\mathbf{E}^{q,p}_{\mu,\nu} - \mathbf{E}\|^2_{0,\Omega} + \|\nabla \times (\mathbf{E}^{q,p}_{\mu,\nu} - \mathbf{E})\,\|^2_{0,\Omega}}}{\sqrt{\|\mathbf{E}\|^2_{0,\Omega} + \|\nabla \times \mathbf{E}\|^2_{0,\Omega}}},$$

where $\mathbf{E}^{q,p}_{\mu,\nu}$ denotes the numerical solution of our Trefftz scheme.

Curves obtained from our simulation, see Fig. 5.7, show the relative error in function of the quantity $kh/(2\pi(q+1))$ which is equivalent to the inverse of the number of points per wavelength. In our case, h is the longest length of micro-faces.

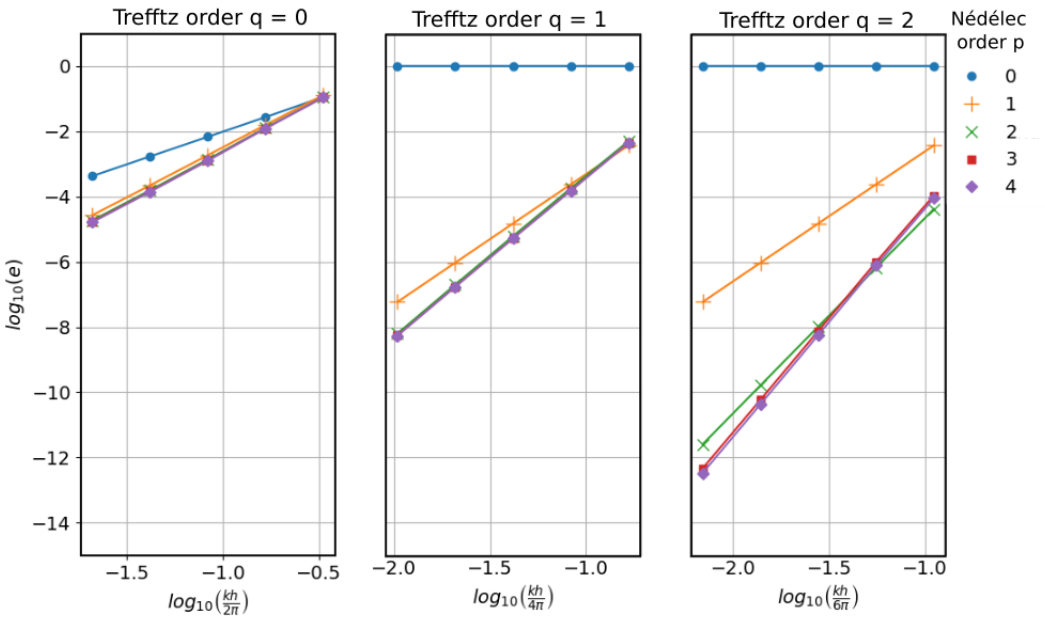

FIGURE 5.7 – Convergence curves in H-curl error of the Trefftz scheme for some local Nédélec approximations.

These plots show that the Trefftz order q and the FEM order p must be carefully chosen. Indeed, if $p < q$, a projection of the "trace" basis functions $\boldsymbol{\varphi}_i$ onto lower-order polynomials occurs leading to a loss in precision and convergence. Consequently, the condition $p \geq q$ must be satisfied for the proposed Trefftz scheme. As a matter of fact, the higher the local Nédélec FEM order, the closer the basis functions are to the exact Maxwell solutions. When p is sufficiently large, the lack of conformity in our local approximation becomes then negligible. Mathematically, the convergence rate of the error induced by the variational crime perpetrated in the discretization of the sesquilinear form should be at least of the same order as the one of the underlying Trefftz method. The bound of this condition, *i.e* $p = q$, denotes a convergence rate similar to Nédélec FEM of order p. In case $p > q$, a super-convergence phenomenon is pointed out. By super-convergence, we mean that the Trefftz method converges at a higher rate than a comparable Nédélec FEM with the same number of degrees of freedom on the boundary. Let us recall that introducing extra degrees of freedom for the Nédélec finite element solver has no extra cost since they are eliminated from the final system. In all plots, highest order curves are close to each other. It reveals a saturation phenomenon. The superposition of curves in Fig. 5.7 demonstrates that the local solver is almost exact in comparison with the Trefftz solver. It brings the optimal choice $p = q + 1$. This is exactly what one should expect from a Trefftz method and motivates their use instead of standard ones.

Evaluation of the pollution error

One of the main motivations of Trefftz type methods is reducing the impact of the pollution effect, by which more classical FEM tend to be limited when the computational domain becomes large in terms of wavelength. We analyze the behaviour of the proposed Trefftz scheme regarding this effect by considering the following problem :

$$\begin{aligned} \nabla \times \nabla \times \mathbf{E} - k_0^2 \mathbf{E} &= 0 \quad && \text{in } \Omega = [0, L] \times [0, 1], \\ \mathbf{n} \times (\mathbf{E} \times \mathbf{n}) &= 0 \quad && \text{if } y = 0 \text{ or } y = 1, \\ \mathbf{n} \times (\mathbf{E} \times \mathbf{n}) &= -1 \quad && \text{if } x = 0, \\ ik_0 \mathbf{n} \times (\mathbf{E} \times \mathbf{n}) - \mathbf{n} \times \nabla \times \mathbf{E} &= 0 \quad && \text{if } x = L, \end{aligned} \tag{5.21}$$

where $k_0 = 2\pi$ and $L = M\lambda_0$ with $\lambda_0 = 2\pi/k_0$ and $M = 10, ..., 200$.

This problem is slightly different than the one initially considered in this paper (see (5.1)). Indeed, in this latter, we have decided to consider only an impedance boundary condition to simplify the presentation. Dirichlet boundary conditions as these ones used in (5.21) can straightforwardly be included in the scheme.

The problem (5.21) models a duct of height 1 and length L where an incoming plane wave is generated at $x = 0$, propagating freely to the right, and finally arriving at a transparent boundary condition on the right side.

In this numerical example, the computational domain for the Trefftz method is decomposed in the set of macro-elements $\mathcal{T} = \{T_m := [m-1, m] \times [0, 1] \ : \ m = 1, ..., M\}$. The mesh used by the global Nédélec FEM is made of the union of Trefftz macro-elements meshes. In each of them, three triangular meshes $\mathcal{T}_\nu(T_m)$ based on spatial discretization steps $h = 1/N$ for $N = 6, 9, 12$ are considered and an "optimal" Nédélec approximation order p is chosen *i.e* $p = q + 1$.

The aim of these simulations is a comparison of the relative L^∞-error

$$e_\infty = 100 \frac{\|\mathbf{E} - \mathbf{E}_h\|_\infty}{||\mathbf{E}||_\infty}$$

induced by the Trefftz and the classical Nédélec FEM. It is known that this error is directly related to numerical dispersion and it is a relevant way to analyze the pollution effect. Here, the analytical solution of (5.21) is $\mathbf{E}(x) = e^{-i2\pi x}\mathbf{e}_y$. Empirically we observe that the relative maximum error as a function of duct length L can be well approximated by a linear interpolation of the form

$$e_{interpolated} = aL + b$$

with $b > 0$ and $a > 0$.

The figure 5.8 (resp. 5.9) shows the results obtained with $q = 1$ (resp. $q = 2$) and a

classical Nédélec FEM of order $p = 1$ (resp. $p = 2$). We can draw the following conclusions. The pollution error is always more important for the FEM. Comparing the curves, we see that both interpolation parameters a and b tend to be roughly one order of magnitude larger for the FEM than for the Trefftz method at comparable meshes and orders.

