N^o d'ordre : 2005-35 Année 2005

THÈSE

présentée à

L'ÉCOLE CENTRALE DE LYON

pour l'obtention du

DIPLÔME DE DOCTORAT

Spécialité : Mathématiques appliquées

soutenue publiquement le 27 octobre 2005

par

Ronan Perrussel

Méthodes multiniveau algébriques pour les éléments d'arête. Application à l'électromagnétisme.

Directeurs de Thèse : François Musy, Laurent Nicolas et Michelle Schatzman

JURY :	M	Patrick Dular,	Président et rapporteur,
	M	Stefan Vandewalle,	Examinateur,
	M	Herbert De Gersem,	Examinateur,
	M	François Musy,	Directeur de thèse,
	M	Laurent Nicolas,	Directeur de thèse,
	Mme	Michelle Schatzman,	Directrice de thèse.

Table des matières

Remerciements **vii**

Notations **ix**

Introduction **xi**
0.1 Objectif . . . xi
0.2 Plan . . . xii
0.3 Contributions originales . . . xiii

1 Modélisation et formulation **1**
1.1 Modélisation des phénomènes électromagnétiques . . . 1
1.1.1 Equations de Maxwell . . . 1
1.1.2 Modèles considérés . . . 2
1.1.3 Conditions aux limites et d'interface . . . 4
1.2 Propriétés remarquables et discrétisation . . . 5
1.2.1 Théorie des réseaux électriques . . . 5
1.2.2 Réseau électrique et discrétisation sur la triangulation d'une surface plane . . . 11
1.2.3 Extension à la dimension trois . . . 15
1.2.4 Formulation variationnelle et discrétisation . . . 15

2 Méthodes itératives pour la résolution des systèmes linéaires **21**
2.1 Notions de base pour les méthodes itératives . . . 21
2.1.1 Méthodes par sous-espaces de Krylov . . . 21
2.1.2 Préconditionnement . . . 23
2.2 Principe des méthodes multiniveau . . . 24
2.2.1 Méthode à deux grilles . . . 24
2.2.2 Méthode multigrille géométrique . . . 26
2.2.3 Principe des méthodes multiniveau algébriques . . . 27
2.3 Particularités pour les équations de Maxwell . . . 28
2.3.1 Difficultés liées à l'opérateur "rot rot" . . . 28
2.3.2 Choix du lisseur . . . 29
2.3.3 Utilisation des lisseurs pour le préconditionnement . . . 30
2.3.4 Cas non-structuré avec des méthodes algébriques . . . 30
2.4 Conclusion . . . 34

3 Base d'approximation grossière par minimisation d'énergie et contraintes **35**
3.1 Introduction . . . 35
3.2 Principe et méthode dans le cas des éléments finis nodaux . . . 36
3.2.1 Formulation du problème d'optimisation . . . 36
3.2.2 Résolution avec des multiplicateurs de Lagrange . . . 37
3.3 Extension directe au cas des éléments finis d'arête . . . 38
3.3.1 Étapes préliminaires et formulation du problème d'optimisation . . . 39
3.3.2 Résolution avec multiplicateurs de Lagrange . . . 41

3.4 Conclusion 44

4 Base d'approximation grossière et résolution de problèmes de flot locaux 45
4.1 Méthode de résolution 45
4.1.1 Notations et principe 45
4.1.2 A propos de la résolution du système linéaire (4.10) 48
4.1.3 Avantages par rapport à la méthode avec les multiplicateurs 49
4.2 Résultats numériques pour des problèmes 2D 49
4.2.1 Analyse du conditionnement de la matrice B^tDB 49
4.2.2 Présentation des problèmes tests 50
4.2.3 Simulations numériques sur le premier problème test 52
4.2.4 Résultats pour le second problème test. 56

5 Conclusion et perspectives 59

A Complément sur les espaces fonctionnels 61

B Préconditionnement à un niveau 63
B.1 Un préconditionneur efficace pour les systèmes linéaires provenant de la méthode des éléments finis pour des problèmes de diffraction 63
B.1.1 Introduction 63
B.1.2 Problem Formulation 63
B.1.3 An efficient preconditioner 64
B.1.4 Mesh quality 65
B.1.5 Numerical results: Efficiency and robustness 66
B.1.6 Conclusion 70
B.2 Préconditionneurs pour des problèmes de diffraction résolus par la méthode des éléments finis 70
B.2.1 Introduction 70
B.2.2 Preconditioners 71
B.2.3 Numerical behavior 72
B.2.4 Conclusion 74

C Construction de fonctions nodales grossières par minimisation d'énergie 75
C.1 Principle of the method 75
C.1.1 Formulation of the optimisation problem 76
C.1.2 Method of resolution 76
C.1.3 Decomposition into subdomains 77
C.1.4 Quantitative information on the different meshes 77
C.2 Dirichlet fine and Dirichlet coarse bases (DFDC) 80
C.3 Neumann fine and Dirichlet coarse bases (NFDC) 81
C.4 Adjusting the matrix at the coarse level 82
C.5 Comparison of the number of iterations and of the conditioning on different meshes 85
C.6 Comparison for other boundary conditions 85
C.7 Some results about the Helmholtz equation 87
C.8 Conclusion 88

D Bases nodale et d'arête grossières compatibles et fonctionnelles d'énergie 89
D.1 Introduction 89
D.2 Definition of the continuous problem and its discretization 90
D.2.1 Formulation 90
D.2.2 Finite element space and properties 90
D.3 Overview of the coarse bases construction 92

D.3.1 Energy minimization problems 92
D.3.2 Solution of Problem D.20 93
D.4 Elements required by the construction 94
D.4.1 Algebraic decomposition into subdomains 94
D.4.2 How to choose the R_n's in order to simplify the computational process and to have a unique coarse edge basis 98
D.4.3 Construction of the index sets L_n 100
D.4.4 Definition of a compatible coarse edge incidence matrix G^H 101
D.5 Numerical experiments 103
D.5.1 Matrix multiplication algorithm 103
D.5.2 Choice of the bilinear form b 103
D.5.3 Structured meshes and constant coefficients 104
D.5.4 Unstructured meshes and varying coefficients 106

E Commutativité entre gradient et prolongement et théorie des graphes 109
E.1 Introduction 109
E.2 Notation and statement of the problem 110
E.3 The essential steps of the proof 111
E.4 Construction of the coarse edge functions 113

Remerciements

Je tiens tout particulièrement à exprimer ma gratitude à mes trois encadrants :
- François Musy pour son extrême disponibilité, que ce soit lors des rédactions de longue haleine, pendant lesquelles il a su apporter rigueur et clarté, ou lors de la préparation des cours et exposés au cours de laquelle il m'a fait partager son expérience pédagogique.
- Laurent Nicolas pour m'avoir accordé toute sa confiance même si parfois je m'éloignais quelque peu des problèmes initiaux. Ceci est aussi illustré par les conférences, séminaires et réunions de groupes de recherche auxquels il m'a donné l'opportunité de participer.
- Michelle Schatzman pour sa connaissance encyclopédique et vivante des Mathématiques mais aussi et surtout pour son enthousiasme et son courage communicatif face aux difficultés.

Je remercie également Andrea Toselli qui, malgré une rentrée très chargée, a accepté sans hésiter d'être rapporteur de mon travail de thèse de doctorat. Patrick Dular, de la même manière, mérite une mention spéciale pour l'ensemble de ses remarques pertinentes concernant la finalisation de ce travail et pour son rôle de président parfaitement rempli le jour de la soutenance. Je suis aussi très reconnaissant à Herbert De Gersem et à Stefan Vandewalle pour leur participation au jury de thèse et pour les idées fort intéressantes qu'ils ont apportées pour approfondir les résultats de ce travail.

Je tiens à exprimer ma gratitude à Mohand Moussaoui qui est la première personne à m'avoir orienté vers ce sujet de thèse et qui a toujours été disponible pour répondre à mes interrogations avec cordialité. Laurent Krähenbühl n'a pas non plus été étranger à ce choix et je l'en remercie vivement.

Je tiens à remercier également tous les membres de l'équipe Maxwell que je n'ai pas déjà cités plus haut : Olivier, Christian et Malek pour toutes les interactions que nous avons eues pendant ces trois années. Un petit mot aussi pour des personnes avec lesquelles j'ai eu des échanges plus brefs mais éclairants ; je pense en particulier ici à Jean-François Maître, Joachim Schöberl, Johannes Kraus et Ulrich Langer.

J'aimerais aussi remercier tous mes amis doctorants ou ex-doctorants, en particulier Riccardo, Thierry, Tuan, Clair, Lucas et Laurent. Je réserve aussi une mention spéciale à Josiane, Philippe et Alice pour leur bonne humeur et leur soutien logistique indispensable à l'aboutissement de ces trois années.

Je suis également reconnaissant à tous les membres du MAPLY et du CEGELY que je n'ai pas cités auparavant et à tous ceux que j'omets, mais qui ont tous contribué au bon déroulement de ce travail.

Enfin, mes plus affectueux remerciements iront à Viviane qui a été mon meilleur soutien durant ces trois années, et ce même quand des milliers de kilomètre nous séparaient. Sa contribution a été importante durant toute cette période que ce soit dans la relecture des mes écrits, l'écoute attentive de mes exposés et les encouragements de tous les instants.

Notations

$\mathbf{E}$	Champ électrique (V/m).
$\mathbf{H}$	Champ magnétique (A/m).
$\mathbf{B}$	Induction magnétique (T).
$\mathbf{D}$	Déplacement électrique (C/m^2).
$\mathbf{J}$	Densité de courant (A/m^2).
$\mathbf{A}$	Potentiel vecteur (Wb/m).
ρ	Densité électrique de charge (C/m^3)
ε	Permittivité diélectrique (F/m).
μ	Perméabilité magnétique (H/m).
σ	Conductivité électrique (S/m).
$\mathrm{L}^2(\Omega)$, $\mathbb{L}^2(\Omega)$	Espace des fonctions et des champs de vecteurs de carré intégrable sur $\mathbb{R}^3$.
∂_t	Dérivation par rapport au temps.
$\times$	Produit vectoriel.

Introduction

La mise au point de *systèmes de communication performants*, la prise en compte des problèmes de *compatibilité électromagnétique* avec les systèmes électriques (voir figure 1(b)) ou avec les systèmes biologiques, ainsi que l'utilisation thérapeutique du rayonnement électromagnétique (voir figure 1(a)) nécessitent l'étude des phénomènes de propagation d'ondes électromagnétiques dès la phase de conception, de façon à réduire les temps et les coûts de développement. Cette étude passe par des codes numériques rapides 3D, fondés sur la discrétisation des équations de Maxwell par des techniques d'éléments finis ou de différences finies, utilisant des maillages spatiaux fixes. Actuellement, la mise en oeuvre de ces méthodes sur des systèmes réalistes se heurte au problème crucial du *rapport précision/coût de calcul.*

Le travail présenté participe à cette recherche d'efficacité. La classe de modèles et le type de discrétisation auxquels nous nous intéressons sont identifiés : ils sont associés au calcul numérique des champs électrique ou magnétique, ou du potentiel vecteur magnétique, en utilisant des discrétisations par la méthode des éléments finis.

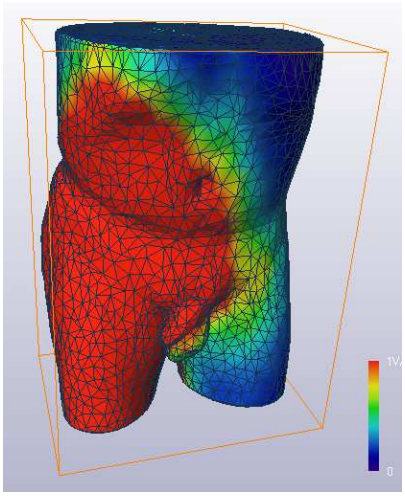
Plus précisément, un exemple important des applications que nous envisageons correspond à l'étude des problèmes de diffraction d'ondes électromagnétiques. Dans ce but, nous calculons le champ électrique (ou magnétique) en utilisant une formulation des équations de Maxwell en régime harmonique. Cette formulation est discrétisée par les éléments finis d'arête du premier ordre. Ces éléments finis sont particulièrement adaptés au calcul en électromagnétisme 3D : ils imposent uniquement la continuité tangentielle à l'interface entre deux éléments, ce qui permet de prendre en compte les propriétés physiques des champ électrique ou magnétique et ils rendent bien compte de la structure particulière du système d'équations de Maxwell.

Pour avoir une précision suffisante, dix noeuds par longueur d'onde sont requis : cela conduit à des systèmes linéaires de 10^4 à 10^6 inconnues pour des géométries classiques. Un traitement efficace de l'étape de résolution, avec tous ses aspects algorithmiques, est donc critique pour optimiser le calcul du champ.

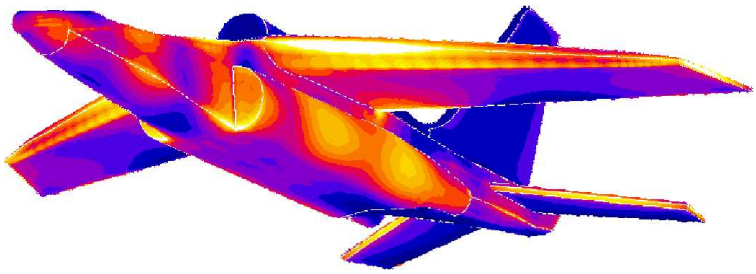
0.1 Objectif

Des *méthodes itératives multiniveau* permettent de résoudre des systèmes linéaires issus de la discrétisation de certaines équations aux dérivées partielles (notamment l'équation de Laplace) de façon très efficace. En particulier, leur coût en temps de calcul et en occupation mémoire est optimal car il est linéaire par rapport au nombre d'inconnues du système linéaire. Elles s'appuient généralement sur une formulation du problème sur une hiérarchie de maillages de finesses distinctes, le plus fin étant celui sur lequel est calculée la solution.

L'objectif de notre travail est d'*obtenir, pour des problèmes de calcul de champ avec l'équation vectorielle des ondes, des algorithmes de résolution dont les performances se rapprochent au mieux des méthodes multiniveau utilisées pour l'équation de Laplace.* En outre, les méthodes recherchées doivent être efficaces pour des maillages *a priori non structurés* c.-à-d. où l'on ne dispose pas d'une hiérarchie de maillages pour résoudre le problème. Dans ce cas, des *méthodes multiniveau algébriques* permettent de générer avec un nombre variable de données sur le problème (uniquement les coefficients de la matrice, ou avec quelques données géométriques supplémentaires) des matrices du problème à plusieurs niveaux. Nous devons cependant adapter ces méthodes aux propriétés des systèmes obtenus lors de la discrétisation des problèmes d'électromagnétisme.



(a) Hyperthermie (échauffement local des tissus) RF (27MHz) pour le traitement de tumeurs profondes. Module du champ électrique.



(b) Illumination (excitation) d'un avion par une onde plane 100Mhz. Répartition de la densité de courant.

FIG. 1 – Deux exemples issus du calcul du champ électrique avec une formulation des équations de Maxwell en régime harmonique.

0.2 Plan

Dans le Chapitre 1, nous présentons quelques aspects de la modélisation et particulièrement de la discrétisation des équations en électromagnétisme ; nous insistons sur les connexions avec la théorie des réseaux et sur les propriétés topologiques de cette discrétisation. Nous formalisons finalement les systèmes linéaires à résoudre et donnons leurs propriétés en fonction des valeurs des paramètres du problème.

Dans le Chapitre 2, nous rappelons quelques notions pour la résolution de systèmes linéaires par méthodes itératives ; nous insistons particulièrement sur le principe et les atouts des méthodes multiniveau. Dans le cas de ces méthodes, nous présentons les particularités liées aux équations de Maxwell et nous introduisons des techniques algébriques qui permettent de travailler avec des maillages non structurés. Les dernières de ces méthodes algébriques mettent en avant une condition de compatibilité qui sert de point de départ à nos développements : elle lie la matrice de prolongement, qui permet de passer d'un niveau grossier à un niveau fin, en éléments finis d'arête à la matrice de prolongement en éléments finis nodaux.

Dans le Chapitre 3, nous présentons la formulation d'un problème de minimisation d'énergie[1] avec contraintes qui permet la mise au point de méthodes multiniveau algébriques pour nos systèmes. La contrainte utilisée correspond à la condition de compatibilité évoquée au chapitre précédent concernant la matrice de prolongement. Une première méthode de résolution du problème de minimisation utilisant des multiplicateurs de Lagrange est proposée. Par ailleurs, nous donnons une condition d'existence d'une solution pour le problème d'optimisation proposé.

Dans le Chapitre 4, nous écrivons la vérification de la contrainte ou condition de compatibilité comme la résolution d'un ensemble de problèmes de flot[2], chacun de petite dimension vis-à-vis de la dimension du système à résoudre. Nous montrons alors comment il est possible de construire, à partir de ces problèmes de flot, une méthode pour résoudre le problème d'optimisation sous contraintes plus simple de mise en oeuvre que celle utilisant les multiplicateurs de Lagrange. Des résultats numériques viennent illustrer les

[1] Cette énergie n'est pas toujours l'énergie au sens physique du problème considéré mais est plus généralement une forme bilinéaire b définie sur l'espace des fonctions solutions telle que $b(\mathbf{E}, \mathbf{E}) > 0$ pour toute fonction solution $\mathbf{E}$ non nulle.

[2] Un exemple simple de problème de flot est la détermination de l'intensité du courant dans un réseau électrique. Le flot correspond alors à l'intensité de ce courant.

performances de la méthode.

Les démonstrations et des simulations numériques complémentaires sont données en Annexe.

0.3 Contributions originales

Divers axes de ce travail apportent une contribution (au moins en partie) originale au calcul numérique en électromagnétisme :

- des méthodes simples à un niveau ont été mises en oeuvre sur des applications réalistes et nous avons montré qu'elles amenaient des gains en terme de temps de calcul [64, 63, 65].
- des méthodes provenant d'applications en éléments finis nodaux pour les méthodes multiniveau, fondées sur des problèmes de minimisation d'énergie sous contraintes, ont été étendues aux éléments finis d'arête. Quelques justifications qualitatives ont été apportées pour cette extension.
- une condition d'existence de matrice de prolongement vérifiant la contrainte de compatibilité a été proposée [59].
- une méthode constructive utilisant des problèmes de flot permet de déterminer une matrice de prolongement vérifiant le problème de minimisation sous contraintes.

Chapitre 1

Modélisation et formulation

Ce chapitre remplit plusieurs objectifs :
- rappeler brièvement les équations de Maxwell en milieu continu et insister sur les modèles auxquels nos méthodes de résolution peuvent s'appliquer ;
- présenter le cadre mathématique dans lequel ces équations vont être approchées numériquement ;
- donner quelques propriétés-clés des espaces d'éléments finis considérés et des matrices issues de la discrétisation des formulations variationnelles.

1.1 Modélisation des phénomènes électromagnétiques

On rappelle les équations régissant les phénomènes électromagnétiques. Des modèles particuliers sont ensuite présentés afin d'introduire quelques cas pour lesquels notre démarche est intéressante.

1.1.1 Equations de Maxwell

Le système d'équations fondamentales s'écrit sous forme différentielle de la manière suivante :

$$\text{Loi de Faraday : } \operatorname{rot}\mathbf{E} + \partial_t \mathbf{B} = 0, \tag{1.1a}$$
$$\text{Loi d'Ampère : } \operatorname{rot}\mathbf{H} - \partial_t \mathbf{D} = \mathbf{J}, \tag{1.1b}$$
$$\text{Loi de Gauss magnétique : } \operatorname{div}\mathbf{B} = 0, \tag{1.1c}$$
$$\text{Loi de Gauss électrique : } \operatorname{div}\mathbf{D} = \rho. \tag{1.1d}$$

Les vecteurs $\mathbf{E}$ et $\mathbf{H}$, $\mathbf{B}$ et $\mathbf{D}$ désignent respectivement les champs électrique et magnétique, l'induction magnétique et le déplacement électrique. Le vecteur $\mathbf{J}$ désigne le vecteur densité de courant et ρ la densité de charge électrique. On déduit aussi de (1.1b) et (1.1d) la relation de conservation de la charge :

$$\operatorname{div}\mathbf{J} + \partial_t \rho = 0, \tag{1.2}$$

On note que la loi de Gauss magnétique (1.1c) peut être déduite de (1.1a) si l'induction magnétique est supposée nulle avant l'instant initial.

Ces équations font apparaître des champs de vecteurs $\mathbf{E}$ et $\mathbf{B}$ découplés des champs de vecteurs $\mathbf{H}$ et $\mathbf{D}$. Pour permettre une modélisation complète des phénomènes, on introduit les équations de comportement qui caractérisent les phénomènes électriques et magnétiques ainsi que la loi d'Ohm qui lie $\mathbf{J}$ à $\mathbf{E}$:

$$\mathbf{D} = \varepsilon\mathbf{E},\ \mathbf{B} = \mu\mathbf{H} \text{ et } \mathbf{J} = \sigma\mathbf{E}. \tag{1.3}$$

Le coefficient ε est la permittivité diélectrique, μ la perméabilité magnétique et σ la conductivité électrique. Ces équations peuvent être non-linéaires ou anisotropes et dans ce cas ε, μ et σ sont des quantités tensorielles. Par simplification, on se placera dans un cadre linéaire et isotrope dans tout le mémoire ; cependant l'exploitation des algorithmes pourra se faire hors de ce cadre. Les lois constitutives (1.3) sont principalement phénoménologiques et donc de nature différente des lois (1.1).

Les équations (1.1) et (1.3) sont rarement considérées dans leur ensemble mais on choisit souvent des modèles simplifiés : certains phénomènes sont négligeables suivant la fréquence ou les matériaux considérés. Dans le paragraphe suivant, on aborde ainsi des problèmes quasistatiques et ceux ne prenant pas en compte les courants de déplacement.

1.1.2 Modèles considérés

Ces modèles font apparaître l'opérateur $\mathrm{rot}\,\mathrm{rot}$ appliqué aux champs de vecteurs $\mathbf{E}$, $\mathbf{H}$ ou $\mathbf{A}$. Le vecteur $\mathbf{A}$ désigne le potentiel vecteur magnétique introduit dans le paragraphe suivant.

Magnétostatique

Dans le cas statique, les dérivations par rapport au temps sont annulées. Les phénomènes électriques, caractérisés par les champs $\mathbf{E}$ et $\mathbf{D}$, et magnétiques, caractérisés par les champs $\mathbf{H}$ et $\mathbf{B}$, sont donc découplés si on ne tient pas compte de la dernière des relations (1.3), en considérant $\mathbf{J}$ comme un terme source. Suivant le contexte, on privilégie une approche électrostatique ou une approche magnétostatique. Les équations de la magnétostatique sont :

$$\mathrm{rot}\,\mathbf{H} = \mathbf{J},\ \mathrm{div}\,\mathbf{B} = 0 \text{ et } \mathbf{B} = \mu\mathbf{H}. \tag{1.4}$$

Pour mener à bien le calcul des grandeurs magnétiques, on s'intéresse principalement à l'utilisation du potentiel vecteur. Si le domaine est contractile[1], on montre qu'il existe un potentiel vecteur magnétique $\mathbf{A}$ tel que $\mathbf{B} = \mathrm{rot}\,\mathbf{A}$. L'équation utilisée pour modéliser le phénomène devient alors :

$$\mathrm{rot}\,\frac{1}{\mu}\,\mathrm{rot}\,\mathbf{A} = \mathbf{J}. \tag{1.5}$$

On peut y ajouter une condition de jauge ou utiliser une méthode de régularisation comme proposée dans [66], afin d'assurer l'unicité de $\mathbf{A}$. La jauge de Coulomb consiste à imposer $\mathrm{div}\,\mathbf{A} = 0$, mais elle conduit à la résolution d'un problème mixte. D'autres jauges sont donc souvent mises en oeuvre directement au niveau discret et reposent sur des algorithmes de graphes ; un inventaire en est réalisé dans [58]. Cependant, le champ magnétique, $\mathbf{B} = \mathrm{rot}\,\mathbf{A}$, demeure la grandeur physique intéressante et elle est indépendante de la jauge choisie.

La figure 1.1 montre l'application d'un modèle magnétostatique pour le calcul du champ magnétique $\mathbf{H}$ dans un bras de suspension d'automobile simulant un contrôle non destructif par magnétoscopie[2].

Courants de Foucault

Le modèle de courants de Foucault correspond au cas où la variation temporelle du déplacement électrique est négligeable, autrement dit on ne considère plus le phénomène de courants de déplacement. On se limite donc au phénomène des courants de Foucault qui sont les courants induits dans les conducteurs par des variations temporelles du flux d'induction magnétique. Les relations à considérer sont les suivantes :

$$\mathrm{rot}\,\mathbf{E} + \partial_t\mathbf{B} = 0,\ \mathrm{rot}\,\mathbf{H} = \mathbf{J},\ \mathrm{div}\,\mathbf{B} = 0,\ \mathbf{B} = \mu\mathbf{H} \text{ et } \mathbf{J} = \mathbf{J}_g + \sigma\mathbf{E}. \tag{1.6}$$

Le terme $\mathbf{J}_g$ modélise une source de courant connue. En utilisant de nouveau le potentiel vecteur introduit dans le cas magnétostatique, l'équation régissant le phénomène s'écrit :

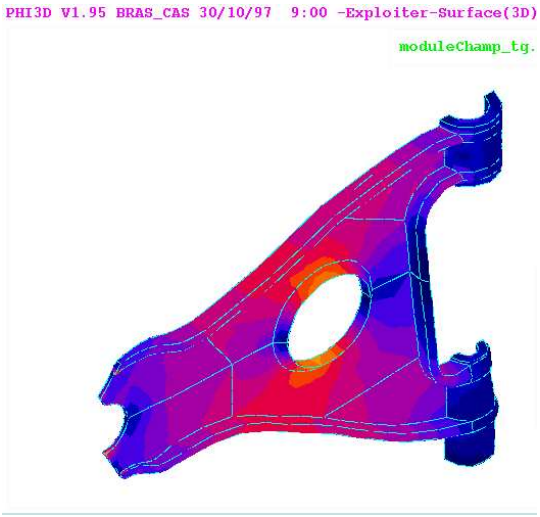
$$\mathrm{rot}\,\frac{1}{\mu}\,\mathrm{rot}\,\mathbf{A} + \sigma\partial_t\mathbf{A} = \mathbf{J}_g. \tag{1.7}$$

On peut faire la même remarque que pour le cas magnétostatique : une condition de jauge est introduite pour assurer l'unicité de $\mathbf{A}$ dans les parties non conductrices ($\sigma = 0$).

La version harmonique de ce modèle est plus couramment utilisée. On suppose pour cela que les termes sources sont harmoniques. Le vecteur potentiel s'écrit sous la forme $\mathbf{A} = \mathrm{Re}\left(\hat{\mathbf{A}}\exp(\mathrm{i}\omega t)\right)$ où $\hat{\mathbf{A}}$

[1] Un domaine est contractile s'il peut être déformé continûment en un point.

[2] Méthode non destructive de contrôle des pièces par poudres magnétiques, réservée exclusivement aux métaux et alliages ferromagnétiques.



Calculs effectués avec une modélisation en magnétostatique pour obtenir le champ magnétique **H** dans un bras de suspension.
On visualise la composante tangentielle du champ magnétique sur la surface de ce bras de suspension.

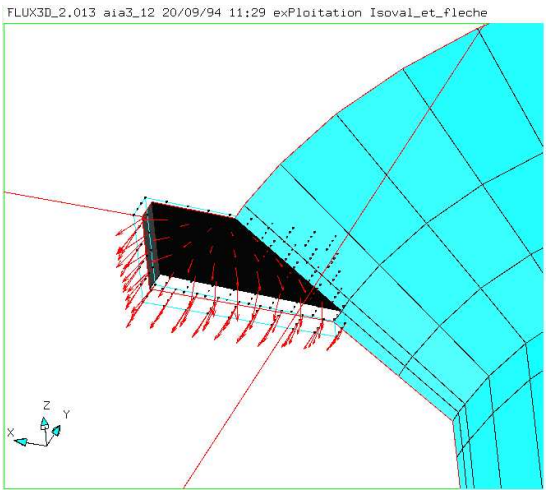
FIG. 1.1 – Contrôle non destructif par magnétoscopie d'un bras de suspension d'automobile.

est une grandeur complexe et ω correspond à la fréquence angulaire. L'équation à considérer prend alors la forme :

$$\text{rot}\,\frac{1}{\mu}\,\text{rot}\,\hat{\mathbf{A}} + \mathrm{i}\sigma\omega\hat{\mathbf{A}} = \hat{\mathbf{J}}_g. \tag{1.8}$$

Il existe aussi des modèles en courants de Foucault où l'inconnue principale est le champ électrique **E** ou le champ magnétique **H** [15, Chapitre 8].

La figure 1.2 montre l'application d'un modèle de courants de Foucault pour évaluer la densité de courant dans le défaut d'un rivet (calculs issus de la thèse de F. Thollon).



Calculs effectués avec une modélisation en courants de Foucault pour obtenir la densité de courant dans le défaut.
On visualise sous forme de flèches le vecteur densité de courant dans le défaut.

FIG. 1.2 – Courants induits dans le défaut d'un rivet.

Equation vectorielle des ondes

L'équation vectorielle des ondes est un modèle qui ne néglige aucun phénomène dans les équations de Maxwell, mais qui cherche à obtenir des formulations ne conservant que le champ électrique **E** ou le champ magnétique **H**.

Elles prennent la forme suivante :

$$\begin{aligned} &\mathrm{rot}\,\frac{1}{\mu}\,\mathrm{rot}\,\mathbf{E} + \varepsilon\partial_t^2\mathbf{E} + \sigma\partial_t\mathbf{E} = < \text{source} >, \\ &\mathrm{rot}\,\frac{1}{\varepsilon}\,\mathrm{rot}\,\mathbf{H} - \partial_t^2\mathbf{H} + \sigma\partial_t\mathbf{H} = < \text{source} >, \end{aligned} \tag{1.9}$$

Le terme $<$ source $>$ est donné par une antenne, une onde incidente ou une source de courant équivalente. On considère plus souvent une version harmonique dans les problèmes de diffraction :

$$\begin{aligned} &\mathrm{rot}\,\frac{1}{\mu}\,\mathrm{rot}\,\hat{\mathbf{E}} - (\omega^2\varepsilon - \mathrm{i}\omega\sigma)\hat{\mathbf{E}} = < \text{source} >, \\ &\mathrm{rot}\,\frac{1}{\varepsilon}\,\mathrm{rot}\,\hat{\mathbf{H}} - (\omega^2\mu - \mathrm{i}\omega\frac{1}{\sigma})\hat{\mathbf{H}} = < \text{source} > . \end{aligned} \tag{1.10}$$

Dans les calculs réalisés en Annexe B, on trouve plusieurs applications, notamment un calcul pour une application en hyperthermie locale[3] par ondes électromagnétiques, du modèle de l'équation vectorielle des ondes.

1.1.3 Conditions aux limites et d'interface

Conditions d'interface

Dans un milieu inhomogène, des conditions d'interface entre deux régions de propriétés différentes[4] doivent être vérifiées :

- la *continuité de la composante tangentielle du champ* électrique $\mathbf{E}$ et s'il n'existe pas de courant d'interface[5] celle du champ magnétique $\mathbf{H}$;
- la *continuité de la composante normale de l'induction* magnétique $\mathbf{B}$ et s'il n'existe pas de charge d'interface[6] celle du *déplacement* électrique $\mathbf{D}$.

Conditions de rayonnement à l'infini

Pour étudier le rayonnement en champ libre (qui est le cas pour l'équation des ondes), il est naturel d'imposer à l'infini la *condition de rayonnement de Sommerfeld* qui est une conséquence de la conservation de l'énergie.

Dans le cas d'un champ scalaire ψ vérifiant l'équation de Helmholtz $\triangle\psi - k^2\psi = 0$ où k est le nombre d'onde, cette condition s'écrit :

$$\lim_{r\to\infty} r^{\frac{d-1}{2}}\Big(\frac{\partial\psi}{\partial r} + ik\psi\Big) = 0,$$

où d est la dimension de l'espace. Dans le cas de l'équation vectorielle des ondes, la variante s'appelle la *condition de Silver-Müller*, mais son principe est identique à celle de Sommerfeld.

Les méthodes d'éléments finis seules ne nous permettant pas de traiter des domaines *non bornés*, l'utilisation des conditions de rayonnement en l'état n'est donc pas envisageable. Pour dépasser cette difficulté, on utilise des conditions locales approchées ou des couplages avec des équations intégrales.

Conditions aux limites

Dans les cas de la magnétostatique et des courants de Foucault, une condition aux limites usuelle est d'imposer un flux magnétique nul sur la surface extérieure du domaine ; en effet, pour se ramener à un domaine de calcul borné, on place généralement le dispositif avec le phénomène à observer dans une boîte dont la dimension est suffisante pour que ses frontières soient assimilables à l'infini. Dans les calculs, on impose alors pour le potentiel vecteur $\mathbf{A} \times \mathbf{n} = 0$ sur la surface concernée où $\mathbf{n}$ désigne la normale à la

[3] Traitement des tumeurs cancéreuses par élévation locale de la température.

[4] Elles découlent des équations de Maxwell. On effectue un bilan de flux au passage de l'interface entre deux matériaux. D'un point de vue mathématique, cela revient à considérer les équations au sens des distributions.

[5] Dans ce cas on peut avoir à vérifier : $\mathbf{n} \times (\mathbf{H}^1 - \mathbf{H}^2) = \mathbf{J}_\mathrm{S}$ sur la surface séparant deux sous-domaines Ω_1 et Ω_2.

[6] Dans ce cas on peut avoir à vérifier : $\mathbf{n} \cdot (\mathbf{D}^1 - \mathbf{D}^2) = \rho_\mathrm{S}$ sur la surface séparant deux sous-domaines Ω_1 et Ω_2.

surface. En présence d'un plan de symétrie (source et géométrie) on peut aussi imposer $1/\mu \operatorname{rot} \mathbf{A} \times \mathbf{n} = 0$ sur ce plan et calculer sur un demi-domaine.

Dans le cas de l'équation vectorielles des ondes, les conditions aux limites fréquemment utilisées sont des conditions de conducteur électrique parfait (respectivement mur magnétique) : $\mathbf{E} \times \mathbf{n} = 0$ (resp. $\mathbf{H} \times \mathbf{n} = 0$), ou des conditions d'impédance : $1/\mu \operatorname{rot} \mathbf{E} \times \mathbf{n} - \delta \mathbf{n} \times (\mathbf{E} \times \mathbf{n}) = 0$ qui traduisent le fait que l'onde peut pénétrer dans le matériau sur de faibles distances.

Ces conditions d'impédance comprennent aussi des conditions absorbantes au premier ordre inspirés de Silver-Müller avec $\delta = \mathrm{i}\omega\sqrt{\varepsilon_0/\mu_0}$ dans le cas harmonique. Elles permettent de ramener la résolution des problèmes de diffraction à des domaines bornés.

1.2 Propriétés remarquables et discrétisation

On rappelle tout d'abord quelques notions sur la théorie des réseaux électriques afin d'introduire progressivement des concepts qui serviront par la suite : notamment les graphes orientés qui seront aussi utiles au Chapitre 4 et des notions de topologie qui apparaissent lors de la discrétisation des problèmes aux limites. Cette partie ne requiert pas de connaissances mathématiques spécialisées ; elle repose sur des modèles électriques simples et familiers.

1.2.1 Théorie des réseaux électriques

La présentation suivante trouve ces fondements, en particulier l'exemple utilisé, dans l'ouvrage [6, Chapitre 2].

La représentation commune d'un réseau électrique est celle d'un système électrique où les différents composants sont modélisés par des *branches* qui sont connectées au niveau de *noeuds* ou *jonctions*. Un chemin fermé dans le réseau est appelé une *boucle*.

Dans la théorie des réseaux, les deux lois de Kirchhoff constituent les lois fondamentales. La *loi des courants de Kirchhoff* affirme que, dans tout réseau électrique, la somme des courants quittant un noeud est égale à zéro à tout instant. Pour le réseau représenté à la figure 1.3, cette loi appliquée au noeud A donne ainsi :

$$i_2 - i_4 - i_3 = 0.$$

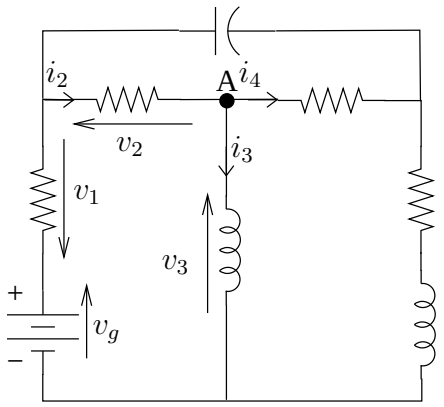


FIG. 1.3 – Exemple de réseau électrique.

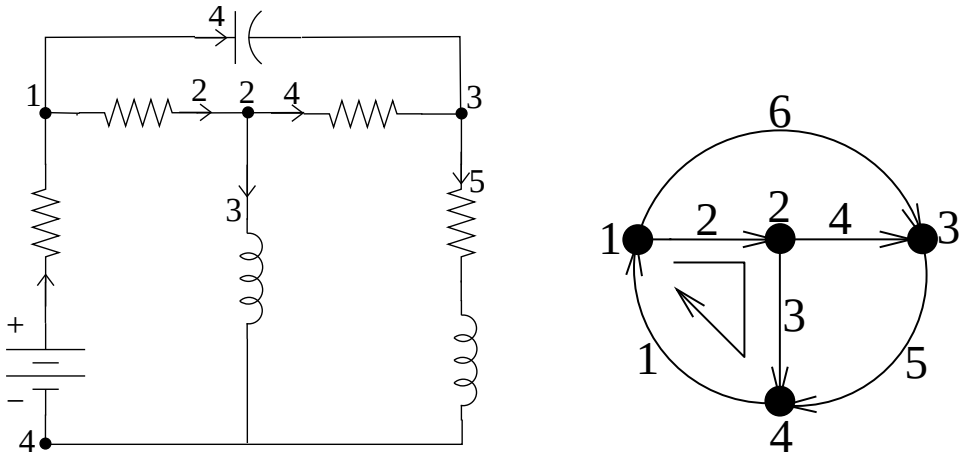
La *loi des tensions de Kirchhoff* affirme que, dans tout réseau électrique, la somme des tensions sur une boucle est réduite à zéro à tout instant. Ainsi en considérant dans le sens des aiguilles d'une montre la boucle constituée par les branches 1, 2 et 3 de la figure 1.3, on peut écrire :

$$v_g - v_1 - v_2 - v_3 = 0.$$

Ces deux lois sont indépendantes des *lois de comportement* qui lient la tension v au courant i dans chaque branche. Elles dépendent uniquement de la *topologie du réseau* ; on peut ainsi définir plus formellement un réseau électrique comme un graphe orienté, sans branche qui relie un noeud à lui-même et où chaque branche est caractérisée par une relation liant la tension v au courant i. Ces variables associées aux branches vérifient alors les deux lois de Kirchhoff.

Lois de Kirchhoff et graphe orienté

Pour le réseau électrique de la figure 1.3, on introduit une numérotation pour les noeuds et les branches comme sur la figure 1.4(a) et on peut lui associer le graphe orienté de la figure 1.4(b). On donne quelques



(a) Numérotation des entités du réseau.

(b) Graphe associé.

FIG. 1.4 – Le réseau de la figure 1.3 et son graphe associé.

définitions de la théorie des graphes nécessaires pour la théorie des réseaux.

Un graphe orienté est défini par un ensemble de noeuds X et un ensemble de couples de $X \times X$ qui sont les arcs orientés, que l'on appelle branche par la suite ; les éléments du couple désignent les extrémités de la branche. Une branche est dite *incidente* en un noeud si cette branche débute ou se termine en ce noeud.

Si l'on considère uniquement un sous-ensemble des noeuds et/ou un sous-ensembles des branches, on peut définir un nouveau graphe qui est un *sous-graphe* du graphe initial.

Un *chemin* joignant deux noeuds du graphe est un sous-graphe particulier qui recense une suite de branches et de noeuds traversés pour joindre les deux extrémités du chemin (on ne passe pas deux fois par la même branche). Une *boucle* (le terme *cycle* est plus couramment utilisé en théorie des graphes) est un chemin particulier où les deux noeuds extrémités coïncident.

Un graphe est dit *connexe* si tout couple de noeuds distincts peut être joint par un chemin. Si n est un noeud, l'ensemble formé par n et par tous les sommets auxquels n peut être lié par un chemin est appelé une *composante connexe* du graphe.

Un *arbre* est un graphe connexe particulier qui ne contient aucune boucle ; supprimer une branche de cet arbre le rend non connexe. On peut montrer qu'un arbre de N noeuds contient toujours exactement $N - 1$ branches [6, p 72].

Dans de nombreuses situations, on est amené à considérer un *arbre couvrant* d'un graphe initial connexe : cet arbre est un sous-graphe du graphe initial avec les mêmes noeuds mais en supprimant quelques branches. Il y a de nombreuses possibilités pour la détermination d'un arbre couvrant : dans le cas du graphe de la figure 1.4(b), on donne deux arbres couvrants particuliers à la figure 1.5. L'ensemble des arêtes non considérées dans l'arbre forme le *coarbre*.

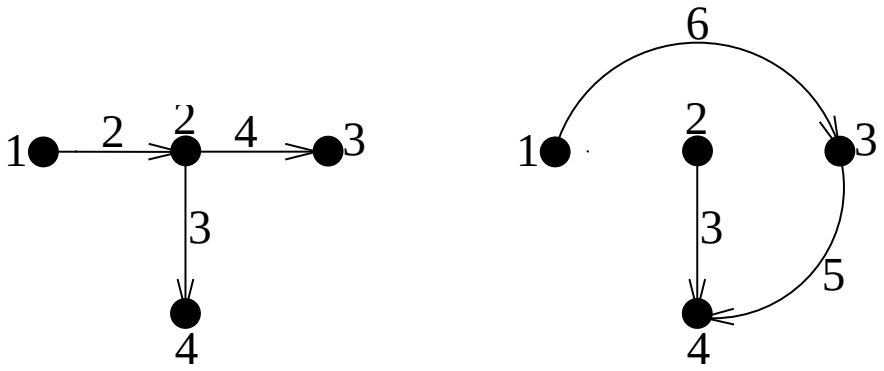


FIG. 1.5 – Deux arbres couvrants pour le graphe de la figure 1.4(b).

Matrice d'incidence branche-noeud et loi des courants de Kirchhoff Pour représenter un graphe, la matrice d'incidence branche-noeud complète G_c du graphe peut être utilisée. Le nombre de colonnes de cette matrice est égal au nombre N de noeuds du graphe et le nombre de lignes au nombre E de branches ; ces entrées sont définies de la manière suivante :

$$(G_c)_{en} = \begin{cases} 1 \text{ si le noeud } n \text{ termine l'arête } e, \\ -1 \text{ si le noeud } n \text{ débute l'arête } e, \\ 0 \text{ sinon.} \end{cases} \tag{1.11}$$

Pour le graphe de la figure 1.4(b), sa transposée s'écrit alors :

$$(G_c)^t = \begin{pmatrix} 1 & -1 & 0 & 0 & 0 & -1 \\ 0 & 1 & -1 & -1 & 0 & 0 \\ 0 & 0 & 0 & 1 & -1 & 1 \\ -1 & 0 & 1 & 0 & 1 & 0 \end{pmatrix}.$$

Si le graphe possède N noeuds et s'il est constitué de P composantes connexes, le rang de cette matrice est égale à $N - P$. Généralement, on supprime un noeud du réseau pour chaque composante connexe, ce noeud fixe une référence et l'on travaille avec une matrice d'incidence branche-noeud réduite. Si l'on oublie le noeud 2 dans le graphe de la figure 1.4(b), la matrice d'incidence réduite donne :

$$G^t = \begin{pmatrix} 1 & -1 & 0 & 0 & 0 & -1 \\ 0 & 0 & 0 & 1 & -1 & 1 \\ -1 & 0 & 1 & 0 & 1 & 0 \end{pmatrix}.$$

La matrice réduite est de rang maximal et la recherche d'un arbre couvrant du graphe permet de déterminer un ensemble de lignes linéairement indépendantes de cette matrice réduite [27, 6, Chapitre 2].

Soit I le vecteur contenant les valeurs des courants dans les différentes branches du réseau, le sens de ceux-ci tenant compte des orientations données aux branches du graphe ; la loi des courants de Kirchhoff peut s'écrire sous forme matricielle et de manière équivalente :

$$(G_c)^t I = 0 \quad \text{ou} \quad G^t I = 0. \tag{1.12}$$

En effet, G^t est déduite de G_c^t en retirant une ligne et donc $\ker G^t \subset \ker G_c^t$. Comme $\operatorname{rang} G^t = \operatorname{rang} G_c^t$, les dimensions de $\ker G^t$ et de $\ker G_c^t$ sont égales et l'inclusion est une égalité.

Matrice de boucles et loi des tensions de Kirchhoff La matrice de boucles complète R_c est aussi intéressante pour l'étude des lois de Kirchhoff. Le nombre de lignes de cette matrice est égal au nombre

de boucles dans le réseau et le nombre de colonnes au nombre E de branches. On fixe une orientation pour chaque boucle et les entrées de cette matrice sont alors définies de la manière suivante :

$$(R_c)_{ef} = \begin{cases} 0 \text{ si la branche } e \text{ n'appartient pas à la boucle } f, \\ 1 \text{ si la branche } e \text{ appartient à la boucle } f \text{ et les orientations coïncident,} \\ -1 \text{ si la branche } e \text{ appartient à la boucle } f \text{ et les orientations ne coïncident pas.} \end{cases} \tag{1.13}$$

Dans le cas du graphe de la figure 1.4(b), on peut recenser les boucles suivantes en donnant la suite des noeuds traversés :

$$\{1,2,4\},\ \{1,3,4\},\ \{1,2,3\},\ \{2,3,4\},\ \{1,2,3,4\},\ \{1,2,4,3\},\ \{1,4,2,3\}.$$

L'orientation induite par la première boucle est représentée sur la figure 1.4(b). Ceci nous conduit à la matrice de boucles complète suivante :

$$R_c = \begin{pmatrix} 1 & 1 & 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 1 & 1 \\ 0 & 1 & 0 & 1 & 0 & -1 \\ 0 & 0 & -1 & 1 & 1 & 0 \\ 1 & 1 & 0 & 1 & 1 & 0 \\ 0 & 1 & 1 & 0 & -1 & -1 \\ -1 & 0 & -1 & 1 & 0 & -1 \end{pmatrix}.$$

Cependant, on doit considérer ici de nombreuses boucles alors que certains sous-ensembles apportent une information équivalente. Ainsi si l'on a déterminé un arbre couvrant du graphe et que l'on ajoute indépendamment et une à une les branches du coarbre, on obtient dans chaque cas un graphe avec une unique boucle, qui est aussi une boucle du graphe initial. On dit que l'on a ainsi déterminé un ensemble de *boucles fondamentales* associé à l'arbre couvrant choisi. Sur la figure 1.6 sont représentées les 3 boucles fondamentales associées au premier arbre couvrant de la figure 1.5. On peut aussi choisir leurs orientations

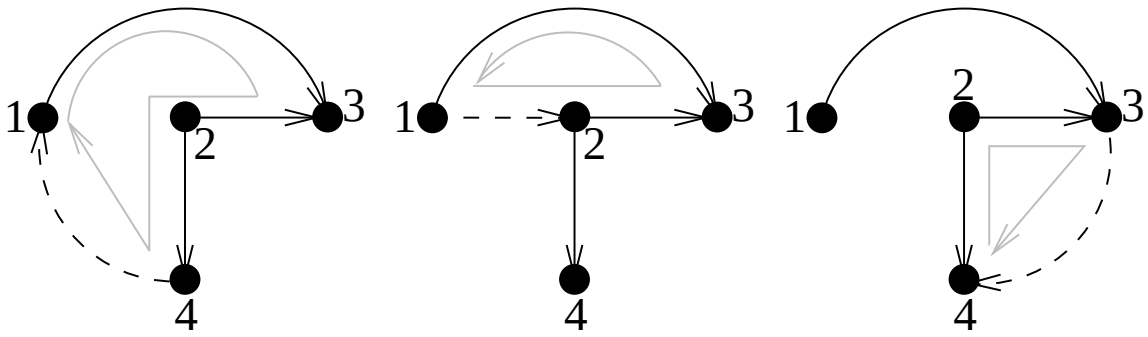


Fig. 1.6 – Ensemble de boucles fondamentales.

pour qu'elles coïncident avec celle de l'arête du coarbre ajoutée. On obtient alors une matrice de boucles fondamentales qui correspond pour le choix présenté à la figure 1.6 à la matrice :

$$R_f = \begin{pmatrix} 1 & 0 & 1 & -1 & 0 & 1 \\ 0 & 1 & 0 & 1 & 0 & -1 \\ 0 & 0 & -1 & 1 & 1 & 0 \end{pmatrix}.$$

Si l'on considère la sous-matrice obtenue en extrayant les colonnes 1, 2 et 5, correspondant aux indices des branches du coarbre, on constate que cette sous-matrice est la matrice identité ce qui nous indique que le rang de R_f est égal au nombre de branches du coarbre, c.-à-d. $E+1-N$. Ce résultat est valable pour tout graphe [27, 6, Chapitre 2]. Plus généralement, on notera R toute matrice de boucles réduite de rang maximal et égal à $E+1-N$ et les boucles choisies sont dites indépendantes.

En outre, on constate simplement, en s'assurant bien entendu que la numérotation des colonnes de $(G_c)^t$ et R_c et de G^t et R soit cohérente, que :

$$(G_c)^t(R_c)^t = 0 \quad \text{et} \quad G^t R^t = 0. \tag{1.14}$$

Ceci signifie que $\mathrm{Im}(R_c)^t \subset \ker(G_c)^t$ et $\mathrm{Im}\, R^t \subset \ker G^t$. Or, on a souligné précédemment que le rang de G_c ou G était $N-1$ (en considérant le graphe connexe) ; en conséquence, la dimension du noyau de $(G_c)^t$ ou G^t est $E+1-N$. Comme le rang de R est égal à $E+1-N$, on a nécessairement $\mathrm{Im}\, R^t = \ker G^t$ et de la même manière $\mathrm{Im}\, R^t = \mathrm{Im}(R_c)^t$ et $\mathrm{Im}(R_c)^t = \ker(G_c)^t$.

Deux informations peuvent être rapidement déduites de ces résultats :

- l'utilisation d'une matrice de boucles réduite contient une information égale à celle de la matrice de boucles complète.
- si le vecteur I_m contient les valeurs des courants de boucles indépendantes, on sait maintenant que tout vecteur I vérifiant la loi des courants de Kirchhoff (1.12) peut s'écrire sous la forme :

$$I = R^t I_m. \tag{1.15}$$

Revenons maintenant à l'expression de la loi des tensions de Kirchhoff avec les nouvelles matrices introduites. Si V est le vecteur contenant les valeurs des tensions dans les différentes branches du réseau, le sens de celles-ci étant cohérent avec les orientations du graphe, la loi des tensions de Kirchhoff s'écrit :

$$R_c V = 0 \quad \text{ou} \quad RV = 0.$$

De la même manière qu'en (1.14), il vient directement que $\mathrm{Im}\, G = \ker R$. Si U est un vecteur contenant les valeurs des potentiels aux noeuds du réseau, on peut ainsi exprimer le vecteur V des tensions dans les branches du réseau vérifiant la loi des tensions de Kirchhoff sous la forme :

$$V = GU. \tag{1.16}$$

Dans ce cas, à cause de la définition de G, la valeur du potentiel du noeud initialement supprimé sert de potentiel de référence et il y a donc uniquement $N-1$ potentiels inconnus.

Formalisation des lois de branches et résolution des systèmes correspondants

Le but de ce paragraphe est de rappeler les notions élémentaires sur les circuits électriques en régime harmonique et les méthodes de résolution des équations. Il se trouve que les systèmes obtenus à la fin de la Sous-section 1.2.4 ou au Chapitre 4 sont, bien que plus larges et plus complexes, fondamentalement de même nature. C'est la raison pour laquelle la résolution est formalisée ici, bien que le système simple donné en exemple ne requiert pas une telle formalisation.

Le type de branches le plus général considéré ici prend la forme donnée à la figure 1.7. La source de tension est placée en série avec un élément passif et la source de courant est en parallèle avec l'ensemble précédent. Pour une question de commodité de représentation, on s'arrangera ainsi toujours pour que les générateurs soient accompagnés par un élément passif. Les tensions aux bornes des différents éléments

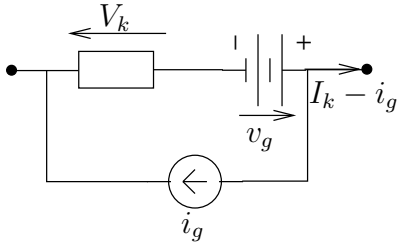


FIG. 1.7 – Branche générique.

passifs et les courants qui les traversent sont les inconnues de notre problème ; ces données sont regroupées dans les vecteurs V et I dont la dimension est égale au nombre de branches. Les vecteurs V_g et I_g

contiennent respectivement les tensions et les courants fournis par les générateurs dans les différentes branches. Les matrices G et R sont des matrices d'incidence branche-noeud et de boucle réduites utilisées pour représenter la topologie du réseau. Avec les lois de Kirchhoff il vient alors :

$$RV = RV_g \quad \text{ou} \quad G^t I = G^t I_g. \tag{1.17}$$

Pour simplifier la suite, on suppose que l'on travaille en régime permanent avec des grandeurs complexes et ainsi pour chaque branche k la relation entre I_k et V_k s'écrit :

$$V_k = Z_k I_k \quad \text{ou} \quad I_k = Y_k V_k. \tag{1.18}$$

où Z_k et Y_k désignent respectivement l'impédance et l'admittance complexe de la branche k. Il n'est pas impossible non plus d'avoir des impédances de couplage entre les différentes branches. On peut ainsi définir une matrice d'impédance Z et une matrice d'admittance $Y = Z^{-1}$. Dans le cas du graphe de la figure 1.4(b), la matrice d'impédance s'écrit :

$$Z = \begin{pmatrix} R_1 & 0 & 0 & 0 & 0 & 0 \\ 0 & R_2 & 0 & 0 & 0 & 0 \\ 0 & 0 & \mathrm{i}\omega L_3 & 0 & 0 & 0 \\ 0 & 0 & 0 & R_4 & 0 & 0 \\ 0 & 0 & 0 & 0 & R_5 + \mathrm{i}\omega L_5 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1/(\mathrm{i}\omega C_6) \end{pmatrix}.$$

Plutôt que de résoudre directement les E équations provenant des lois de Kirchhoff (1.17) plus les E relations de comportement (1.18), on résout généralement un nombre réduit d'équations dites de noeuds ou de boucles.

Equations de boucles Pour se ramener à un système d'équations de boucles, on remplace le vecteur des tensions V dans la loi des tensions de Kirchhoff en utilisant les lois de comportement ; on obtient :

$$RZI = RV_g.$$

Ensuite, on utilise la relation (1.15) afin de faire apparaître un nombre réduit d'inconnues, les courants de boucle, contenus dans le vecteur I_m. Le système d'équations à résoudre prend alors la forme suivante :

$$(RZR^t)I_m = R(V_g - ZI_g).$$

On remonte ensuite facilement aux grandeurs intéressantes :

$$I = R^t I_m + I_g \quad \text{et} \quad V = ZI.$$

Equations de noeuds Pour se ramener à un système d'équations de noeuds, on remplace le vecteur des courant I dans la loi des courants de Kirchhoff en utilisant les lois de comportement ; on obtient :

$$G^t YV = G^t I_g.$$

Ensuite, on utilise la relation (1.16) afin de faire apparaître un nombre réduit d'inconnues, les potentiels au noeuds, contenus dans le vecteur U. Le système d'équations à résoudre prend alors la forme suivante :

$$(G^t YG)U = G^t(I_g - YV_g).$$

On remonte ensuite facilement aux grandeurs intéressantes :

$$V = GU + V_g \quad \text{et} \quad I = YV.$$

Remarques sur la construction des équations de réseau

Pour résumer une partie de la démarche de la théorie des réseaux, on peut relever les points suivants :
- Des grandeurs physiques ont été associées aux différentes entités, noeuds, branches et boucles, définies sur le graphe du réseau. Dans la loi des tensions de Kirchhoff, ces grandeurs sont respectivement les potentiels, les tensions et les forces électromotrices (fem) de boucles ; dans la loi des courants de Kirchhoff, ce sont respectivement les courants de noeud, de branche et de boucle.
- Le passage d'un type de grandeur à l'autre est assuré simplement par la matrice d'incidence G et de boucles R ; il est donc uniquement lié à la topologie du réseau. Dans la loi des tensions de Kirchhoff, on peut représenter la structure algébrique de la manière suivante :

$$\mathbb{C}^N \xrightarrow{G} \mathbb{C}^E \xrightarrow{R} \mathbb{C}^F \tag{1.19}$$

où $\mathbb{C}^N$, $\mathbb{C}^E$ et $\mathbb{C}^F$ désignent respectivement les espaces de vecteurs à coefficients complexes de dimension N, E et F. De même pour la loi des courants de Kirchhoff, on obtient :

$$\mathbb{C}^N \xleftarrow{G^t} \mathbb{C}^E \xleftarrow{R^t} \mathbb{C}^F. \tag{1.20}$$

Ces deux suites sont dites exactes car elles possèdent une propriété remarquable, à savoir $\operatorname{Im} G = \ker R$ et donc $\operatorname{Im} R^t = \ker G^t$.
- On peut coupler ces deux suites en introduisant la matrice d'impédance et/ou d'admittance définissant une bijection entre les valeurs des potentiels V et les valeurs des courants I :

$$\begin{array}{ccccc} \mathbb{C}^N & \xrightarrow{G} & \mathbb{C}^E & \xrightarrow{R} & \mathbb{C}^F \\ & & Z \uparrow \downarrow Y & & \\ \mathbb{C}^N & \xleftarrow{G^t} & \mathbb{C}^E & \xleftarrow{R^t} & \mathbb{C}^F \end{array} \tag{1.21}$$

On fait apparaître ainsi une dualité entre la suite (1.19) et la suite (1.20).

1.2.2 Réseau électrique et discrétisation sur la triangulation d'une surface plane

Pour faire le lien entre la théorie des réseaux et la discrétisation par éléments finis sur une triangulation d'un domaine plan, on introduit quelques notions sur les graphes planaires. Les graphes planaires sont ceux pour lesquels il existe une représentation dans le plan de sorte que les sommets soient des points distincts, les branches des courbes simples et deux branches ne se rencontrent pas en dehors de leurs extrémités [12] ; voir la figure 1.8 pour la représentation de graphes planaire et non planaire.

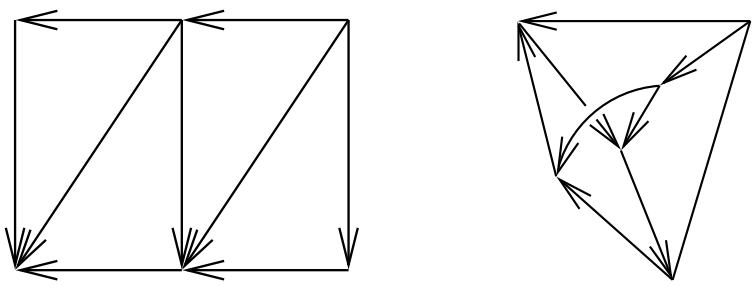


FIG. 1.8 – Graphes planaire et non planaire.

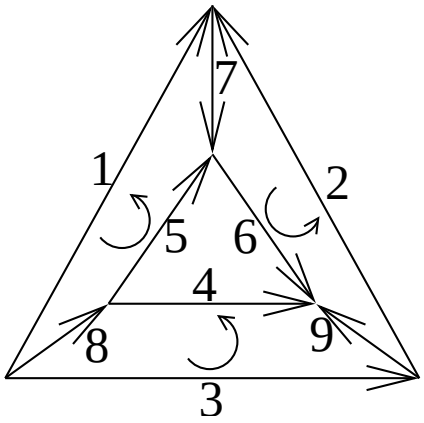
Dans le cas de graphes planaires, les branches séparent le plan en petites régions appelée *mailles* ; une maille est une région délimitée par une boucle ne contenant pas de branches, cette boucle est le *contour* de la maille. L'ensemble des contours des mailles forme une famille de boucles indépendantes [12] et peuvent

servir pour la définition des lignes d'une matrice de boucle réduite B. Dans le graphe de la figure 1.4(b), les contours des mailles sont les boucles $\{1,2,4\}$, $\{2,3,4\}$ et $\{1,2,3\}$.

On peut toujours définir le graphe dual d'un graphe planaire de la manière suivante :

- à chaque maille du graphe primal est associé un unique noeud du graphe dual. Pour le représenter on peut le placer au barycentre de la maille primale.
- à chaque branche du graphe primal est associée une unique branche du graphe dual. L'orientation des arêtes du graphe dual est fixée de manière à ce que la matrice d'incidence branche-noeud du graphe dual soit égale à la transposée de la matrice de boucles du graphe primal.

Pour que l'analogie soit plus complète avec une triangulation d'un domaine plan, on doit considérer une classe de graphes (hypergraphes) ayant à la fois des branches (ou arêtes) et des faces, qui seront un sous-ensemble des mailles. La matrice d'incidence face-arête se substitue alors à la matrice de boucles. Elle peut-être formée à partir de la matrice des contours, en retirant les lignes correspondant aux contours des mailles qui ne sont pas des faces. A la figure 1.9, on représente un graphe avec les faces et leurs orientations. R' est la matrice des contours associée au graphe planaire et R la matrice d'incidence face-arête. On vérifie $\operatorname{rang} R = \operatorname{rang} R' - 1$ et en conséquence, $\ker R' = \operatorname{Im} G \subsetneq \ker R$ et le défaut est de dimension 1. Ce type de graphe convient à la représentation de la triangulation d'un domaine plan avec



Matrice de contours :

$$R' = \begin{pmatrix} -1 & 0 & 0 & 0 & 1 & 0 & -1 & 1 & 0 \\ 0 & 1 & 0 & 0 & 0 & 1 & 1 & 0 & -1 \\ 0 & 0 & 1 & -1 & 0 & 0 & 0 & -1 & 1 \\ 0 & 0 & 0 & -1 & 1 & 1 & 0 & 0 & 0 \end{pmatrix}$$

Matrice d'incidence face-arête :

$$R = \begin{pmatrix} -1 & 0 & 0 & 0 & 1 & 0 & -1 & 1 & 0 \\ 0 & 1 & 0 & 0 & 0 & 1 & 1 & 0 & -1 \\ 0 & 0 & 1 & -1 & 0 & 0 & 0 & -1 & 1 \end{pmatrix}$$

FIG. 1.9 – Représentation d'un graphe et de ses faces.

ses noeuds, ses arêtes et ses faces.

Utilisation des concepts de la théorie des réseaux

Considérons la triangulation d'un domaine plan Ω. Soient $\mathcal{N}$, $\mathcal{E}$ et $\mathcal{F}$ les ensembles de noeuds, d'arêtes et de faces du maillages de cardinal respectif N, E et F.

Une méthode de discrétisation naturelle en électromagnétisme revient à remplacer le milieu continu par un "réseau" s'appuyant sur le graphe associé à la triangulation [18] ; ce graphe fait bien entendu intervenir la notion de noeuds, d'arêtes mais aussi de faces comme indiqué précédemment.

On utilise alors les suites mises en évidence pour la théorie des réseaux. Les grandeurs physiques intervenant dans l'équivalent de la suite (1.19) sont les potentiels électriques associés aux noeuds, les forces électromotrices associées aux arêtes et les flux magnétiques associés aux faces. On représente cette suite sous la forme :

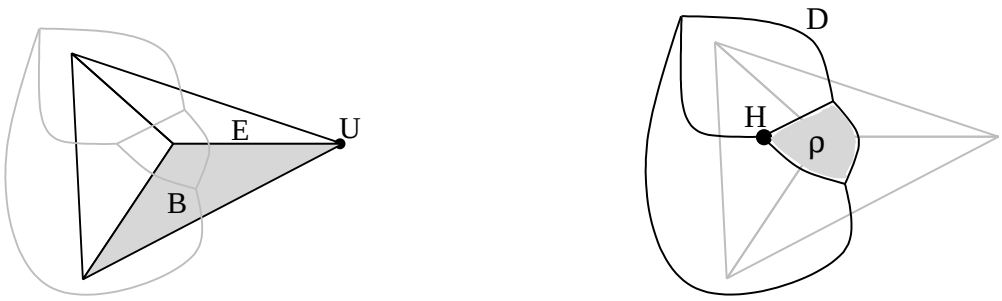
$$\mathrm{U} \in \mathbb{C}^N \xrightarrow{G} \mathrm{E} \in \mathbb{C}^E \xrightarrow{R} \mathrm{B} \in \mathbb{C}^F, \tag{1.22}$$

vérifiant $RG = 0$.

Les grandeurs physiques associées au graphe dual et intervenant dans l'équivalent de la suite (1.20) sont les circulation du champ magnétique associées aux noeuds du graphe dual, les flux de déplacement électrique (ou les courants) associés aux arêtes et les densités de charge associées aux éléments du graphe dual. On a représenté ces grandeurs avec celle du graphe primal à la figure 1.10. On peut représenter cette suite sous la forme :

$$\rho \in \mathbb{C}^N \xleftarrow{G^t} \mathrm{D} \in \mathbb{C}^E \xleftarrow{R^t} \mathrm{H} \in \mathbb{C}^F. \tag{1.23}$$

Pour compèter ce modèle, il manque l'introduction et la discrétisation des lois de comportement afin de pouvoir définir une bijection M_ε, de rôle équivalent à la matrice d'admittance Y, entre les forces électromotrices et les courants de déplacements pour coupler les suites de manière équivalente à (1.21).



(a) Graphe primal (b) Graphe dual

FIG. 1.10 – Graphes primal et dual - Grandeurs physiques associées aux composantes des graphes.

Une représentation complète de cette dualité peut prendre alors la forme du diagramme suivant :

$$\begin{array}{ccccc} \mathrm{U} \in \mathbb{C}^N & \xrightarrow{G} & \mathrm{E} \in \mathbb{C}^E & \xrightarrow{R} & \mathrm{B} \in \mathbb{C}^F \\ & & \downarrow M_\varepsilon & & \\ \rho \in \mathbb{C}^N & \xleftarrow{G^t} & \mathrm{D} \in \mathbb{C}^E & \xleftarrow{R^t} & \mathrm{H} \in \mathbb{C}^F. \end{array} \tag{1.24}$$

Le parallèle avec le milieu continu est lié au fait que l'on peut écrire une relation de la même forme que (1.22) liant fonctions, champs de vecteurs et opérateurs différentiels rot et grad :

$$\text{potentiel sur } \Omega : u \xrightarrow{\text{grad}} \text{champ de vecteur sur } \Omega : \mathbf{E} \xrightarrow{\text{rot}} \text{densité : } b. \tag{1.25}$$

On a de façon analogue $\mathrm{rot}\,\mathrm{grad} = 0$. Ce type de suite définit un complexe de De Rham au sens suivant :

Définition 1.1. *Soit $V_0, \ldots, V_n$ une famille d'espaces vectoriels et $\mathrm{A}_i : V_{i-1} \mapsto V_i,\ i = 1 \ldots n$ une famille d'applications linéaires.*

La suite :

$$V_0 \xrightarrow{\mathrm{A}_1} V_1 \xrightarrow{\mathrm{A}_2} \ldots \xrightarrow{\mathrm{A}_n} V_n \tag{1.26}$$

est appelée complexe de De Rham.

Cette suite est dite exacte au niveau de V_i si l'on a :

$$\mathrm{Im}(\mathrm{A}_i) = \ker(\mathrm{A}_{i+1}).$$

On peut définir aussi l'espace quotient :

$$\mathrm{C}_i \equiv \frac{\ker(\mathrm{A}_{i+1})}{\mathrm{Im}(\mathrm{A}_i)}$$

appelé groupe de cohomologie de De Rham associée à V_i[7].

Si on définit aussi pour des fonctions et des champs de vecteurs suffisament réguliers les applications :

$$\Pi^0 : u \mapsto \mathrm{U} \in \mathbb{C}^N \text{ de composantes } u(n),\ n \in \mathcal{N},$$
$$\Pi^1 : \mathbf{E} \mapsto \mathrm{E} \in \mathbb{C}^E \text{ de composantes } \int_e \mathbf{E} \cdot \mathrm{d}l,\ e \in \mathcal{E},$$
$$\Pi^2 : b \mapsto \mathrm{B} \in \mathbb{C}^F \text{ de composantes } \int_f b \mathrm{d}s,\ f \in \mathcal{F},$$

[7]Le terme cohomologie désigne la branche de la topologie algébrique qui s'est donnée pour tâche d'étudier des suites telles que (1.26).

on peut alors écrire le diagramme commutatif suivant :

$$\begin{array}{ccccc} C^\infty(\Omega,\mathbb{C}) & \xrightarrow{\text{grad}} & C^\infty(\Omega,\mathbb{C})^2 & \xrightarrow{\text{rot}} & C^\infty(\Omega,\mathbb{C}) \\ \downarrow \Pi^0 & & \downarrow \Pi^1 & & \downarrow \Pi^2 \\ \mathbb{C}^N & \xrightarrow{G} & \mathbb{C}^E & \xrightarrow{R} & \mathbb{C}^F. \end{array} \tag{1.27}$$

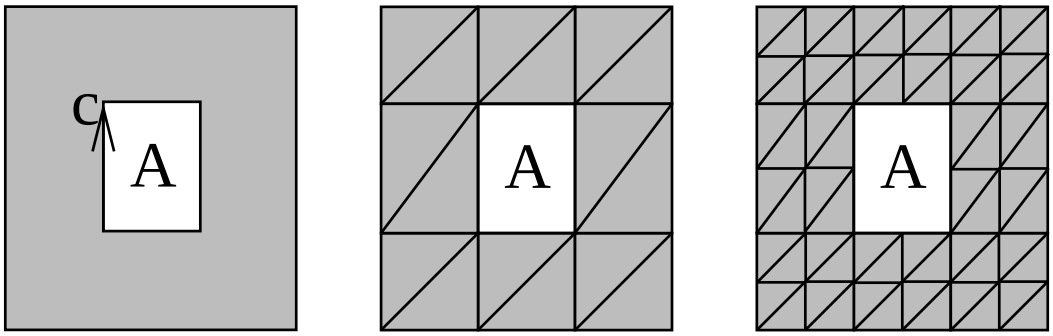
Le terme commutatif signifie que l'on peut suivre des chemins différents mais que le vecteur obtenu ne dépend que du point de départ et du point d'arrivée ; on a par exemple :

$$\forall u \in C^\infty(\Omega,\mathbb{C}),\ G\Pi^0 u = (u(m)-u(n))_{e=(n,m)\in\mathcal{E}} = \left(\int_e \operatorname{grad} u \cdot \mathrm{d}l\right)_{e\in\mathcal{E}} = \Pi^1 \operatorname{grad} u. \tag{1.28}$$

La représentation (1.27) fait apparaître que le complexe (1.22) est une représentation discrète du complexe (1.25). Par exemple, si l'on cherche à caractériser un champ électrique en électrostatique, c.-à-d. tel que $\operatorname{rot}\mathbf{E}=0$, on écrit que la circulation du champ électrique sur les frontières de toutes les faces est réduite à zéro c.-à-d. $R\mathrm{E}=0$, pour E représentation discrète du champ $\mathbf{E}$. On obtient l'analogue de l'écriture de la loi des tensions de Kirchhoff $RV=0$. Cependant il apparaît une difficulté nouvelle à savoir que la relation $R\mathrm{E}=0$ n'implique pas toujours $\mathrm{E}=G\mathrm{U}$ comme illustré dans la suite.

Invariants topologiques

Considérons le domaine de la figure 1.11(a) et la triangulation donnée par la figure 1.11(b). Comme



(a) Domaine simple présentant un trou.

(b) Une première triangulation du domaine.

(c) Raffinement du maillage initial.

FIG. 1.11 – Domaine troué et une triangulation possible.

indiqué au début de cette Sous-section, la matrice d'incidence face-arête R se déduit de la matrice des contours R' en supprimant la ligne correspondant au contour c entourant la maille A. La matrice R vérifie $\operatorname{rang} R = \operatorname{rang} R' - 1$ et on a donc seulement $\operatorname{Im} G \subsetneq \ker R$.

Quels que soient la finesse et le type de maillage, on peut utiliser par exemple le maillage raffiné figure 1.11(c), le défaut de rang entre $\ker R$ et $\operatorname{Im} G$ sera toujours de dimension 1 et caractérisé par le contour c. Ce défaut passe aussi au niveau continu, c.-à-d. $\mathbf{E}$ ne s'écrit pas nécéssairement comme un gradient sous la seule condition $\operatorname{rot}\mathbf{E}=0$. Bien que les espaces considérés ne sont plus de dimension finie dans le cas continu, le défaut peut être caractérisé par un sous-espace de dimension 1. Ce sous-espace, qui est plus généralement un groupe est le *groupe de cohomologie* associé à $C^\infty(\Omega,\mathbb{C})^2$ pour le complexe introduit dans le diagramme (1.27) ; sa dimension est caractéristique de la topologie du domaine, elle correspond au nombre de trous dans le domaine de dimension donnée. Plus précisément, elle est aussi égale au nombre de lignes orientées comme c qui ne peuvent être la frontière d'une surface incluse dans le domaine [37].

On voit ainsi que pour une triangulation correcte du domaine (qui n'oublie aucune particularité topologique), la suite discrète (1.22) aura les mêmes invariants topologiques que la suite continue (1.25) et conserve en ce sens les informations liées à la géométrie du domaine.

1.2.3 Extension à la dimension trois

En dimension trois, on peut ajouter un étage au complexe (1.27). Avec des espaces de fonctions régulières, elle s'écrit :

$$C^\infty(\Omega,\mathbb{C}) \xrightarrow{\text{grad}} C^\infty(\Omega,\mathbb{C})^3 \xrightarrow{\text{rot}} C^\infty(\Omega,\mathbb{C})^3 \xrightarrow{\text{div}} C^\infty(\Omega,\mathbb{C}) \tag{1.29}$$

Au niveau discret, en plus des ensembles $\mathcal{N}$, $\mathcal{E}$ et $\mathcal{F}$, on doit considérer l'ensemble $\mathcal{T}$ de cardinal T des éléments du maillage. Une grandeur supplémentaire définie par une intégrale de volume doit être introduite, à savoir la quantité de charge magnétique (électrique) dans un élément t. On définit une nouvelle matrice d'incidence volume-face D qui permet de relier un vecteur de $\mathbb{C}^F$ associé aux faces du maillage à un vecteur de $\mathbb{C}^T$ associé aux éléments du maillage. Ces entrées sont définies de la manière suivante :

$$D_{tf} = \begin{cases} 0 \text{ si } f \text{ n'est pas une face de l'élément } t, \\ 1 \text{ si } f \text{ est une face de l'élément } t \text{ et que les orientations coïncident,} \\ -1 \text{ si } f \text{ est une face de l'élément } t \text{ et que les orientations ne coïncident pas.} \end{cases} \tag{1.30}$$

La suite discrète a alors la forme suivante :

$$\mathbb{C}^N \xrightarrow{G} \mathbb{C}^E \xrightarrow{R} \mathbb{C}^F \xrightarrow{D} \mathbb{C}^T. \tag{1.31}$$

Comme pour la modélisation 2D, si le maillage considéré respecte la topologie du domaine, la suite (1.31) aura les mêmes invariants topologiques que la suite (1.29). On peut aussi de nouveau écrire un diagramme commutatif tel que (1.27) avec cette fois :

$$\begin{aligned} &\Pi^0 : u \mapsto \mathrm{u} \in \mathbb{C}^N \text{ de composantes } u(n),\ n \in \mathcal{N}, \\ &\Pi^1 : \mathbf{E} \mapsto \mathrm{E} \in \mathbb{C}^E \text{ de composantes } \int_e \mathbf{E}\cdot \mathrm{d}l,\ e \in \mathcal{E}, \\ &\Pi^2 : \mathbf{B} \mapsto \mathrm{B} \in \mathbb{C}^F \text{ de composantes } \int_f \mathbf{B}\cdot \mathrm{d}s,\ f \in \mathcal{F}. \\ &\Pi^3 : \rho_m \mapsto \mathrm{r_m} \in \mathbb{C}^T \text{ de composantes } \int_t \rho_m \mathrm{d}v,\ t \in \mathcal{T}. \end{aligned} \tag{1.32}$$

Le second groupe de cohomologie associé à $C^\infty(\Omega,\mathbb{C})^3$ traduit les défauts de représentation d'une induction magnétique $\mathbf{B}$, vérifiant $\operatorname{div}\mathbf{B} = 0$, par un potentiel vecteur $\mathbf{A}$ sous la forme $\mathbf{B} = \operatorname{rot}\mathbf{A}$. Sa dimension est égale au nombre de cavités du domaine. Cependant, si S est la frontière d'une cavité alors la quantité $\int_S \mathbf{B}\cdot \mathrm{d}s$ peut être supposée nulle et la représentation de $\mathbf{B}$ par un rotationnel est toujours possible. Des difficultés apparaissent seulement quand il s'agit de prendre en compte les conditions aux limites pour le potentiel vecteur et que l'on utilise une fonction de flot sur cette frontière [37, Chapitre 1 ex. 1.15].