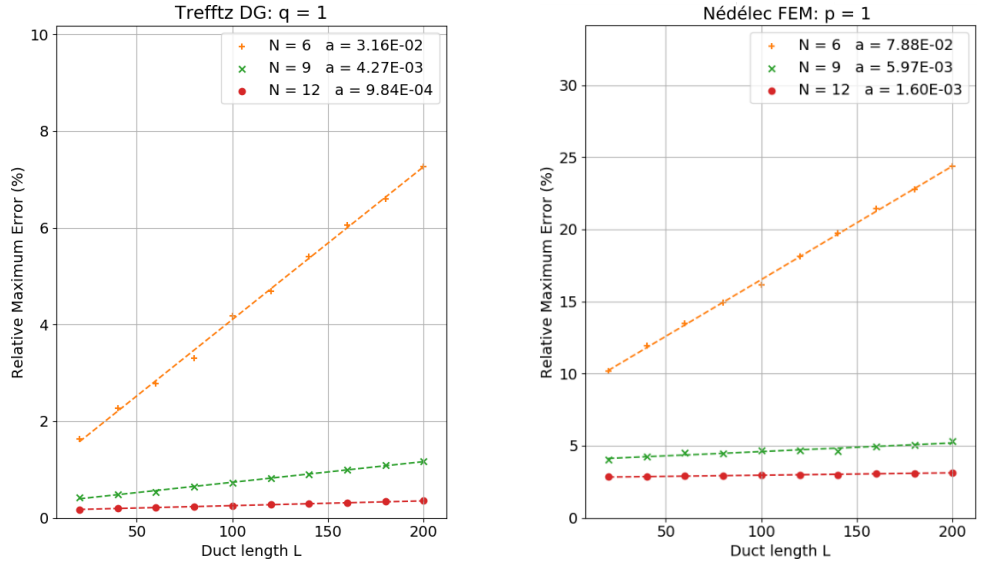

FIGURE 5.8 – Relative maximum order induced by the Trefftz DG method for $(q, p) = (1, 2)$ and the classical Nédélec FEM of order $p = 1$ in function of L.

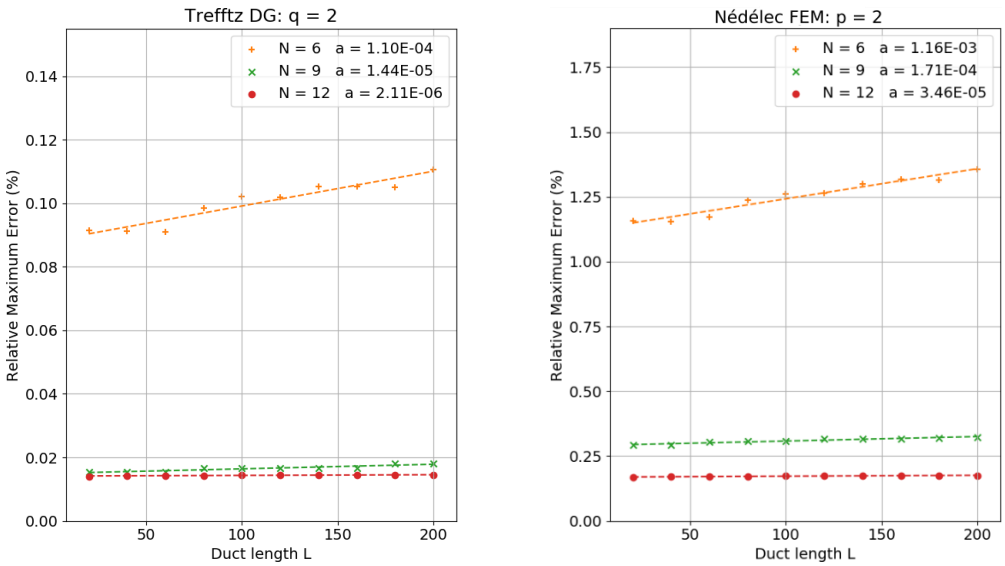

FIGURE 5.9 – Relative maximum order induced by the Trefftz DG method for $(q, p) = (2, 3)$ and the classical Nédélec FEM of order $p = 2$ in function of L.

5.4.2 Illustrative examples

We consider three examples to point out Trefftz method's accuracy and flexibility. Two of them consist of the computation of scattered field by bounded obstacles, see Fig 5.11 and Fig 5.12. The other one is a heterogeneous duct problem. In all simulations, the domain is decomposed in a set $\mathcal{T}$ of macro-elements T. Triangular meshes $\mathcal{T}_\nu(T)$ are constructed by using spatial discretization steps h_ν. From now on, we set the Trefftz order $q = 2$ and the Nédélec FEM order $p = 3$.

Duct propagation with variable dielectric parameters

In this example, we consider a duct problem with variable dielectric parameters defined by :

$$\begin{array}{rcll} \nabla \times \left(\frac{1}{\mu_r}\nabla \times \mathbf{E}\right) - \varepsilon_r k_0^2 \mathbf{E} &=& 0 & \text{in } \Omega = [0,6]\times[0,1], \\ \mathbf{n}\times(\mathbf{E}\times\mathbf{n}) &=& 0 & \text{if } y = 0 \text{ or } y = 1, \\ \mathbf{n}\times(\mathbf{E}\times\mathbf{n}) &=& -1 & \text{if } x = 0, \\ ik_0\, Y\, \mathbf{n}\times(\mathbf{E}\times\mathbf{n}) \;-\; \mathbf{n}\times\left(\nabla\times\frac{1}{\mu_r}\mathbf{E}\right) &=& 0 & \text{if } x = 6, \end{array} \tag{5.22}$$

where $k_0 = 3\pi$, $Y = \sqrt{\varepsilon_r/\mu_r}$,

$$\varepsilon_r(x,y) = \begin{cases} 1 & \text{if } 0 < x < 2 \\ \frac{1}{2} & \text{if } 2 < x < 4 \\ 2 & \text{if } 4 < x < 6 \end{cases} \text{ and } \mu_r(x,y) = \begin{cases} 1 & \text{if } 0 < x < 2 \\ 2 & \text{if } 2 < x < 4 \\ 2 & \text{if } 4 < x < 6 \end{cases}.$$

For this simulation, the Trefftz approximation is characterized by the following table :

Ω	$[0,6]\times[0,1]$
$\mathcal{T}$	$\{T_m := [m, m+2]\times[0,1] \,:\, m = 0,1,2\}$
h_ν	$1/12$
(q,p)	$(2,3)$

We recall the definition of the relative wave travelling speed and the relative impedance associated to a dielectric media :

$$c_r = \frac{1}{\sqrt{\mu_r\varepsilon_r}} \text{ and } Z_r = \sqrt{\frac{\mu_r}{\varepsilon_r}}.$$

Now, we use the first macro-element (in blue in Fig. 5.10) where $c_r = Z_r = 1$ as a reference. In the second macro-element (in red in Fig. 5.10), we have $c_r = 1$ and $Z_r = 2$ and physically, we must observe, in comparison to the first region, a doubling of the wave amplitude without modification of the wavenumber. In the third macro-element (in yellow in Fig. 5.10), we have $c_r = 1/2$ and $Z_r = 1$ and this situation leads to a doubling of the wavenumber without modification of the wave amplitude. Blue, red and yellow macro-elements in figure 5.10 are the meshes of the Nédélec local solver.

The figure 5.10 shows the imaginary part of the numerical solution computed by our Trefftz scheme. We observe that the desired physical behaviour imposed with amplitude and frequency changes by (5.22), is correctly reproduced.

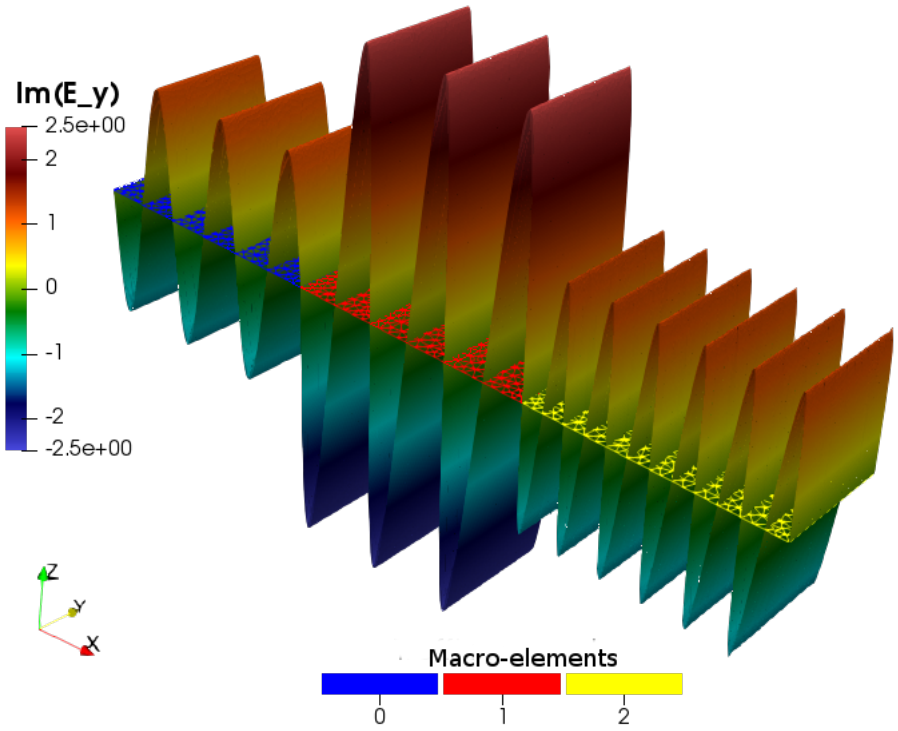

FIGURE 5.10 – A plane wave generated on the left edge (macro-element 0) propagating towards the right (towards macro-element 2). On macro-element 0 : reference plane wave. On macro-element 1 : twice the amplitude. On macro-element 2 : twice the wavenumber.

Scattering by a perfectly conducting disk and a L-shaped obstacle

In this part, we present two test-cases which model the scattering of an incident plane wave by perfectly conducting obstacles in an unbounded domain. Meshes $\mathcal{T}_\nu(T)$ are constructed from two spatial discretization steps h_ν^{obs}, close to the obstacle, and h_ν^{ext}, close to the exterior boundary of the computational domain. We set the dielectric parameters to $\varepsilon_r = \mu_r = 1$ and the wavenumber to $k_0 = 10\pi$.

The first obstacle is smooth and convex and corresponds to a disk $D(0, R)$. The perfectly conducting character of this latter is taken into account by imposing the homogeneous Dirichlet boundary condition, $\mathbf{n} \times \mathbf{E} = 0$ on the circle $C(0, R)$.

For this simulation, the Trefftz approximation is characterized by the following table :

Ω	$([-1, 1] \times [-1, 1]) \setminus D(0, R)$, with $R = 0.4$
$\mathcal{T}$	4 macro-elements T (bottom right of Fig. 5.11)
h_ν^{obs}	$1/30$
h_ν^{ext}	$1/15$
(q, p)	$(2, 3)$

The scattered and total electric fields computed by the Trefftz method are represented in Fig. 5.11. We observe that the method accurately computes the electromagnetic wave propagation in such situation. In particular, the "shadow" region in the total field on the right side of the disk is well restored. Moreover, the non-convex macro-elements such as those used in this example are well supported.

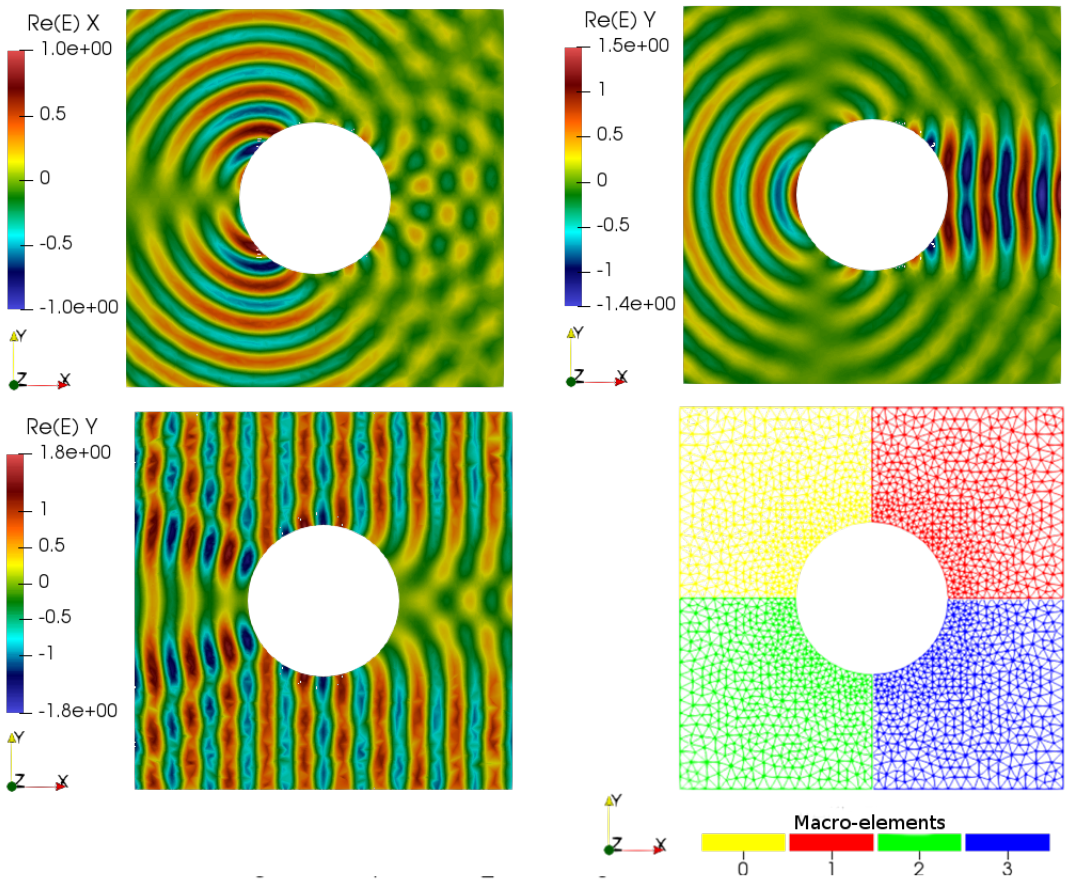

FIGURE 5.11 – On top left : scattered field (x-component). On top right : scattered field (y-component). On bottom left : total field (y-component). On bottom right : colour macro-elements with embedded FEM triangular mesh, refined close to the circle.

The second experiment uses the L-shaped obstacle Ω_L described in Fig. 5.12, which has a length and height of 1, and a thickness of 0.2. It leads to a non-convex and non-smooth problem whose approximation is defined by the following table :

Ω	$([-1,1]\times[-1,1])\setminus\Omega_L$
$\mathcal{T}$	12 macro-elements T (top of Fig. 5.12)
h_ν^{obs}	$1/30$
h_ν^{ext}	$1/15$
(q,p)	$(2,3)$

The numerical solution represented in the figures on bottom of Fig. 5.12 shows that the proposed Trefftz method seems to work well in presence of singularities and to accurately

compute the interference phenomenon which takes place in the trapping region induced by non-convexity of the obstacle according to classic electromagnetic theory. Moreover, this example gives a good idea of the method flexibility in terms of computational domain partitioning by using macro-elements.