1.2.4 Formulation variationnelle et discrétisation

La formulation variationnelle du problème conduit à introduire un cadre fonctionnel adéquat pour étudier l'existence et l'unicité des solutions continues et la convergence des discrétisations par éléments finis. Ces aspects de convergence ainsi que des aspects beaucoup généraux d'analyse fonctionnelle pour les équations de Maxwell peuvent être trouvés dans l'ouvrage [57].

Formulation variationnelle

Soit Γ_{D} la portion du domaine Ω sur laquelle on impose des conditions aux limites essentielles et Γ_{I} celle sur laquelle on impose des conditions d'impédance. Les différents modèles (1.5), (1.7), (1.8), (1.9) et (1.10) conduisent à une formulation forte de la forme :

$$\begin{cases} \text{Trouver } \mathbf{E} \text{ tel que :} \\ \operatorname{rot} \nu \operatorname{rot} \mathbf{E} + \gamma \mathbf{E} = f, \text{ dans } \Omega, \\ \mathbf{E} \times \mathbf{n} = g \text{ sur } \Gamma_{\mathrm{D}} \text{ et } \nu \operatorname{rot} \mathbf{E} \times \mathbf{n} = \delta \mathbf{n} \times (\mathbf{E} \times \mathbf{n}) \text{ sur } \Gamma_{\mathrm{I}}. \end{cases} \tag{1.33}$$

Le coefficient ν est toujours strictement positif. Le coefficient γ peut être :
- nul dans le cas statique (à moins que l'on utilise une régularisation) ;
- positif dans le cas du régime transitoire ;
- imaginaire pur dans le problème des courants de Foucault en régime harmonique ;
- à partie imaginaire parfois non nulle et à partie réelle négative dans le cas de l'équation vectorielle des ondes en régime harmonique.

Enfin, le coefficient δ est supposé non nul uniquement si l'on utilise des conditions aux limites absorbantes ; il est alors imaginaire pur.

Si l'on intègre sur le domaine Ω le produit scalaire de l'équation aux dérivées partielles par un champ test suffisament régulier et que l'on applique la formule de Green pour l'opérateur rot (voir Annexe A), on aboutit à une formulation intégrale de la forme :

$$\int_\Omega \nu \operatorname{rot} \mathbf{E} \cdot \operatorname{rot} \mathbf{E}' + \int_\Omega \gamma \mathbf{E} \cdot \mathbf{E}' + \int_{\Gamma_{\mathrm{D}}} \nu (\operatorname{rot} \mathbf{E} \times \mathbf{n}) \cdot \mathbf{E}' + \int_{\Gamma_{\mathrm{I}}} \delta (\mathbf{E} \times \mathbf{n}) \cdot (\mathbf{E}' \times \mathbf{n}) = \int_\Omega f \cdot \mathbf{E}'.$$

Si l'on veut étudier mathématiquement cette formulation variationnelle, il faut que toutes les intégrales soient définies. En particulier, il est nécessaire que $\mathbf{E}$, $\mathbf{E}'$, $\operatorname{rot} \mathbf{E}$ et $\operatorname{rot} \mathbf{E}'$ soient de carré intégrable, c.-à-d. appartiennent à $\mathbb{L}^2(\Omega)$ (voir Annexe A). On introduit donc l'espace des champs d'énergie finie $\mathbb{H}(\mathrm{rot}, \Omega)$:

$$\mathbb{H}(\mathrm{rot}, \Omega) = \{\mathbf{E} \in \mathbb{L}^2(\Omega) \ / \ \operatorname{rot} \mathbf{E} \in \mathbb{L}^2(\Omega)\}. \tag{1.34}$$

Pour vérifier les conditions aux limites essentielles sur Γ_D, on cherche la solution dans le sous-espace affine $\mathbb{H}_{\Gamma_{\mathrm{D}}}(\mathrm{rot}, \Omega)$ de $\mathbb{H}(\mathrm{rot}, \Omega)$:

$$\mathbb{H}_{\Gamma_{\mathrm{D}}}(\mathrm{rot}, \Omega) = \{\mathbf{E} \in \mathbb{H}(\mathrm{rot}, \Omega) \ / \ \mathbf{E} \times \mathbf{n} = g \text{ sur } \Gamma_{\mathrm{D}}\}. \tag{1.35}$$

Les champs tests sont pris dans l'espace vectoriel $\mathbb{H}_{\Gamma_{\mathrm{D}},0}(\mathrm{rot}, \Omega)$:

$$\mathbb{H}_{\Gamma_{\mathrm{D}},0}(\mathrm{rot}, \Omega) = \{\mathbf{E} \in \mathbb{H}(\mathrm{rot}, \Omega) \ / \ \mathbf{E} \times \mathbf{n} = 0 \text{ sur } \Gamma_{\mathrm{D}}\}. \tag{1.36}$$

On aboutit finalement à la formulation variationnelle :

$$\begin{cases} \text{Trouver } \mathbf{E} \in \mathbb{H}_{\Gamma_{\mathrm{D}}}(\mathrm{rot}, \Omega) \text{ tel que } a(\mathbf{E}, \mathbf{E}') = F(\mathbf{E}'), \ \forall \mathbf{E}' \in \mathbb{H}_{\Gamma_{\mathrm{D}},0}(\mathrm{rot}, \Omega), \\ \text{avec } a(\mathbf{E}, \mathbf{E}') = \displaystyle\int_\Omega \nu \operatorname{rot} \mathbf{E} \cdot \overline{\operatorname{rot} \mathbf{E}'} + \int_\Omega \gamma \mathbf{E} \cdot \overline{\mathbf{E}'} + \int_{\Gamma_{\mathrm{I}}} \delta (\mathbf{E} \times \mathbf{n}) \cdot (\overline{\mathbf{E}'} \times \mathbf{n}). \end{cases} \tag{1.37}$$

Le terme F représente la forme linéaire sur $\mathbb{H}_{\Gamma_{\mathrm{D}},0}(\mathrm{rot}, \Omega)$ définie par :

$$F(\mathbf{E}') = \int_\Omega f \cdot \overline{\mathbf{E}'}. \tag{1.38}$$

Remarque 1.1. *En dehors de* $\mathbb{H}(\mathrm{rot}, \Omega)$*, d'autres espaces jouent un rôle important dans la discrétisation des équations de Maxwell :*
- *pour les potentiels, il s'agit de l'espace* $\mathrm{H}^1(\Omega)$ *des fonctions de carré intégrable et dont le gradient est de carré intégrable.*
- *pour les champs de vecteurs densité de flux, on introduit aussi l'espace* $\mathbb{H}(\mathrm{div}, \Omega)$ *des champs de vecteurs de carré intégrables et dont la divergence est de carré intégrable.*

Ayant introduit ces espaces, on peut définir une suite de même caractéristique que la suite (1.29) *:*

$$\mathrm{H}^1(\Omega) \xrightarrow{\mathrm{grad}} \mathbb{H}(\mathrm{rot},\Omega) \xrightarrow{\mathrm{rot}} \mathbb{H}(\mathrm{div},\Omega) \xrightarrow{\mathrm{div}} \mathrm{L}^2(\Omega)\ . \tag{1.39}$$

Elle possède en particulier les mêmes invariants topologiques que les suites (1.29) *et* (1.31)*. Les applications linéaires* Π^0*,* Π^1*,* Π^2 *et* Π^3 *introduites en* (1.32) *ne sont par contre pas définies pour toutes les fonctions et champs de vecteurs de chacun de ces espaces. Cependant, sur les domaines de définitions* $D(\Pi^i)$ *de ces applications, on a aussi le diagramme commutatif suivant :*

$$\begin{array}{ccccccc} D(\Pi^0)\cap \mathrm{H}^1(\Omega) & \xrightarrow{\mathrm{grad}} & D(\Pi^1)\cap \mathbb{H}(\mathrm{rot},\Omega) & \xrightarrow{\mathrm{rot}} & D(\Pi^2)\cap \mathbb{H}(\mathrm{div},\Omega) & \xrightarrow{\mathrm{div}} & \mathbb{L}^2(\Omega) \\ \downarrow{\scriptstyle \Pi^0} & & \downarrow{\scriptstyle \Pi^1} & & \downarrow{\scriptstyle \Pi^2} & & \downarrow{\scriptstyle \Pi^3} \\ \mathbb{C}^N & \xrightarrow{\quad G \quad} & \mathbb{C}^E & \xrightarrow{\quad R \quad} & \mathbb{C}^F & \xrightarrow{\quad D \quad} & \mathbb{C}^T. \end{array} \tag{1.40}$$

Espaces d'éléments finis

Soit $\mathcal{T}$ un maillage par éléments finis tétraédriques du domaine Ω de $\mathbb{R}^3$. On introduit quatre espaces d'éléments finis $\mathcal{W}^i(\mathcal{T})$ que l'on peut relier au moyen des opérateurs de l'analyse vectorielle classique :

$$\mathcal{W}^0(\mathcal{T}) \xrightarrow{\mathrm{grad}} \mathcal{W}^1(\mathcal{T}) \xrightarrow{\mathrm{rot}} \mathcal{W}^2(\mathcal{T}) \xrightarrow{\mathrm{div}} \mathcal{W}^3(\mathcal{T})\ . \tag{1.41}$$

Les espaces $\mathcal{W}^i(\mathcal{T})$ sont les espaces d'éléments finis définis de la façon suivante :

- $\mathcal{W}^0(\mathcal{T})$ est construit à partir de l'élément P_1-Lagrange. Sur chaque élément, une fonction est approchée par une fonction affine ; chaque degré de liberté σ^0_n est lié à un noeud n de coordonnées x_n du maillage et associe à une fonction g la valeur $\sigma^0_n(g) = g(x_n)$. La base duale correspondante est notée $(w^0_n)_{n\in\mathcal{N}}$ et est incluse dans $\mathrm{H}^1(\Omega)$.
- $\mathcal{W}^1(\mathcal{T})$ est construit à partir de l'élément d'arête d'ordre 1 incomplet [62]. Sur chaque élément, une fonction est approchée par une fonction de la forme $a \times x + b$ avec a et b dans $\mathbb{C}^3$; chaque degré de liberté σ^1_e est lié à une arête orientée e du maillage de vecteur tangent t et associe à un champ de vecteurs $\mathbf{E}$ la valeur $\int_e \mathbf{E}\cdot t$. La base duale correspondante est notée $(w^1_e)_{e\in\mathcal{E}}$ et est incluse dans $\mathbb{H}(\mathrm{rot},\Omega)$.
- $\mathcal{W}^2(\mathcal{T})$ est construit à partir de l'élément de Raviart-Thomas [62]. Sur chaque élément, une fonction est approchée par une fonction de la forme $cx + d$ avec c dans $\mathbb{C}$ et d dans $\mathbb{C}^3$; chaque degré de liberté σ^2_f est lié à une face orientée f du maillage de normale $\mathbf{n}_f$ et associe à un champ de vecteurs $\boldsymbol{\eta}$ la valeur $\int_f \boldsymbol{\eta}\cdot\mathbf{n}_f$. La base duale correspondante est notée $(w^2_f)_{f\in\mathcal{F}}$ et est incluse dans $\mathbb{H}(\mathrm{div},\Omega)$.
- $\mathcal{W}^3(\mathcal{T})$ est construit à partir de l'élément P_0. Sur chaque élément, une fonction est approchée par une constante ; chaque degré de liberté σ^3_t est lié à un volume t du maillage et associe à une fonction h la valeur $\int_t h$. La base duale correspondante est notée $(w^3_t)_{t\in\mathcal{T}}$ et est incluse dans $\mathrm{L}^2(\Omega)$.

Cette suite d'espaces a été reconnue par Bossavit [14] comme un complexe d'éléments de Whitney [82]. On dispose d'applications pour les relier aux complexes (1.31) et (1.39) ; en particulier, pour chacun des espaces $\mathcal{W}^i(\mathcal{T})$, on peut définir une application interpolation $p^i_{\mathcal{T}}$ qui à un vecteur de $\mathbb{C}^{n_i}$, où n_i est la dimension de $\mathcal{W}^i(\mathcal{T})$, associe une fonction dans $\mathcal{W}^i(\mathcal{T})$:

$$\begin{aligned} p^i_{\mathcal{T}} : \mathbb{C}^{n_i} &\to \mathcal{W}^i(\mathcal{T}) \\ X &\mapsto \sum_{n=1}^{n_i} X_n w^i_n. \end{aligned} \tag{1.42}$$

On obtient le diagramme commutatif suivant qui complète le diagramme (1.40) :

$$\begin{array}{ccccccc} D(\Pi^0)\cap \mathrm{H}^1(\Omega) & \xrightarrow{\mathrm{grad}} & D(\Pi^1)\cap \mathbb{H}(\mathrm{rot},\Omega) & \xrightarrow{\mathrm{rot}} & D(\Pi^2)\cap \mathbb{H}(\mathrm{div},\Omega) & \xrightarrow{\mathrm{div}} & \mathbb{L}^2(\Omega) \\ \downarrow \Pi^0 & & \downarrow \Pi^1 & & \downarrow \Pi^2 & & \downarrow \Pi^3 \\ \mathbb{C}^N & \xrightarrow{G} & \mathbb{C}^E & \xrightarrow{R} & \mathbb{C}^F & \xrightarrow{D} & \mathbb{C}^T \\ \downarrow p_{\mathcal{T}}^0 & & \downarrow p_{\mathcal{T}}^1 & & \downarrow p_{\mathcal{T}}^2 & & \downarrow p_{\mathcal{T}}^3 \\ \mathcal{W}^0(\mathcal{T}) & \xrightarrow{\mathrm{grad}} & \mathcal{W}^1(\mathcal{T}) & \xrightarrow{\mathrm{rot}} & \mathcal{W}^2(\mathcal{T}) & \xrightarrow{\mathrm{div}} & \mathcal{W}^3(\mathcal{T}). \end{array} \tag{1.43}$$

Le complexe de Whitney (1.41) a des groupes de cohomologie de même dimension que la suite discrète (1.31) et que la suite (1.39). Ces propriétés signifient que les informations topologiques essentielles sont conservées de manière systématique et automatique ; en particulier, la condition inf-sup, condition suffisante pour montrer la convergence de discrétisation par éléments finis, est toujours vérifiée [3, 2].

En pratique, on doit souvent considérer des conditions aux limites essentielles sur une portion de la frontière Γ_{D}. Les propriétés de la suite (1.41) et le diagramme de compatibilité (1.43) restent valides "si l'on oublie" les entités géométriques situées sur cette portion de la frontière. En particulier les matrices d'incidence ne doivent pas prendre en compte les entités géométriques de cette portion de frontière.

Un point important qui sera repris dans la construction algébrique d'une suite d'espaces d'éléments finis présentée dans la Section 2.3.4 est la commutativité du diagramme au premier niveau qui s'écrit :

$$\mathrm{grad}\, p_{\mathcal{T}}^0 = p_{\mathcal{T}}^1 G. \tag{1.44}$$

En d'autres termes la matrice d'incidence arête-noeud G du maillage est la représentation discrète du gradient en tant qu'opérateur de l'espace $\mathcal{W}^0(\mathcal{T})$ dans l'espace $\mathcal{W}^1(\mathcal{T})$.

Discrétisation du problème

Soit $\mathcal{T}$ une triangulation du domaine Ω. L'espace des éléments finis d'arête d'ordre 1 incomplet $\mathcal{W}^1(\mathcal{T})$ est utilisé pour discrétiser la formulation (1.37). On introduit l'espace $\mathcal{W}^1_{\Gamma_{\mathrm{D}},0}(\mathcal{T})$, sous-espace des fonctions de $\mathcal{W}^1(\mathcal{T})$ appartenant à $\mathbb{H}_{\Gamma_{\mathrm{D}},0}(\mathrm{rot},\Omega)$ et un champ de $\mathbb{H}_{\Gamma_{\mathrm{D}}}(\mathrm{rot},\Omega)$ vérifiant $\mathbf{E}\times\mathbf{n} = g$ sur Γ_{D}, noté $\mathbf{E}_{\Gamma_{\mathrm{D}}}$. Le problème discret à résoudre s'écrit alors :

$$\left\{ \begin{array}{l} \text{Trouver } \tilde{\mathbf{E}}_h = \mathbf{E}_{\Gamma_{\mathrm{D}}} + \mathbf{E}_h \text{ avec } \mathbf{E}_h \in \mathcal{W}^1_{\Gamma_{\mathrm{D}},0}(\mathcal{T}) \text{ tel que :} \\ a(\mathbf{E}_h, \mathbf{E}'_h) = F(\mathbf{E}'_h) - \displaystyle\int_\Omega \nu\, \mathrm{rot}\, \mathbf{E}_{\Gamma_{\mathrm{D}}} \cdot \overline{\mathrm{rot}\, \mathbf{E}'_h} - \int_\Omega \gamma \mathbf{E}_{\Gamma_{\mathrm{D}}} \cdot \overline{\mathbf{E}'_h},\ \forall \mathbf{E}'_h \in \mathcal{W}^1_{\Gamma_{\mathrm{D}},0}(\mathcal{T}). \end{array} \right. \tag{1.45}$$

Dans la base $(w_e^1)_{e=1,\dots,E^h}$ de $\mathcal{W}^1_{\Gamma_{\mathrm{D}},0}(\mathcal{T})$, le problème discret se ramène à un système linéaire de la forme suivante :

$$Au = b, \text{ avec } A = S_\nu + M_\gamma + M_{\delta,\Gamma_{\mathrm{I}}}, \tag{1.46}$$

où :

- u contient les composantes de l'inconnu $\mathbf{E}_h$ dans la base des $(w_e^1)_{e=1,\dots,E^h}$;
- b contient les contributions des différents termes sources, c.-à-d. :

$$b_e = F(w_e^1) - \int_\Omega \nu\, \mathrm{rot}\, \mathbf{E}_{\Gamma_{\mathrm{D}}} \cdot \mathrm{rot}\, w_e^1 - \int_\Omega \gamma \mathbf{E}_{\Gamma_{\mathrm{D}}} \cdot w_e^1;$$

- les coefficients de S_ν prennent les valeurs $\int_\Omega \nu\, \mathrm{rot}\, w_j^1 \cdot \mathrm{rot}\, w_i^1$;
- les coefficients de M_γ prennent les valeurs $\int_\Omega \gamma w_j^1 \cdot w_i^1$;
- les coefficients de $M_{\delta,\Gamma_{\mathrm{I}}}$ prennent les valeurs $\int_{\Gamma_{\mathrm{I}}} \delta (w_j^1 \times \mathbf{n}) \cdot (w_i^1 \times \mathbf{n})$;

La matrice S_ν est symétrique semi-définie positive. Si les e_i sont les vecteurs de la base canonique de $\mathbb{R}^E$, on note que $w_i^1 = p_{\mathcal{T}}^1 e_i$; en utilisant la relation $\mathrm{rot}\, p_{\mathcal{T}}^1 = p_{\mathcal{T}}^2 R$ issue du diagramme commutatif (1.43) où R désigne la matrice d'incidence face-arête il vient :

$$\int_\Omega \mathrm{rot}\, w_j^1 \cdot \mathrm{rot}\, w_i^1 = \int_\Omega \nu (p_{\mathcal{T}}^2 R e_j) \cdot (p_{\mathcal{T}}^2 R e_i) = \sum_{l=1}^{E} R_{lj} \Big(\sum_{k=1}^{F} \big(\int_\Omega \nu w_l^2 \cdot w_k^2 \big) R_{ki} \Big),$$

soit $S_\nu = R^t M_\nu R$ où M_ν est la matrice de composante $\int_\Omega \nu w_l^2 \cdot w_k^2$. La matrice M_ν étant symétrique définie positive (SDP), le noyau de S_ν coïncide avec le noyau de R. En conséquence, on aura $S_\nu G = 0$ où G est la matrice d'incidence arête-noeud.

La matrice M_γ est aussi symétrique. Elle est :

- nulle dans le cas d'un problème de magnétostatique sans régularisation. La matrice A n'est pas alors inversible. Cependant, si le second membre est compatible ($G^t b = 0$), il est possible d'utiliser le gradient conjugué qui va converger vers une solution [44]. Une autre possibilité pour résoudre le système (1.46) est de travailler sur une sous-matrice inversible en utilisant des techniques de graphe (détermination d'un arbre couvrant du graphe, voir [58]).
- définie positive dans le cas du régime transitoire et donc A est SDP ;
- imaginaire pure dans le cas des courants de Foucault en régime harmonique. La matrice A est alors inversible.
- à partie réelle définie négative dans le cas de l'équation des ondes en régime harmonique, on obtient alors une matrice globale A indéfinie.

Ce type de matrice généralise le type de systèmes présenté pour la théorie des réseaux.

Les équations physiques et leurs discrétisations ont été présentées pour quelques modèles. Elles aboutissent à l'écriture de systèmes linéaires dont les matrices ont des caractéristiques :

- qui prennent en compte certaines propriétés du complexe de Whitney et donc de la géométrie du domaine.
- qui dépendent des valeurs des coefficients ν, γ et δ donc du modèle choisi.

La mise au point d'algorithmes de résolution efficaces devra tenir compte au mieux de ces deux aspects.

Chapitre 2

Méthodes itératives pour la résolution des systèmes linéaires

Le calcul de l'approximation numérique des solutions d'une équation aux dérivées partielles par éléments finis ou différences finies conduit généralement à la résolution de systèmes linéaires, ce qui est souvent la partie la plus coûteuse en temps de calcul et en espace mémoire. Cependant, des gains de performance sont possibles si sont prises en compte les propriétés intrinsèques du problème et de sa discrétisation.

Les méthodes directes, telles que Gauss, factorisation LU, conduisent à des solveurs souvent robustes mais elles ne peuvent pas prendre en compte les spécificités du problème de départ. On peut seulement tirer profit du caractère creux des matrices provenant de discrétisation par éléments finis ou différences finies [26]. Au contraire, des méthodes itératives multiniveau, telles que la méthode multigrille géométrique [74, 40] ou la méthode multigrille algébrique [17, 67], permettent de concevoir des algorithmes très performants mais spécifiques à certaines classes d'équations aux dérivées partielles.

On rappelle tout d'abord dans ce chapitre quelques notions de base concernant les méthodes itératives de résolution des systèmes linéaires, telles que les méthodes par sous-espaces de Krylov et le préconditionnement. Ensuite, les méthodes multiniveau géométrique puis algébrique sont introduites. Enfin, on présente les particularités à prendre en compte pour résoudre la classe de problèmes définie par (1.33) au Chapitre 1 pour obtenir des méthodes multiniveau performantes.

2.1 Notions de base pour les méthodes itératives

Les méthodes itératives consistent à construire à partir d'un vecteur initial, une suite de vecteurs qui converge vers la solution du système. Ce sont très souvent des méthodes par sous-espaces de Krylov couplées à des techniques de préconditionnement.

2.1.1 Méthodes par sous-espaces de Krylov

La résolution de systèmes linéaires reposent, dans la plupart des codes de calcul, sur l'utilisation de méthodes par sous-espaces de Krylov dont les représentants les plus connus sont la méthode du gradient conjugué pour les matrices symétriques définies positives et la méthode GMRES. Ceci est dû à leur efficacité sur une grande variété de problèmes. Nous décrivons ici le principe général de ces méthodes ; voir [68] pour une présentation générale et [35] pour une présentation avec application en électromagnétisme.

Principe de fonctionnement

Soit A la matrice (n, n) du système à résoudre, v un vecteur quelconque, on note :

$$K^m(A, v) = \text{vect}(v, Av, Av^2, ..., Av^{m-1}).$$

le *sous-espace de Krylov* engendré par A et v.

Pour un vecteur initial x_0 donné, une méthode itérative par sous-espaces de Krylov consiste à chercher x_m, $m^{\text{ième}}$ itéré, dans le sous-espace affine $x_0 + K^m(A, r_0)$ où $r_0 = b - Ax_0$ est le résidu initial. On peut, en conséquence, écrire ce $m^{\text{ième}}$ itéré sous la forme $x_m = x_0 + V_m y_m$ où V_m est la matrice contenant les vecteurs d'une base de $K^m(A, v)$, $\{v_1, ..., v_m\}$ et y_m un vecteur de coefficients. Les différences entre les méthodes reposent alors sur :

- le choix de la base $\{v_1, ..., v_m\}$,
- la méthode utilisée pour déterminer les coefficients de y_m.

Construction de V_m

Deux constructions de la base $\{v_1, ..., v_m\}$ de $K^m(A, r_0)$ peuvent être distinguées : l'une utilise l'algorithme d'Arnoldi, l'autre l'algorithme de biorthogonalisation de Lanczos.

Algorithme d'Arnoldi Le vecteur v_{m+1} est obtenu en orthogonalisant Av_m, au sens du produit scalaire canonique de $\mathbb{R}^n$, par rapport à tous les vecteurs de la base de $K^m(A, r_0)$ générés auparavant, puis en normalisant le résultat obtenu. A chaque itération, une *base orthonormale* du sous-espace de Krylov $K^m(A, r_0)$ est ainsi obtenue.

Algorithme de biorthogonalisation de Lanczos La base $\{v_1, ..., v_m\}$ de $K^m(A, r_0)$ doit respecter une relation de bi-orthogonalité : $(v_i, w_j) = \delta_{i,j}\ \forall i, j$ par rapport à la base $\{w_1, ..., w_m\}$ d'un second espace de Krylov L_m. Ce sous-espace L_m est engendré par A^T et un vecteur $\tilde{r}_0$.

Dans un des cas qui nous intéressent ici, matrice symétrique à coefficients complexes, un bon choix d'espace L_m est $K^m(A^T, \overline{r_0}) = K^m(A, \overline{r_0})$, mais d'autres choix peuvent être faits.

Construction de y_m

Plusieurs procédés permettent, une fois la base construite, de déterminer les coefficients de y_m. Deux types de techniques existent : les procédés de projection ou ceux de minimisation de résidu dans $K^m(A, r_0)$.

Méthode de Ritz-Galerkin C'est une méthode de projection : on choisit y_m de manière à rendre $r_m = b - Ax_m$, le $m^{\text{ième}}$ résidu, orthogonal au sous-espace de Krylov $K^m(A, r_0)$.

Méthode de Petrov-Galerkin C'est aussi une méthode de projection. Elle propose de déterminer y_m de manière à rendre r_m orthogonal au second sous-espace de Krylov L_m.

Minimisation de résidu Il s'agit de minimiser $\|r_m\|_2 = \sqrt{\sum_i r_m^i \overline{r_m^i}}$ dans le sous-espace $K^m(A, r_0)$. Cette minimisation implique de résoudre à chaque itération un problème aux moindres carrés.

On peut alors préférer résoudre le problème modifié proposé par Freund [31] de quasi-minimisation du résidu. Quelques exemples de méthodes par sous-espaces de Krylov sont regroupés dans le tableau 2.1 avec une classification suivant les approches décrites ci-dessus.

	Algorithme d'Arnoldi	Algorithme de Lanczos
Projection de Ritz-Galerkin	CG, FOM [68, Chapitre 6]	
Projection de Petrov-Galerkin		BiCGCR [23] COCG [75]
(Quasi) Minimisation du résidu	CR, GMRES [68, Chapitre 6]	QMR [31, 30]

TAB. 2.1 – Classification de méthodes par sous-espaces de Krylov.

Les méthodes COCG (Conjugate Orthogonal Conjugate Gradient), BiCGCR (BiConjugate Gradient Conjugate Residual) et QMR (Quasi-minimum Residual) du tableau 2.1 sont adaptées pour résoudre des systèmes linéaires symétriques à coefficients complexes. Nous les avons comparées dans [64] sur des problèmes régis par l'équation vectorielle des ondes en régime harmonique (1.10) ; on peut retrouver le

contenu et les résultats de cet article en Section B.1. On peut noter que la méthode COCG donne les meilleurs résultats sur les problèmes testés.

2.1.2 Préconditionnement

Les méthodes par sous-espaces de Krylov sont des algorithmes dont on peut optimiser la vitesse de convergence grâce à des méthodes de préconditionnement.

On peut définir le *nombre de conditionnement* en norme 2[1] pour la matrice A de la manière suivante :

$$K(A) = \|A\|_2 \|A^{-1}\|_2.$$

Ce nombre de conditionnement est toujours supérieur ou égal à 1. Il influe fortement sur la vitesse de convergence de la plupart des algorithmes itératifs.

Par exemple, si l'on utilise l'algorithme du *gradient conjugué* pour résoudre un système de matrice A *symétrique définie positive*, on montre que [68, Chapitre 6] :

$$\|u - u_k\|_A \leq 2\left(\frac{\sqrt{K(A)} - 1}{\sqrt{K(A)} + 1}\right)^k \|u - u_0\|_A. \tag{2.1}$$

où u_0 et u_k désignent respectivement l'itéré initial et le k$^{\text{ième}}$ itéré, $\|.\|_A$ désigne la *norme d'énergie* c.-à-d. la norme induite par le produit scalaire $(A.,.)$. On constate que plus le nombre de conditionnement est proche de 1, plus la convergence de l'algorithme est rapide.

La discrétisation par différences finies ou par éléments finis des problèmes auxquels on s'intéresse conduit à des matrices dont le nombre de conditionnement augmente comme le carré de l'inverse du pas de discrétisation lorsque l'on raffine la maillage [7]. Cela se traduit en pratique par une augmentation du nombre d'itérations pour résoudre le système avec une précision fixée et un temps de calcul prohibitif pour les très grands systèmes pour lesquels il est donc nécessaire d'utiliser des techniques de préconditionnement.

L'idée du préconditionnement est d'introduire un problème équivalent où la matrice du système à résoudre est *mieux conditionnée* (nombre de conditionnement plus faible) :

$$\text{Résoudre } \hat{A}\hat{x} = \hat{b}, \text{ où } \hat{A} = C^{-t}AC^{-1}, \hat{x} = Cx, \hat{b} = C^{-t}b.$$

La matrice $M = C^tC$ est dite *matrice de préconditionnement*. Elle est déterminée de manière à ce que $K(\hat{A}) \ll K(A)$.

Dans la pratique, l'introduction d'un préconditionneur nécessite simplement de résoudre, à l'intérieur de la boucle d'itération classique de la méthode par sous-espaces de Krylov, un système de matrice M. Il est donc important que ce système soit d'un *coût faible* à résoudre.

Les méthodes les plus classiques de préconditionnement sont des factorisations partielles [52] de la matrice du système où l'on construit explicitement une factorisation de la matrice M ou un pas d'une itération linéaire standard telle que les méthodes de Jacobi avec relaxation, Gauss-Seidel symétrique ou plus généralement SSOR [68, Chapitre 10].

On utilise aussi des méthodes itératives moins simples de mise en oeuvre et qui sont définies par décomposition d'espace et corrections par sous-espaces [1, 83]. Cette classe regroupe la plupart des méthodes de décomposition de domaines et multiniveau. Le choix d'utiliser ces méthodes en tant que préconditionneur plutôt que solveur du système conduit à des algorithmes qui convergent plus rapidement et qui peuvent être plus robutes. Par exemple, pour le régime harmonique, la gamme de fréquences où la méthode de résolution est performante s'étend ; voir les exemples numériques dans le cas de multigrille et des équations de Maxwell dans [10, 36].

[1] $\|A\|_2 = \max_{x \in \mathbb{C}^n} \left\{ \frac{\|Ax\|_2}{\|x\|_2} \right\} = \sqrt{\rho(\bar{A}^t A)}$ où ρ désigne le rayon spectral de la matrice. Pour une matrice hermitienne définie positive, cela correspond ainsi au rapport de la valeur propre la plus grande grande sur la plus petite.

2.2 Principe des méthodes multiniveau

On souhaite résoudre un problème aux limites sur un domaine Ω, discrétisé par une méthode d'éléments finis ou de différences finies. On se propose d'expliquer les grandes lignes de la méthode multigrille pour la résolution numérique de ce problème. Dans un premier temps on présente la méthode à deux grilles, avant d'étendre l'analyse à la mise au point d'un algorithme multigrille complet. Les grands principes des méthodes multiniveau algébriques viendront compléter ces notions.

2.2.1 Méthode à deux grilles

Sur le domaine d'étude Ω, on définit deux maillages distincts :
- un maillage grossier $\mathcal{T}_H$ de paramètre H qui désigne le diamètre maximal des éléments du maillage,
- un maillage fin $\mathcal{T}_h$ obtenu par raffinement du maillage grossier $\mathcal{T}_H$ de paramètre h.

On prend généralement $H = 2h$. Après avoir discrétisé le problème par différences finies ou par éléments finis, on doit résoudre sur le maillage fin un système linéaire de la forme :

$$A_h u_h = b_h. \tag{2.2}$$

On souhaiterait le résoudre en s'appuyant sur les informations apportées par le maillage grossier sur lequel le système linéaire s'écrit :

$$A_H u_H = b_H.$$

Pour cela, il faut définir :
- un *opérateur de prolongement* P qui permet de passer d'une solution formulée sur le maillage grossier à une solution formulée sur le maillage fin,
- un *opérateur de restriction* R qui assure l'opération inverse (on utilise souvent P^t).

La méthode se décompose alors en 3 étapes :

1. Le *prélissage* consiste à appliquer une itération linéaire au vecteur u_h (ex : 1 ou 2 itérations de Gauss-Seidel). Matriciellement, cela s'écrit :

$$u_h \leftarrow u_h + M_h^{-1}\big(A_h u_h - b_h\big). \tag{2.3}$$

   Cette étape a pour but de lisser l'erreur (supprimer la composante oscillante de l'erreur, voir fig. 2.1 pour $-\triangle u + u = 0$ sur le carré unité avec condition de Neumann au bord). La matrice M_h, ou de façon équivalente la méthode de lissage, doit être choisie de façon à atteindre cet objectif. Après le prélissage, la composante lisse de l'erreur est dominante, donc la restriction par R de l'erreur à la grille grossière fait perdre peu d'information.

2. La *correction* consiste à résoudre sur la grille grossière (par une méthode directe généralement) :

$$A_H \theta_H = R(b_h - A_h u_h). \tag{2.4}$$

   Le résultat obtenu permet alors de *corriger* l'itéré :

$$u_h \leftarrow u_h + P\theta_H. \tag{2.5}$$

   L'étape de correction consiste à réduire la composante lisse de l'erreur qui est la seule visible sur la grille grossière.

3. Le *postlissage* effectue le même travail que l'étape de prélissage sur la solution corrigée.

Cette méthode est itérative, les étapes précédentes sont donc répétées tant que la précision désirée n'est pas atteinte (voir fig. 2.2).

Cependant, n'utiliser que deux grilles limite l'intérêt de la méthode, la résolution sur la grille grossière pouvant conduire à un système matriciel qui reste de grande dimension.

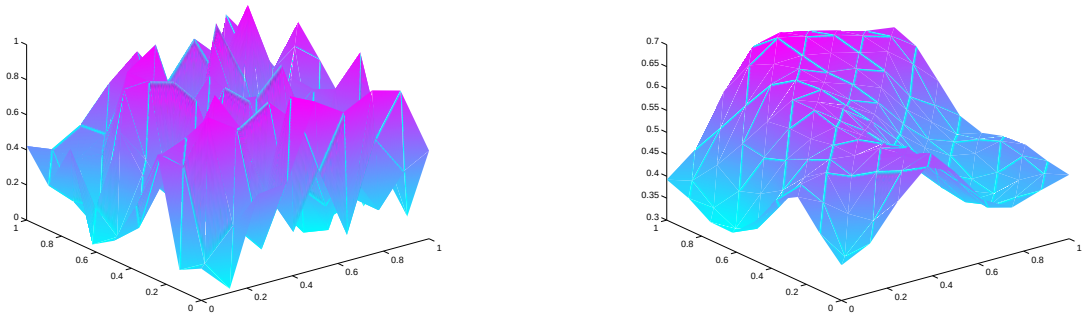


Fig. 2.1 – Erreur initiale et lissée après 2 itérations de Gauss-Seidel.

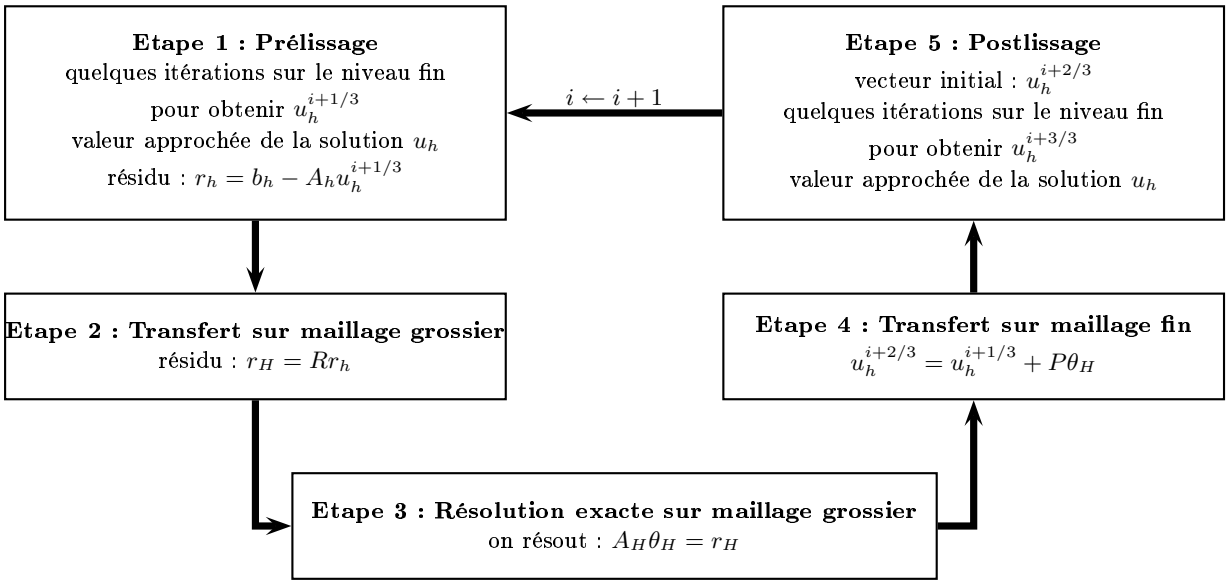


Fig. 2.2 – Représentation schématique de la méthode itérative deux grilles.

2.2.2 Méthode multigrille géométrique

En partant de l'approche deux grilles, on peut définir de manière récursive un algorithme dit *multigrille.*

On se donne une hiérarchie de maillages $\mathcal{T}_1, ..., \mathcal{T}_j$ de paramètres respectifs $h_1, ..., h_j$ telle que $\mathcal{T}_k$ soit un raffinement de $\mathcal{T}_{k-1}$, $\forall k \in \{1, \ldots, j\}$.