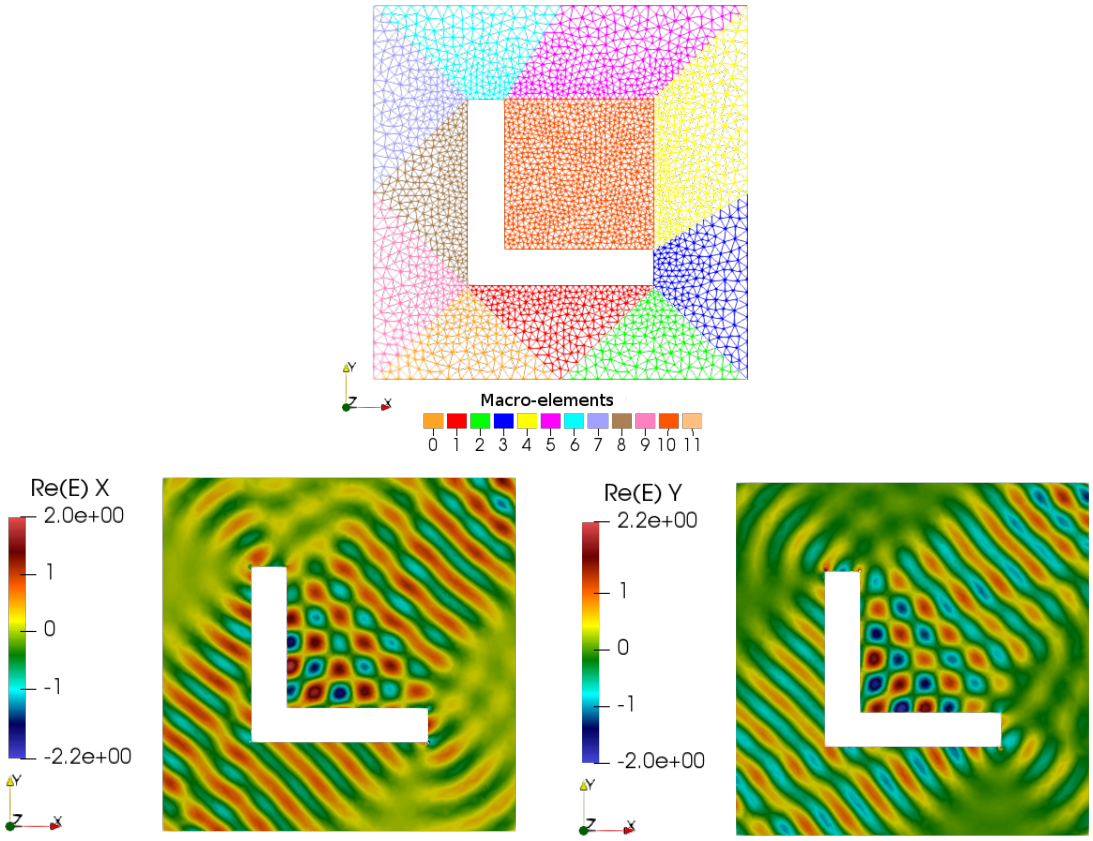

FIGURE 5.12 – On the top : $\mathcal{T}$ and $\mathcal{T}_\nu(T)$ associated to the L-shaped computational domain. On the bottom : x and y- components of the real part of the numerical solution.

5.5 Conclusion

In this paper, we have presented a Trefftz method associated to a Nédélec FEM approximation for a two dimensional Maxwell problem. As highlighted in our numerical analysis, a super-convergence phenomenon appears. This crucial aspect points out the "performance" of such a method. In a similar way, the propagation of electromagnetic waves through domains with obstacles emphasizes its robustness. This encourages us to consider a three dimensional implementation. However, these two dimensional cases are not really representative of three dimensional simulations which are more complex to implement.

Trefftz methods provide a particularly adapted framework for enriched Galerkin method

in the context of multi-scale modeling as presented by Dauge et al. at the Enumath 2019 conference. This question is of particular importance for Maxwell equations. We are rather convinced that the proposed formulation is perfectly adapted to this type of problems.

The iterative solution of Trefftz problems is not dealt with in this paper. This seems to be a good alternative to efficient preconditioner like in [97]. Testing GMRES solver [91] seems a good question for a future research program but is mostly relevant for three dimensional problems.

Conclusion

Dans le Chapitre 1, nous avons introduit deux méthodes numériques que nous jugions représentatives (en termes de coût mémoire) des méthodes standard existantes. La première, la méthode d'EF, basée sur le problème de Maxwell d'ordre 2, permet de simuler des ondes électromagnétiques sur des domaines (composés de cubes dans notre cas) allant jusqu'à $\mathcal{D}_\Omega^{\max} = 27\lambda$. La seconde est une méthode de GD dérivée à partir du problème de Maxwell d'ordre 1. La formulation variationnelle associée a été discrétisée sur un domaine maillé en tétraèdres. Cette deuxième méthode a eu finalement des capacités limitées en termes de mémoire face à la première méthode : $\mathcal{D}_\Omega^{\max} = 9\lambda$. Les méthodes d'EF et de GD que nous avons présentées dans le Chapitre 1 sont alors inutilisables dans le cadre de grandes scènes de calcul. Toutefois, nos implémentations auraient pu être davantage optimisées. De plus, dans le cas où un nombre très important de seconds membres est considéré, des méthodes directes performantes resteront, quoi qu'il arrive, plus compétitives que des méthodes itératives. Leurs performances sont alors à ne pas négliger.

Dans le Chapitre 2, nous avons développé une méthode de Trefftz directe, reposant sur une factorisation LU. Cette méthode est basée sur une formulation variationnelle et sur un espace discret. Ce dernier est composé d'ondes planes dans notre cas, bien qu'elles causent des problèmes de conditionnement. Nous avons présenté deux approches de construction des formulations variationnelles Trefftz. Dans les deux cas, il s'agit d'une perturbation de la formule de réciprocité que nous avons introduite. Cette dernière provient notamment de la formule des travaux virtuels et transcrit la physique du phénomène étudié. D'une part, une approche par l'ajout de formes consistantes a été présentée. Cela nous a permis d'écrire les formulations Trefftz sous la forme d'un tableau de coefficients pour lesquels nous avons donné des critères pour assurer la coercivité de la formulation. D'autre part, nous avons expliqué la notion de traces numériques. Ces dernières peuvent aussi être utilisées pour perturber la formule de réciprocité et mener à des formulations variationnelles Trefftz bien posées. Il s'agit d'une généralisation de traces de fonctions continues au cas discontinu (ainsi adaptées au GD). Plus précisément, les traces numériques de Riemann, dans le cas homogène, et les traces numériques upwind, dans le cas hétérogène, ont été employées. Ces dernières ont été obtenues grâce à un choix de paramètres permettant à la fois d'assurer la coercivité et de contrôler la continuité de la solution numérique. Ces formulations Trefftz, définies sur le squelette du maillage, permettent d'obtenir une méthode de faible coût mémoire et ainsi de simuler des ondes électromagnétiques sur des domaines de taille $\mathcal{D}_\Omega^{\max} = 35\lambda$. Cependant, nous avons observé que l'utilisation d'une factorisation LU reste un frein pour considérer des tailles de domaine supérieures. Une méthode de Trefftz directe est donc seulement adaptée pour traiter des scènes de calcul de taille intermédiaire.

Dans le Chapitre 3, nous avons mis en place une méthode de Trefftz itérative menant à l'implémentation du code GoTEM3. Nous avons débuté par la dérivation des traces numériques de Cessenat-Després. Elles ont conduit à une nouvelle formulation variationnelle Trefftz UWVF consistante, coercive et équivalente à la formulation upwind construite dans le Chapitre 2. L'UWVF nous a permis de mettre en oeuvre un solveur de type Jacobi. Ce dernier n'a pas de bonnes propriétés de convergence lorsque le spectre de la matrice devient trop étalé. Nous avons alors choisi de développer deux méthodes de Krylov. Elles sont toutes les deux définies sur des espaces de Krylov. La première est une méthode de GMRES. Elle consiste à minimiser la norme du résidu associé au système matriciel. Elle est très généralement utilisée avec une stratégie de *restart*, permettant de diminuer son coût mémoire. La seconde méthode présentée dans le Chapitre 3, une méthode de Krylov Galerkin, emploie elle aussi le *restart* et est définie sur un espace de Krylov. Contrairement à la première, son cadre de travail de Galerkin donne de bonnes propriétés pour prouver sa convergence dans le cas d'un préconditionnement de type Cessenat-Després. Nous espérions alors obtenir de meilleurs résultats de convergence en appliquant la méthode de Krylov Galerkin pour résoudre le problème de Trefftz UWVF (préconditionné ou non). Or, il s'est avéré que la méthode de Krylov Galerkin que nous avons développée a un temps de calcul et un coût mémoire plus élevé que l'algorithme de GMRES du CERFACS©. Elle converge cependant en moins d'itérations que la méthode de GMRES, ce qui révèle le potentiel de cette méthode. De plus, nous avons aussi observé que la méthode de GMRES UWVF avec préconditionnement de Cessenat-Després donne un gain en termes de nombre d'itérations face à une méthode non préconditionnée. Finalement, nous avons construit une méthode de Trefftz itérative basée sur un algorithme GMRES. En choisissant un nombre de vecteur adapté dans l'espace de Krylov, nous atteignons une taille de domaine $\mathcal{D}_\Omega^{\max} = 150\lambda$ (pour 1To de mémoire).