Les notations $P_{k-1\to k}$ et $R_{k\to k-1}$ désignent les opérateurs de transfert (prolongement et restriction) nécessaires au passage du niveau k au niveau $k-1$. La matrice A_k correspond à la discrétisation du problème sur le maillage $\mathcal{T}_k$.

On peut alors formuler l'Algorithme 2.1 qui correspond à une itération de multigrille. On prend généralement le paramètre γ égal à 1, ce qui correspond à un V-cycle, ou 2, ce qui correspond à un W-cycle.

Procédure MGC(l : **entier**, u_l : **vecteur**, b_l : **vecteur**)
  Si $(l = 1)$ **Alors**
    $u_1 \leftarrow A_1^{-1} b_1$;
  Sinon
    $u_l \leftarrow \text{liss}(u_l, b_l)$; *[prélissage]*
    $\theta_{l-1} \leftarrow 0$;
    Pour i **de** 1 **à** γ **faire**
      MGC($l-1$, θ_{l-1}, $R_{l\to l-1}(b_l - A_l u_l)$) ;
    Fin Pour
    $u_l = u_l + P_{l-1\to l}\theta_{l-1}$;
    $u_l \leftarrow \text{liss}(u_l, b_l)$; *[postlissage]*
  Fin Si
Fin

Algorithme 2.1: Algorithme multigrille.

Pour les problèmes dont l'opérateur est du type Laplacien, l'algorithme multigrille avec un lisseur de Gauss-Seidel conduit à un solveur *optimal.* Le terme optimal signifie que la complexité[2] de l'algorithme et la quantité de données à stocker pour sa mise en oeuvre (en pratique temps CPU et occupation mémoire) varient linéairement avec le nombre d'inconnues sur le maillage le plus fin. C'est le meilleur résultat que l'on puisse obtenir pour la résolution de systèmes linéaires.

Cependant, afin d'obtenir ce résultat pour d'autres classes de problèmes, comme ceux qui nous intéressent où apparaît l'opérateur rot rot, il est nécessaire d'adapter les différentes composantes de la méthode. De façon générale, il s'agit de déterminer :

- un *lisseur efficace* qui atténue la composante oscillante de l'erreur,
- une *hiérarchie de niveaux* et d'opérateurs de prolongement et de restriction qui s'accordent aux caractéristiques du problème à résoudre (anisotropie ou inhomogénéité par exemple).

Concernant la définition des "grilles" il existe deux approches :

- une approche géométrique : on construit une hiérarchie de maillages sur le domaine de résolution ce qui implique de disposer d'un mailleur adapté. On parlera de méthode multigrille géométrique dans la suite.
- une approche algébrique : seul le maillage du domaine sur lequel on résout le problème est requis. On construit algébriquement les opérateurs de transfert et les matrices correspondant à des discrétisations de l'opérateur initial à différents niveaux. On parlera de méthode multiniveau algébrique dans la suite.

[2]Nombre d'opérations arithmétiques élémentaires pour effectuer la tâche.

2.2.3 Principe des méthodes multiniveau algébriques

Matrice de prolongement pour le multigrille géométrique

Dans le cas où l'on dispose d'une hiérarchie de maillages emboîtés, la méthode multigrille géométrique peut être mise en oeuvre et les opérateurs de transfert entre niveaux sont définis de manière naturelle à partir de l'inclusion des espaces d'éléments finis [19, Chapitre 6].

En effet, si $(w_i^H)_{i=1,\dots,N^H}$ désigne la base d'éléments finis de l'espace V_H sur la grille grossière et $(w_i^h)_{i=1,\dots,N^h}$ la base d'éléments finis de l'espace V_h sur la grille fine, l'inclusion de V_H dans V_h implique qu'il existe une matrice P telle que :

$$w_i^H = \sum_{j=1}^{N^h} P_{ji} w_j^h, \ \forall i \in \{1, \dots, N^H\}. \tag{2.6}$$

Une fonction de composantes données par le vecteur x^H dans la base grossière a alors ses composantes données par Px^H dans la base fine. La matrice P apparaît donc comme une matrice de prolongement naturelle. On montre aussi que la matrice A_H du problème discrétisé sur la grille grossière est reliée à la matrice A_h du problème sur la grille fine par la relation de Galerkin : $A_H = P^t A_h P$.

Mise en oeuvre des méthodes multiniveau algébriques

Lorsque l'on ne dispose pas d'une suite de maillages emboîtés mais uniquement du maillage sur lequel on désire résoudre le problème, on doit générer algébriquement des niveaux grossiers à partir de niveaux plus fins. Il s'agit donc, à partir d'un ensemble de variables fines et d'une matrice de discrétisation qui relie ces variables fines, de définir des stratégies pour déterminer un ensemble de variables grossières, un opérateur de prolongement entre les variables grossières et fines et une matrice de discrétisation au niveau grossier.

Une introduction aux méthodes multiniveau algébriques est donnée dans [78]. On peut décrire brièvement le principe de l'une d'entre elles dans le cas où la matrice de discrétisation A_h est symétrique.

Tout d'abord, on utilise un algorithme pour déterminer un ensemble de noeuds ω_C dits noeuds maîtres qui vont représenter les variables grossières. Les noeuds restants, dits esclaves, sont regroupés dans l'ensemble ω_F. Pour faire cette répartition des noeuds, on définit un graphe valué en utilisant les coefficients non nuls de la matrice au niveau fin : si a_{ij} est non nul avec j différent de i, le noeud i est connecté au noeud j. Le poids $|a_{ij}|$ est affecté à l'arête qui relie le noeud i au noeud j. Le principe est alors de construire un ensemble maximal de noeuds maîtres qui ne soient pas fortement connectés entre eux. A chaque noeud est affecté un poids qui mesure le nombre de voisins fortement connectés. Le noeud de plus fort poids est placé dans l'ensemble ω_C et ses voisins dans ω_F. Ce noeud et ses voisins sont supprimés du graphe et le poids des noeuds restants est réestimé. L'algorithme se poursuit jusqu'à ce que tous les noeuds soient répartis. Le voisinage des noeuds fortement connectés à un noeud donné est déterminé suivant un critère qui peut être $|a_{ij}| \geq \theta \sqrt{|a_{ii} a_{jj}|}$ [76], $|a_{ij}| \geq \theta \max_{k \neq l} |a_{kl}|$ [67], $|a_{ij}| \geq \theta$ [49], pour un paramètre θ à ajuster. Un exemple de simulation est donné figure 2.3.

L'algorithme de sélection des noeuds grossiers de Ruge et Stüben [67, 78, Chapitre 4] est utilisé sur la matrice issue de la discrétisation de $-\triangle u + u$ sur le carré unité avec conditions de Neumann au bord et le maillage ci-contre.
Les noeuds maîtres sont représentés par des disques pleins.

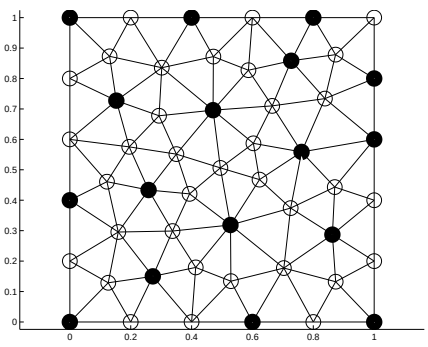


FIG. 2.3 – Sélection des noeuds maîtres.

Afin de construire une matrice de prolongement, on introduit pour chaque noeud esclave d'indice i un ensemble $\omega_{C,i}$ qui contient des noeuds de ω_C voisins de i. Différentes techniques peuvent être ensuite utilisées pour la détermination des coefficients d'interpolation [78]. La matrice de prolongement P la plus simple à construire est celle qui préserve les fonctions constantes en passant du niveau grossier au niveau fin, c.-à-d. :

$$P_{ij} = \begin{cases} 1 \text{ si } i = j \in \omega_C, \\ \dfrac{1}{|\omega_{C,i}|} \text{ si } i \in \omega_F \text{ et } j \in \omega_{C,i}, \\ 0 \text{ sinon.} \end{cases} \tag{2.7}$$

où la notation $|Z|$ désigne le nombre d'éléments dans un ensemble fini Z.

Enfin, la matrice au niveau grossier A_H est généralement construite à partir de la matrice A_h au niveau fin en utilisant un *produit de Galerkin* : $A_H = P^t A_h P$.

Remarque 2.1. *Pour la détermination des variables grossières, une approche simple basée uniquement sur la place des éléments non nuls de la matrice au niveau fin est proposée dans [9]. A l'opposé des stratégies plus sophistiquées prenant en compte des informations géométriques ou les matrices élémentaires calculées pendant l'assemblage du problème sont données dans [20, 39].*

2.3 Particularités pour les équations de Maxwell

L'utilisation d'une méthode multiniveau pour la résolution des équations de Maxwell nécessite le choix de lisseurs adaptés prenant en compte les caractéristiques propres à l'opérateur rot rot.

2.3.1 Difficultés liées à l'opérateur "rot rot"

Les lisseurs classiques (Jacobi avec relaxation, Gauss-Seidel par points) atténuent fortement la partie d'énergie élevée de l'erreur ; celle-ci correspond pour une forme bilinéaire définie positive aux composantes modales associées aux valeurs propres de modules les plus grands.

Pour un opérateur comme le Laplacien, cette partie coïncide avec la partie "oscillante" d'où l'efficacité de ces lisseurs.

Au contraire pour les opérateurs avec un terme en rot rot, il existe des vecteurs fortement oscillants et de faible énergie. Revenons à la classe de problèmes définie par la formulation (1.33). Plaçons nous dans le cas où ν et γ sont strictement positifs et δ nul. La forme bilinéaire a dans la formulation variationnelle associée donnée en (1.37) définit un produit scalaire qui induit une norme dite d'énergie. Considérons alors dans le cube unité le vecteur champ :

$$\mathbf{E} = \begin{pmatrix} 0 & \sin(n\pi x) & 0 \end{pmatrix}^t. \tag{2.8}$$

L'énergie de ce champ sur le cube unité est donnée par :

$$a(\mathbf{E},\mathbf{E}) = \int_{[0;1]^3} \nu \operatorname{rot}\mathbf{E} \cdot \operatorname{rot}\mathbf{E} + \int_{[0;1]^3} \gamma \mathbf{E}\cdot\mathbf{E} = (n\pi)^2 \frac{\nu}{2} + \frac{\gamma}{2}. \tag{2.9}$$

Son énergie est d'autant plus forte qu'il est oscillant. Considérons maintenant le vecteur champ à rotationnel nul et de même amplitude :

$$\tilde{\mathbf{E}} = \begin{pmatrix} \sin(n\pi x) & 0 & 0 \end{pmatrix}^t. \tag{2.10}$$

Son énergie est égale à $\gamma/2$ et ne dépend pas de n. Bien qu'ils aient des fréquences angulaires identiques, l'énergie de ces deux champs est donc très différente pour des valeurs de n élevées. Ce phénomène est d'autant plus amplifié que ν est prépondérant devant γ.

De façon plus générale, l'existence d'un espace de grande dimension (infinie dans le cas continu et proportionnelle au nombre d'inconnues dans le cas discret) de champs de vecteurs à rotationnel nul et d'énergie "quasi-nulle" peut expliquer que le lisseur de Gauss-Seidel par points ne fonctionne plus pour l'opérateur rot rot [41].

2.3.2 Choix du lisseur

Il existe deux références principales qui proposent un lisseur efficace pour les équations de Maxwell : Hiptmair [41] et Arnold, Falk et Winther [4].

La justification théorique de l'efficacité de ces deux lisseurs est faite pour la classe de problèmes (1.33) dans le cas où ν et γ sont strictement positifs et lorsqu'ils sont utilisés dans une méthode multigrille géométrique. Elle s'appuie sur la décomposition de Helmholtz discrète de l'espace $\mathcal{W}^1(\mathcal{T})$ des éléments d'arête d'ordre 1 :

$$\mathcal{W}^1(\mathcal{T}) = \operatorname{grad} \mathcal{W}^0(\mathcal{T}) \oplus \mathcal{D}(\mathcal{T}). \tag{2.11}$$

L'espace $\operatorname{grad} \mathcal{W}^0(\mathcal{T})$ est l'espace des champs de $\mathcal{W}^1(\mathcal{T})$ à rotationnel nul. L'espace $\mathcal{D}(\mathcal{T})$ est composé de vecteurs à divergence nulle faiblement, c.-à-d. :

$$A \in \mathcal{D}(\mathcal{T}) \;\Rightarrow\; \int_\Omega A \cdot \operatorname{grad} \phi = 0, \;\forall \phi \in \mathcal{W}^0(\mathcal{T}).$$

En ce qui concerne leur mise en oeuvre, cette décomposition n'apparaît explicitement que dans le lisseur de Hiptmair.

Sur le système $Au = b$, le lisseur de Hiptmair effectue à chaque pas successivement :

- une itération de Gauss-Seidel sur ce système,
- une itération de Gauss-Seidel sur un système de matrice $A_\phi = G^t A G$ et de second membre $\rho = G^t(b - Au)$ où G est la matrice d'incidence arête-noeud du maillage ; ce vecteur correspond à la divergence discrète de $b - Au$. La matrice A_ϕ provient de la discrétisation du problème dans l'espace $\operatorname{grad} \mathcal{W}^0(\mathcal{T})$, elle coïncide avec la matrice d'un problème en potentiel scalaire.

L'algorithme 2.2 décrit ce lisseur en détail. L'indice l fait référence au niveau l sur lequel l'opération est effectuée. La multiplication par la matrice G, qui correspond au calcul d'un gradient discret, permet le passage d'un vecteur de degrés de liberté nodaux ψ à un vecteur de degrés de liberté en arête.

Procédure liss(u_l : **vecteur**, b_l : **vecteur**)
  itération de Gauss-Seidel sur $A_l u_l = b_l$;
  $\rho_l \leftarrow (G^l)^t (b_l - A_l u_l)$;
  $\psi_l \leftarrow 0$;
  itération de Gauss-Seidel sur $A_{l,\phi} \psi_l = \rho_l$;
  $u_l \leftarrow u_l + G^l \psi_l$;
Fin

Algorithme 2.2: Algorithme du lisseur défini par Hiptmair.

Arnold, Falk and Winther ne profitent pas aussi explicitement de la décomposition de Helmholtz dans la construction de leur lisseur. Ils proposent d'utiliser un préconditionneur de Schwarz avec recouvrement, dont le principe est rappelé en Annexe B Section B.2. L'espace d'éléments finis d'arête est écrit comme une somme non directe de sous-espaces :

$$\mathcal{W}^1_{\Gamma_D}(\mathcal{T}) = \sum_{n \in \mathcal{N}} \mathcal{W}^{1,n}_{\Gamma_D}(\mathcal{T}), \tag{2.12}$$

où le sous-espace $\mathcal{W}^{1,n}_{\Gamma_D}(\mathcal{T})$ est l'espace d'éléments d'arête engendré par les fonctions de base $(w^1_e)_e$ associées aux arêtes e ayant le noeud n en commun. On corrige l'erreur sur chacun de ces sous-espaces lors d'une itération du préconditionneur.

Remarque 2.2. *La démonstration de l'optimalité de la méthode multigrille avec les lisseurs de Hiptmair et de Arnold, Falk et Winther est faite dans [36] pour les équations de Maxwell harmoniques. Le cas de la permittivité non réelle et des conditions aux limites absorbantes n'est cependant pas traité.*

Remarque 2.3. *Des problèmes en magnétostatique dans [73, 22, 69], en courants de Foucault et en régime transitoire dans [70] et en régime harmonique pour l'équation vectorielle des ondes dans [10] sont résolus à l'aide d'une méthode multigrille géométrique qui utilise l'un ou l'autre de ces deux lisseurs.*

2.3.3 Utilisation des lisseurs pour le préconditionnement

Les lisseurs de Hiptmair et de Arnold, Falk et Winther sont aussi des méthodes itératives de résolution qui en tant que telles peuvent être utilisées comme méthode de préconditionnement. Nous les avons testées dans [63], [64] et [65] sur l'équation vectorielle des ondes en régime harmonique, pour des géométries et des données non triviales (avion, modélisation d'un tronc humain).

Ces résultats sont repris en Annexe B. Une variante de l'algorithme 2.2 est testée en Sections B.1 et B.2 et l'algorithme de Arnold, Falk et Winther en Section B.2 uniquement. Leur efficacité est comparée au préconditionneur SSOR couramment utilisé car simple de mise en oeuvre. Deux enseignements peuvent être tirés de ces simulations numériques :

- le gain en temps de calcul est important pour les deux méthodes et l'effort de programmation est faible pour l'algorithme décrit par Hiptmair. Le tableau 2.2 reporte une comparaison entre différentes méthodes de résolution pour un calcul de diffraction par un cylindre 3D ; COCG-Helmholtz désigne la méthode COCG préconditionnée par l'algorithme décrit par Hiptmair. Les temps de calcul pour cette méthode sont à peu près divisés par deux par rapport à celles utilisant un préconditionnement SSOR.
- le comportement asymptotique du nombre d'itérations de la méthode préconditionnée en fonction du nombre d'inconnues du problème est amélioré. Une estimation sous la forme CN^{α} où N est le nombre d'inconnues est faite dans l'Annexe B.2. Les valeurs de α se situent entre 0.17 et 0.32 pour les deux lisseurs et entre 0.34 et 0.42 pour le préconditionnement SSOR. Cependant, on est encore loin du cas optimal correspondant à $\alpha = 0$.

Nombre de degrés de liberté	84 385	153 293	256 121	392 524
QMR - SSOR	2 215	5 341	9 807	19 735
COCG - SSOR	1 862	4 433	8 408	17 175
BiCGCR - SSOR	2 395	5 214	10 364	22 264
COCG - Helmholtz				
Résolution	1 108	2 312	4 447	7 981
Préparation du préconditionneur	3	5	8	11

TAB. 2.2 – Temps de calcul pour la diffraction sur un cylindre 3D.

2.3.4 Cas non-structuré avec des méthodes algébriques

Les premières méthodes multiniveau algébriques pour les équations de Maxwell ont été proposées dans les travaux de Beck [8, 9] et de Reitzinger et Schöberl [66]. Leurs approches divergent essentiellement dans la définition des opérateurs de transfert (prolongement et restriction).

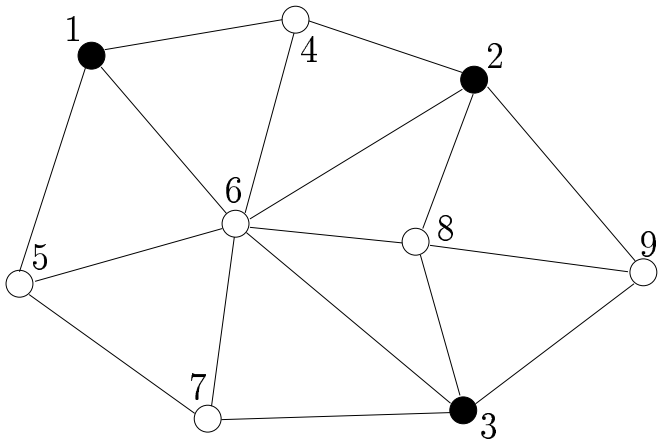
Méthode proposée par Beck

Les problèmes traités par Beck sont des problèmes en régime temporel ou harmonique mais *basses fréquences* [8] utilisant une discrétisation par éléments finis d'arête.

S'appuyant sur la décomposition de Helmholtz discrète (2.11), Beck construit un préconditionneur multiniveau en deux phases :

- l'une calcule une correction uniquement dans le sous-espace $\mathrm{grad}\,\mathcal{W}^0(\mathcal{T})$ de $\mathcal{W}^1(\mathcal{T})$ en considérant un système dont la matrice est celle en potentiel scalaire introduite dans le lisseur de Hiptmair : $A_\phi = G^t A G$ avec G la matrice d'incidence arête-noeud du maillage fin.
- l'autre calcule une correction dans l'espace d'éléments finis d'arête $\mathcal{W}^1(\mathcal{T})$.

Pour calculer la correction dans l'espace $\mathrm{grad}\,\mathcal{W}^0(\mathcal{T})$, il utilise un V-cycle multigrille sur le système de matrice A_ϕ avec des itérations de Gauss-Seidel par points pour le pré- et le postlissage. Pour la détermination des matrices de restriction, les coefficients non nuls de la matrice A_ϕ lui permettent de définir un graphe comme décrit Sous-section 2.2.3. Il utilise ce graphe pour sélectionner les noeuds maîtres mais sans tenir compte des valeurs des coefficients de la matrice. Il définit alors les matrices de restriction nodal fin vers nodal grossier de la façon suivante : la valeur des noeuds maîtres se transfère avec un poids



$$P^t = \begin{pmatrix} 1 & 0 & 0 & 1/2 & 1 & 1/3 & 0 & 0 & 0 \\ 0 & 1 & 0 & 1/2 & 0 & 1/3 & 0 & 1/2 & 1/2 \\ 0 & 0 & 1 & 0 & 0 & 1/3 & 1 & 1/2 & 1/2 \end{pmatrix}.$$

FIG. 2.4 – Représentation des noeuds maîtres et esclaves ainsi que la matrice de restriction P^t.

de 1, la valeur d'un noeud esclave s connecté avec n_s noeuds maîtres est transférée avec un poids de $1/n_s$ vers chaque noeud maître connecté avec s (fig. 2.4). Cela conduit à des matrices simples du même type que (2.7).

Pour calculer la correction dans l'espace d'éléments finis d'arête dans son ensemble, il utilise aussi un V-cycle multigrille avec des itérations de Gauss-Seidel par points pour le pré- et le postlissage mais pas directement sur le système à résoudre. En effet, il ne définit pas d'opérateurs de transfert directement entre espaces d'éléments finis d'arêtes. Il introduit une base vectorielle d'éléments finis P_1 nodaux comme base annexe et réécrit le problème à résoudre dans cette base ; le V-cycle s'effectue sur ce système annexe. Pour revenir aux variables initiales, il faut définir une matrice de transfert. Plus précisément, dans l'espace annexe, un vecteur $\mathbf{E}$ s'écrit :

$$\mathbf{E} = \sum_{i=1}^{d} \sum_{n=1}^{N^h} w_n^0 E_{n_i} \mathbf{e}_i, \tag{2.13}$$

où d est la dimension de l'espace physique ($\mathbb{R}^2$ ou $\mathbb{R}^3$), N^h le nombre de noeuds du maillage, $\mathbf{e}_i$ le $i^{\text{ème}}$ vecteur de la base canonique de $\mathbb{R}^d$ et w_n^0 désigne une fonction de base des éléments finis nodaux. La matrice de transfert Q de la base vectorielle annexe à la base d'éléments finis d'arête est définie de la manière suivante :

$$\begin{aligned} &\text{Soit } c \text{ l'arête d'indice } e, \text{ de vecteur unitaire tangent } \mathbf{t} \text{ et de longueur } l_c, \\ &\forall n \in \{1, \ldots, N^h\},\ \forall i \in \{1, \ldots, d\}, \quad Q_{en_i} = \int_c w_n^0 \mathbf{e}_i \cdot \mathbf{t} \mathrm{d}l = \frac{1}{2}(\mathbf{e}_i \cdot \mathbf{t}) l_c. \end{aligned} \tag{2.14}$$

La matrice du problème dans cette base annexe s'écrit alors $\tilde{A} = Q^t A Q$ où A est la matrice du système en éléments finis d'arête.

Procédure beckprecond(x : **vecteur**, r : **vecteur**)
[x : résidu préconditionné.]
[r : résidu non préconditionné.]
$x_\phi \leftarrow 0$, $x \leftarrow 0$;
V-cycle sur le sytème $A_\phi x_\phi = G^t r$;
$x \leftarrow x + G x_\phi$;
1 itération de Gauss-Seidel en descente sur $Ax = r$
$\tilde{x} \leftarrow 0$;
V-cycle sur le sytème $\tilde{A}\tilde{x} = Q^t(r - Ax)$;
$x \leftarrow x + Q\tilde{x}$;
1 itération de Gauss-Seidel en remontée sur $Ax = r$
$x_\phi \leftarrow 0$;
V-cycle sur le sytème $A_\phi x_\phi = G^t(r - Ax)$;
$x \leftarrow x + G x_\phi$;
Fin

Algorithme 2.3: Algorithme du préconditionnement proposé par Beck.

Le passage d'inconnues d'arête à inconnues vectorielles nodales crée un problème pour la prise en compte des conditions aux limites : naturelles en éléments finis d'arêtes, elles ne sont pas adaptées à la base de vecteurs annexes[3]. Beck propose des solutions pour résoudre ce problème ; voir [8, Section 5].

L'algorithme 2.3 résume cette méthode de préconditionnement. Cette technique donne des résultats satisfaisants et quasi-optimaux pour les problèmes considérés par Beck.

Méthode proposée par Reitzinger et Schöberl

Reitzinger et Schöberl proposent une méthode multiniveau qui est l'analogue algébrique de l'algorithme multigrille géométrique avec les lisseurs de Hiptmair ou Arnold et collaborateurs. A la différence de Beck, en plus de niveaux d'espace d'éléments finis nodaux, ils définissent des niveaux d'espace d'éléments finis d'arête. Notamment, ils proposent une technique pour déduire d'un ensemble d'arêtes fines un ensemble d'arêtes grossières.

Ils définissent des opérateur de transfert entre les niveaux fin et grossier de façon à vérifier une relation de commutativité. Cette relation peut se représenter par le diagramme commutatif suivant :

$$\begin{array}{ccc} \mathbb{C}^{N^H} & \xrightarrow{G^H} & \mathbb{C}^{E^H} \\ \downarrow{\scriptstyle \alpha} & & \downarrow{\scriptstyle \beta} \\ \mathbb{C}^{N^h} & \xrightarrow{G^h} & \mathbb{C}^{E^H} \end{array} \tag{2.15}$$

ou de manière plus concise :

$$G^h \alpha = \beta G^H. \tag{2.16}$$

avec :
- α l'opérateur de prolongement pour les éléments finis nodaux,
- β l'opérateur de prolongement pour les éléments finis d'arêtes,
- G^h et G^H les représentations discrètes du gradient dans les espaces aux niveaux fin et grossier.

Dans ce but, ils introduisent une matrice annexe B d'inconnues nodales telles que la matrice d'incidence arête-noeud du graphe associé à B soit la matrice d'incidence arête-noeud du graphe associé au maillage fin. Le choix $B = A_\phi$ matrice du problème en potentiel scalaire est une possibilité mais d'autres sont envisageables, voir [48] à ce sujet.

Reitzinger et Schöberl forment tout d'abord une partition des noeuds du maillage à partir de la matrice B en deux étapes :

[3]Les opérateurs de trace sur la frontière sont effectivement différents : composante tangentielle pour $\mathbb{H}(\mathrm{rot}, \Omega)$, toutes les composantes pour $\mathbb{H}^1(\Omega)$.

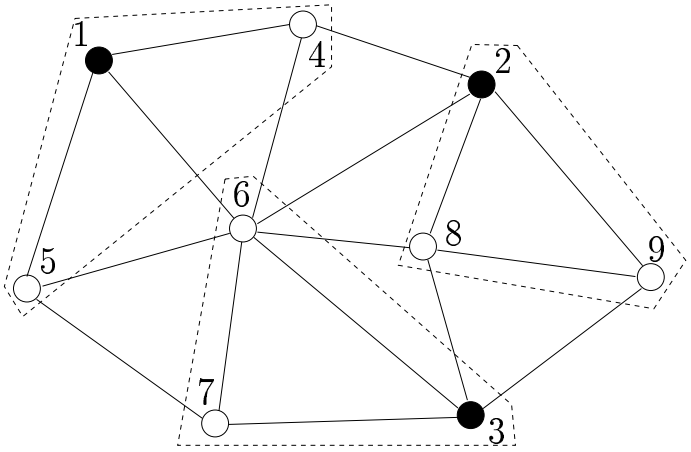
– des noeuds maîtres qui vont définir les variables grossières sont sélectionnés par des algorithmes communément utilisés pour des problèmes du type Laplacien (algorithme de Ruge et Stüben par exemple [67]),
– des aggrégats sont formés en associant de manière unique les noeuds restants à un noeud maître.

Ainsi, chaque noeud de la grille fine est associé de *manière unique* à un noeud grossier (voir fig. 2.5), ce qui permet de définir l'application ind :

$$\mathrm{ind} : \omega_h^{\mathrm{nod}} \to \omega_H^{\mathrm{nod}}.$$

où ω_h^{nod} désigne l'ensemble des indices des noeuds du maillage fin et ω_H^{nod} ceux du maillage grossier. La matrice de prolongement ne contient que des 0 et des 1 et est définie de la manière suivante :

$$\alpha_{pn} = \begin{cases} 1 & \text{si } p \in \omega_h^{\mathrm{nod}} \text{ tel que } n = \mathrm{ind}(p), \\ 0 & \text{sinon.} \end{cases} \tag{2.17}$$



$$\alpha^t = \begin{pmatrix} 1 & 0 & 0 & 1 & 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 1 & 1 \\ 0 & 0 & 1 & 0 & 0 & 1 & 1 & 0 & 0 \end{pmatrix}.$$

FIG. 2.5 – Représentation des noeuds maîtres et esclaves ainsi que la matrice de restriction α^t.

Reitzinger et Schöberl définissent alors un graphe grossier en s'imposant la règle suivante : une arête grossière relie l'aggrégat n à l'aggrégat m si au moins une arête fine relie une variable fine de l'aggrégat n à une variable fine de l'aggrégat m.

Soit ω_h^{edg} l'ensemble des indices des arêtes du maillage fin et ω_H^{edg} ceux des arêtes du maillage grossier. Soit i l'indice de l'arête fine liant le noeud p au noeud q (noté $i = \overline{pq}^h \in \omega_h^{\mathrm{edg}}$) et e l'indice de l'arête grossière liant le noeud n au noeud m (noté $e = \overline{nm}^H \in \omega_H^{\mathrm{edg}}$), la matrice de prolongement β est définie de la manière suivante :

$$\beta_{ie} = \begin{cases} 1 & \text{si } e = \overline{\mathrm{ind}(p)\,\mathrm{ind}(q)}^H, \\ -1 & \text{si } e = \overline{\mathrm{ind}(q)\,\mathrm{ind}(p)}^H, \\ 0 & \text{sinon.} \end{cases} \tag{2.18}$$

Les matrices de prolongement α et β définit par (2.17) et (2.18) vérifient la relation de commutativité (2.16).

Reitzinger et Schöberl appliquent leur méthode plutôt à des problèmes de *magnétostatique* mais on trouve aussi quelques applications en régime harmonique basse-fréquence [48]. Son utilisation s'est depuis répandue dans la communauté du calcul en électromagnétisme [53, 55, 54, 56, 81, 80].

Remarque 2.4. *Concernant la relation de commutativité* (2.15), *une analogie en terme de réseaux électriques peut être faite comme cela a été décrit au Chapitre 1. Les graphes orientés aux niveaux fin et grossier peuvent être considérés respectivement comme un réseau fin et une représentation de ce réseau à un niveau grossier. Si l'on connaît alors une distribution de potentiels U^H aux noeuds du réseau grossier, il y a deux moyens de connaître les forces électromotrices dans chacune des branches du réseau fin :*

- *première possibilité : on déduit de U^H le potentiel résultant aux noeuds du réseau fin u^h par l'opération αU^H et on calcule alors la distribution des forces électromotrices dans les branches du réseau fin par l'opération $G^h u^h$.*
- *seconde possibilité : on calcule d'abord les forces électromotrices dans les branches du réseau grossier E^H par l'opération $G^H U^H$ et on en déduit la distribution dans les branches du réseau fin par l'opération βE^H.*

La relation (2.16) *signifie que, quelles que soient les potentiels aux noeuds du réseau grossier, les deux calculs des forces électromotrices dans le réseau fin doivent coïncider.*

Méthode proposée par Bochev et ses collaborateurs

La vérification de la relation de commutativité (2.15) n'est pas suffisante pour assurer l'efficacité optimale de la méthode multiniveau de Reitzinger et Schöberl. C'est pourquoi Bochev et ses collaborateurs [13] ont proposé des modifications dans la définition des opérateurs de transfert.

Bochev et ses collaborateurs ont considéré uniquement les cas où la matrice du système à résoudre est symétrique définie positive [13, 43].

Leur méthode reprend celle de Reitzinger et Schöberl en modifiant la définition des matrices de prolongement α et β. Les améliorations proposées s'effectuent dans deux directions.

La première concerne la matrice de prolongement β :

- une première matrice simple $\hat{\beta}$ vérifiant la relation de commutativité (2.15) est définie : celle de Reitzinger et Schöberl, par exemple,
- Dans le but de minimiser l'énergie de la base des fonctions d'arête grossière, cette matrice est "lissée" en utilisant une itération de Jacobi avec relaxation :

$$\beta = (I - \eta D^{-1} S_\nu)\hat{\beta}, \tag{2.19}$$

où I est la matrice identité, η un coefficient de relaxation, S_ν désigne la partie "$\mathrm{rot}\,\nu\,\mathrm{rot}$" de la matrice du système à résoudre, et D est la matrice diagonale contenant les coefficients diagonaux de S_ν.

La condition de commutativité (2.15) reste vérifiée après le lissage. Cependant la matrice β est moins creuse ce qui augmente la complexité de l'algorithme.

La seconde amélioration concerne le respect de la relation de commutativité pour d'autres matrices de prolongement α que celle proposée par Reitzinger et Schöberl. Une méthode par moindres carrés est utilisée dans [13].

2.4 Conclusion

Quelques notions importantes pour les méthodes itératives de résolution des systèmes linéaires ont été décrites. Après une présentation du principe des méthodes multiniveau, des méthodes spécifiques pour les équations de Maxwell discrétisées par les éléments finis d'arêtes ont été présentées.

Dans le cas des méthodes multiniveau algébrique, la relation de commutativité (2.15) mise en exergue par Reitzinger et Schöberl dans leur méthode joue un rôle important. Son respect conduit à des méthodes efficaces et elle permet de conserver au niveau grossier certaines propriétés des espaces d'éléments finis de Whitney décrit dans le Chapitre 1. Cette relation de commutativité est le point de départ de la méthode multiniveau algébrique que nous proposons dans le chapitre suivant associée à une technique de minimisation d'énergie de la base d'éléments finis.

Chapitre 3

Base d'approximation grossière par minimisation d'énergie et contraintes

3.1 Introduction

Afin de construire une méthode multigrille géométrique robuste pour des problèmes avec des opérateurs du type $\mathrm{div}(\sigma\,\mathrm{grad})$ discrétisés par éléments finis P_1 nodaux, Wan, Chan et Smith proposent dans [79] de définir des bases de fonctions sur chaque maillage, en résolvant un problème de minimisation d'énergie avec contraintes.

Pour construire algébriquement une famille de base d'éléments finis d'arête dans la méthode algébrique que nous proposons dans la suite, nous en avons repris en les adaptant les trois idées suivantes :

- *minimisation d'énergie* : dans le cas où la forme bilinéaire du problème permet de définir une norme d'énergie associée, les fonctions de base au niveau grossier doivent avoir l'énergie la plus faible possible ; cette idée proposée initialement dans [76] et reprise dans [51, 79, 84] découle des estimations en norme d'énergie de la convergence données dans [16]. On trouve des justifications dans [79].
- *recouvrement limité* des supports des fonctions de bases : le caractère creux des matrices de prolongement et des matrices aux niveaux grossiers est lié à ce recouvrement ; dans le cas où il n'est pas contrôlé, cela accroît la complexité algorithmique de la méthode.
- *préservation du sous-espace des fonctions d'énergie quasi-nulle* : les fonctions d'énergie quasi-nulle (par exemple les fonctions constantes dans le cas du Laplacien) doivent être conservées au niveau grossier car elles ne sont pas traitées par les lisseurs. Pour les éléments finis d'arête et l'opérateur $\mathrm{rot}\,\mathrm{rot}$, nous l'exprimons sous la forme d'une condition de compatibilité des opérateurs de transfert, plus précisément l'égalité (2.15) mise en évidence par Reitzinger et Schöberl dans leur méthode.

Les principes de la minimisation d'énergie pour la construction de la base d'approximation grossière à partir de la base d'approximation fine en éléments finis nodaux sont donnés dans l'Annexe C, ainsi que des simulations numériques sur le Laplacien et l'opérateur de Helmholtz. L'Annexe D décrit en détail la méthode que nous proposons pour les éléments finis d'arête. La méthode est testée sur des géométries simples : rectangle en dimension 2 et cube en dimension 3, pour des problèmes appartenant à la classe définie par la formulation (1.33) dans le cas où δ est nul et ν et γ sont strictement positifs.

Dans ce chapitre, on se propose essentiellement d'introduire, pour les éléments finis nodaux puis d'arête, le formalisme mathématique qui permet de montrer comment, à l'aide de multiplicateurs de Lagrange, les problèmes de minimisation d'énergie avec contraintes se ramènent à la résolution de systèmes linéaires.

On donne aussi, dans le cas des éléments finis d'arête, un résultat qui assure une solution unique au problème de minimisation sous la forme d'une condition sur un graphe d'incidence arête-noeud introduit au niveau grossier.

3.2 Principe et méthode dans le cas des éléments finis nodaux

Plaçons nous dans le cas d'un problème avec un opérateur de Laplace discrétisé par une méthode d'éléments finis P_1 nodaux, sur un domaine Ω de $\mathbb{R}^2$ ou de $\mathbb{R}^3$. On rencontre ce type de problème en électrostatique ou en magnétostatique dans l'air. Sur le maillage initial, la base d'éléments finis nodaux est notée $(w_p^{0,h})_{p=1,\dots,N^h}$.

On cherche à définir une base grossière $(w_n^{0,H})_{n=1,\dots,N^H}$ en vue de définir une méthode multiniveau.

3.2.1 Formulation du problème d'optimisation

On découpe le domaine Ω en sous-domaines $(\Omega_n^H)_{n=1,\dots,N^H}$ se recouvrant de façon que chaque sous-domaine ne soit pas totalement recouvert par ses voisins.

On cherche alors la base grossière comme solution du problème suivant, où sont formalisées les trois idées énoncées dans l'introduction :

$$\begin{cases} \text{Trouver } (w_n^{0,H})_{n=1,\dots,N^H} \text{ minimisant } \displaystyle\sum_{n=1}^{N^H} \int_\Omega \operatorname{grad} w_n^{0,H} \cdot \operatorname{grad} w_n^{0,H} \text{ sous les contraintes :} \\ \displaystyle\sum_{m=1}^{N^H} w_m^{0,H}(x) = 1,\ \forall x \in \overline{\Omega} \text{ et } \operatorname{supp}(w_n^{0,H}) \subset \overline{\Omega_n^H},\ \forall n \in \{1,\dots,N^H\}. \end{cases} \tag{3.1}$$

Pour formaliser algébriquement le problème (3.1), on cherche plutôt les fonctions de la base grossière comme combinaisons linéraires des fonctions de la base fine :

$$\forall n \in \{1,\dots,N^H\},\ w_n^{0,H} = \sum_{p=1}^{N^h} \alpha_{pn} w_p^{0,h}. \tag{3.2}$$

Mathématiquement, cela signifie que l'espace grossier engendré est inclus dans l'espace d'éléments finis initial.