Dans le Chapitre 4, nous avons développé un nouveau préconditionneur "global", mettant en jeu les trois dimensions du domaine et assurant une convergence plus rapide de la solution numérique au fur et à mesure des itérations GMRES. En effet, ce préconditionneur a fait très largement ses preuves face à celui de Cessenat-Després. Nous avons pu observer des diminutions d'un facteur 2 en temps de calcul, et d'un facteur 4 en nombre d'itérations. Dans ce même chapitre, nous avons aussi proposé une stratégie de réduction de la base d'ondes planes afin d'écarter celles qui apportent peu d'informations à la description de la solution numérique. Cela a mené à l'introduction d'un problème de Trefftz GMRES UWVF réduit pour lequel le coût de stockage de l'espace de Krylov est alors diminué vis à vis de celui sans réduction. Des gains, à la fois en temps d'exécution, en nombre d'itérations et en mémoire, ont été observés. Enfin, la stratégie permettant de passer de $\mathcal{D}_\Omega^{\max} = 150\lambda$ à $\mathcal{D}_\Omega^{\max} = 370\lambda$ est le désassemblage de la matrice Trefftz. Cela a été rendu possible grâce à la structure cartésienne du maillage. C'est en grande partie grâce à cette mise en oeuvre que nous avons

simulé une onde électromagnétique à l'échelle d'un cas industriel. Il s'agit d'un écho radar sur un porte-avion considéré parfaitement conducteur, de taille $24 \times 61 \times 154$ mètres (pour une longueur d'onde d'un mètre), avec une taille d'élément $h = 0.25$ et $R_{\partial\Omega} = 0$, voir la Figure 5.13. Sur cette figure, nous observons l'ombre sous le bateau et le champ proche électromagnétique généré par l'onde incidente. Les grandes dimensions de cette expérience numérique et les coûts associés (voir le Tableau 5.1) témoignent de la robustesse du code GoTEM3.

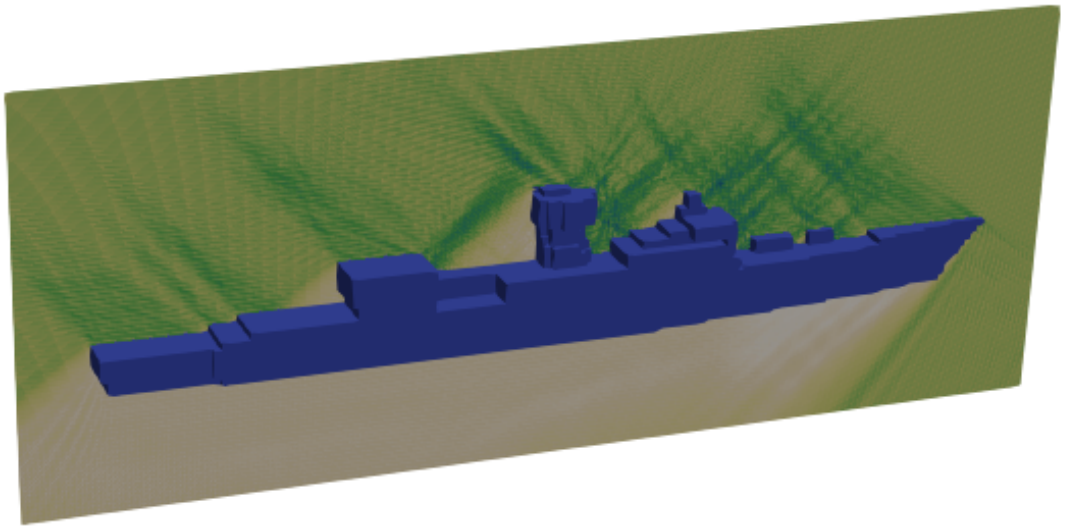

FIGURE 5.13 – Visualisation de l'amplitude d'une onde électromagnétique frappant le bateau de type porte-avion.

#elem	N_{red}	N	$\#\text{ddl}_{\text{red}}$	#ddl
$> 14.4 \times 10^6$	46	52	$> 0.663 \times 10^9$	$> 0.75 \times 10^9$
N_{kry}	Nb Itérations	Coût (Go)	Temps (h)	Résidu GMRES $e^{\text{prec}}_{N_{\text{kry}}}$
10	800	389	28.6	3.8×10^{-2}

TABLE 5.1 – Résultats associés à l'expérience numérique du bateau, voir la Figure 5.13.

Nous avons finalement proposé une méthode Quasi-Trefftz pour pallier les problèmes de conditionnement causés par le caractère numériquement lié des ondes planes. La démarche consiste à résoudre localement les équations de Maxwell grâce à une méthode d'EF de Nédélec d'ordre élevé, où l'ordre est noté p. En particulier, nous avons mis en évidence une relation de quasi-optimalité entre l'ordre q de la méthode de Trefftz et celui de la méthode d'EF utilisée : $p = q + 1$. De plus, l'erreur de dispersion numérique sur la solution Trefftz est

inférieure à celle de la solution d'EF, alors que cette dernière est pourtant déjà connue pour être de bonne qualité.

Perspectives

Le travaux présentés dans cette thèse autour du développement d'un solveur de Trefftz amènent à plusieurs perspectives de recherche. Nous les classons en trois thèmes.

Amélioration de la formulation :

Tout d'abord, nous pensons qu'une approche Quasi-Trefftz possède toutes les caractéristiques pour envisager de traiter des scènes de calcul tridimensionnelles plus complexes. La mise en place de cette approche nécessite le développement d'un solveur local rapide pour construire les fonctions de base. Pour cela, différents solveurs peuvent être utilisés. Nous proposons d'étudier une approche appelée Différences Spectrales. Elle consisterait dans notre cas à chercher la solution, sous forme de polynômes élément par élément, d'un système multi-second membres des équations fortes de Maxwell associé au calcul des fonctions de base.

La seconde amélioration est de dériver une formulation à partir d'opérateurs de transmission plus précis. En effet, la méthode de Trefftz est un décomposeur de domaine naturel qui s'appuie sur une approximation des opérateurs de transmission/réflexion. Cette dernière a un impact sur le nombre d'itérations nécessaires au solveur itératif pour converger à une précision donnée. Nous envisageons d'étudier deux approches : par des conditions d'ordre élevé ou par un apprentissage automatique.

Amélioration de la méthode itérative et de préconditionnement :

Plusieurs stratégies peuvent être étudiées afin d'améliorer la vitesse de convergence des solveurs de type Krylov. Premièrement, il serait intéressant d'employer la version flexible d'un algorithme GMRES. Cette option permet d'alterner les préconditionneurs au fur et à mesure des itérations et ainsi d'accélérer la convergence. De plus, nous pourrions implémenter notre propre variante flexible dans GoTEM3. Il s'agirait d'augmenter la taille de l'espace réduit au cours des itérations. Cette tactique permettrait d'accélérer les premières itérations sans détériorer la précision finale. Par ailleurs, la recherche de différents préconditionneurs globaux est aussi un enjeu majeur comme pour tous les problèmes de propagation d'ondes.

Implémentation et HPC :

Dans le but de tenir compte de la courbure des obstacles, nous pourrions premièrement avoir recours à des maillages hybrides qui consistent à utiliser des tétraèdres et des prismes au voisinage de la frontière. Cette tâche très complexe nécessitera une coopération avec des équipes spécialisées. Deuxièmement, l'extension de la stratégie de désassemblage à des problèmes hétérogènes est en cours. Elle permettra de considérer des milieux de propagation hétérogènes constitués de différents matériaux diélectriques ou même magnétiques. Par ailleurs, l'amélioration des performances de GoTEM3 passera par une implémentation parallèle hybride OpenMP/MPI. Il ne s'agit pas seulement de diminuer les temps d'exécution mais de pouvoir considérer des cas numériques de l'ordre du PetaOctet. Le passage de la mémoire partagée à la mémoire distribuée est donc un changement de paradigme notamment dans le cadre d'une méthode de Krylov.

BIBLIOGRAPHIE

[1] M. AINSWORTH : Dispersive and dissipative behaviour of high order discontinuous Galerkin finite element methods. Journal of Computational Physics, 198:106–130, 2004.

[2] M. AINSWORTH : Dispersive properties of high–order Nédélec/edge element approximation of the time–harmonic Maxwell equations. Philosophical Transactions of the Royal Society of London. Series A : Mathematical, Physical and Engineering Sciences, 362(1816):471–491, 2004.

[3] M. AINSWORTH, P. MONK et W. MUNIZ : Dispersive and dissipative properties of discontinuous Galerkin finite element methods for the second-order wave equation. Journal of Scientific Computing, 27(1–3):5–40, 2006.

[4] P. AMESTOY, I. DUFF et J.-Y. L'EXCELLENT : Mumps multifrontal massively parallel solver version 2.0, 1998.

[5] P. R. AMESTOY, A. BUTTARI, J.-Y. L'EXCELLENT et T. A. MARY : Bridging the gap between flat and hierarchical low-rank matrix formats : The multilevel block low-rank format. SIAM Journal on Scientific Computing, 41(3):A1414–A1442, 2019.

[6] I. BABUŠKA et J. M. MELENK : The partition of unity method. International journal for numerical methods in engineering, 40(4):727–758, 1997.

[7] N. BALIN : Etude de méthodes de couplage pour la résolution des équations de Maxwell. Application au calcul de la signature radar d'aéronefs par hybridation de méthodes exactes et asymptotiques. Thèse de doctorat, Institut National des Sciences Appliquées de Toulouse, 2005.

[8] A. BARKA et N. DOUCHIN : Asymptotic simplifications for hybrid BEM/GO/PO/PTD techniques based on a generalized scattering matrix approach. Computer Physics Communications, 183(9):1928–1936, 2012.

[9] H. BARUCQ, A. BENDALI, J. DIAZ et S. TORDEUX : Local strategies for improving the conditioning of the plane-wave ultra-weak variational formulation. Journal of Computational Physics, 441:110449, 2021.

[10] H. BARUCQ, A. BENDALI, M. FARES, V. MATTESI et S. TORDEUX : A symmetric DG formulation based on a local boundary element method for the solution of the Helmholtz equation. Journal of Computational Physics, 330:1069–1092, 2017.

[11] M. BEBENDORF : Approximation of boundary element matrices. Numerische Mathematik, 86(4):565–589, 2000.

[12] A. BONITO, J.-L. GUERMOND et F. LUDDENS : Regularity of the Maxwell equations in heterogeneous media and Lipschitz domains. Journal of Mathematical Analysis and applications, 408(2):498–512, 2013.

[13] P. BONNET, X. FERRIERES, B. MICHIELSEN, P. KLOTZ et J. ROUMIGUIÈRES : Finite-volume time domain method. San Diego, CA : Academic Press, 1999.

[14] O. P. BRUNO et L. A. KUNYANSKY : A fast, high-order algorithm for the solution of surface scattering problems : basic implementation, tests, and applications. Journal of Computational Physics, 169(1):80–110, 2001.

[15] A. BUFFA, M. COSTABEL et D. SHEEN : On traces for H(curl) in Lipschitz domains. Journal of Mathematical Analysis and Applications, 276(2):845–867, 2002.

[16] A. BUFFA et R. HIPTMAIR : Galerkin boundary element methods for electromagnetic scattering. In Topics in computational wave propagation, p. 83–124. Springer, 2003.

[17] A. BUFFA et P. MONK : Error estimates for the ultra weak variational formulation of the Helmholtz equation. Mathematical Modelling and Numerical Analysis, 42(6):925–940, 2008.

[18] A. BUFFA et I. PERUGIA : Discontinuous Galerkin approximation of the Maxwell eigenproblem. SIAM Journal on Numerical Analysis, 44(5):2198–2226, 2006.

[19] E. BURMAN, H. WU et L. ZHU : Linear continuous interior penalty finite element method for Helmholtz equation with high wave number : one-dimensional analysis. Numerical Methods for Partial Differential Equations, 32(5):1378–1410, 2016.

[20] F. CAKONI, D. COLTON et P. MONK : The electromagnetic inverse-scattering problem for partly coated Lipschitz domains. Proceedings of the Royal Society of Edinburgh Section A : Mathematics, 134(4):661–682, 2004.

[21] O. CESSENAT : Application d'une nouvelle formulation variationnelle aux équations d'ondes harmoniques. Problèmes d'Helmholtz 2D et de Maxwell 3D. Thèse de doctorat, University of Paris XI Dauphine, 1996.

[22] O. CESSENAT et B. DESPRÉS : Application of an ultra weak variational formulation of elliptic PDE to the two-dimensional Helmholtz problem. SIAM J. Num. Analysis, 35(1):255–299, 1998.

[23] P. G. CIARLET, P. G. CIARLET, B. MIARA et J.-M. THOMAS : Introduction to numerical linear algebra and optimisation. Cambridge University Press, 1989.

[24] B. COCKBURN, J. GOPALAKRISHNAN et R. LAZAROV : Unified hybridization of discontinuous Galerkin, mixed, and continuous Galerkin methods for second order elliptic problems. SIAM Journal on Numerical Analysis, 47(2):1319–1365, 2009.

[25] P.-H. COCQUET et M. J. GANDER : How large a shift is needed in the shifted Helmholtz preconditioner for its effective inversion by multigrid ? SIAM Journal on Scientific Computing, 39(2):A438–A478, 2017.

[26] G. COHEN et S. PERNET : Finite element and discontinuous Galerkin methods for transient wave equations. Springer, 2017.

[27] G. C. COHEN : Higher-order numerical methods for transient wave equations, vol. 5. Springer, 2002.

[28] F. COLLINO, S. GHANEMI et P. JOLY : Domain decomposition method for harmonic wave propagation : a general presentation. Computer methods in applied mechanics and engineering, 184(2-4):171–211, 2000.

[29] F. COLLINO, P. JOLY et M. LECOUVEZ : Exponentially convergent non overlapping domain decomposition methods for the Helmholtz equation. ESAIM : Mathematical Modelling and Numerical Analysis, 54(3):775–810, 2020.

[30] S. CONGREVE, J. GEDICKE et I. PERUGIA : Numerical investigation of the conditioning for plane wave discontinuous Galerkin methods. In European Conference on Numerical Mathematics and Advanced Applications, p. 493–500. Springer, 2017.

[31] M. COSTABEL : A remark on the regularity of solutions of Maxwell's equations on Lipschitz domains. Mathematical Methods in the Applied Sciences, 12(4):365–368, 1990.

[32] M. COSTABEL, M. DAUGE et S. NICAISE : Singularities of Maxwell interface problems. ESAIM : Mathematical Modelling and Numerical Analysis, 33(3):627–649, 1999.

[33] E. DARVE : The fast multipole method : numerical implementation. Journal of Computational Physics, 160(1):195–240, 2000.

[34] E. DARVE et P. HAVÉ : Efficient fast multipole method for low-frequency scattering. Journal of Computational Physics, 197(1):341–363, 2004.

[35] B. DESPRÉS : Sur une formulation variationnelle ultra-faible. Comptes Rendus de l'Académie des Sciences, Série I 318:939–944, 1994.

[36] V. DOLEAN, H. FOL, S. LANTERI et R. PERRUSSEL : Solution of the time-harmonic Maxwell equations using discontinuous Galerkin methods. Journal of computational and applied mathematics, 218(2):435–445, 2008.

[37] V. DOLEAN, M. J. GANDER, S. LANTERI, J.-F. LEE et Z. PENG : Effective transmission conditions for domain decomposition methods applied to the time-harmonic curl–curl Maxwell's equations. Journal of computational physics, 280:232–247, 2015.

[38] M. DURUFLE : Intégration numérique et éléments finis d'ordre élevé appliqués aux équations de Maxwell en régime harmonique. Thèse de doctorat, Citeseer, 2006.