Pour décrire la contrainte sur les supports, on introduit les ensembles d'indices L_n des noeuds du maillage initial appartenant au sous-domaine Ω_n^H. Pour assurer que le support de $w_n^{0,H}$ soit inclus dans Ω_n^H, on impose :

$$p \notin L_n \implies \alpha_{pn} = 0. \tag{3.3}$$

On définit alors le vecteur α_n constitué des composantes "non nulles" α_{pn}, c.-à-d. celles pour p dans L_n.

Il est commode d'introduire un opérateur de projection P_n associé à $L_n = \{p_1,\dots,p_{|L_n|}\}$ défini de la manière suivante :

$$\begin{aligned} &P_n : \mathbb{R}^{N^h} \to \mathbb{R}^{|L_n|}, \text{ tel que :} \\ &\forall x \in \mathbb{R}^{N^h},\ (P_n x)_k = x_{p_k},\ \forall k \in \{1,\dots,|L_n|\}. \end{aligned} \tag{3.4}$$

Seules les composantes indexées par L_n sont conservées par cette projection. Il est alors facile de voir que l'opérateur transposé s'écrit :

$$\begin{aligned} &P_n^t : \mathbb{R}^{|L_n|} \to \mathbb{R}^{N^h}, \text{ tel que :} \\ &\forall y \in \mathbb{R}^{|L_n|},\ (P_n^t y)_i = 0 \text{ si } i \notin L_n \text{ et } y_k \text{ si } i = p_k. \end{aligned} \tag{3.5}$$

Compte tenu que la base initiale vérifie $\forall x \in \overline{\Omega}, \displaystyle\sum_{p=1}^{N^h} w_p^{0,h}(x) = 1$, la condition $\displaystyle\sum_{n=1}^{N^H} w_n^{0,H}(x) = 1$ s'écrit :

$$\forall p \in \{1,\dots,N^h\},\ \sum_{n=1}^{N^H} \alpha_{pn} = 1, \tag{3.6}$$

soit $\displaystyle\sum_{n=1}^{N^H} P_n^t \alpha_n = \mathbf{1}_{N^h \times 1}$, où $\mathbf{1}_{N^h \times 1}$ désigne le vecteur à N^h composantes toutes égales à 1.

Le problème (3.1) s'écrit finalement sous la forme matricielle suivante :

$$\begin{cases} \text{Trouver } (\alpha_n)_{n=1,\dots,N^H} \text{ minimisant } \displaystyle\sum_{n=1}^{N^H} \alpha_n^t P_n K P_n^t \alpha_n \text{ sous la contrainte :} \\ \displaystyle\sum_{n=1}^{N^H} P_n^t \alpha_n = \mathbf{1}_{N^h \times 1}. \end{cases} \tag{3.7}$$

K est la matrice du problème discrétisé sur le maillage initial définie par $K_{pq} = \int_\Omega \operatorname{grad} w_p^{0,h} \cdot \operatorname{grad} w_q^{0,h}$.

3.2.2 Résolution avec des multiplicateurs de Lagrange

Ce problème de minimisation sous contraintes peut être écrit comme un problème de point-selle en introduisant un vecteur de multiplicateurs de Lagrange μ appartenant à $\mathbb{R}^{N^h}$, pour prendre en compte les N^h contraintes égalités.

A partir des vecteurs α_n, on définit par blocs le vecteur α de la façon suivante :

$$\alpha = \begin{pmatrix} \alpha_1 \\ \vdots \\ \alpha_{N^H} \end{pmatrix}. \tag{3.8}$$

On pose $K_n = P_n K P_n^t$, matrice extraite de K en conservant les lignes et les colonnes d'indice dans L_n. C'est aussi la matrice du problème discrétisé sur le sous-domaine Ω_n^H.

Le problème de minimisation sous contraintes (3.7) peut se mettre sous la forme :

$$\begin{cases} \text{Trouver le point-selle } (\alpha_c, \mu_c) \text{ du Lagrangien } L \text{ défini par :} \\ L(\alpha, \mu) = \dfrac{1}{2} \displaystyle\sum_{n=1}^{N^H} \alpha_n^t K_n \alpha_n + \mu^t \left(\sum_{n=1}^{N^H} P_n^t \alpha_n - \mathbf{1}_{N^h \times 1} \right). \end{cases} \tag{3.9}$$

Le point critique de L doit ainsi vérifier les équations suivantes :

$$\begin{cases} K_n \alpha_n + P_n \mu_c = 0, \ \forall n \in \{1, \dots, N^H\}, \\ \displaystyle\sum_{n=1}^{N^H} P_n^t \alpha_n - \mathbf{1}_{N^h \times 1} = 0. \end{cases} \tag{3.10}$$

Le vecteur μ_c des multiplicateurs de Lagrange est donc solution du système :

$$\left(\sum_{n=1}^{N^H} P_n^t K_n^{-1} P_n \right) \mu_c = -\mathbf{1}_{N^h \times 1}.$$

En pratique, on résout donc le système d'équations (3.10) de la manière suivante :

- d'abord on détermine le vecteur μ_c des multiplicateurs de Lagrange avec une méthode itérative sur le système :

$$C\mu = -\mathbf{1}_{N^h \times 1}, \quad \text{où } C = \sum_{n=1}^{N^H} P_n^t K_n^{-1} P_n.$$

- ensuite on remonte à la valeur de α en résolvant les N^H systèmes :

$$K_n \alpha_n = -P_n \mu_c.$$

La matrice C est *symétrique définie positive* car tous les K_n le sont. Pour déterminer μ_c, on peut donc employer la méthode du gradient conjugué qui a l'avantage de ne pas demander la construction de la matrice C, construction trop coûteuse en temps et en espace mémoire. Seuls des produits matrice-vecteur de la forme $(P_n^t K_n^{-1} P_n)\mu$ sont nécessaires. La factorisation de toutes les matrices K_n est donc requise pour minimiser les coûts. Elle est aussi utile pour le calcul des α_n à partir du calcul des multiplicateurs. Le produit matrice-vecteur $C\mu$ est détaillé dans l'algorithme 3.1.

Pour n **de** 1 **à** N^H **faire**
  $b_n \leftarrow P_n \mu$
Fin Pour
Pour n **de** 1 **à** N^H **faire**
  résoudre à l'aide des factorisations $K_n x_n = b_n$
Fin Pour
Calculer finalement $\sum_{n=1}^{N^H} P_n^t x_n$

Algorithme 3.1: Multiplication matrice-vecteur $C\mu$.

Les multiplications par P_n et P_n^t sont peu coûteuses car ces matrices ne contiennent que $|L_n|$ coefficients non-nuls.

Remarque 3.1. *Dans la cas de problèmes dont l'opérateur contient des coefficients fortement non homogènes, le conditionnement de la matrice C est altéré. Il est alors nécessaire de préconditionner la méthode du gradient conjugué. La matrice $\sum_{n=1}^{N^H} P_n^t K_n P_n$ est proposée comme matrice de préconditionnement dans [79, 84].*

Remarque 3.2. *Dans [51], le même problème de minimisation d'énergie sous contraintes est utilisé pour calculer une base d'approximation grossière. La résolution n'est pas alors réalisée par l'utilisation d'un Lagrangien mais par l'utilisation d'une méthode de gradient projeté. Les auteurs montrent que le premier pas de ce nouvel algorithme correspond à la méthode multiniveau algébrique par aggrégation et lissage (*smoothing aggregation*) qu'ils avaient précédemment développée [76].*

D'autres techniques telles que la méthode AMGe [20, 47] permettent d'obtenir des solutions quasi-optimales pour le problème de minimisation (3.7).

3.3 Extension directe au cas des éléments finis d'arête

Notre objectif est d'introduire un problème similaire à (3.1) pour le cas des éléments finis d'arête. Les principes de *minimisation d'énergie* et de *restriction des supports* des fonctions d'approximation grossière se transposent facilement du cas nodal.

La difficulté est la prise en compte du sous-espace des fonctions d'énergie quasi-nulle. Dans le cas de l'espace d'éléments finis d'arête, ce sous-espace correspond à celui engendré par les gradients des fonctions de base d'éléments finis nodaux, dont la dimension est proportionnelle aux nombre d'arêtes (avec un facteur 1/6 en dimension 3). Donc à la différence du Laplacien avec les constantes, on ne peut pas exiger une *préservation du sous-espace d'énergie quasi-nulle* au niveau grossier sauf à ne pas réduire significativement le nombre d'inconnues du niveau grossier par rapport au nombre d'inconnues du niveau fin.

On souhaite utiliser les lisseurs de Hiptmair et de Arnold, Falk et Winther car ces techniques ont montré leur validité pour le multigrille géométrique. Leur caractéristique particulière est de lisser la composante de l'erreur dans le noyau de l'opérateur rotationnel, c.-à-d. dans l'espace d'énergie quasi-nulle. Pour que l'étape de correction sur la grille grossière soit efficace, il faut alors que les fonctions lisses du noyau du rotationnel soient bien approchées dans l'espace d'arête grossier. Dans le cas de la méthode multigrille géométrique, ceci est automatiquement vérifié grâce aux propriétés des espaces de Whitney pour lesquels toute fonction définie sur la grille fine et de rotationnel nul peut être bien approchée par

une fonction définie sur la grille grossière et de rotationnel nul. Dans le cas d'une méthode algébrique, il faudrait donc disposer d'une bonne représentation du noyau de l'opérateur rotationnel à tous les niveaux, en s'assurant cependant que la dimension du noyau décroît proportionnellement au nombre d'inconnues sur chaque niveau.

Reitzinger et Schöberl ont proposé une solution dans [66]. Ils contruisent des opérateurs de prolongement α pour les éléments finis nodaux et β pour les éléments finis d'arête qui vérifient la relation de commutativité (2.16) $\beta G^H = G^h \alpha$ où G^h et G^H sont les représentations discrètes du gradient dans les espaces aux niveaux fin et grossier. Cette relation assure que le gradient de toute fonction nodale grossière appartient à l'espace d'arête grossier et donc que le noyau du rotationnel reste bien représenté dans l'espace d'arête grossier. Les opérateurs α et β qu'ils proposent permettent en plus de conserver une analogie avec le cas géométrique : le noyau du rotationnel est exactement l'image du gradient à tous les niveaux.

Leur méthode est cependant liée au choix d'un type de matrice α particulier. Nous avons décidé de retenir la relation de commutativite (2.16) comme point de départ pour nos développements.

3.3.1 Étapes préliminaires et formulation du problème d'optimisation

On dispose d'une base nodale fine $(w_p^{0,h})_{p=1,\dots,N^h}$, d'une base d'arête fine $(w_i^{1,h})_{i=1,\dots,E^h}$ ainsi que d'une matrice d'incidence arête-noeud G^h de dimension $E^h \times N^h$.

On cherche une base nodale grossière $(w_n^{0,H})_{n=1,\dots,N^H}$ et une base d'arête grossière $(w_e^{1,H})_{e=1,\dots,E^H}$ respectivement sous les formes suivantes :

$$w_n^{0,H} = \sum_{p=1}^{N^h} \alpha_{pn} w_p^{0,h}, \ \forall n \in \{1,\dots,N^H\}, \tag{3.11a}$$

$$w_e^{1,H} = \sum_{i=1}^{E^h} \beta_{ie} w_i^{1,h}, \ \forall e \in \{1,\dots,E^H\}. \tag{3.11b}$$

Comme dans la méthode multigrille géométrique, ces coefficients fourniront les matrices de prolongement nodale α et d'arête β.

Tout d'abord, les fonctions nodales grossières $(w_n^{0,H})_{n=1,\dots,N^H}$ sont calculées comme solution du problème de minimisation sous contraintes (3.1) défini précédemment. Les coefficients de α vérifient notamment :

$$\forall p \in \{1,\dots,N^h\}, \ \sum_{n=1}^{N^H} \alpha_{pn} = 1, \tag{3.12}$$

que l'on peut qualifier de propriété de partition de l'unité et :

$$\forall n \in \{1,\dots,N^H\}, \ p \notin L_n \implies \alpha_{pn} = 0. \tag{3.13}$$

Les L_n sont des ensembles de noeuds fins que l'on introduit pour limiter le support des fonctions nodales grossières $w_n^{0,H}$ dans un domaine Ω_n^H.

Pour calculer ensuite β, on définit un graphe grossier représenté par sa matrice d'incidence arête-noeud G^H de taille $E^H \times N^H$.

On limite alors les supports des fonctions d'arête grossières $w_e^{1,H}$ en introduisant des ensembles d'indices d'arête $(I_e)_{e=1,\dots,E^H}$ de manière à assurer que le support des $w_e^{1,H}$ soit inclus dans $\Omega_n^H \cap \Omega_m^H$ avec n et m les extrémités de l'arête grossière e. On impose ainsi :

$$\forall e \in \{1,\dots,E^H\}, \ i \notin I_e \implies \beta_{ie} = 0. \tag{3.14}$$

On définit alors le vecteur β_e constitué des composantes "non nulles" β_{ie}, c.-à-d. celles pour i dans I_e.

Pour e dans $\{1,\dots,E^H\}$ avec $I_e = \{i_1,\dots,i_{|I_e|}\}$, on définit l'opérateur de projection qui conserve uniquement les composantes indexées par I_e :

$$\begin{aligned} &Q_e : \mathbb{R}^{E^h} \to \mathbb{R}^{|I_e|}, \text{ tel que :} \\ &\forall x \in \mathbb{R}^{E^h}, \ (Q_e x)_k = x_{i_k}, \ \forall k \in \{1,\dots,|I_e|\}. \end{aligned} \tag{3.15}$$

Comme annoncé précédemment, β doit vérifier la contrainte $\beta G^H = G^h \alpha$ qui s'écrit aussi :

$$\sum_{e=1}^{E^H} \beta_{ie} G^H_{en} = \sum_{p=1}^{N^h} G^h_{ip} \alpha_{pn}, \ \forall i \in \{1, \dots, E^h\}, \ \forall n \in \{1, \dots, N^H\}. \tag{3.16}$$

soit avec les opérateurs Q^t_e :

$$\sum_{e=1}^{E^H} G^H_{en} Q^t_e \beta_e = G^h \alpha_{\bullet n}, \ \forall n \in \{1, \dots, N^H\}. \tag{3.17}$$

La notation $\alpha_{\bullet n}$ désigne la $n^{\text{ième}}$ colonne de la matrice α.

Pour définir un critère de minimisation sur l'espace d'éléments finis d'arête, il nous faut introduire une norme associée au problème à résoudre. Pour la classe des problèmes qui nous intéressent définie par (1.33), la forme bilinéaire associée définie en (1.37) est un produit scalaire dans le cas où ν et γ sont strictement positifs et on peut chercher à minimiser :

$$\sum_{e=1}^{E^H} a(w^{1,H}_e, w^{1,H}_e), \tag{3.18}$$

soit de façon équivalente :

$$\sum_{e=1}^{E^H} \beta^t_e Q_e A Q^t_e \beta_e, \tag{3.19}$$

où A est la matrice du problème au niveau fin avec $A_{ij} = a(w^{1,h}_j, w^{1,h}_i)$. De façon générale, on cherchera à minimiser :

$$\sum_{e=1}^{E^H} \beta^t_e A_e \beta_e, \tag{3.20}$$

où les matrices A_e sont des matrices symétriques semi-définies positives de taille $|I_e| \times |I_e|$.

Le problème à résoudre s'écrit alors sous la forme :

$$\begin{cases} \text{Trouver } (\beta_e)_{e=1,\dots,E^H} \text{ minimisant } \displaystyle\sum_{e=1}^{E^H} \beta^t_e A_e \beta_e \text{ sous les contraintes :} \\ \displaystyle\sum_{e=1}^{E^H} G^H_{en} Q^t_e \beta_e = G^h \alpha_{\bullet n}, \ \forall n \in \{1, \dots, N^H\}. \end{cases} \tag{3.21}$$

C'est un problème qui a donc $\sum_{e=1}^{E^H} |I_e|$ inconnues et $E^h N^H$ contraintes. On note cependant qu'un certain nombre de contraintes sont implicitement vérifiées. Par exemple, chaque fois qu'il existe une arête fine d'indice i n'appartenant à aucun support d'une fonction d'arête grossière, plus précisément s'il n'existe pas de e tel que $i \in I_e$; l'égalité

$$\sum_{e=1}^{E^H} \beta_{ie} G^H_{en} = \sum_{p=1}^{N^h} G^h_{ip} \alpha_{pn}, \ \forall n \in \{1, \dots, N^H\}.$$

se réduit alors à $0 = 0$.

On est donc amené à réduire le nombre de contraintes à imposer en introduisant des ensembles $(J_n)_{n=1,\dots,N^H}$ qui contiennent les indices des équations (3.17) que l'on veut imposer. Les J_n pouvant être vides, nous introduisons l'ensemble :

$$F = \{n \in \{1, \dots, N^H\}, \ J_n \neq \emptyset\}. \tag{3.22}$$

De manière analogue à Q_e, on définit pour n dans F un opérateur de projection associé à J_n :

$$R_n : \mathbb{R}^{E^h} \to \mathbb{R}^{|J_n|}. \tag{3.23}$$

Le problème à résoudre possède moins de contraintes que le problème initial et prend la forme suivante :

$$\begin{cases} \text{Trouver } (\beta_e)_{e=1,\dots,E^H} \text{ minimisant } \displaystyle\sum_{e=1}^{E^H} \beta_e^t A_e \beta_e \text{ sous les contraintes :} \\ R_n \left(\displaystyle\sum_{e=1}^{E^H} G_{en}^H Q_e^t \beta_e \right) = R_n G^h \alpha_{\bullet n}, \ \forall n \in F. \end{cases} \tag{3.24}$$

La quantité $R_n G^h \alpha_{\bullet n}$ ayant été préalablement calculée, on la note ξ_n par souci de simplicité.

Remarque 3.3. *Une minimisation qui couple à la fois les problèmes nodal et d'arête a initialement été envisagée mais les résultats obtenus nous ont conduit à préférer une méthode découplée car aussi performante et moins coûteuse ; voir [60, Sous-section 4.1].*

3.3.2 Résolution avec multiplicateurs de Lagrange

Description de la méthode

On définit quelques notations pour pouvoir décrire plus simplement la méthode de résolution : les vecteurs ξ, β, ρ et les matrices D et T. Les composantes du vecteur ρ sont les multiplicateurs de Lagrange relatifs aux contraintes dans (3.24).

Les vecteurs ξ et ρ appartiennent à $\mathbb{R}^M$ avec $M = \sum_{n\in F} |J_n|$. Si l'on suppose que $F = \{i_1, \dots, i_{|F|}\}$, ξ et ρ peuvent être décrits par blocs :

$$\xi = \begin{pmatrix} \xi_{i_1} \\ \vdots \\ \xi_{i_{|F|}} \end{pmatrix}, \quad \rho = \begin{pmatrix} \rho_{i_1} \\ \vdots \\ \rho_{i_{|F|}} \end{pmatrix} \text{ avec } \xi_{i_n}, \rho_{i_n} \in \mathbb{R}^{|J_n|}. \tag{3.25}$$

De manière analogue, le vecteur β appartient à $\mathbb{R}^{\tilde{M}}$, avec $\tilde{M} = \sum_{e=1}^{E^H} |I_e|$, et peut être défini par blocs. La matrice D est diagonale par blocs et ces blocs sont les matrices A_e, $e \in \{1 \dots E^H\}$. Enfin, la matrice T envoie un vecteur de $\mathbb{R}^M$ vers $\mathbb{R}^{\tilde{M}}$ de la manière suivante :

$$T : \rho \mapsto \begin{pmatrix} Q_1 \left(\displaystyle\sum_{n\in F} G_{1n}^H R_n^t \rho_n \right) \\ \vdots \\ Q_{E^H} \left(\displaystyle\sum_{n\in F} G_{E^H n}^H R_n^t \rho_n \right) \end{pmatrix}. \tag{3.26}$$

Résoudre le problème d'optimisation (3.24) devient alors équivalent à la résolution de :

$$\begin{cases} \text{Trouver le point critique } (\beta_c, \rho_c) \in \mathbb{R}^{\tilde{M}} \times \mathbb{R}^M \text{ du Lagrangien } \mathcal{L} \text{ défini par :} \\ \mathcal{L}(\beta, \rho) = \dfrac{1}{2} \beta^t D \beta + \rho^t (\xi - T^t \beta). \end{cases} \tag{3.27}$$

Le point critique vérifie alors le système linéaire suivant :

$$\begin{pmatrix} D & -T \\ T^t & 0 \end{pmatrix} \begin{pmatrix} \beta \\ \rho \end{pmatrix} = \begin{pmatrix} 0 \\ \xi \end{pmatrix} \tag{3.28}$$

Ce système linéaire peut être résolu en deux étapes :

- tout d'abord, le vecteur ρ_c des multiplicateurs de Lagrange est calculé par une méthode itérative sur le système :

$$T^t D^{-1} T \rho = \xi. \tag{3.29}$$

  Les matrices A_e étant par hypothèse symétriques définies positives, la matrice $T^t D^{-1} T$ dans (3.29) est au moins symétrique semi-définie positive.
- on peut alors revenir au calcul de β_c en résolvant :

$$D\beta = T\rho_c. \tag{3.30}$$

Comme dans le cas nodal, la matrice $T^t D^{-1} T$ n'est pas assemblée et l'on peut mener de manière efficace la multiplication matrice-vecteur ; des détails sont donnés en Annexe D Sous-section D.5.1.

Propriétés du problème

Dans l'Annexe E, on montre que si le graphe grossier construit à partir des noeuds grossiers possède suffisament d'arêtes, l'application T^t est surjective dans le cas où l'on impose toutes les contraintes, c.-à-d. dans le cas où $J_n = \{1, \ldots, E^h\}$ pour tout n. Ceci implique que le problème (3.21) admet une unique solution. D'autre part avec les mêmes conditions sur le graphe grossier, on montre dans l'Annexe D Corollaire D.2 qu'il est possible de choisir les ensembles $(J_n)_{n=1,\ldots,N^H}$ de manière à ce que le problème (3.24) ait une solution unique et soit équivalent au problème (3.21). La matrice $T^t D^{-1} T$ est alors symétrique définie positive et l'on peut résoudre le système par la méthode du gradient conjugué.

On propose dans les Sous-section D.4.3 et D.4.4 une méthode de décomposition du domaine Ω en sous-domaines Ω_n^H et la définition d'un graphe grossier associé qui vérifie les conditions du corollaire.

On se propose ici de rappeler uniquement l'énoncé du Corollaire D.2 et de l'illustrer sur un exemple simple. Dans ce but, on introduit d'abord quelques notations :

- des ensembles $\tilde{C}_i$ définis par

$$\forall i \in \{1, \ldots, E^h\},\ \tilde{C}_i = \left\{ n \in \{1, \ldots, N^H\} \ / \ i = \overline{pq}^h \text{ avec } p \in L_n \text{ ou } q \in L_n \right\}; \tag{3.31}$$

  $\tilde{C}_i$ est l'ensemble des indices de noeuds grossiers tels que le support de $w_n^{0,H}$ contienne une extrémité de l'arête fine i.
- des ensembles $\tilde{I}_i$ définis par

$$\forall i \in \{1, \ldots, E^h\},\ \tilde{I}_i = \left\{ e \in \{1, \ldots, E^H\} \ / \ i \in I_e \right\}; \tag{3.32}$$

  $\tilde{I}_i$ est l'ensemble des indices d'arêtes grossières telles que le support de $w_e^{1,H}$ contienne l'arête fine i.
- des ensembles $\tilde{J}_i$ définis par

$$\forall i \in \{1, \ldots, E^h\},\ \tilde{J}_i = \left\{ n \in \{1, \ldots, N^H\} \ / \ i \in J_n \right\}; \tag{3.33}$$

  $\tilde{J}_i$ est l'ensemble des indices de noeuds grossiers pour lesquels on impose la contrainte donnée par (3.16) pour un indice d'arête fine i fixé.

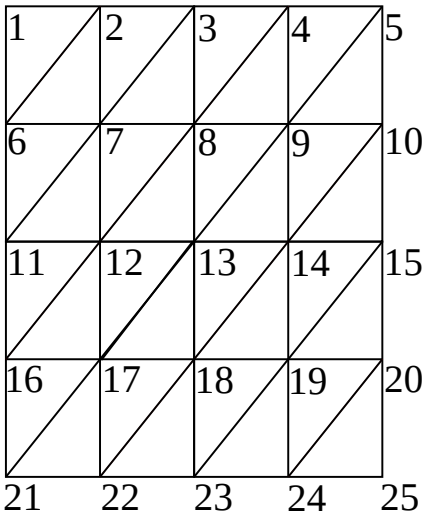
Le corollaire s'écrit en terme de sous-graphes locaux de la façon suivante :

> *Pour $\tilde{J}_i = \tilde{C}_i \setminus \{m\}$, $i = 1, \ldots, E^h$, T est injective si et seulement si pour tout i dans $\{1, \ldots, E^h\}$ le sous-graphe dont les sommets sont indexés par $\tilde{C}_i$ et les arêtes par $\tilde{I}_i$ est connexe.*

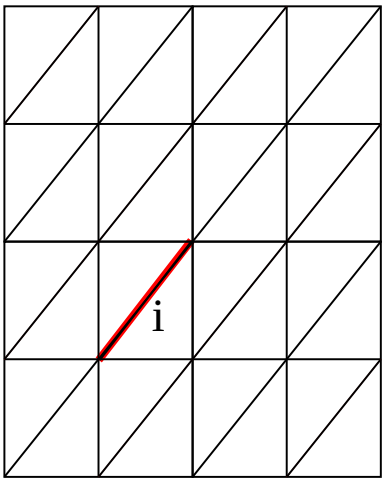
L'égalité $\tilde{J}_i = \tilde{C}_i \setminus \{m\}$ s'interprète de la façon suivante : pour une arête fine d'indice fixé i, on n'impose pas les contraintes dont le numéro grossier correspond à un sous-domaine Ω_n^H ne contenant pas l'arête. En outre, parmi les contraintes restantes, on peut en supprimer une de façon arbitraire. En particulier, si l'arête fine i n'appartient qu'à un seul sous-domaine Ω_n^H alors aucune contrainte associée à cette arête n'est imposée.

Exemple On considère le maillage de la figure 3.1(a) où les noeuds sont numérotés. Le domaine Ω peut s'écrire comme la réunion des sous-domaines $(\Omega_n^H)_{n=1,\dots,4}$ représentés à la figure 3.2 ; ils servent à limiter les supports de quatre fonctions nodales qui définissent un niveau grossier. Les ensembles L_n donnant les indices des éléments non nuls dans la $n^{\text{ième}}$ colonne de la matrice de prolongement nodal α s'écrivent alors :

$$\begin{aligned} &L_1 = \{1,2,3,6,7,8,11\}; \quad L_2 = \{3,4,5,8,9,10,13,14,15\}; \\ &L_3 = \{11,12,13,16,17,18,21,22,23\}; \quad L_4 = \{14,15,19,20,23,24,25\}. \end{aligned} \tag{3.34}$$

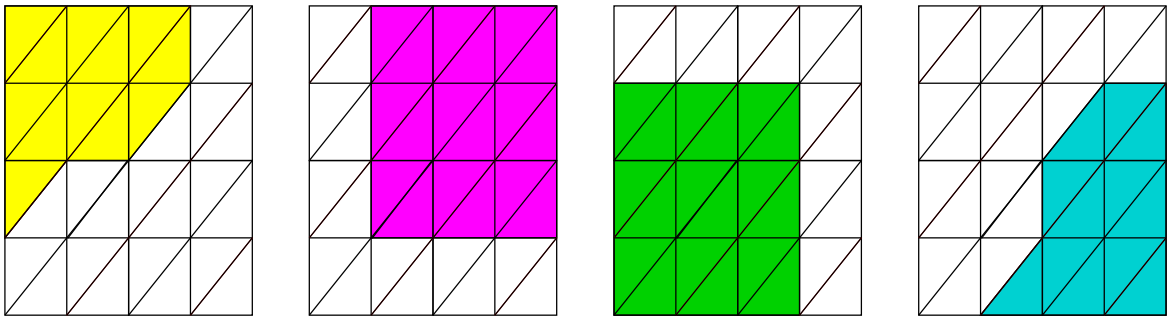


(a) Maillage fin et numérotation des noeuds



(b) Arête fine i considérée

Fig. 3.1 – Le maillage fin et sa numérotation et une arête fine considérée.



(a) Ω_1^H (b) Ω_2^H (c) Ω_3^H (d) Ω_4^H

Fig. 3.2 – Les sous-domaines $(\Omega_n^H)_{n=1,\dots,4}$.

Nous regardons le sous-graphe $\mathcal{S}^{H,i}$ pour l'arête i représentée à la figure 3.1(b) en considérant deux graphes grossiers différents, voir figure 3.3. Nous donnons tout d'abord une définition plus précise des ensembles I_e : pour une arête grossière e qui joint un noeud grossier n à un noeud grossier m, I_e est l'ensemble des indices d'arêtes fines appartenant strictement à l'intersection de Ω_n^H et de Ω_m^H.

Le premier graphe grossier est donné à la figure 3.3(a) et on obtient dans ce cas pour l'arête i :

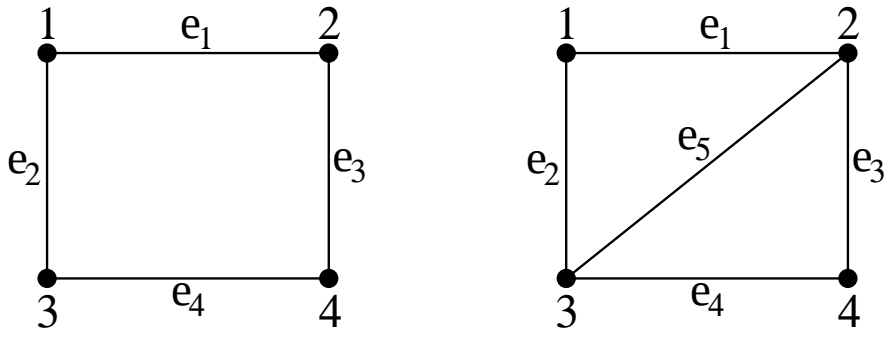
$$\tilde{C}_i = \{2,3\} \text{ et } \tilde{I}_i = \emptyset.$$

Le sous-graphe $\mathcal{S}^{H,i}$ ne vérifie pas la condition de connexité ; le graphe grossier ne possède alors pas assez d'arêtes pour vérifier la relation de compatibilité.

Pour le second graphe grossier donné à la figure 3.3(b), on a ajouté une arête entre les noeuds grossiers 2 et 3 et dans ce cas :

$$\tilde{C}_i = \{2,3\} \text{ et } \tilde{I}_i = \{e_5\}.$$

Le sous-graphe $\mathcal{S}^{H,i}$ vérifie alors la condition de connexité. On peut vérifier que cela est aussi le cas pour toutes les arêtes fines et que le graphe grossier possède suffisament d'arêtes.



(a) Premier graphe grossier (b) Second graphe grossier

Fig. 3.3 – Deux graphes grossiers envisagés.

3.4 Conclusion

Les trois exigences affichées dans l'introduction pour la construction algébrique de base d'approximation grossière ont été intégrées dans la formulation (3.21) en éléments finis d'arête. Celle-ci est équivalente à la formulation (3.24) en supprimant judicieusement certaines contraintes. Cette formulation admet une solution unique à condition que le graphe grossier possède suffisament d'arêtes.

Les matrices de prolongement obtenues par la résolution du problème d'optimisation conduisent à des méthodes multigrilles algébriques ayant des vitesses de convergence similaires à la méthode multigrille géométrique lorsque l'on compare les méthodes sur des maillages emboîtés ; les simulations numériques sont reportées dans l'Annexe D Sous-section D.5 pour différents choix de critère de minimisation, c.-à-d. de matrices A_e.

Cependant, le coût de calcul des différentes matrices de prolongement est important devant le coût de résolution du système lui-même, surtout en dimension 3, ce qui en limite l'intérêt pratique. On peut cependant envisager d'utiliser cette technique pour traiter des problèmes instationnaires avec des schémas implicites. En effet, ceux-ci demandent une résolution de système à chaque pas de temps alors qu'une seule phase de construction des matrices de prolongement est nécessaire.

Une des raisons du coût de la construction de la base grossière est l'introduction des multiplicateurs de Lagrange pour résoudre le problème de minimisation sous contraintes. Le système à résoudre contient alors $M+\tilde{M}$ inconnues où M est le nombre d'éléments non-nuls de la matrice de prolongement β et $\tilde{M}$ le nombre de contraintes imposées. De plus, elle demande la factorisation des sous-matrices A_e qui apparaissent dans le critère d'énergie.

Dans le chapitre suivant, nous proposons une autre technique pour résoudre le problème (3.21). Elle est basée sur la résolution de problèmes de flots locaux et permet de s'affranchir des multiplicateurs de Lagrange et d'éviter la factorisation des matrices.

Chapitre 4

Base d'approximation grossière et résolution de problèmes de flot locaux

Durant l'analyse de la résolution du problème d'optimisation avec multiplicateurs de Lagrange, on a soulevé l'importance de sous-graphes du graphe grossier associés à chaque arête fine pour assurer une solution au problème de minimisation d'énergie. Ces sous-graphes vont aussi nous permettre d'introduire des problèmes de flots locaux dont la résolution fournira une matrice β' vérifiant la condition de compatibilité $\beta' G^H = G^h \alpha$ ainsi qu'une base de $M - \tilde{M}$ matrices $\beta^{i,k}$ vérifiant $\beta^{i,k} G^H = 0$. La solution du problème de minimisation est alors cherchée sous la forme :

$$\beta = \beta' + \sum_{(i,k)} \theta_{i,k} \beta^{i,k}. \tag{4.1}$$

On est ainsi ramené à résoudre un système linéaire de dimension $M - \tilde{M}$ pour calculer les coefficients $\theta_{i,k}$.

4.1 Méthode de résolution

4.1.1 Notations et principe

Pour chaque arête fine i, on note $\mathcal{S}^{H,i}$ le sous-graphe du graphe grossier défini comme suit : les sommets de $\mathcal{S}^{H,i}$ sont les sommets du graphe grossier qui sont indexés par les éléments de $\tilde{C}_i$ et les arêtes de $\mathcal{S}^{H,i}$ sont les arêtes du graphe grossier qui sont indexées par les éléments de $\tilde{I}_i$. On note $\bar{G}^{H,i}$ la matrice d'incidence arête-noeud du sous-graphe $\mathcal{S}^{H,i}$.

On rappelle que :

- $\tilde{C}_i$ est l'ensemble des indices de noeuds grossiers tels que le support de la fonction nodale grossière $w_n^{0,H}$ contienne une extrémité de l'arête fine i,
- $\tilde{I}_i$ est l'ensemble des indices d'arêtes grossières telles que le support de la fonction d'arête grossière $w_e^{1,H}$ contienne l'arête fine i.

Ces définitions sont illustrées sur la figure 4.1. A gauche est représentée une partition des noeuds du maillage qui définissent des ensembles $(H_n)_{n=1,\dots,9}$. Les ensembles L_n sont définis comme les noeuds fins de H_n auxquels on a rajouté tous leurs voisins. Le domaine Ω_n^H est alors la réunion des supports des fonctions nodales fines dont l'indice est dans L_n. C'est la méthode de décomposition en sous-domaines grossiers à partir d'une partition des noeuds fins proposée dans l'Annexe D et utilisée dans les simulations numériques.

Au centre figure le graphe grossier associé à cette décomposition. Deux noeuds grossiers n et m sont reliés si une arête fine a une extrémité dans H_n et l'autre dans H_m. C'est aussi la technique qui est proposée dans l'Annexe D et qui assure que les sous-graphes $\mathcal{S}^{H,i}$ sont connexes. Une méthode plus simple consisterait à connecter deux noeuds grossiers n et m dès que l'intersection de Ω_n^H et Ω_m^H est non vide mais cela conduirait à obtenir un nombre d'arêtes grossières plus important.

A droite sont représentés les sous-graphes $\mathcal{S}^{H,i}$ et $\mathcal{S}^{H,j}$ associés à deux arêtes fines i (en vert) et j (en rouge). Pour l'arête fine i (en vert) dont le sous-graphe $\mathcal{S}^{H,i}$ est repris à la figure 4.2, la matrice $\bar{G}^{H,i}$

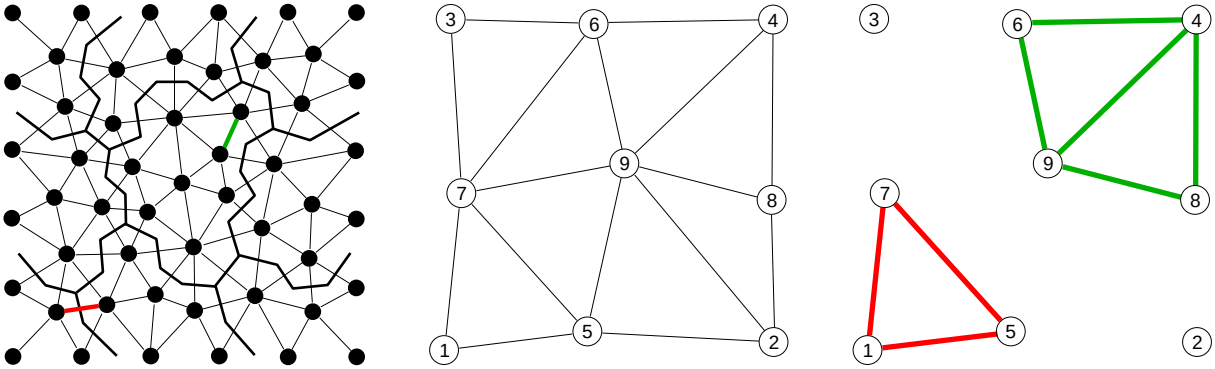


Fig. 4.1 – Partition, graphe grossier et les sous-graphes $\mathcal{S}^{H,i}$ et $\mathcal{S}^{H,j}$.

s'écrit :

$$\bar{G}^{H,i} = \begin{pmatrix} 1 & -1 & 0 & 0 \\ 0 & -1 & 1 & 0 \\ -1 & 0 & 1 & 0 \\ -1 & 0 & 0 & 1 \\ 0 & 0 & -1 & 1 \end{pmatrix}. \tag{4.2}$$

On se place dans le cas où tous les sous-graphes $\mathcal{S}^{H,i}$ sont connexes qui est une condition nécessaire pour que le problème de minimisation admette toujours une solution.

On rappelle que, par construction, le coefficient β_{ie} de la matrice β est nul si i n'appartient pas à I_e ou de façon équivalente si e n'appartient pas à $\tilde{I}_i$. On désigne par $\tilde{F}$ l'ensemble des indices i d'arêtes fines tels que $\tilde{I}_i$ soit non vide et pour i appartenant à $\tilde{F}$, on définit le vecteur β^i qui contient les coefficients β_{ie} de β *a priori* non nuls.

On note ξ la matrice $G^h\alpha$ de dimension $E^h \times N^H$ qui intervient dans la relation de compatibilité. Pour chaque indice d'arête fine i, on construit un vecteur ξ^i obtenu en extrayant les composantes de la $i^{\text{ième}}$ ligne de ξ dont les indices sont dans $\tilde{C}_i$.

Dans l'Annexe E, on montre que la matrice β vérifie :

$$\beta G^H = \xi \tag{4.3}$$

si et seulement si les vecteurs $(\beta^i)_{i\in\tilde{F}}$ sont solutions des problèmes de flot suivants :

$$(\bar{G}^{H,i})^t \beta^i = \xi^i. \tag{4.4}$$

On cherche la solution de chacun des systèmes linéaires (4.4) sous la forme :

$$\beta^i = (\beta^i)' + (\beta^i)'' \tag{4.5}$$

où $(\beta^i)'$ est une solution particulière du système et $(\beta^i)''$ une composante vérifiant $(\bar{G}^{H,i})^t(\beta^i)'' = 0$.

Une solution particulière $(\beta^i)'$ de (4.4) peut être déterminée en utilisant un arbre couvrant du sous-graphe $\mathcal{S}^{H,i}$. On rappelle qu'un arbre couvrant d'un graphe est un graphe minimal connexe qui relie tous les noeuds du graphe. On peut toujours en déterminer un car $\mathcal{S}^{H,i}$ est connexe par hypothèse. Pour calculer $(\beta^i)'$, on fixe à zéro les composantes dans le coarbre, c.-à-d. pour les arêtes qui ne sont pas dans l'arbre couvrant, et on supprime une équation dans le problème de flot ; on obtient un système linéaire dont la matrice est inversible. Dans l'exemple de l'arbre couvrant de la figure 4.2, on supprime les lignes 3 et 5 de la matrice $\bar{G}^{H,i}$ et par exemple la colonne numéro 4. On transpose la matrice 3×3 obtenue et la matrice du système à résoudre est :

$$\begin{pmatrix} 1 & 0 & -1 \\ -1 & -1 & 0 \\ 0 & 1 & 0 \end{pmatrix}.$$

Le nombre d'arêtes du coarbre fournit la dimension k_i du noyau de $(\bar{G}^{H,i})^t$. Si le noyau de $(\bar{G}^{H,i})^t$ est non réduit à zéro, il est possible d'en déterminer simplement une base notée $((\beta^{i,k})'')_{1\leq k\leq k_i}$. Pour l'obtenir, on cherche dans le sous-graphe k_i cycles indépendants (appelées boucles fondamentales dans le Chapitre 1, Section 1.2.1). Les composantes de $(\beta^{i,k})''$ prennent alors la valeur 0 si l'arête ne fait pas partie du cycle et -1 ou 1 sinon, suivant le sens du parcours du cycle. Dans l'exemple de la figure 4.2 la dimension du noyau est 2 et les cycles indépendants sont donnés par :

$$\begin{pmatrix}-1, & 1, & -1, & 0, & 0\end{pmatrix}^t \text{ et } \begin{pmatrix}0, & 0, & 1, & -1, & 1\end{pmatrix}^t.$$

Le vecteur $(\beta^i)''$ peut alors s'écrire comme une combinaison linéaire des $(\beta^{i,k})'' : (\beta^i)'' = \sum_{k=1}^{k_i} \theta_{i,k}(\beta^{i,k})''$.

Pour éviter de référer à des sommes systématiquement nulles, on définit :

$$\tilde{\tilde{F}} = \{i \in \tilde{F} \ / \ k_i > 0\}. \tag{4.6}$$

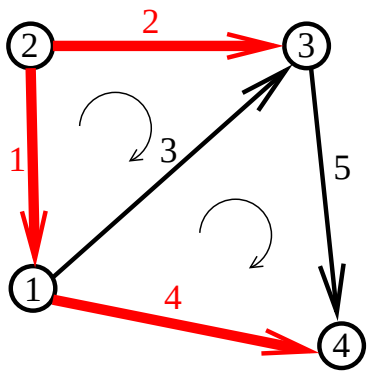


FIG. 4.2 – Sous-graphe $\mathcal{S}^{H,i}$ (voir figure 4.1). Un arbre couvrant est en trait fort rouge. Pour les deux cycles indépendants, l'orientation est représentée.

Pour chaque arête fine d'indice i, on définit l'opérateur de projection $\tilde{Q}_i$ qui extrait d'un vecteur de $\mathbb{R}^{E^H}$ les composantes d'indice dans $\tilde{I}_i$.

On construit la matrice β' de dimension $E^h \times E^H$ à partir des solutions $(\beta^i)'$. Les lignes i telles que i n'appartienne pas à $\tilde{F}$ sont nulles. Pour i dans $\tilde{F}$, la ligne i est formée des composantes du vecteur $\tilde{Q}_i^t(\beta^i)'$.

A partir du vecteur de base $(\beta^{i,k})''$, on définit une matrice $\beta^{i,k}$ de dimension $E^h \times E^H$ de la même façon.

La forme générale d'une matrice vérifiant (4.3) qui s'écrit aussi en transposant l'égalité $(G^H)^t\beta^t = (G^h\alpha)^t$ prend alors la forme :

$$\beta = \beta' + \sum_{i\in\tilde{\tilde{F}}} \sum_{k=1}^{k_i} \theta_{i,k}\beta^{i,k}.$$

En effet, il est évident que les matrices $\beta^{i,k}$ sont indépendantes. De plus, on note que les cycles indépendants des sous-graphes sont aussi des cycles indépendants pour le graphe grossier, ce qui implique que les $\beta^{i,k}$ sont dans le noyau de $(G^H)^t$. En conséquence β vérifie :

$$(G^H)^t\beta^t = (G^H)^t(\beta')^t = (G^h\alpha)^t.$$

Les coefficients $\theta_{i,k}$ sont ainsi les degrés de liberté qu'il suffit de calculer pour déterminer complètement la matrice β dans le problème de minimisation.

On rappelle le problème de minimisation à résoudre :

$$\text{Trouver } (\beta_e)_{e=1,\ldots,E^H} \text{ minimisant } \sum_{e=1}^{E^H} \beta_e^t A_e \beta_e \text{ sous la contrainte : } \beta G^H = G^h\alpha, \tag{4.7}$$

où β_e contient les composantes non nulles de la $e^{\text{ième}}$ colonne de β. Si on définit $\bar{\beta}$ le vecteur regroupant les différents vecteurs β_e et D la matrice bloc diagonale contenant les matrices A_e, le critère à minimiser s'écrit alors $\bar{\beta}^t D\bar{\beta}$.

De la même manière, on définit les vecteurs $\bar{\beta}'$ et $\bar{\beta}^{i,k}$ qui vérifient encore la relation :

$$\bar{\beta} = \bar{\beta}' + \sum_{i\in\tilde{\tilde{F}}} \sum_{k=1}^{k_i} \theta_{i,k}\bar{\beta}^{i,k}. \tag{4.8}$$

On introduit le vecteur Θ qui contient les coefficients $\theta_{i,k}$. Le problème de minimisation s'écrit :

$$\text{Trouver } \Theta \text{ minimisant } \sum_{(i,k)(j,l)} \theta_{i,k}\theta_{j,l}(\bar{\beta}^{i,k})^t D(\bar{\beta}^{j,l}) + 2\sum_{(i,k)} \theta_{i,k}(\bar{\beta}')^t D(\bar{\beta}^{i,k}) + (\bar{\beta}')^t D\bar{\beta}', \tag{4.9}$$

soit encore en introduisant la matrice B dont les vecteurs colonnes sont les $\bar{\beta}^{i,k}$ pour i dans $\tilde{\tilde{F}}$ et k dans $\{1,\ldots,k_i\}$:

$$\text{Trouver } \Theta \text{ minimisant } (B\Theta)^t D(B\Theta) + 2(B\Theta)^t D\bar{\beta}' + (\bar{\beta}')^t D\bar{\beta}'.$$

Le minimum de cette forme quadratique est donné par la solution du système linéaire :

$$(B^t DB)\Theta = -B^t D\bar{\beta}'. \tag{4.10}$$

Remarque 4.1. *Le problème de flot* (4.4) *correspond aussi à l'écriture de la loi des courants de Kirchhoff pour un réseau dont la topologie est décrite par* $\bar{G}^{H,i}$.

Remarque 4.2. *Dans le cas de la méthode définie par Reitzinger et Schöberl [66], les sous-graphes se réduisent à un sommet ou à une arête unique. La résolution des problèmes de flot donne alors l'unique matrice* β *vérifiant* (4.3) *mais il n'y a aucun degré de liberté pour la minimisation.*

4.1.2 A propos de la résolution du système linéaire (4.10)

Comme les matrices A_e sont symétriques définies positives, la matrice $B^t DB$ est aussi symétrique définie positive et on peut utiliser le gradient conjugué pour résoudre le système (4.10). La méthode ne demande pas de construire explicitement la matrice du système car elle effectue uniquement des produits matrice-vecteur.

La matrice B calculée à partir des solutions des problèmes de flot est creuse et ne contient que des 1 et des -1. Seuls ses éléments non nuls nécessitent d'être stockés. L'algorithme 4.1 décrit le produit $B^t DBx$ en trois étapes.

$y \leftarrow Bx$;
Pour e **de** 1 **à** E^H **faire**
  $z_e = A_e x_e$
Fin Pour
$x \leftarrow B^t z$;

Algorithme 4.1: Multiplication matrice-vecteur $B^t DBx$.

L'avantage notable par rapport à la résolution avec des multiplicateurs est qu'il n'est pas nécessaire de factoriser les matrices A_e d'où un gain en temps et en espace de stockage.

Remarque 4.3. *On peut envisager de calculer la matrice* $B^t DB$ *dans le cas où les matrices* A_e *sont extraites d'une matrice globale* A. *En effet, on vérifie que l'élément d'indice* $((i,k),(j,l))$ *de la matrice* $B^t DB$ *s'écrit :*

$$(\bar{\beta}^{i,k})^t D\bar{\beta}^{j,l} = p_{ikjl}A_{ij}. \tag{4.11}$$

où A_{ij} *est le coefficient d'indice* (i,j) *de la matrice globale.*

Le coefficient p_{ikjl} *est égal à :*

$$\sum_{e \in \tilde{I}_i \cap \tilde{I}_j} \mathrm{or}(e, C_{i,k})\mathrm{or}(e, C_{j,l}).$$

La notation $C_{i,k}$ *désigne le cycle de numéro* k *dans le sous-graphe grossier* $\mathcal{S}^{H,i}$*. Le terme* $\mathrm{or}(e, C_{i,k})$ *vaut 0 si l'arête grossière* e *n'appartient pas au cycle et* 1 *ou* -1 *si son orientation coïncide ou non avec l'orientation du cycle.*

4.1.3 Avantages par rapport à la méthode avec les multiplicateurs

De nombreux avantages apparaissent avec cette approche :
- le calcul des problèmes de flot (4.4) avec les arbres couvrants est simple et peu coûteux tant que le nombre de noeuds dans les sous-graphes reste faible.
- on maîtrise directement la précision sur la solution du problème de minimisation alors que dans le cas précédent on ajuste la précision sur les multiplicateurs ;
- après chaque pas du gradient conjugué, l'itéré définit une matrice de prolongement qui vérifie la condition de compatibilité (4.3) ;
- Si la dimension du système linéaire (4.10) est jugée trop importante, on peut volontairement ne pas considérer certains degrés de liberté pour la minimisation, tout en conservant la vérification de la condition de compatibilité (4.3) ;
- Dans tous les essais numériques réalisés, le produit par B^tDB est nettement moins coûteux que celui par $T^tD^{-1}T$, en particulier à cause de la différence importante de coût entre la multiplication par D et celle par D^{-1}. Il n'y a pas de factorisation des matrices A_e et pas de résolution des systèmes locaux.

4.2 Résultats numériques pour des problèmes 2D

4.2.1 Analyse du conditionnement de la matrice B^tDB

La convergence de la méthode du gradient conjugué pour résoudre le système (4.10) peut s'évaluer à partir de son nombre de conditionnement ; voir le rappel en Sous-section 2.1.2 du Chapitre 2. Une majoration grossière est donnée par l'inégalité suivante :

$$\mathrm{cond}(B^tDB) \leq \mathrm{cond}(B^tB)\,\mathrm{cond}(D). \tag{4.12}$$

Pour construire les vecteurs $\bar{\beta}^{i,k}$ à partir des matrices $\beta^{i,k}$, on peut regrouper les éléments non nuls de chaque ligne plutôt que de chaque colonne comme il a été fait dans le paragraphe précédent. La matrice B est alors diagonale par blocs de blocs diagonaux $(B^{H,i})^tB^{H,i}$ où, pour i dans $\tilde{\tilde{F}}$, les colonnes de $B^{H,i}$ sont les vecteurs $(\beta^{i,k})''_{1\leq k\leq k_i}$ obtenus à partir des cycles indépendants du sous-graphe $\mathcal{S}^{H,i}$. La matrice B^tB est alors diagonale par blocs et son spectre vérifie :

$$\mathrm{Sp}(B^tB) = \bigcup_{i\in\tilde{\tilde{F}}} \mathrm{Sp}((B^{H,i})^tB^{H,i}). \tag{4.13}$$

On obtient l'égalité :

$$\mathrm{cond}(B^tB) = \frac{\lambda_{\max}(B^tB)}{\lambda_{\min}(B^tB)} = \frac{\max_{i\in\tilde{\tilde{F}}}\lambda_{\max}((B^{H,i})^tB^{H,i})}{\min_{i\in\tilde{\tilde{F}}}\lambda_{\min}((B^{H,i})^tB^{H,i})}. \tag{4.14}$$

Le nombre de noeuds et d'arêtes dans chacun des sous-graphes est *a priori* indépendant de la dimension globale du problème et donc on peut conjecturer que le conditionnement de B^tB l'est aussi. On le vérifie d'ailleurs numériquement dans les tableaux 4.2 et 4.8.

En outre, comme la matrice D est bloc-diagonale avec les matrices de blocs diagonaux $(A_e)_{e=1,\ldots,E_H}$, on a aussi :

$$\mathrm{Sp}(D) = \bigcup_{e=1}^{E^H} \mathrm{Sp}(A_e). \tag{4.15}$$

Par conséquent, son nombre de conditionnement vérifie l'égalité :

$$\mathrm{cond}(D) = \frac{\max_{e\in\{1,\dots,E_H\}} \lambda_{\max}(A_e)}{\min_{e\in\{1,\dots,E_H\}} \lambda_{\min}(A_e)} \tag{4.16}$$

Le conditionnement de la matrice D peut être fortement dépendant de la dimension globale lorsque les matrices A_e qui définissent le critère d'énergie sont des sous-matrices du problème à résoudre. Pour illustrer ce point, on peut faire l'analyse simplifiée suivante. Pour un maillage de paramètre H sur un domaine Ω_e, la matrice de discrétisation sur le problème (1.33) pour les cas où ν et γ sont strictement positifs obtenue avec les éléments d'arête s'écrit en omettant les termes de bord :

$$A_e^H = S_\nu^H + M_\gamma^H. \tag{4.17}$$

On déforme par une homothétie de rapport h/H le domaine Ω_e et son maillage. La matrice de discrétisation sur ce nouveau domaine est de même dimension que A_e^H. Elle s'écrit aussi :

$$A_e^h = S_\nu^h + M_\gamma^h.$$

Pour des problèmes 2D, on vérifie la relation :

$$A_e^h = \left(\frac{H}{h}\right)^2 S_\nu^H + M_\gamma^H,$$

et pour des problèmes 3D la relation :

$$A_e^h = \left(\frac{H}{h}\right) S_\nu^H + \left(\frac{h}{H}\right) M_\gamma^H,$$

d'où on tire :

$$\mathrm{cond}(A_e^h) \approx \left(\frac{H}{h}\right)^2 \mathrm{cond}(A_e^H).$$

Ainsi si l'on considère une matrice D avec des problèmes locaux directement issus de la forme bilinéaire du problème, on peut s'attendre à obtenir un nombre de conditionnement évoluant en $1/h^2$ quand le diamètre maximal des éléments du maillage h diminue, ce qui nuit à la vitesse de convergence de la méthode du gradient conjugué.

Pour conserver un conditionnement stable de la matrice D, on peut envisager de prendre une matrice locale du type :

$$A_e = S_\nu + \upsilon \max(\mathrm{diag}(S_\nu))\,\mathrm{Id},$$

où υ est un petit paramètre (0.01 par exemple). Pour que la matrice A_e soit symétrique définie positive, il faut que ce paramètre soit strictement positif. En pratique on améliore aussi la convergence du gradient conjugué, par rapport au choix de matrices A_e de la forme (4.17) avec ν strictement positif, en prenant υ égale à 0. Utiliser des problèmes locaux de ce type permet en outre de mettre en oeuvre la méthode de minimisation lorsque la forme bilinéaire associée au problème à résoudre ne définit pas un produit scalaire (dès que ν n'est pas strictement positif par exemple).

4.2.2 Présentation des problèmes tests

On rappelle la forme du système linéaire que l'on souhaite résoudre par les méthodes multiniveau :

$$Au = b, \text{ avec : } A = S_\nu + M_\gamma + M_{\delta,\Gamma_\mathrm{I}}. \tag{4.18}$$

La matrice S_ν correspond à la discrétisation de l'opérateur $\mathrm{rot}\,\nu\,\mathrm{rot}$, la matrice M_γ est la matrice de masse et $M_{\delta,\Gamma_\mathrm{I}}$ est la contribution des conditions aux limites absorbantes.

Les problèmes tests ont pour but de valider les méthodes multiniveau algébriques que l'on propose et de comparer leurs performances à d'autres méthodes.

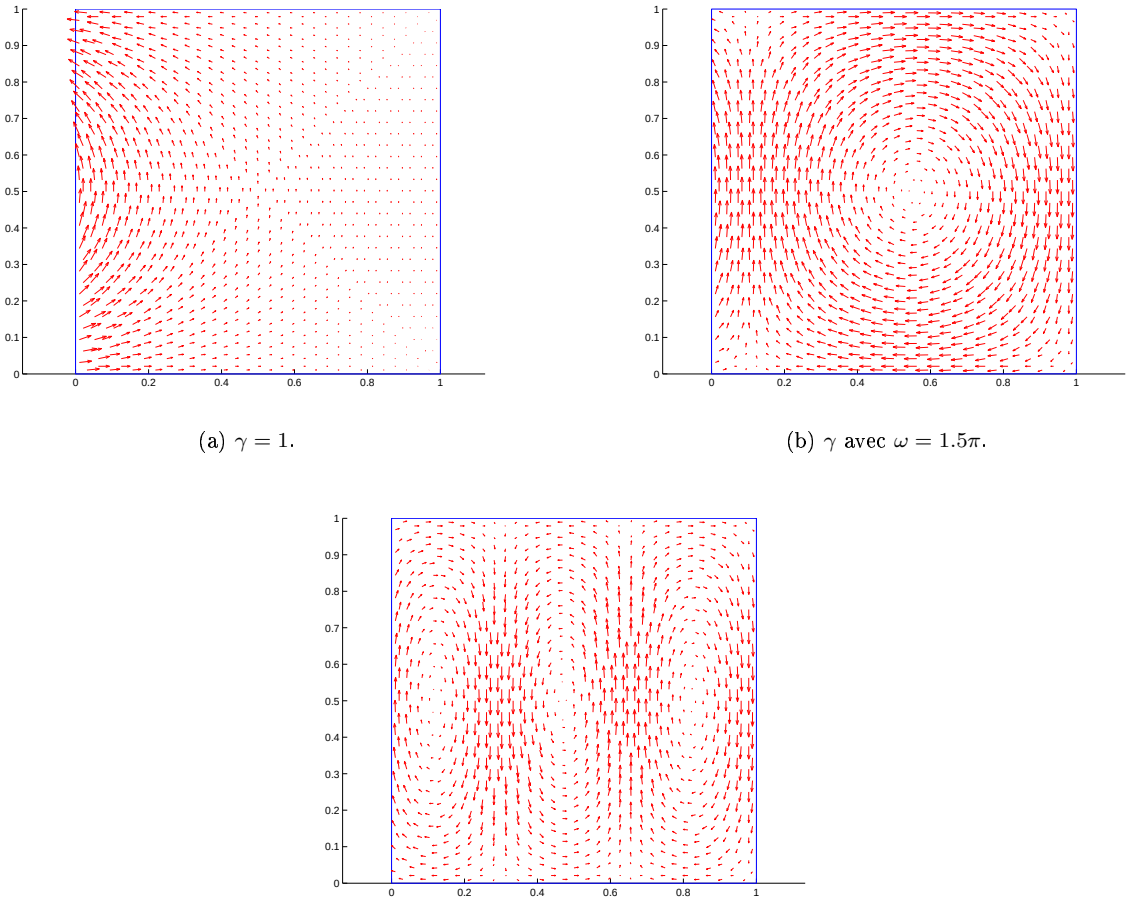
Premier problème test : Géométrie régulière et coefficients constants

Le premier problème test a été utilisé par Beck dans la note [8]. Le domaine considéré est le carré unité $\Omega =]0;1[^2$. Sur la portion latérale gauche de la frontière du domaine, c.-à-d. celle joignant le point de coordonnées $(0,0)$ au point de coordonnées $(0,1)$, on impose la composante tangentielle du champ électrique $\mathbf{E}$: $\mathbf{E}_y = \sin(\pi y)$. Sur la partie restante on laisse des conditions aux limites naturelles, à savoir $\nu \operatorname{rot} \mathbf{E} \times n = 0$.

Le coefficient ν est pris constant égal à 1 sur le domaine Ω. Le coefficient γ sera aussi constant sur le domaine et considéré pour trois valeurs différentes :

- égal à 1 pour considérer un système simple où le caractère défini positif de la matrice du système à résoudre est assuré.
- égal à $-\omega^2$, où ω prend successivement les valeurs 1.5π et 3π. Ceci revient à considérer un problème en régime harmonique. Dans ce cas, on perd le caractère défini positif.

Les solutions obtenues avec ces différentes valeurs de γ sont reproduites à la figure 4.3.



(a) $\gamma = 1$.

(b) γ avec $\omega = 1.5\pi$.

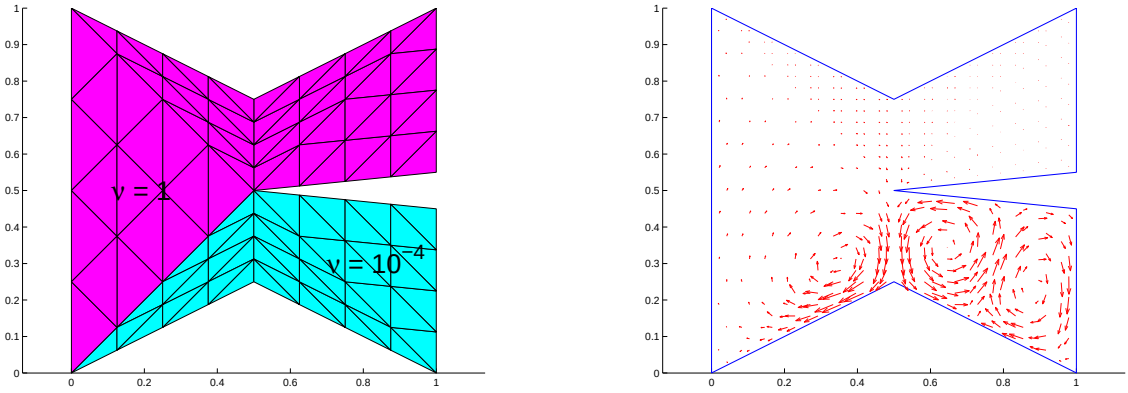
(c) γ avec $\omega = 3\pi$.

FIG. 4.3 – Solutions obtenues pour le premier cas test avec les différentes valeurs de γ.

Second problème test : Géométrie non régulière et coefficient ν non constant

Ce second problème a aussi été utilisé par Beck dans la note [8] : on a déformé le carré unité de l'exemple précédent pour créer des singularités géométriques. On distingue deux régions dans le domaine de calcul ; dans l'une de ces régions $\nu = 1$ et dans l'autre $\nu = 10^{-4}$. Si l'on pense à un calcul en

champ électrique ou en potentiel vecteur magnétique, cela peut correspondre au passage de l'air, où la perméabilité relative μ_r vaut 1, à un matériau ferromagnétique, où μ_r serait égale à 10^4. Le coefficient γ est constant sur le domaine et égal à $-\omega^2$ avec une valeur de ω fixée à 0.05π. Les conditions aux limites sont identiques au premier problème : la composante tangentielle du champ est imposée sur la portion latérale gauche de la frontière et des conditions aux limites naturelles sont utilisées sur le reste de la frontière. Une représentation du domaine ainsi que de la solution obtenue est donnée à la figure 4.4.



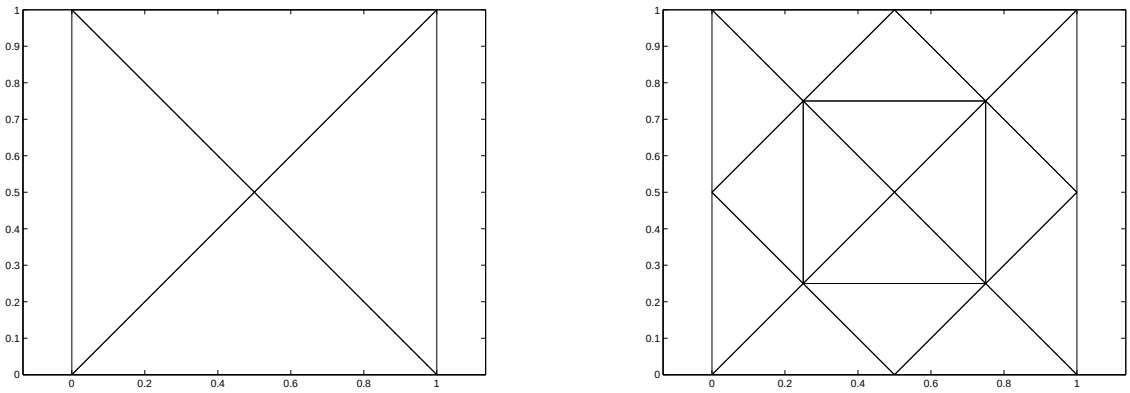
(a) Géométrie et caractéristique + maillage structuré.

(b) Représentation de la solution.

Fig. 4.4 – Géométrie et solution obtenue pour le second problème test.

4.2.3 Simulations numériques sur le premier problème test

On part d'un maillage simple et structuré sur le carré unité, voir figure 4.5(a), que l'on le raffine ensuite de manière régulière, voir figure 4.5(b) pour le premier raffinement. Ces maillages sont notés τ_i^h où l'indice i désigne le nombre de raffinements successifs depuis le maillage initial τ_0^h.



(a) Maillage le plus grossier τ_0^h.

(b) Premier raffinement τ_1^h.

Fig. 4.5 – Maillage initial (gauche) et premier raffinement (droite).

L'intérêt du maillage structuré est de permettre la comparaison des résultats de la méthode algébrique avec les résultats d'une méthode multigrille géométrique standard [40]. Cependant, les méthodes

multiniveau algébriques mises en oeuvre dans les simulations qui suivent, ignorent le caractère structuré du maillage.

La matrice annexe nodale utilisée pour la définition du graphe grossier est une matrice proposée par Reitzinger et Kaltenbacher [48]. Elle s'écrit $B = \mathrm{Id} + \tilde{B}$ où $\tilde{B}$ est obtenue par assemblage de matrice élémentaire sur le maillage fin. Pour un élément t de sommets $(\mathbf{x}_j)_{j=1,\dots,N_t}$ et de coefficient local ν_t, les éléments extra-diagonaux de la matrice élémentaire $\tilde{B}_t$ sont donnés par :

$$(\tilde{B}_t)_{ij} = \frac{-\nu_t}{\|\mathbf{x}_j - \mathbf{x}_i\|_2} \text{ si } i \neq j, \text{ et } (\tilde{B}_t)_{ii} = -\sum_{j \neq i} (\tilde{B}_t)_{ij}. \tag{4.19}$$

Le graphe déterminé à partir des éléments non nuls de la matrice B coïncide exactement avec le graphe défini par le maillage dans le cas d'un maillage par triangles ou par tétraèdres. Cette matrice permet d'avoir une représentation des variations du coefficient ν et des distortions du maillage. Reitzinger et Kaltenbacher proposent ces coefficients pour obtenir une M-matrice, au sens où elle vérifie les propriétés suivantes :

$$B_{ii} > 0 \ \forall i, \quad B_{ij} \leq 0 \ \forall i \neq j, \quad \text{les entrées de } B^{-1} \text{ sont positives ou nulles.} \tag{4.20}$$

Cette propriété est souvent requise dans les méthodes multiniveau algébriques [78, 38] pour les heuristiques de choix des noeuds définissant les variables grossières.

L'algorithme que nous avons choisi dans nos tests pour réaliser la partition des noeuds est l'algorithme d'aggrégation présenté dans [76, 78, page 63]. Il fait intervenir un paramètre θ permettant de savoir si deux noeuds sont fortement connectés, les auteurs de [76] suggèrent de choisir ce paramètre égal à $0.08 \times (1/2)^{(\text{numéro niveau}-1)}$ où le niveau le plus fin est numéroté 1. Cette valeur est utilisée dans les expériences numériques sauf précision contraire.

Résolution du problème de minimisation

La tableau 4.1 compare le nombre d'inconnues $(M - \tilde{M})$ pour le problème de minimisation au niveau le plus fin à la taille du système linéaire à résoudre (E^h, approximativement le nombre d'arêtes sur le maillage fin). Le nombre d'inconnues pour la minimisation représentent approximativement les 2/3 du

	τ_2^h	τ_3^h	τ_4^h	τ_5^h	τ_6^h	τ_7^h
$M - \tilde{M}$	62	253	1016	4081	16355	65479
E^h	100	392	1552	6176	24640	98432

TAB. 4.1 – Nombre d'inconnues pour les problèmes de minimisation et le système à résoudre — 2D, maillages structurés.

nombre d'inconnues du problème.

Pour poser le problème de minimisation, différents choix de matrices A_e ont été testés :

- des matrices extraites de la matrice A qui discrétise l'opérateur du problème sur le maillage fin ; ce choix sera noté A dans les tableaux de résultats,
- des matrices extraites de S_ν, la partie provenant de la discrétisation de l'opérateur $\mathrm{rot}\,\nu\,\mathrm{rot}$; ce choix sera noté S_ν dans les tableaux de résultats. Ceci définit uniquement des matrices semi-définies positives, mais on verra que cela ne nuit pas aux résultats en pratique.
- des matrices extraites de la matrice $A + GM_\phi^{-1}G^t$ où M_ϕ désigne une matrice de masse en éléments nodaux avec condensation de masse et G la matrice d'incidence arête-noeud ; ce choix sera noté $A + GM_\phi^{-1}G^t$ dans les tableaux de résultats.
- les matrices sont toutes égales à l'identité ; ce choix sera noté Id.

La solution du problème de minimisation est calculée par une méthode de gradient conjugué non préconditionnée. L'algorithme s'arrête quand la norme du résidu a été divisée par 10^3. Dans le tableau 4.2 est reporté le nombre d'itérations requis pour atteindre le critère d'arrêt lors de la résolution du problème de minimisation au niveau le plus fin (système linéaire (4.10)).

Pour les quatre choix des matrices A_e, le nombre d'itérations est indépendant de la dimension globale du système.

	τ_2^h	τ_3^h	τ_4^h	τ_5^h	τ_6^h	τ_7^h
A	24	19	12	12	12	12
$A + G^h M_\phi^{-1} (G^h)^t$	9	8	8	8	8	8
S_ν	14	12	12	12	12	12
Id	3	1	1	1	1	1

Tab. 4.2 – Nombre d'itérations pour le problème de minimisation (division du résidu par 10^3) — 2D, maillages structurés.

Résolution du système linéaire (4.18)

Le système linéaire est résolu par l'algorithme du gradient conjugué préconditionné dans le cas où A est symétrique définie positive. On utilise l'algorithme COCG [75] préconditionné dans le cas où A n'est pas définie positive. L'algorithme s'arrête lorsque la norme du résidu a été divisée par 10^{10}. Les comportements de différents préconditionnements multiniveau vont être comparés à deux méthodes à un niveau inspirées respectivement des lisseurs de Arnold, Falk et Winther et de Hiptmair (voir Chapitre 2 et Annexe B Sous-section B.2.2).

Les algorithmes de préconditionnement multiniveau testés sont de deux types :
- une itération d'un V-cycle avec une étape de prélissage et une étape de postlissage assurées par le lisseur de Arnold, Falk et Winther,
- un algorithme issu des travaux de Hiptmair mais avec un agencement des corrections sur les sous-espaces proposé par Beck pour lequel on donne quelques détails à la suite.

Cet algorithme proposé par Beck est constitué d'une suite de 3 V-cycles. Deux V-cycles sont réalisés sur le problème en potentiel scalaire de matrice $A_\phi = G^t A G$ qui encadre un V-cycle sur le problème initial de matrice A. Un pas du préconditionneur correspond à l'algorithme 4.2.

Procédure precond(x : **vecteur**, r : **vecteur**)
[x : résidu préconditionné.]
[r : résidu non préconditionné.]
$x_\phi \leftarrow 0$, $x \leftarrow 0$;
V-cycle sur le sytème $A_\phi x_\phi = G^t r$;
$x \leftarrow x + G x_\phi$;
V-cycle sur le sytème $Ax = r$;
$x_\phi \leftarrow 0$;
V-cycle sur le sytème $A_\phi x_\phi = G^t (r - Ax)$;
$x \leftarrow x + G x_\phi$;
Fin

Algorithme 4.2: Algorithme de préconditionnement utilisé pour les simulations.

Pour les 3 V-cycles de l'algorithme, on utilise :
- pour le prélissage, un pas de l'algorithme de Gauss-Seidel par points en *descente* ;
- pour le postlissage, un pas de l'algorithme de Gauss-Seidel par points en *remontée*.

Pour les V-cycles en potentiel scalaire, la résolution sur la grille grossière est remplacée par un pas de Gauss-Seidel symétrique, la matrice grossière obtenue n'étant généralement pas inversible.

Dans les tableaux qui vont suivre :
- la mention 1 niveau Arnold ou Hiptmair réfère aux algorithmes à un niveau.
- la mention *géométrique* indique l'utilisation d'une méthode multigrille classique, où les opérateurs de prolongement sont les opérateurs de transfert canoniques [19, Chapter 6],
- les mentions A, $A + G^h M_\phi^{-1} (G^h)^t$, S_ν et Id réfèrent à l'utilisation de méthode multiniveau avec les matrices de prolongement obtenues par minimisation d'énergie.

- la mention méthode RS réfère à la méthode proposée par Reitzinger et Schöberl [66],
- la mention sans minimisation réfère à l'utilisation d'une matrice de prolongement en arête compatible issue de la résolution des problèmes de flot mais non soumis au processus de minimisation.

Cas où $\gamma = 1$ Le tableau 4.3 compare le nombre d'itérations sur différents maillages fins définis par τ_i^h. L'algorithme utilisé pour la méthode multiniveau algébrique est le V-cycle avec le lisseur de Arnold. Les résultats avec l'algorithme de préconditionnement 4.2 sont reportés dans le tableau 4.4.

	τ_2^h	τ_3^h	τ_4^h	τ_5^h
A	11	14	17	20
$A + G^h M_\phi^{-1} (G^h)^t$	11	15	20	25
S_ν	11	14	17	20
Id	11	15	20	25
sans minimisation	11	16	22	31
méthode RS	11	18	28	44
géométrique	7	7	8	8
1 niveau Arnold	17	31	59	116

TAB. 4.3 – Nombre d'itérations du gradient conjugué préconditionné (division du résidu par 10^{10}) — 2D, maillages structurés.

	τ_2^h	τ_3^h	τ_4^h	τ_5^h	τ_6^h	τ_7^h
A	15	20	22	24	27	30
$A + G^h M_\phi^{-1} (G^h)^t$	17	22	32	44	67	96
S_ν	15	20	22	24	27	30
Id	17	22	32	40	57	72
sans minimisation	18	28	45	66	105	163
méthode RS	18	29	50	83	146	260
géométrique	9	9	9	9	9	9
1 niveau Hiptmair	22	42	83	157	298	488

TAB. 4.4 – Nombre d'itérations du gradient conjugué préconditionné (division du résidu par 10^{10}) — 2D, maillages structurés.

Les algorithmes multiniveau algébriques correspondant aux choix A et S_ν donnent des résultats quasi-optimaux au sens où le nombre d'itérations dépend faiblement de la taille du système à résoudre. Cependant la méthode multigrille géométrique, qui à la différence de nos méthodes algébriques bénéficient pleinement de l'aspect structuré du maillage, fournit le meilleur préconditionnement. Le choix correspondant à Id donne de moins bons résultats mais a néanmoins un meilleur comportement que la méthode proposée par Reitzinger et Schöberl et que la méthode à un niveau. La méthode qui calcule une matrice de prolongement par résolution des problèmes de flot sans minimisation, et donc à faible coût, donne des résultats intéressants.

Dans la suite, pour les préconditionneurs multiniveau géométrique et algébriques on utilisera l'algorithme 4.2.

Cas du régime harmonique : $\gamma = -\omega^2$. Pour les valeurs de γ négatives, correspondant à un calcul en régime harmonique, on se limite à tester deux méthodes multiniveau algébriques à savoir celles qui correspondent aux choix S_ν et Id. Les matrices de prolongement issues des méthodes de minimisation vont alors être identiques au cas précédent où $\gamma = 1$ car la valeur de γ n'intervient ni dans le choix des noeuds, ni dans le processus de minimisation.

Le tableau 4.5 compare le nombre d'itérations entre les différents préconditionneurs pour le cas $\omega = 1.5\pi$. Le nombre d'inconnues sur la grille la plus grossière est donné entre parenthèses.

Les performances des deux méthodes multiniveau algébriques avec minimisation d'énergie correspondant à S_ν et Id se rapprochent de celles de la méthode multiniveau géométrique. Ces performances sont peu affectées par le nombre d'inconnues sur la grille grossière, qui est parfois petit compte-tenu de la fréquence angulaire ω. Pour la méthode multigrille géométrique, la grille la plus grossière est donnée par τ_1^h.