[39] M. EL BOUAJAJI, B. THIERRY, X. ANTOINE et C. GEUZAINE : A quasi-optimal domain decomposition algorithm for the time-harmonic Maxwell's equations. Journal of Computational Physics, 294:38–57, 2015.

[40] M. EMBREE : How descriptive are GMRES convergence bounds ? 1999.

[41] O. G. ERNST et M. J. GANDER : Why it is difficult to solve Helmholtz problems with classical iterative methods. Numerical analysis of multiscale problems, p. 325–363, 2012.

[42] V. ERVIN : RTK and BDMK on triangles. Computers & Mathematics with Applications, 64(8):2765–2774, 2012.

[43] C. FARHAT, R. TEZAUR et P. WEIDEMANN-GOIRAN : Higher-order extensions of a discontinuous Galerkin method for mid-frequency Helmholtz problems. International journal for numerical methods in engineering, 61(11):1938–1956, 2004.

[44] V. FRAYSSÉ, L. GIRAUD, S. GRATTON et J. LANGOU : A set of GMRES routines for real and complex arithmetics. URL http ://www. cerfacs. fr/algor/Softs/GMRES/index. html, 1997.

[45] H. S. FURE, S. PERNET, M. SIRDEY et S. TORDEUX : A discontinuous Galerkin Trefftz type method for solving the two dimensional Maxwell equations. SN Partial Differential Equations and Applications, 1(4):1–25, 2020.

[46] G. GABARD : Discontinuous Galerkin methods with plane waves for time-harmonic problems. Journal of Computational Physics, 225:1961–1984, 2007.

[47] P. GAMALLO et R. J. ASTLEY : A comparison of two Trefftz-type methods : The ultra-weak variational formulation and the least-squares method, for solving shortwave 2-

D Helmholtz problems. International Journal for Numerical Methods in Engineering, 71:406–432, 2007.

[48] C. GITTELSON et R. HIPTMAIR : Dispersion analysis of plane wave discontinuous methods. International Journal for Numerical Methods in Engineering, 98(5):313–323, 2014.

[49] C. J. GITTELSON, R. HIPTMAIR et I. PERUGIA : Plane wave discontinuous Galerkin methods : analysis of the h-version. ESAIM : Mathematical Modelling and Numerical Analysis, 43(2):297–331, 2009.

[50] J. S. HESTHAVEN et T. WARBURTON : Nodal discontinuous Galerkin methods : algorithms, analysis, and applications. Springer Science & Business Media, 2007.

[51] R. HIPTMAIR : Finite elements in computational electromagnetism. Acta Numerica, 11:237–339, 2002.

[52] R. HIPTMAIR, A. MOIOLA et I. PERUGIA : Plane wave discontinuous Galerkin methods for the 2D Helmholtz equation : analysis of the p-version. SIAM Journal on Numerical Analysis, 49(1):264–284, 2011.

[53] R. HIPTMAIR, A. MOIOLA et I. PERUGIA : Stability results for the time-harmonic Maxwell equations with impedance boundary conditions. Mathematical Models and Methods in Applied Sciences, 21(11):2263–2287, 2011.

[54] R. HIPTMAIR, A. MOIOLA et I. PERUGIA : A survey of Trefftz methods for the Helmholtz equation. In Building bridges : connections and challenges in modern approaches to numerical partial differential equations, p. 237–279. Springer, 2016.

[55] R. HIPTMAIR, A. MOIOLA, I. PERUGIA et C. SCHWAB : Approximation by harmonic polynomials in star-shaped domains and exponential convergence of Trefftz hp-dgfem. Mathematical Modelling and Numerical Analysis, 48:727–752, 2014.

[56] C. HOFREITHER : A Non-standard Finite Element Method using Boundary Integral Operators. Thèse de doctorat, J. Kepler University, Linz, 2012.

[57] C. HOFREITHER, U. LANGER et S. WEISSER : Convection-adapted BEM-based FEM. ZAMM-Journal of Applied Mathematics and Mechanics/Zeitschrift für Angewandte Mathematik und Mechanik, 96(12):1467–1481, 2016.

[58] L. HÖRMANDER : On the uniqueness of the cauchy problem. Mathematica Scandinavica, p. 213–225, 1958.

[59] P. HOUSTON, I. PERUGIA, A. SCHNEEBELI et D. SCHÖTZAU : Interior penalty method for the indefinite time-harmonic Maxwell equations. Numerische Mathematik, 100(3): 485–518, 2005.

[60] T. HUTTUNEN, M. MALINEN et P. MONK : Solving Maxwell's equations using the ultra weak variational formulation. Journal of Computational Physics, 223(2):731–758, 2007.

[61] P. M. I. PERUGIA, D. Schötzau : Stabilized interior penalty methods for the time-harmonic Maxwell equations. Comput. Methods Appl. Mech. Engrg., 191:4675–4697, 2002.

[62] F. IHLENBURG et I. BABUSKA : Finite element solution of the Helmholtz equation with high wave number – part i : the h-version of the FEM. Computers Math. Applic., 30(9):9–37, 1995.

[63] F. IHLENBURG et I. BABUSKA : Finite element solution of the Helmholtz equation with high wave number – part ii : the h-p version of the FEM. SIAM J. Numer. Anal., 34(1):315–358, 1997.

[64] L.-M. IMBERT-GERARD : Amplitude-based generalized plane waves : New quasi-Trefftz functions for scalar equations in two dimensions. SIAM Journal on Numerical Analysis, 59(3):1663–1686, 2021.

[65] L.-M. IMBERT-GÉRARD, A. MOIOLA et P. STOCKER : A space-time quasi-Trefftz DG method for the wave equation with piecewise-smooth coefficients. arXiv preprint arXiv :2011.04617, 2020.

[66] J. M. JIN : The Finite Element Method in Electromagnetics, Second Edition. John Wiley & Sons, New York, 2002.

[67] E. KITA et N. KAMIYA : Trefftz method : an overview. Advances in Engineering software, 24(1-3):3–12, 1995.

[68] S. KURZ, O. RAIN et S. RJASANOW : The adaptive cross-approximation technique for the 3D boundary-element method. IEEE transactions on Magnetics, 38(2):421–424, 2002.

[69] J. LABAT : Modélisation multi-échelle de la diffraction des ondes électromagnétiques par de petits obstacles. Thèse de doctorat, Pau, 2019.

[70] J. LIESEN et P. TICHÝ : Convergence analysis of krylov subspace methods. GAMM-Mitteilungen, 27(2):153–173, 2004.

[71] J. LIESEN et P. TICHÝ : The worst-case GMRES for normal matrices. BIT Numerical mathematics, 44(1):79–98, 2004.

[72] T. LUOSTARI, T. HUTTUNEN et P. MONK : Improvements for the ultra weak variational formulation. International Journal for Numerical Methods in Engineering, 94(6): 598–624, 2013.

[73] W. MCLEAN : Strongly Elliptic Systems and Boundary Integral Equations. Cambridge University Press, Cambridge university press, 2000.

[74] J. M. MELENK, A. PARSANIA et S. SAUTER : General DG-methods for highly indefinite Helmholtz problems. J. Sci. Comput., 57:536–581, 2013.

[75] A. MODAVE, X. ANTOINE et C. GEUZAINE : An efficient domain decomposition method with cross-point treatment for Helmholtz problems. In CSMA 2019-14e Colloque National en Calcul des Structures, 2019.

[76] A. MOIOLA : Trefftz-discontinuous Galerkin methods for time-harmonic wave problems. Thèse de doctorat, ETH Zurich, 2011.

[77] A. MOIOLA, R. HIPTMAIR et I. PERUGIA : Plane wave approximation of homogeneous Helmholtz solutions. Z. Angew. Math. Phys., 62:809–837, 2011.

[78] P. MONK : Finite Element Methods for Maxwell's Equations. Numerical Analysis and Scientific Computations. Clarendon Press, 2003.

[79] D. MORO, N. NGUYEN et J. PERAIRE : A hybridized discontinuous Petrov–Galerkin scheme for scalar conservation laws. International journal for numerical methods in engineering, 91(9):950–970, 2012.

[80] C.-D. MUNZ, R. SCHNEIDER et U. VOSS : A finite-volume method for the Maxwell equations in the time domain. SIAM Journal on Scientific Computing, 22(2):449–475, 2000.