	τ_2^h	τ_3^h	τ_4^h	τ_5^h	τ_6^h	τ_7^h
S_ν	44 (8)	42 (28)	47 (8)	63 (30)	85 (8)	97 (30)
Id	36 (8)	32 (28)	46 (8)	55 (30)	82 (8)	98 (30)
sans minimisation	37 (7)	39 (26)	59 (7)	80 (29)	151 (7)	220 (28)
méthode RS	27 (8)	47 (28)	72 (8)	107 (30)	191 (10)	339 (30)
géométrique	24	21	25	26	41	83
1 niveau	35	58	111	226	394	X

Tab. 4.5 – Nombre d'itérations de COCG préconditionné (division du résidu par 10^{10}) — 2D, maillages structurés. La notation X signifie que le critère d'arrêt n'est pas satisfait après 500 itérations.

Le tableau 4.6 regroupe les résultats pour le cas $\omega = 3\pi$. Dans les méthodes multiniveau algébriques seul le niveau le plus grossier est autorisé à avoir moins de 150 inconnues arête. Entre parenthèses est donné le nombre d'inconnues sur la grille la plus grossière.

	τ_3^h	τ_4^h	τ_5^h	τ_6^h	τ_7^h
S_ν	170 (28)	124 (104)	133 (30)	283 (106)	287 (30)
Id	96 (28)	59 (104)	104 (30)	104 (106)	185 (30)
sans minimisation	110 (26)	84 (100)	153 (29)	213 (103)	354 (29)
méthode RS	73 (28)	103 (104)	192 (30)	268 (106)	X (30)
géométrique	41	43	44	58	–
1 niveau	96	184	347	738	1436

Tab. 4.6 – Nombre d'itérations de COCG préconditionné (division du résidu par 10^{10}) — 2D, maillages structurés.

On constate que le comportement de toutes les méthodes testées se détériore légèrement mais que la hiérarchie des méthodes est conservée.

4.2.4 Résultats pour le second problème test.

Résolution du problème de minimisation

La tableau 4.7 compare le nombre d'inconnues pour les problèmes de minimisation aux différents niveaux à la taille du système linéaire à résoudre au niveau le plus fin. Le rapport des deux nombres d'inconnues est là aussi proche de 2/3.

	τ_2^h	τ_3^h	τ_4^h	τ_5^h	τ_6^h
$M - \tilde{M}$	111	445	1784	7153	28643
E^h	181	699	2743	10863	43231

Tab. 4.7 – Nombre d'inconnues pour les problèmes de minimisation et le système à résoudre — 2D, maillages structurés.

La méthode et le critère d'arrêt pour la résolution du problème de minimisation sont identiques à ceux proposés pour le premier problème. Considérant un cas harmonique, on travaille aussi uniquement avec les choix Id et S_ν. Dans le tableau 4.8 est reporté le nombre d'itérations requis pour atteindre le critère d'arrêt.

	τ_2^h	τ_3^h	τ_4^h	τ_5^h	τ_6^h
S_ν	19	23	22	22	21
Id	2	3	1	1	1

TAB. 4.8 – Nombre d'itérations pour le problème de minimisation (division du résidu par 10^3) — 2D, maillages structurés.

Résolution du système (4.18)

La méthode et le critère d'arrêt pour la résolution du système linéaire (4.18) sont identiques à ceux utilisés pour le cas harmonique du premier problème test.

Le tableau 4.9 donne l'évolution du nombre d'itérations de la méthode COCG préconditionné par les différentes méthodes en fonction du maillage le plus fin.

On constate que le nombre d'itérations évolue peu dans le cas des méthodes avec minimisation d'énergie. En outre, la tendance globale pour les méthodes avec minimisation, lorsque le nombre d'inconnues augmente, est meilleure que pour les méthodes sans minimisation d'énergie. L'intérêt des méthodes multiniveau algébrique par rapport à une méthode à un niveau apparaît plus nettement sur cet exemple.

Pour les méthodes multiniveau, on voit aussi ici apparaître une dépendance vis-à-vis de la taille du niveau le plus grossier indiquée entre parenthèses ; la résolution réclame un nombre accru d'itérations lorsque la grille au niveau le plus bas est trop grossière.

	τ_2^h	τ_3^h	τ_4^h	τ_5^h	τ_6^h
S_ν	65 (15)	53 (6)	45 (15)	62 (6)	75 (15)
Id	71 (15)	74 (6)	52 (15)	82 (6)	66 (15)
sans minimisation	67 (14)	38 (49)	54 (14)	78 (49)	132 (15)
méthode RS	66 (15)	84 (6)	73 (15)	118 (6)	186 (15)
géométrique	–	21	25	33	49
1 niveau	101	100	229	397	754

TAB. 4.9 – Nombre d'itérations de COCG préconditionné (division du résidu par 10^{10}) — 2D, maillages structurés.

Chapitre 5

Conclusion et perspectives

Bilan

Tout d'abord, nous avons présenté des principes de modélisation mathématique et de discrétisation des problèmes de issus l'électromagnétisme, ainsi que des méthodes itératives de résolution de systèmes linéaires, en insistant particulièrement sur les algorithmes multiniveau.

Ensuite, partant de ces fondements, nous avons étudié les algorithmes multiniveau utilisés pour la résolution de systèmes linéaires issus de la méthode des éléments finis d'arête. Nous avons alors introduit une méthode de construction d'opérateurs de prolongement pour ces algorithmes multiniveau. Cette technique de construction d'opérateurs de prolongement s'appuie sur un problème de minimisation d'énergie avec la contrainte de vérifier une condition de compatibilité entre les fonctions d'arête et les fonctions nodales. Pour résoudre ce problème d'optimisation, deux approches sont décrites : l'une utilise les multiplicateurs de Lagrange et l'autre s'appuie sur la résolution d'une suite de problèmes de flot sur des graphes.

Les méthodes multiniveau utilisant les opérateurs de prolongement construits avec ces techniques présentent un comportement quasi-optimal pour le nombre d'itérations dans les cas les plus simples.

Perspectives

La première étape consiste à implémenter les algorithmes proposés dans ce manuscrit dans un code de calcul afin de tester de manière extensive leurs propriétés sur des applications réalistes 3D.

On peut aussi envisager pour la poursuite directe des travaux de ce manuscrit au moins deux voies :

- l'introduction d'autres techniques de construction du graphe grossier. Je pense ici notamment à des techniques proposant l'agglomération d'éléments pour définir une topologie grossière [46].
- dans les cas de problèmes difficiles, on peut aussi envisager d'optimiser des matrices de prolongement déjà construites et vérifiant la condition de compatibilité. En particulier, on peut penser aux bases provenant des méthodes multigrille géométriques qui ne sont pas nécessairement robustes vis-à-vis de variations importantes des paramètres du problème.

Un autre aspect de la poursuite de ces travaux concerne la prise en compte des spécificités de l'équation des ondes. Pour cela, on peut envisager :

- d'utiliser des lisseurs spécifiques pour l'équation des ondes. On pourra se servir notamment comme point de départ d'algorithme utilisant la méthode GMRES[1] comme lisseur pour l'équation de Helmholtz [28].
- de coupler la méthode multiniveau avec des méthodes de décomposition de domaines. On pourra se servir notamment des méthodes de Schwarz sans recouvrement avec conditions de transmission optimisées pour l'équation des ondes [34].
- d'incorporer la prise en compte des milieux ouverts avec des méthodes telles que les couches parfaitement absorbantes [11] ou des formulations intégrales.

[1] Méthode par sous-espaces de Krylov évoquée à la Sous-section 2.1.1.

Annexe A

Complément sur les espaces fonctionnels

L'étude mathématique des équations de Maxwell nécessite l'introduction d'*espaces de Sobolev*[1]. Soit Ω un domaine suffisamment régulier de $\mathbb{R}^3$ de frontière Γ. Le vecteur normal unitaire sortant du domaine sera notée $\mathbf{n}$.

$\mathrm{L}^2(\Omega)$ et $\mathrm{H}^1(\Omega)$

L'espace $\mathrm{L}^2(\Omega)$ est l'espace des fonctions de Ω à valeurs dans $\mathbb{C}$ et de carré intégrable. C'est un espace de Hilbert muni du produit scalaire : $(f,g) \mapsto \int_\Omega f\overline{g}$.

L'espace $\mathrm{H}^1(\Omega)$ est un sous-espace de $\mathrm{L}^2(\Omega)$. Le gradient des fonctions de cet espace est lui aussi de carré intégrable. C'est un espace de Hilbert muni du produit scalaire :

$$(u,v)_{\mathrm{H}^1(\Omega)} = \int_\Omega u\overline{v} + \int_\Omega \operatorname{grad} u \cdot \overline{\operatorname{grad} v}. \tag{A.1}$$

$\mathbb{L}^2(\Omega)$, $\mathbb{H}(\mathrm{rot},\Omega)$ et $\mathbb{H}(\mathrm{div},\Omega)$

Pour nos problèmes, il est aussi nécessaire d'introduire des espaces fonctionnels qui concernent plus particulièrement les champs de vecteurs. L'espace $\mathbb{L}^2(\Omega)$ est l'analogue de $\mathrm{L}^2(\Omega)$ pour les champs de vecteurs :

$$\mathbb{L}^2(\Omega) = \{\mathbf{V} : \Omega \to \mathbb{C}^3 / \int_\Omega \|\mathbf{V}\|^2 < +\infty\}. \tag{A.2}$$

Le symbole $(.,.)$ s'il n'y a pas plus de précisions désigne le produit scalaire sur $\mathbb{L}^2(\Omega)$ défini par : $(f,g) = \int_\Omega \sum_{i=1}^3 f_i\overline{g_i}$ et qui donne à $\mathbb{L}^2(\Omega)$ une structure d'espace de Hilbert. Sa norme est notée : $\|.\|_{\mathbb{L}^2(\Omega)}$.

Si l'on considère un système avec une énergie électromagnétique finie, on cherchera les solutions admissibles pour les champs $\mathbf{E}$ et $\mathbf{H}$ dans l'espace $\mathbb{H}(\mathrm{rot},\Omega)$. Cet espace apparaît de façon naturelle lors de la formulation faible du problème. Il est défini de la manière suivante :

$$\mathbb{H}(\mathrm{rot},\Omega) = \{\mathbf{E} \in \mathbb{L}^2(\Omega) \ / \ \operatorname{rot}\mathbf{E} \in \mathbb{L}^2(\Omega)\} \tag{A.3}$$

Cet espace de Hilbert est muni du produit scalaire :

$$(\mathbf{E},\mathbf{E}')_{\mathbb{H}(\mathrm{rot},\Omega)} = (\mathbf{E},\mathbf{E}') + (\operatorname{rot}\mathbf{E}, \operatorname{rot}\mathbf{E}') \tag{A.4}$$

dont la norme induite est :

$$\|\mathbf{E}\|_{\mathbb{H}(\mathrm{rot},\Omega)} = (\|\mathbf{E}\|^2_{\mathbb{L}^2(\Omega)} + \|\operatorname{rot}\mathbf{E}\|^2_{\mathbb{L}^2(\Omega)})^{\frac{1}{2}} \tag{A.5}$$

D'autre part, nous nous intéressons ici à des problèmes avec conditions aux limites et il est donc nécessaire de savoir quel *type de conditions* a un sens mathématique dans $\mathbb{H}(\mathrm{rot},\Omega)$. Pour cela nous disposons du théorème de trace suivant :

[1] Espaces de Sobolev : Espaces idéaux pour la formulation faible des problèmes issus d'équations aux dérivées partielles.

Théorème 1.1. *L'application* $\gamma_t : \mathbf{E} \mapsto \mathbf{E} \times \mathbf{n}$ *définie sur* $D(\Omega)^3$ [2] *peut être prolongée par continuité en une application linéaire et continue de* $\mathbb{H}(\mathrm{rot}, \Omega)$ *dans* $\mathbb{H}^{-\frac{1}{2}}(\Gamma)$ *(voir [21] pour des précisions sur cet espace) de norme 1.*

Autre point important pour les formulations variationnelles, on peut définir l'analogue de la formule de Green sur l'espace $\mathbb{H}(\mathrm{rot}, \Omega)$:

$$\begin{aligned} &\forall \mathbf{E} \in \mathbb{H}(\mathrm{rot}, \Omega),\ \forall \boldsymbol{\eta} \in \mathbb{H}^1(\Omega) = \left\{ \boldsymbol{\eta} \in \mathbb{L}^2(\Omega) \ /\ \frac{\partial \eta_i}{\partial x_j} \in \mathbb{L}^2(\Omega)\ \forall i, j \in \{1, \dots 3\} \right\}, \\ &(\mathrm{rot}\, \mathbf{E}, \boldsymbol{\eta}) - (\mathbf{E}, \mathrm{rot}\, \boldsymbol{\eta}) = (\mathbf{E} \times \mathbf{n}, \boldsymbol{\eta})_{\mathrm{L}^2(\Gamma)} \end{aligned} \tag{A.6}$$

Ces résultats sont cohérents avec les conditions aux limites et d'interface que nous avons identifiées Sous-section 1.1.3.

Nous pourrions définir de la même manière les propriétés de l'espace $\mathbb{H}(\mathrm{div}, \Omega)$:

$$\mathbb{H}(\mathrm{div}, \Omega) = \{ \boldsymbol{\eta} \in \mathbb{L}^2(\Omega) \ /\ \mathrm{div}\, \boldsymbol{\eta} \in \mathrm{L}^2(\Omega) \} \tag{A.7}$$

C'est un espace naturel pour définir les champs $\mathbf{B}$ et $\mathbf{D}$.

[2] $D(\Omega)^3$ désigne l'ensemble des fonctions de $C^\infty(\Omega)^3$ à support compact.

Annexe B

Préconditionnement à un niveau

B.1 Un préconditionneur efficace pour les systèmes linéaires provenant de la méthode des éléments finis pour des problèmes de diffraction

An efficient preconditioner for linear systems issued from the Finite Element Method for scattering problems[1]

Ronan Perrussel, Laurent Nicolas, François Musy

Abstract. *An efficient preconditioner for systems issued from the finite element discretization of time harmonic Maxwell's equations with absorbing boundary conditions is presented. It is based on the Helmholtz decomposition of the electromagnetic field and its discrete counterpart. It is compared to a classical preconditioner on both simple and realistic problems. Its behaviour is also evaluated on meshes showing different characteristics.*

B.1.1 Introduction

Electromagnetic scattering problems are classically modeled using time harmonic Maxwell's equations with Silver-Müller conditions [25]. The numerical solution of these equations leads to complex and symmetric matrices. To solve these systems, Krylov subspace methods may be used: BiCGCR [24], symmetric QMR [32] or COCG [75]. Classical preconditioning methods are implemented in order to accelerate the convergence of these iterative algorithms: SSOR, incomplete Cholesky factorization [52],... An efficient preconditioner based on the Helmholtz decomposition was previously proposed for simple eddy currents problems [8]. The aim of this paper is to test its efficiency on realistic scattering problems and its robustness on meshes with different characteristics.

The problem formulation is first given. The preconditioner based on the Helmholtz decomposition is then described. Numerical results are finally presented.

B.1.2 Problem Formulation

This work deals with time harmonic Maxwell's equations and Silver-Müller conditions. The following finite element formulation, with the incomplete first order edge elements [41] on the domain Ω (space $\mathbf{Q}_h$), can be written (for the electric field $\mathbf{E}$ here):

$$\begin{aligned} &\text{Find } \mathbf{E} \text{ in } \mathbf{Q}_{h,\Gamma_d} \text{ such that:} \\ &a(\mathbf{E},\mathbf{E}') = \mathbf{F}(\mathbf{E}') \quad \forall \mathbf{E}' \in \mathbf{Q}_{h,\Gamma_d}, \\ &\text{with } a(\mathbf{E},\mathbf{E}') = \int_\Omega \frac{1}{\mu} \operatorname{rot} \mathbf{E} \cdot \operatorname{rot} \overline{\mathbf{E}'} + \mathrm{i} \int_{\Gamma_a} \frac{1}{\mu} \|\mathbf{k}\| (\mathbf{E} \times \mathbf{n}) \cdot (\overline{\mathbf{E}'} \times \mathbf{n}) - \omega^2 \int_\Omega \tilde{\varepsilon} \mathbf{E} \cdot \overline{\mathbf{E}'}, \end{aligned} \tag{B.1}$$

[1] Article publié dans IEEE Trans. on Mag., 40(2), 2004; voir [64].

ω denotes the angular frequency, $\mathbf{k}$ the wave vector, $\mathbf{n}$ the boundary normal direction, $\tilde{\varepsilon}$ the complex-valued permittivity, μ the permeability, $\mathbf{F}(\mathbf{E}')$ the source term (incident plane wave), Γ_a the absorbing boundary, Γ_d the perfect electric conductor boundary ($\mathbf{E} \times \mathbf{n} = 0$ on Γ_d). The formulation space of the problem is defined as:

$$\mathbf{Q}_{h,\Gamma_d} = \{\mathbf{E}' \in \mathbf{Q}_h \ / \ \mathbf{E}' \times \mathbf{n} = 0 \text{ on } \Gamma_d\}.$$

Essential characteristics of this formulation are:

- the kernel of the rot operator ($\{\mathbf{E},\ \mathrm{rot}(\mathbf{E}) = 0\}$) is of infinite dimension,
- the sesquilinear form a is not hermitian,
- the operator's spectrum has eigenvalues with positive and negative real parts, the sesquilinear form a is therefore indefinite.

The linear system $Ax = b$ to be solved is complex-valued, symmetric and indefinite. These characteristics are essential for the choice of solving methods.

B.1.3 An efficient preconditioner

Classical solving methods are adapted to the operator gradient and deal badly with the kernel of the rot operator. Following the Helmholtz decomposition, the electric field $\mathbf{E}$ or the magnetic field $\mathbf{H}$ can be decomposed into two components [41]:

$$\mathbf{E} = \mathbf{E}_s \oplus^{\perp} \mathrm{grad}\, \phi \tag{B.2}$$

where:

- $\oplus^{\perp}$ means the orthogonal sum for the scalar product in the square-integrable functions space,
- grad ϕ is a static component with ϕ a scalar potential. It belongs to the kernel of the curl operator; it is the orthogonal projection on the kernel,
- $\mathbf{E}_s$ is a propagation component called solenoidal component. It is divergence-free, because the decomposition is orthogonal.

The rot operator has a dissymmetric behaviour on these components [41]. The decomposition's discrete counterpart in $\mathbf{Q}_h$ is of practical importance:

$$\mathbf{E}_h = \mathbf{E}_{s,h} \oplus^{\perp} \mathrm{grad}\, \phi_h \tag{B.3}$$

where:

- $\mathbf{E}_h$ belongs to the first incomplete order edge element space $\mathbf{Q}_h$,
- ϕ_h belongs to the first order nodal element space N_h.

Since the equation B.1 with $\mathbf{E} = \mathrm{grad}\, \phi_h$ and $\mathbf{E}' = \mathrm{grad}\, \phi'_h$ gives:

$$a(\mathrm{grad}\, \phi_h, \mathrm{grad}\, \phi'_h) = -\omega^2 \int_{\Omega} \tilde{\varepsilon}\, \mathrm{grad}\, \phi_h \cdot \mathrm{grad}\, \phi'_h + \mathrm{i} \int_{\Gamma_a} \frac{1}{\mu} \|\mathbf{k}\| (\mathrm{grad}\, \phi_h \times \mathbf{n}) \cdot (\mathrm{grad}\, \phi'_h \times \mathbf{n}). \tag{B.4}$$

the existence of the scalar potential ϕ_h enables to consider an auxiliary problem. The SSOR preconditioner has shown to be efficient for this secondary problem (issued from the laplacian operator with specific boundary conditions) [5].

For the implementation, a practical operator to transfer potential representation in the space N_h to the field space $\mathbf{Q}_h$ is required. Its construction uses the definition of the degrees of freedom (dof) which are: $\int_e \mathrm{grad}\, \phi_h \cdot \mathbf{t}$ on each edge e of the mesh for the space $\mathbf{Q}_h$, the values on each vertex for the space N_h. The expression of this operator G for a mesh T_h is then issued from the relation for a edge e:

$$\underbrace{\int_{\mathbf{x}_{init}}^{\mathbf{x}_{final}} \mathrm{grad}\, \phi_h \cdot \mathbf{t}}_{\text{edge element dof}} = \underbrace{\phi_h(\mathbf{x}_{final})}_{\text{nodal element dof}} - \underbrace{\phi_h(\mathbf{x}_{init})}_{\text{nodal element dof}} \tag{B.5}$$

matrix	A	A_ϕ	G
Number of non-zeros	$96m_n$	$13m_n$	$12m_n$

Table B.1: Number of non-zeros entries for each matrix. m_n nb of nodes.

where $\mathbf{x}_{init}$ and $\mathbf{x}_{final}$ are the extremities of the edge e and $\mathbf{t}$ the tangential vector to e. The global relation $\{\text{dof}(\text{grad}\,\phi_h)\} = G\{\text{dof}(\phi_h)\}$ defines the searched operator G as a sparse matrix with exactly 2 non-zero elements per line: 1 and -1 respectively for the last and first node of each edge.

With this operator, the matrix for the auxiliary problem can be assembled by a Galerkin product: $A_\Phi = G^T A G$ where A is the edge elements matrix. The numerical cost of this assembly is roughly equivalent to 4 matrix/vector products with A. It can be neglected in comparison with the numerical solving cost (Table B.1).

Once these elements defined, the algorithm of the preconditioning method can be written (Fig. B.1). Note that this method should be incorporated in an iterative solver (COCG, QMR, BiCGCR,...). The preconditioning operation simply consists in transforming the residual r into a preconditioned one by solving a linear system $Mg = r$ of reduced numerical costs (M is not necessarily assembled like here).

The cost of one Gauss-Seidel iteration with a matrix is directly linked to its number of non-zeros entries (nnz). The cost of our preconditioner is then a direct function of the nnz of the A, A_ϕ and G matrices.

The approximative nnz in each matrix is given in Table B.1. It is evaluated with [50] and practical estimations with the test problems:

$$\begin{aligned} &\text{nnz}(A) = m_e + 2*(3*m_f) + 6*m_t, \text{nnz}(A_\phi) = m_n + 2*m_e, \\ &\text{nnz}(G) = 2*m_e, \text{ with } \frac{m_e}{m_n} \approx 6,\ \frac{m_f}{m_n} \approx 10 \text{ and } \frac{m_t}{m_n} \approx 5. \end{aligned} \tag{B.6}$$

m_n, m_e, m_f and m_t are respectively the number of nodes, edges, faces and tetrahedral elements in the mesh. It indicates the overcost in terms of memory requirement and supplementary matrix/vector products due to G and A_ϕ. Comparing to classical SSOR, it roughly doubles the preconditioning time.

Solve $Mg = r$, g_Φ refers to the potential part of g.

1. $g \leftarrow 0$, $g_\phi \leftarrow 0$
2. γ forward Gauss-Seidel on $A_\phi g_\phi = G^T r$
3. $g \leftarrow g + G g_\phi$
4. symmetric Gauss-Seidel on $Ag = r$
5. $g_\phi \leftarrow 0$
6. γ backward Gauss-Seidel on $A_\phi g_\phi = G^T(r - Ag)$
7. $g \leftarrow g + G g_\phi$

Figure B.1: One iteration of the preconditioning algorithm using the Helmholtz decomposition. Generally $\gamma = 1$ or 2.

B.1.4 Mesh quality

An intrinsic mesh quality cannot be defined, since it depends on the physical problem being modeled [33]. Different criteria can be considered: good precision, well-conditioned problem...

For the Poisson equation, it was previously shown that the conditioning number of the matrix is linked to the shape of each element and the mesh uniformity [71]. By extension, two parameters to qualify tetrahedron shape and mesh uniformity are used for this formulation:

- The ratio of inscribed ρ_{ins} over circumscribed ρ_{circ} sphere radius with a normalizing factor is used to evaluate tetrahedron shape:

$$\rho = 3\frac{\rho_{\text{ins}}}{\rho_{\text{circ}}}. \tag{B.7}$$

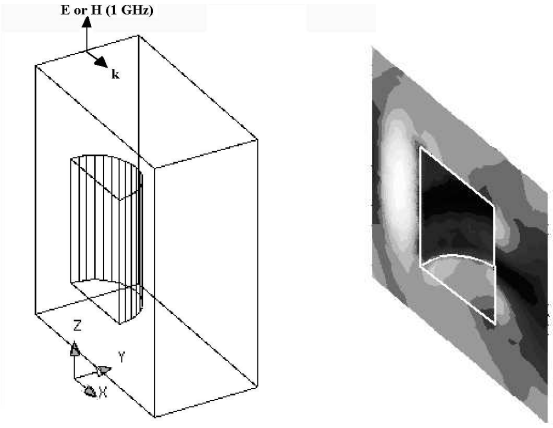


Figure B.2: An incident plane wave on a 3D cylinder.

This ratio equals 1 if the tetrahedron is regular and decreases to zero if the tetrahedron is fully degenerated.

- The ratio between the largest and the smallest volume of tetrahedra is used to measure the uniformity.

B.1.5 Numerical results: Efficiency and robustness

Three kinds of comparisons with classical solvers are implemented to test the efficiency of the Helmholtz decomposition preconditioner:

- by increasing the number of degrees of freedom of a given problem,
- by analysing the influence of the mesh quality,
- by computing two realistic problems.

Number of degrees of freedom

A 1GHZ plane wave scattered by a 3D cylinder is studied (Fig. B.2). From Fig. B.3, it is shown how the number of iterations evolves with the number of degrees of freedom (dof) for the four implemented solving methods: 3 solvers (COCG, BiCGCR, QMR) with SSOR preconditioning, and a COCG solver with the Helmholtz decomposition preconditioner. Table B.2 gives the corresponding CPU times. Here, COCG is the fastest classical solver with SSOR preconditioning. Consequently in the following, only the results with a COCG solver are analysed. The Helmholtz decomposition preconditioner needs roughly three times fewer iterations and half the CPU time.

Number of dofs	84 385	153 293	256 121	392 524
QMR - SSOR	2 215	5 341	9 807	19 735
COCG - SSOR	1 862	4 433	8 408	17 175
BiCGCR - SSOR	2 395	5 214	10 364	22 264
COCG - Helmholtz solving	1 108	2 312	4 447	7 981
assembling preconditioner	3	5	8	11

Table B.2: Comparison of CPU time (s) for the 3D cylinder.

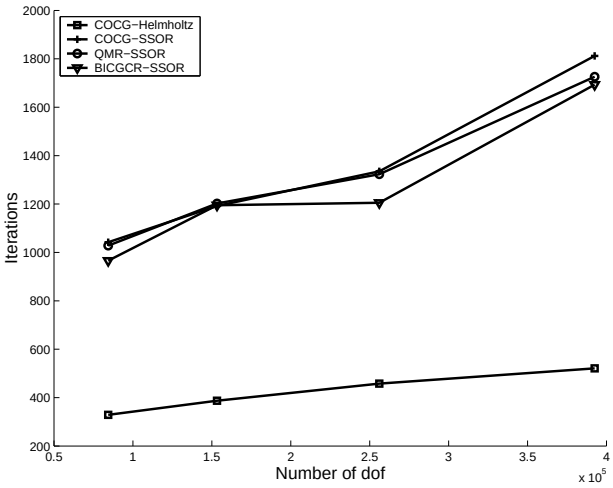


Figure B.3: Number of iterations against the number of dof.

Quality of the mesh

The influence of the quality of the mesh is tested on the 3D cylinder problem. Mean shape ratio and uniformity of two different meshes are evaluated on this problem (Table B.3). The ratio $\frac{\lambda}{h_{max}}$ is also given, where λ is the wavelength and h_{max} the length of the longest edge in the mesh.

Meshes	ρ_{mean}	volume ratio	$\frac{\lambda}{h_{max}}$
mesh 1 463 213 elem.	0.795	77.2	9.2
mesh 2 63 565 elem.	0.784	8 682	0.95

Table B.3: Shape ratio and uniformity for two meshes.

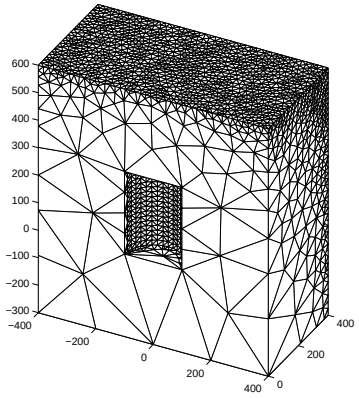


Figure B.4: Lack of uniformity for mesh 2.

The mean shape ratio is equivalent for both meshes. The main difference is concerning uniformity: mesh 2 (Fig. B.4) is less uniform than mesh 1.

In Table B.4, the influence of the quality of the mesh on the convergence is illustrated. The convergence is greatly slowed for mesh 2. The effect is significant even with the Helmholtz decomposition

preconditioner. However, it is largely more robust than the SSOR preconditioner, which does not converge after 8000 iterations.

Meshes		COCG-Helmholtz	COCG-SSOR
mesh 1	iter.	590	1 582
557539 dof	CPU (s)	14 690	21 603
mesh 2	iter.	1 400	$> 8\ 000$
72 229 dof	CPU (s)	4 935	

Table B.4: CPU time and iterations for two meshes.

The ratio $\frac{\lambda}{h_{max}}$ is less than 1 in mesh 2. However a mean of 10 nodes per wavelength is necessary to correctly discretize the wave equation. Note that the physical validity of this discretization is doubtful. The influence of this ratio is tested on the mesh 2 by reducing the frequency of the incident wave, which leads to increase the ratio $\frac{\lambda}{h_{max}}$. Fig. B.5 shows the number of iterations as a function of the frequency for both solvers. Table B.5 presents corresponding CPU times. In the considered frequency band, both

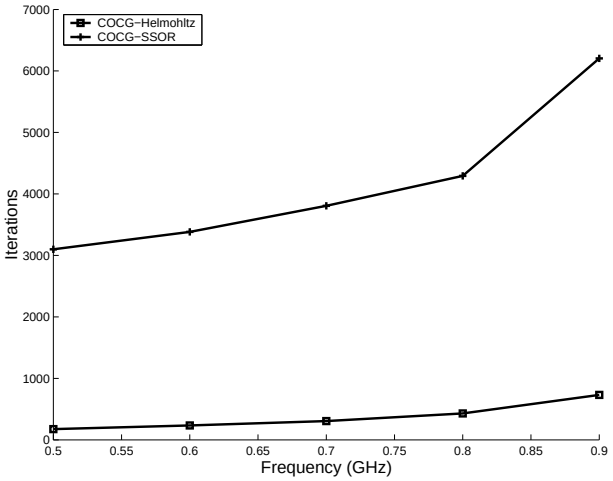


Figure B.5: Number of iterations against the frequency.

solvers converge. Obviously, the Helmholtz decomposition preconditioner performs better than the SSOR preconditioner and is less sensitive to the ratio $\frac{\lambda}{h_{max}}$.

Realistic problems

The efficiency of the Helmholtz decomposition preconditioner is observed on two realistic problems. In the first problem (Fig. B.6), the electric field due to a RF source is computed inside a human body during an hyperthermia treatment [72]. The second problem models (Fig. B.7) an airplane illuminated by a plane wave [77]. The Helmholtz decomposition preconditioner shows its efficiency in both cases (Table B.1.5), more particularly in the hyperthermia case, for which COCG-SSOR did not converge after 8000 iterations.

Frequencies (GHz)	0.5	0.7	0.9
$\frac{\lambda}{h_{max}}$	1.9	1.36	1.06
COCG-SSOR	5 894	6 425	7 225
COCG-Helm.	612	816	1 044

Table B.5: CPU time (s) function of frequencies - mesh 2.

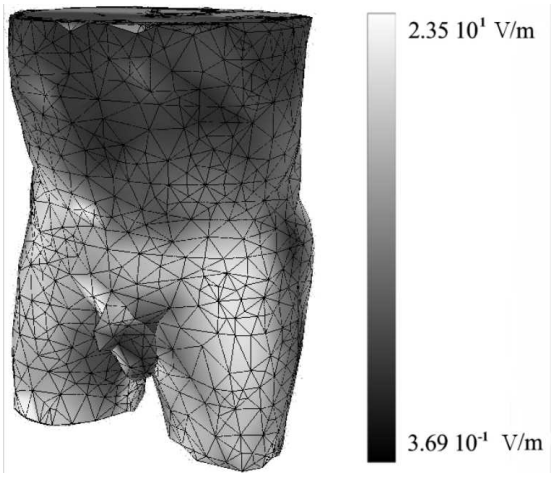


Figure B.6: Hyperthermia RF (27MHz) for treating deep tumours - Magnitude of the electric field.

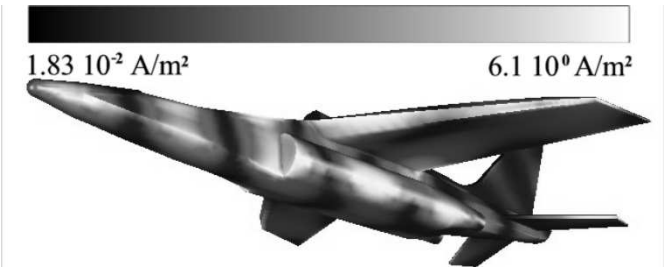


Figure B.7: Illumination of a plane by a 100MHz plane wave - Magnitude of the current density.

Problems	Hyperthermia	Plane
Number of dofs	202 701	574 151
COCG - SSOR		
Number of iterations	> 8000	2 890
CPU time(s)		40 096
COCG - Helmholtz		
Number of iterations	474	1 210
CPU time(s)	3 362	26 320

Table B.6: CPU time and iterations for two realistic problems.

B.1.6 Conclusion

A preconditioner based on the Helmholtz decomposition has been developed for scattering problems. This method is efficient because well adapted to the rot operator. The robustness has been tested on a non-uniform mesh. Its efficiency has been evaluated on realistic problems. Furthermore it is simple to implement and requires only a light overcost.

B.2 Préconditionneurs pour des problèmes de diffraction résolus par la méthode des éléments finis

Preconditioners for finite element method in scattering problems[2].

FRANÇOIS MUSY, LAURENT NICOLAS, RONAN PERRUSSEL AND MICHELLE SCHATZMAN

ABSTRACT. *A comparison of preconditioners for systems produced by the finite element discretization of time harmonic Maxwell's equations with absorbing boundary conditions is presented. This comparison is based on the asymptotic behavior of the resolution of test problems; the geometry is fixed, but we change parameters, such as the formulation or the characteristics of the medium. The methods under consideration are a classical SSOR and one-level multiplicative Schwarz preconditioners drawn from previous works by several authors on efficient solvers for edge finite element and Maxwell's equations.*

B.2.1 Introduction

In order to deal with electromagnetic scattering problems, we compute electric ($\mathbf{E}$) or magnetic ($\mathbf{H}$) field using time harmonic Maxwell's equations on a domain Ω. For instance, in the $\mathbf{E}$-field formulation, it gives:

$$\mathrm{rot}\,\frac{1}{\mu}\,\mathrm{rot}\,\mathbf{E} - \omega^2\varepsilon\mathbf{E} + \mathrm{i}\omega\sigma\mathbf{E} = < \text{source} > \text{ on } \Omega. \tag{B.8}$$

where μ denotes the magnetic permeability, ε the dielectric permittivity, σ the conductivity and ω the pulsation. The source term might be an incident plane wave, an antenna...

Furthermore, electromagnetic waves are often considered in free space.

To represent the infinite domain, we impose Silver-Müller conditions on the boundary, which are absorbing to the first order:

$$\mathrm{rot}\,\mathbf{E}\times\mathbf{n} = -\mathrm{i}\omega\sqrt{\varepsilon\mu}\mathbf{n}\times(\mathbf{E}\times\mathbf{n}) \text{ on } \partial\Omega \ (\mathbf{n} \text{ unit normal on } \partial\Omega). \tag{B.9}$$

Any finite element discretization of the boundary value problem (B.8) + (B.9) leads to a system whose matrix is complex-valued and symmetric. Here, we use Nédélec lowest order edge element [62].

Edge elements have become generally accepted for three-dimensional field formulation in electromagnetism, because of their quality and reliability. We chose the lowest degree elements for simplicity and ease of programming. This choice has a drawback: we need at least 10 elements per wave-length, and more if the wave number $\omega\sqrt{\varepsilon\mu}$ is high (see [45]).

In addition, since the Silver-Müller condition is absorbing only at lowest order, the boundary has to be kept at sufficiently large distances from the "interesting" part of the domain of integration (see [29]).

Therefore, solving large sparse systems with roughly 10^5 unknowns is required. To overcome this difficulty, we use Krylov subspace methods with preconditioners: this critical point is the purpose of this paper.

In a first part, we detail the one-level multiplicative Schwarz preconditioners. Then, we study the numerical behavior of geometrically identical test problems: a plane wave scattered by a cylinder; we vary the parameters: boundary conditions ($\mathbf{E}$ or $\mathbf{H}$ field), media characteristics...

[2] Acte de la conférence ECCOMAS 2004; voir [65]

B.2.2 Preconditioners

The algorithms studied are one-level multiplicative Schwarz preconditioners of two different kinds. The first type is the SSOR smoother; the second type is made out of smoothers from multigrid techniques, and more precisely:

- the smoother of Arnold et al. [4],
- the smoother of Hiptmair [41].

It must be observed that we do *not* use a multigrid algorithm, but only the smoother on the finest grid, i.e. the available grid.

Schwarz multiplicative algorithm

Let us recall the definition of a Schwarz multiplicative algorithm. Q is the discretization space chosen for the formulation; its choice leads to solve a system $Ax = b$. The decomposition of Q into a not necessarily direct sum is considered:

$$Q = \sum_{i=1}^{m} Q_i. \tag{B.10}$$

For each Q_i, a projection P_i from Q to Q_i and an operator $A_i = P_i A P_i^t$ are associated.

Let r be the residual at a given stage of a Krylov subspace method. The non symmetrized preconditioning step, which from r maps to g, is:

$g \leftarrow 0$
for $i = 1 : m$
  $e_i \leftarrow P_i(r - Ag)$
  $y_i \leftarrow A_i^{-1} e_i$
  $g \leftarrow g + P_i^t y_i$
end.

$\mathcal{M}\{A, (P_i)_{i\in\mathcal{I}}\}$ denotes the operator which from r maps to g. $\mathcal{I}$ is the *ordered* set $\{1, \ldots, m\}$ and the notation $(P_i)_{i\in\mathcal{I}}$ represents an ordered sequence of projections.

The indices might as well have been sorted in the opposite order; the set of indices would be then $\bar{\mathcal{I}} = \{m, m-1, \ldots, 1\}$ leading to an operator $\mathcal{M}\{A, (P_i)_{i\in\bar{\mathcal{I}}}\}$.

The symmetrized method corresponds to the choice of index set:

$$\tilde{\mathcal{I}} = \{1, \ldots, m-1, m, m-1, \ldots, 1\} \tag{B.11}$$

and the operator of one iteration step is $\mathcal{M}\{A, (P_i)_{i\in\tilde{\mathcal{I}}}\}$.