[81] J.-C. NÉDÉLEC : Mixed finite elements in $\mathbb{R}^3$. Numerische Mathematik, 35(3):315–341, 1980.

[82] J.-C. NÉDÉLEC : A new family of mixed finite elements in $\mathbb{R}^3$. Numerische Mathematik, 50(1):57–81, 1986.

[83] J.-C. NÉDÉLEC : Acoustic and electromagnetic equations : integral representations for harmonic problems, vol. 144. Springer, 2001.

[84] N. C. NGUYEN, J. PERAIRE et B. COCKBURN : Hybridizable discontinuous Galerkin methods for the time-harmonic Maxwell's equations. Journal of Computational Physics, 230(19):7151–7175, 2011.

[85] E. PAROLIN, D. HUYBRECHS et A. MOIOLA : Stable approximation of Helmholtz solutions by evanescent plane waves. arXiv preprint arXiv :2202.05658, 2022.

[86] S. PERNET, N. SERDIUK, M. SIRDEY et S. TORDEUX : Discontinuous Galerkin method based on Riemann fluxes for the time domain Maxwell System. Thèse de doctorat, INRIA Bordeaux-Sud-Ouest, 2021.

[87] E. PERREY-DEBAIN : Plane wave decomposition in the unit disc : convergence estimates and computational aspects. Journal of Computational and Applied Mathematics, 193(1):140–156, 2006.

[88] B. PLUYMERS, B. VAN HAL, D. VANDEPITTE et W. DESMET : Trefftz-based methods for time-harmonic acoustics. Archives of Computational Methods in Engineering, 14(4):343–381, 2007.

[89] M. H. PROTTER : Unique continuation for elliptic equations. Num. 7. Mathematics Division, Office of Scientific Research, US Air Force, 1959.

[90] L. G. RAMOS, O. SETE et R. NABBEN : Preconditioning the Helmholtz equation with the shifted laplacian and faber polynomials. 2021.

[91] Y. SAAD : Iterative Methods for Sparse Linear Systems. PWS Pblishing Company, Boston, 1996.

[92] Y. SAAD et M. H. SCHULTZ : GMRES : a Generalized Minimal RESidual algorithm for solving nonsymmetric linear systems. SIAM Journal on scientific and statistical computing, 7(3):856–869, 1986.

[93] S. A. SAUTER et C. SCHWAB : Boundary Element Methods. Springer-Verlag, Berlin-Heidelberg, 2011.

[94] T. B. A. SENIOR et J. L. VOLAKIS : Approximate Boundary Conditions in Electromagnetics. IEE Electromagnetic Waves Series 41. IEE Press, New York, 1995.

[95] K. Y. SZE et G. H. LIU : Hybrid-Trefftz six-node triangular finite element models for Helmholtz problems. Computational Mechanics, 46(6):455–470, 2010.

[96] A. TAFLOVE : Computational Electrodynamics. The Finite-Difference Time-Domain Method. Artech house, Inc., Norwood, MA 02062, 1995.

[97] A. VION et C. GEUZAINE : Double sweep preconditioner for optimized Schwarz methods applied to the Helmholtz problem. J. Comput. Phys., 266:171–190, 2014.

[98] D. WANG, R. TEZAUR, J. TOIVANEN et C. FERHAT : Overview of the discontinuous enrichment method, the ultra-weak variational formulation, and the partition of unity method for the acoustic scattering in the medium frequency regime and performance comparisons. International Journal for Numerical Methods in Engineering, 89:403–417, 2012.

[99] N. ZERBIB : Méthodes de Sous-Structuration et de Décomposition de Domaine pour la Résolution des Équations de Maxwell : Application au Rayonnement d'antenne dans un Environnement Complexe. Thèse de doctorat, National Institute for Applied Sciences (INSA), INSA Toulouse, 2006.

[100] K. ZHAO, M. N. VOUVAKIS et J.-F. LEE : The adaptive cross approximation algorithm for accelerated method of moments computations of EMC problems. IEEE transactions on electromagnetic compatibility, 47(4):763–773, 2005.

TABLE DES FIGURES

LISTE DES TABLEAUX

Résumé : La simulation d'ondes électromagnétiques en trois dimensions intervient dans de nombreuses applications civiles et militaires et met très souvent en jeu la résolution de très grands systèmes linéaires. La mémoire nécessaire pour la factorisation LU de la matrice croît très rapidement avec la taille du domaine de calcul de telles sortes que les méthodes de type EF ou GD classiques sont inutilisables. Cela conduit naturellement à employer une méthode itérative. Dans cette thèse, nous développons GoTEM3, un solveur Trefftz itératif HPC basé sur des espaces de Krylov.

Les méthodes de Trefftz peuvent être interprétées comme des méthodes de Galerkin Discontinues dont les fonctions de base sont des solutions locales des équations aux dérivées partielles étudiées. Les formulations variationnelles Trefftz sont présentées sous le point de vue de formes consistantes ou de traces numériques. Ces dernières sont obtenues alternativement, pour les milieux homogènes, par un solveur de Riemann, et dans le cas général des milieux hétérogènes, par un problème de Cessenat-Després ou upwind. Elles conduisent toutes à des formulations équivalentes et coercives. Un algorithme itératif reposant sur l'UWVF de Cessenat-Després mène à un problème de point fixe dont la matrice est contractante. Toutefois, cette propriété n'est parfois plus vérifiée numériquement à cause des erreurs d'arrondis. Nous mettons alors en place un solveur GMRES et un solveur de type Krylov Galerkin dans GoTEM3. Les fonctions de base employées sont des ondes planes et peuvent devenir linéairement dépendantes numériquement. Un nouveau préconditionneur global, au sens où il implique les trois dimensions du domaine, permet d'obtenir une solution numérique précise avec nettement moins d'itérations qu'un préconditionneur de Cessenat-Després. L'amélioration du conditionnement passe aussi par une stratégie de réduction de la base d'ondes planes, conduisant à des diminutions du temps d'exécution et du coût mémoire. Ce dernier aspect est particulièrement optimisé avec un désassemblage de la matrice, rendu possible grâce au caractère cartésien du maillage. Ainsi, GoTEM3 simule les ondes électromagnétiques sur des domaines contenant plus d'un milliard de degrés de liberté.

Mot-clés : Équations de Maxwell, Haute fréquence, Méthode de Trefftz, Méthode itérative, Ondes planes, Espace de Krylov, GMRES, Préconditionnement, Réduction de base, Désassemblage.

Abstract : Three-dimensional electromagnetic waves simulation is used in several civil and military applications. It often involves large linear systems leading to memory cost and computation time issues. When using a LU factorisation, the memory needed to invert the matrix increases very quickly with the domain size. In this thesis, we develop GoTEM3 : a HPC Trefftz iterative solver.

Trefftz methods can be interpreted as Discontinuous Galerkin methods whose basis functions are local solutions of the studied partial differential equations. Trefftz variational formulations are presented following either the consistant forms or the numerical traces point of view. The latter are obtained, in the homogeneous case, thanks to a Riemann solver, and in the heterogeneous case, thanks to a Cessenat-Després or an upwind problem. They all lead to equivalent and coercive variational formulations. A fixed point algorithm based on Cessenat-Després UWVF is written and uses a contractant matrix. However, this property might not be numerically satisfied due to rounding errors. We then derive two Trefftz iterative solvers based on Krylov spaces : a GMRES solver and a Krylov Galerkin solver. In this thesis, basis functions are plane waves and can become numerically dependant. A new global preconditioner, implying the three directions of the domain, outperforms the Cessenat-Després preconditioner by using significantly fewer iterations. Conditioning enhancement is also performed thanks to a plane waves basis reduction, leading to both time and memory gains. This latter aspect is strongly improved by the use of a free-matrix strategy achieved thanks to the cartesian structure of the mesh. GoTEM3 is therefore a code simulating electromagnetic waves on domains containing more than one billion of degrees of freedom.

Keywords : Maxwell equations, High frequency, Trefftz method, Iterative method, Plane waves, Krylov space, GMRES, Preconditioning, Basis reduction, Free-matrix.