These notations enable us to describe the preconditioning methods. In what follows, $\mathbf{w}_e$ denotes the shape function of the edge element indexed by e and ϕ_v the shape function of the nodal element indexed by v. The whole set of edge indices in initial numbering is noticed $\mathcal{E}$ and of nodal indices $\mathcal{V}$. With these notations, $\mathbf{Q} = \text{vect}\{\mathbf{w}_e,\ e \in \mathcal{E}\}$.

Description of preconditioners

SSOR $\mathbf{Q}_e$ is the one-dimensional space spanned by the function $\mathbf{w}_e$ and in case of a unit relaxation parameter, SSOR is an iterative method of operator $\mathcal{M}\{A, (\mathbf{Q}_e)\}_{e\in\tilde{\mathcal{E}}}$. We have tried other relaxation parameters, which do not give better results in term of computation time.

Arnold, Falk and Winther $\mathbf{Q}_v = \text{vect}\{\mathbf{w}_e \in \mathbf{Q},\ \text{vertex}\ \ v \text{ is one extremity of } e\}$. We apply a symmetrized algorithm whose operator is $\mathcal{M}\{A, (\mathbf{Q}_v)_{v\in\tilde{\mathcal{V}}}\}$.

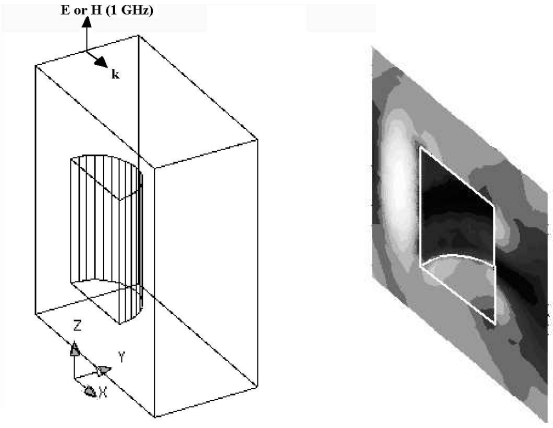


Figure B.8: An incident plane wave on a 3D PEC cylinder, **H**-formulation.

Hiptmair In addition to $\mathbf{Q}_e$, we define a subspace N_v which is spanned by $\operatorname{grad}\phi_v$ for v running over the indices of nodal elements. G is the node-edge incidence matrix; G maps $\mathcal{V}$ to $\mathbf{Q}$ and its transpose maps $\mathbf{Q}$ to $\mathcal{V}$. The matrix A_ϕ is $G^t A G$.

The algorithm of the smoother used as a preconditioner, which from r maps to g, is the following:

$g \leftarrow 0$, $g_\phi \leftarrow 0$
forward Gauss-Seidel on $A_\phi g_\phi = G^t r$
$g \leftarrow g + G g_\phi$
symmetric Gauss-Seidel on $Ag = r$
$g_\phi \leftarrow 0$
backward Gauss-Seidel on $A_\phi g_\phi = G^t(r - Ag)$
$g \leftarrow g + G g_\phi$

The corresponding operator is (Id represents the identity operator):

$$\mathcal{N} + G\mathcal{M}\{A_\phi, (N_v)_{v\in\bar{\mathcal{V}}}\}G^t\big(\mathrm{Id} - A\mathcal{N}\big) \qquad \text{(B.12)}$$
$$\text{with } \mathcal{N} = \mathcal{M}\{A, (\mathbf{Q}_e)_{e\in\bar{\mathcal{E}}}\}(\mathrm{Id} - AG\mathcal{M}\{A_\phi, (N_v)_{v\in\mathcal{V}}\}G^t).$$

B.2.3 Numerical behavior

Scattering problem

A 1GHZ plane wave scattered by a 3D cylinder is studied (Fig. B.8). The cylinder has a length of 300mm and a radius of 100mm. The distance between the bounding box and the cylinder is 300mm. As the cylinder is surrounded by vacuum, this represents a distance of λ ($\lambda = \frac{2\pi}{\omega\sqrt{\varepsilon_0\mu_0}}$ is the wavelength) from scatterer to absorbing boundary.

We can take into account the variation of different parameters:

- formulation: electric field **E** or magnetic **H** (implies distinct boundary conditions on perfect electric conductor),
- type of material: perfect electric conductor, dielectric or conducting dielectric cylinder.

Solvers

Conjugate Orthogonal Conjugate Gradient [75] is used as a Krylov subspace method. This method (very similar to classical Conjugate Gradient) is adapted to systems with symmetric complex-valued matrix. To solve this type of matrices, other Krylov subspace methods could be used, but the really critical question here is the choice of preconditioner (for comparison between a few methods see [64]).

The stopping criterion for the iterative methods is $\|r\|_2 \leq 10^{-6}$ where r is the actual residual of the iterative method.

Convergence results

We use *unstructured* meshes with increasing number of degrees of freedom (d.o.f.). First, the evolution of the number of iterations against the number of d.o.f. is given in the different test cases. Then the results are gathered to estimate numerically asymptotic behaviors.

PEC cylinder

On the PEC surface, distinct formulations correspond to distinct boundary conditions:

- for $\mathbf{H}$, the boundary condition is natural i.e. $\mathrm{rot}\,\mathbf{H} \times \mathbf{n} = 0$;
- for $\mathbf{E}$, the boundary condition is essential i.e. $\mathbf{E} \times \mathbf{n} = 0$. Thus, we have slightly fewer unknowns.

On the Fig. B.9 and B.10, the number of d.o.f. versus the number of iterations for both formulations is shown. The evolutions seems quite similar.

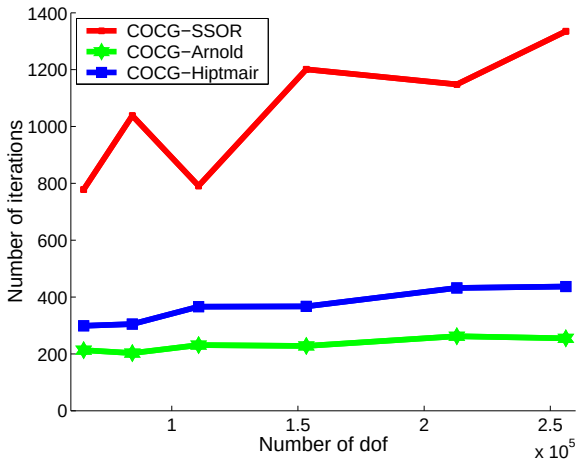


Figure B.9: $\mathbf{H}$-formulation.

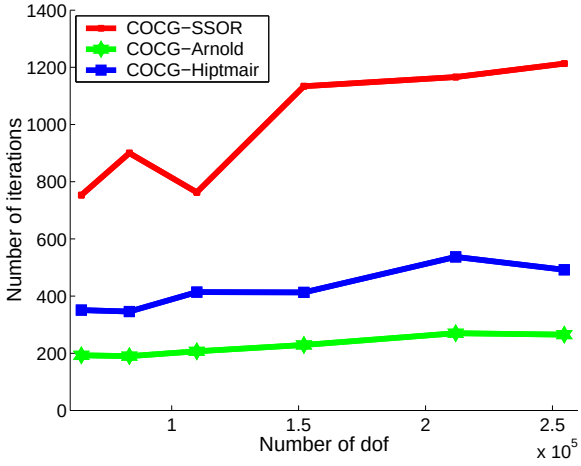


Figure B.10: $\mathbf{E}$-formulation.

Dielectric cylinder

We consider here two different types of dielectric cylinders. One is insulating (called also perfect dielectric); the other is conducting.

In the perfect dielectric, the interesting medium parameter is the permittivity. Here, the value $2\varepsilon_0$ is used (ε_0 corresponds to vacuum value). The electric phenomenon to take into account (related to the permittivity) is the displacement current.

In the conducting dielectric, both displacement currents and eddy currents are existing. One must pay attention to the skin depth ($\delta = \sqrt{\frac{2}{\omega\sigma\mu}}$) which is the distance from the material surface where the major part of the current density is concentrated. In this case, the permittivity is equal to $2\varepsilon_0$ and the conductivity σ to $0.04\mathrm{S.m}^{-1}$, leading to $\delta = 89.2\mathrm{mm}$; this skin depth is compatible with the dimensions of the cylinder.

On the Fig. B.11 et B.12, the number of d.o.f. versus the number of iterations for both kinds of dielectrics is shown. Evolutions are also quite similar. Convergence is slightly faster in the conducting case.

Asymptotic behaviors

We are looking for an asymptotic evolution linking the number of iterations required to solve the system at a given precision and the number of d.o.f. in the problem. Previous results for the laplacian (see [5])

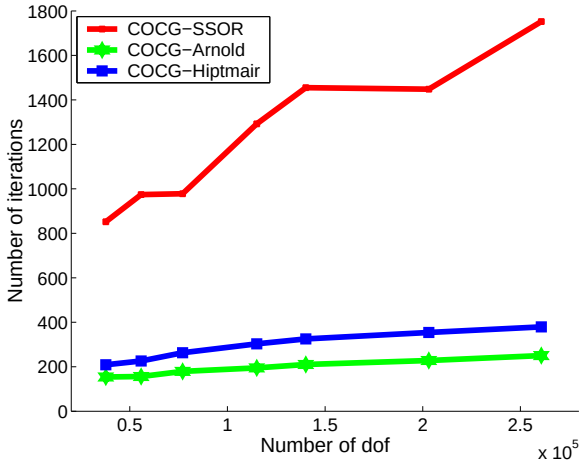


Figure B.11: Perfect dielectric.

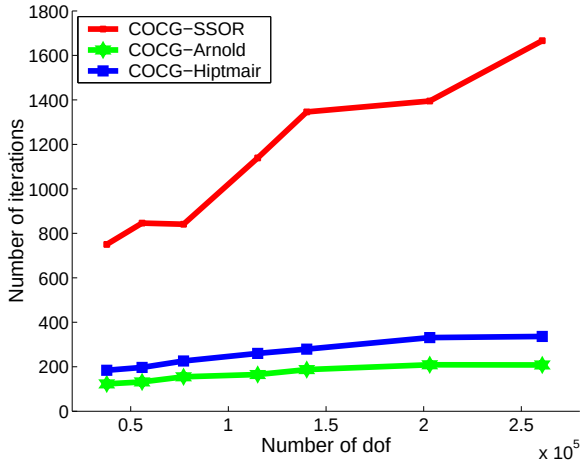


Figure B.12: Conducting dielectric.

leads us to consider a relation of the form:

$$\text{number of iterations} = C(\text{number of d.o.f.})^{\alpha}. \tag{B.13}$$

The computation results obtained previously are used. Least-squares method is applied to estimate numerically C and α from a $\log - \log$ representation.

	PEC / **H**		PEC / **E**		perfect dielectric		cond. dielectric	
	α	C	α	C	α	C	α	C
COCG - without prec.			0.6	4.9			0.58	7
COCG - SSOR	0.34	18	0.36	14	0.37	16.5	0.42	8
COCG - Arnold	0.17	32	0.28	8.5	0.26	9.3	0.28	6.5
COCG - Hiptmair	0.3	11	0.31	11.5	0.32	6.9	0.32	6

The coefficient α is roughly divided by two between COCG without preconditioning and COCG preconditioned by the methods of Arnold or Hiptmair. Differences between behaviors of preconditioners are less visible but it can be observed that:

- for each test cases, the preconditioner of Arnold has the best asymptotic behavior,
- the preconditioner of Hiptmair is also better than the classical SSOR,
- C is generally higher for SSOR than for the other preconditioners.

Although this is not a matter for concern here, let us put these results in perspective with estimates of the computational time: the cost of one iteration is higher for non-classical preconditioners but the reduction in the number of iterations is sufficiently important and therefore the total computational cost is reduced (see [64]).

B.2.4 Conclusion

The preconditioners of Arnold et al and Hiptmair entail better convergence for the test cases than classical SSOR. Moreover, as simple one-level preconditioners, they are relatively easy to implement, particularly the algorithm of Hiptmair. Of course, if it is possible to use a grid hierarchy to solve the problem, the multilevel version using these smoothers leads to optimal or quasi-optimal methods as is shown in [36].

Appendix C

Construction de fonctions nodales grossières par minimisation d'énergie

Energy-minimising construction of coarse nodal elements for multilevel methods[1]

Two-level algebraic methods are introduced for solving linear systems coming from a finite element discretization.

Two methods are considered: one takes into account the boundary conditions, the other does not. Indeed, the choice of the coarse level seems to be sensitive to the presence of Dirichlet boundary conditions.

A simple problem is considered on an elementary geometry. The strong formulation of the test case is:

$$\begin{cases} -\triangle u = f \text{ in } \Omega =]0;1[\times]0;1[, \\ \quad u = 0 \text{ on } \partial\Omega, \\ \quad f = \pi^2 \sin(\pi x)\sin(\pi y). \end{cases} \tag{C.1}$$

The weak formulation is the following:

$$\begin{cases} \text{To find } u \in \mathrm{H}_0^1(\Omega) \text{ such that:} \\ \displaystyle\int_\Omega \operatorname{grad} u \cdot \operatorname{grad} v = \int_\Omega fv \quad \forall v \in \mathrm{H}_0^1(\Omega). \end{cases} \tag{C.2}$$

C.1 Principle of the method

An unstructured mesh is given on the domain Ω where a discretization by P_1 finite elements is used.

In the two-level method proposed for solving the linear system, the main point is the construction of the coarse basis. This basis must satisfy the following constraints:

- the basis must span a subspace of the initial P_1 finite element space;
- the support of every coarse basis function is a subdomain Ω_i of Ω. Ω_i is defined so that the matrix at the coarse level has a very regular structure. This regular structure will be exhibited later;
- Given the supports of these coarse functions, an optimisation problem is solved to compute them; it consists in minimising the energy of the coarse basis under appropriate constraints.

[1] Extrait du rapport d'avancement de Septembre 2004; voir [60].

C.1.1 Formulation of the optimisation problem

Suppose that Ω is decomposed into overlapping subdomains $(\Omega_i)_{i=1,\dots,d_H}$ and that each subdomain is not completely overlapped by its neighbours. $(I_i)_{i=1,\dots,d_H}$ denotes the set of nodes belonging to each subdomain Ω_i.

Let $(\Phi_i)_{i=1,\dots,d_H}$ be the coarse basis functions. This basis is the solution of the following problem:

$$\begin{cases} \text{To find } (\Phi_i)_{i=1,\dots,d_H} \text{ minimising } \displaystyle\sum_{i=1}^{d_H} a(\Phi_i,\Phi_i) \text{ under the constraints:} \\ \displaystyle\sum_{j=1}^{d_H} \Phi_j(x) = 1,\ \forall x \in \overline{\Omega} \text{ and } \mathrm{supp}(\Phi_i) \subset \Omega_i,\ \forall i \in \{1,\dots,d_H\}. \end{cases} \tag{C.3}$$

Observe that $\mathrm{H}^1_0(\Omega)$ is equipped with an energy norm by the bilinear form $a(u,v) = \int_\Omega \mathrm{grad}\, u \cdot \mathrm{grad}\, v$; it would be a semi-norm if Neumann conditions were assigned, and then we would work in $\mathrm{H}^1(\Omega)$, as is the case in [79].

Let $(\phi_i)_{i=1..d_h}$ be the fine basis on the initial mesh. The coarse space is included in the fine one, and thus it is possible to write:

$$\forall i \in \{1,\dots,d_H\},\ \Phi_i = \sum_{j=1}^{d_h} \phi_j \alpha_{ji} \qquad \text{with } \alpha_{ji} = 0 \text{ if } j \notin I_i \text{ (support constraints).} \tag{C.4}$$

The prolongation operator for the two-level method is denoted by α, which is a $d_h \times d_H$ matrix.

In order to describe the problem in an algebraic fashion, a matrix P_j is introduced, with j varying from 1 to d_H. Denoting by e_i the i-th vector of the canonical basis of $\mathbb{R}^{d_h}$, the rows of P_j are the vector e_i^t whose index i belongs to I_j. The dimension of P_j is $n_j \times d_h$ where n_j is the number of nodes in I_j.

Then, d_H vectors $(\alpha_i)_{i=1\dots d_H}$ of respective dimension n_i can be defined so that: $\alpha_i = P_i \alpha_{\bullet i}$. Here and in what follows, $\alpha_{\bullet i}$ denotes the column vector $(\alpha_{ji})_j$. Let K be the matrix whose elements are $a(\phi_j,\phi_i)$ and let $K_i = P_i K P_i^t$; its dimension is $n_i \times n_i$.

Then, the problem takes the matrix form:

$$\begin{cases} \text{To find } (\alpha_i)_{i=1,\dots,d_H} \text{ minimising } \displaystyle\sum_{i=1}^{d_H} \alpha_i^t K_i \alpha_i \text{ under the constraints:} \\ \displaystyle\sum_{i=1}^{d_H} P_i^t \alpha_i = 1_{d_h \times 1}. \end{cases} \tag{C.5}$$

The notation $1_{d_h \times 1}$ denotes a column vector with d_h components equal to 1.

C.1.2 Method of resolution

This optimisation problem with constraints is solved by introducing Lagrange multipliers.

An iterative method is used to compute the multiplier vector. The principle of the resolution is sketched.

Let us define:

$$\alpha = \begin{pmatrix} \alpha_1 \\ \vdots \\ \alpha_{d_H} \end{pmatrix}, \quad Q = \mathrm{diag}(K_i,\ i = 1 \dots d_H), \quad B = \begin{pmatrix} P_1 \\ \vdots \\ P_{d_H} \end{pmatrix}, \quad \gamma = 1_{d_h \times 1}. \tag{C.6}$$

The notation $\mathrm{diag}(A_i,\ i = 1 \dots n)$ refers to a block-diagonal matrix, whose diagonal blocks are the A_i's, $i = 1 \dots n$. Denoting the Lagrange multiplier vector by $\mu \in \mathbb{R}^{d_h}$, the optimisation problem takes the form:

$$\begin{cases} \text{To find the saddle-point } (\alpha_c, \mu_c) \text{ of the Lagrangian } \mathcal{L} \text{ defined by:} \\ \mathcal{L}(\alpha,\mu) = \dfrac{1}{2}\alpha^t Q \alpha + \mu^t (B^t \alpha - \gamma). \end{cases} \tag{C.7}$$

The critical point of $\mathcal{L}$ must then satisfy the following equations:

$$\begin{cases} Q\alpha_c = -B\mu_c \\ B^t\alpha_c = \gamma. \end{cases} \tag{C.8}$$

This system can be solved in the following way:

- first, compute the multiplier vector μ_c by an iterative method applied to the system:

$$B^tQ^{-1}B\mu = -\gamma, \text{ i.e. } \sum_{i=1}^{d_H} P_i^t K_i^{-1} P_i \mu = -\gamma.$$

- then obtain α by solving:

$$Q\alpha = -B\mu_c, \text{ i.e. } \alpha_i = -K_i^{-1}P_i\mu_c, \ \forall i = 1 \dots d_H.$$

All the K_i's are symmetric positive definite, and so is $B^tQ^{-1}B$. Then the system for μ can be solved by the conjugate gradient method. The factorisation of every K_i, which is a local representation of the matrix of the problem, is needed for the matrix-vector product. This product is implemented as follows:

1. for $i = 1 \dots d_H$ compute $b_i \leftarrow R_i\mu$,
2. for $i = 1 \dots d_H$ solve $K_i x_i = b_i$,
3. finally compute $\sum_{i=1}^{d_H} P_i^t x_i$.

Of course, the multiplications by P_i and P_i^t can be computed very fast because these matrices only contain n_i non-zero coefficients. Observe that the matrix-vector product in the conjugate gradient algorithm does not need to assemble the matrix, which is too expensive in computational time and memory.
The computation of the α_i's can be simply obtained from the factorisations of the local matrices K_i.

C.1.3 Decomposition into subdomains

The subdomain decomposition is constructed as follows (see Fig. C.1):

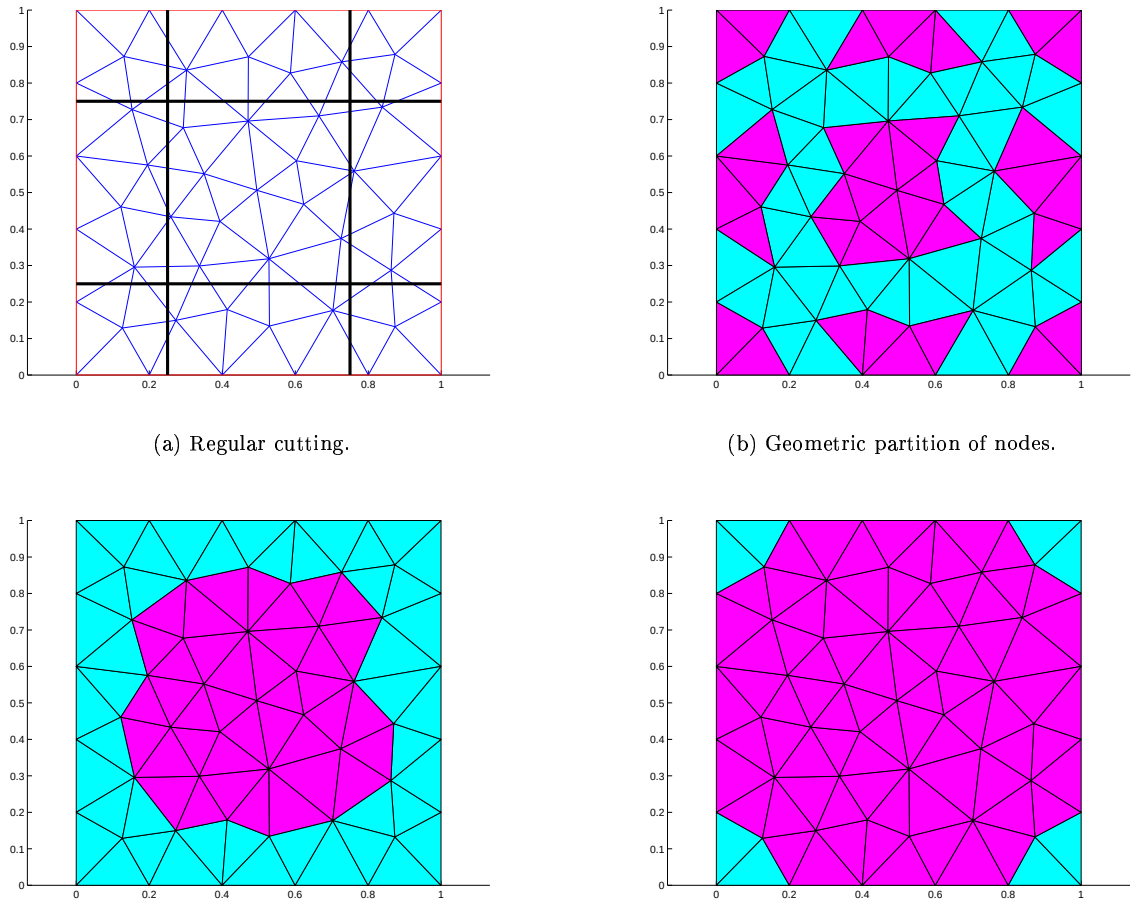
1. the domain Ω is cut according to regular geometric shapes;
2. the nodes are partitioned according to the geometric partition;
3. each node set is extended by taking all its nearest neighbours, in order to increase the overlap between subdomains. Thus, d_H overlapping node sets $(I_i)_{i=1,\dots,d_H}$ have been created;
4. the subdomain Ω_i is the union of all the elements which have a node of the set I_i, as one of their vertices.

C.1.4 Quantitative information on the different meshes

The results obtained for the algorithms are compared on the three meshes shown in Fig. C.2. The number of nodes and elements of these meshes, are given in Table C.1.

	Number of elements	Number of nodes
Mesh C.2(a)	312	177
Mesh C.2(b)	1248	665
Mesh C.2(c)	4992	2577

Table C.1: Quantitative information on meshes.



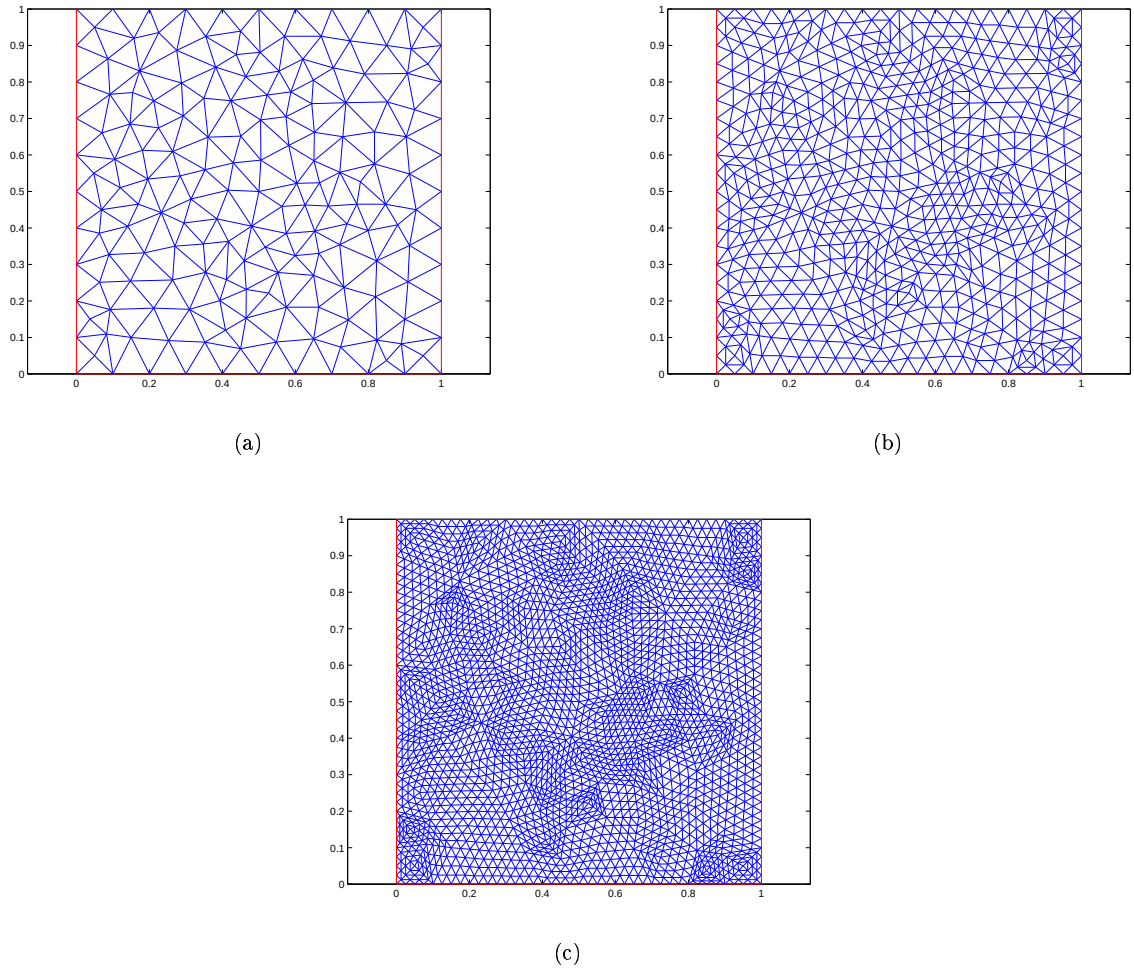
(a) Regular cutting.

(b) Geometric partition of nodes.

(c) Central set: overlapping by extension to the nearest neighbours.

(d) Central set: corresponding subdomain.

Figure C.1: Decomposition into subdomains. Observe that this mesh is very coarse, and consequently, the central subdomain Ω_i is close to the whole domain Ω.



(a) (b)

(c)

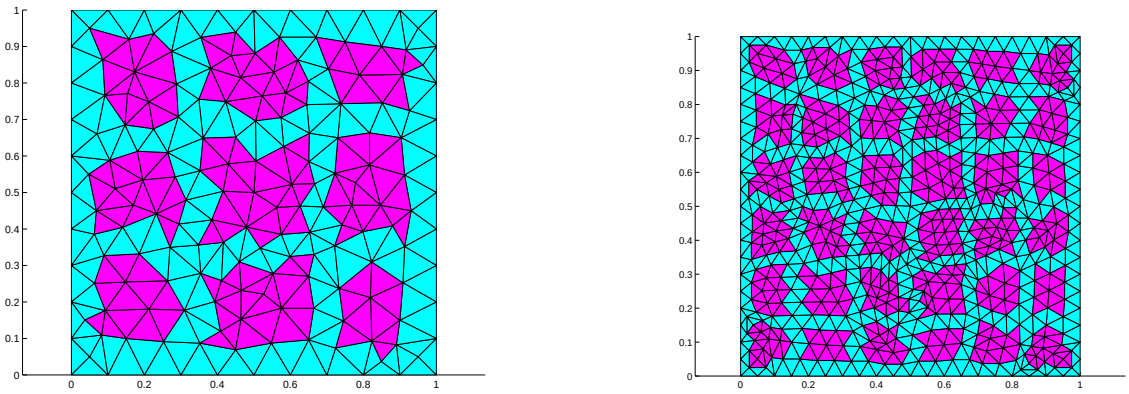
Figure C.2: Mesh C.2(a): $h_{\max} < 0.1$, meshes C.2(b) and C.2(c) are obtained by regular refinement.

C.2 Dirichlet fine and Dirichlet coarse bases (DFDC)

In this case, the constraint is imposed only on the interior nodes. Moreover, every coarse basis function is a linear combination of the fine basis functions from the finite element space included in $\mathrm{H}_0^1(\Omega)$.

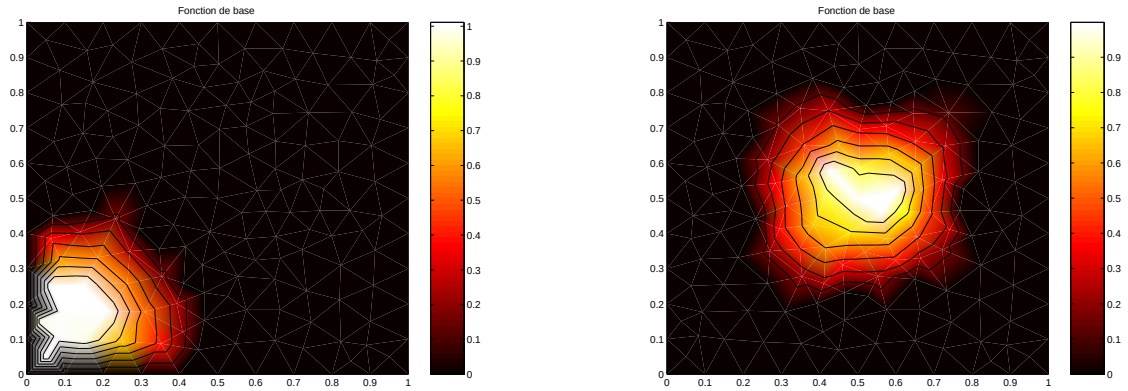
The domain Ω is subdivided into squares of length: $H = 1/\sqrt{d_H}$. The lexicographical numbering is used for the subdomains.

In Fig. C.3, the partitions obtained with meshes C.2(a) and C.2(b) are presented. After solving the optimisation problem, one obtains the coarse basis functions. Two examples are shown in Fig. C.4, they are obtained from mesh C.2(a).



(a) (b)

Figure C.3: Node partitions for meshes C.2(a) and C.2(b).



(a) (b)

Figure C.4: Basis functions on the boundary and in the interior of the domain, generated with Dirichlet boundary conditions.

In Fig. C.4(a), observe the large gradient of the basis function on the boundary where Dirichlet conditions are imposed: the isovalue curves are very close. This generates large diagonal coefficients, though the problem is homogeneous. This is demonstrated on the stiffness matrix (C.9) of the coarse

level for mesh C.2(a) and for the coefficients $(1,1)$, $(3,3)$, $(7,7)$ and $(9,9)$.

$$\begin{pmatrix} 11.3 & -0.31 & 0 & -0.54 & -0.46 & 0 & 0 & 0 & 0 \\ -0.31 & 7.17 & -0.22 & -0.46 & -0.85 & -0.45 & 0 & 0 & 0 \\ 0 & -0.22 & 9.88 & 0 & -0.37 & -0.24 & 0 & 0 & 0 \\ -0.54 & -0.46 & 0 & 7.56 & -0.67 & 0 & -0.29 & -0.20 & 0 \\ -0.46 & -0.85 & -0.37 & -0.67 & 4.99 & -0.97 & -0.46 & -0.98 & -0.24 \\ 0 & -0.45 & -0.24 & 0 & -0.97 & 7.03 & 0 & -0.4 & -0.22 \\ 0 & 0 & 0 & -0.29 & -0.46 & 0 & 10.22 & -0.54 & 0 \\ 0 & 0 & 0 & -0.20 & -0.98 & -0.4 & -0.54 & 6.71 & -0.38 \\ 0 & 0 & 0 & 0 & -0.24 & -0.22 & 0 & -0.38 & 10.81 \end{pmatrix} \tag{C.9}$$

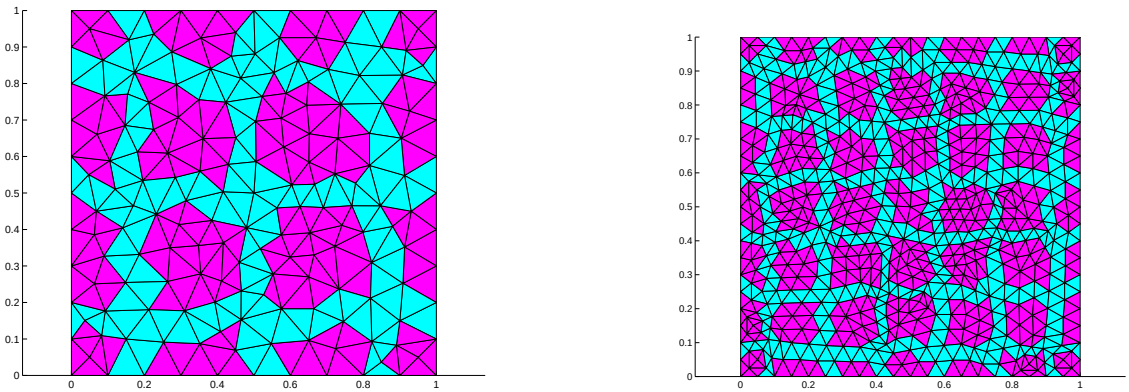
If we only look at its non-zero entries, we can see that the matrix (C.9) has a regular structure, which is comparable to that which would be obtained with a nine-point stencil and a lexicographical numbering. Moreover, all the off-diagonal entries are negative.

C.3 Neumann fine and Dirichlet coarse bases (NFDC)

In order to avoid the boundary effects that appeared in the previous method and to enforce on the whole domain $\overline{\Omega}$ the constraint $\sum_i \Phi_i(x) = 1$, we will define coarse elements satisfying a Neumann condition on the boundary of the large domain. Moreover we adopt a different cutting strategy (compare Fig. C.3 and Fig. C.5) and we discard the coarse elements on the boundary.

The decomposition is indeed adapted to this NFDC method; the coarse size is the same as in the previous method but boundary elements have normal size $H/2$; we choose $H = (\sqrt{d_H} - 1)^{-1}$. With this choice, all the fine basis functions contribute to coarse elements. The lexicographical numbering is kept for the subdomains.

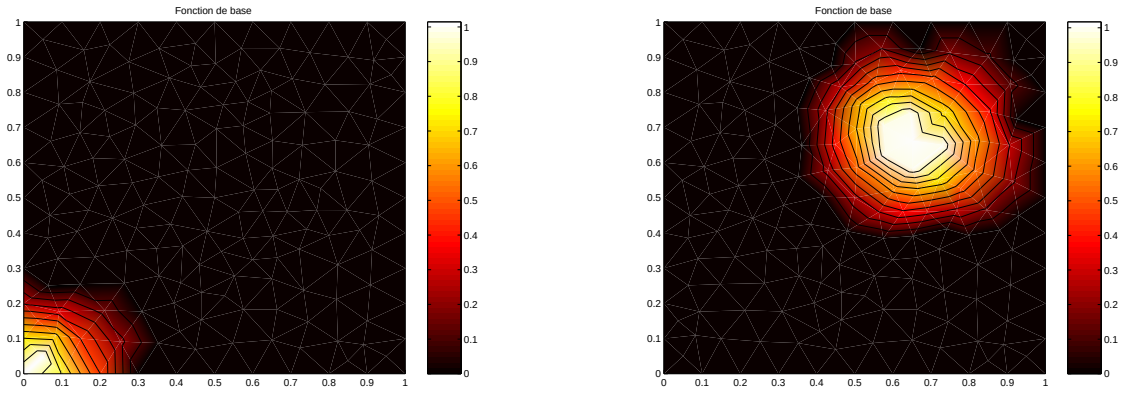
In Fig. C.5, the partitions obtained with meshes C.2(a) and C.2(b) are presented. After solving the optimisation problem, one obtains the coarse basis functions. Two examples are shown in Fig. C.6, corresponding to a computation on mesh C.2(a).



(a) (b)

Figure C.5: Node partitions for meshes C.2(a) and C.2(b).

Fig. C.6(a) demonstrates the isovalue lines of a basis function whose support intersects the boundary. In comparison with Fig. C.4(a), we observe that the gradients here are smaller. However, this basis function is not used in the coarse basis; indeed, on one hand, Dirichlet conditions are imposed on the original problem and on the other hand, we have enough basis functions.



(a) (b)

Figure C.6: Basis functions on the boundary and in the interior of the domain, generated with Neumann boundary conditions.

In the case of partition C.5(a), only four coarse basis functions remain after removing the basis functions whose support intersects the boundary. The coarse stiffness matrix corresponding to this choice of coarse basis functions is:

$$\begin{pmatrix} 5 & -1.06 & -0.8 & -0.47 \\ -1.06 & 4.82 & -0.21 & -1.01 \\ -0.8 & -0.21 & 4.38 & -0.82 \\ -0.47 & -1.01 & -0.82 & 4.99 \end{pmatrix} \tag{C.10}$$

Observe that the variation of coefficients is much smoother than in matrix (C.9). In Fig. C.7, we display the regular structure of the non-zero entries in the coarse matrix used for solving the problem on mesh C.2(b) with partition C.5(b).

C.4 Adjusting the matrix at the coarse level

The construction of the coarse basis is only the first step towards our aim. It allows us to obtain a matrix with a regular pattern of non-zero elements. Moreover, we observe that every point receives contributions only from its eight direct neighbours and that the similar connectivity relations (diagonal, lateral neighbours) yield coefficient of the same order, which was one of the aims of this construction.

The idea is then to build a mean nine-point stencil and to assemble a new coarse matrix using only this stencil; we can even rewrite it on the old matrix because the pattern is preserved. We conjecture that the new coarse matrix is spectrally equivalent to the old one.

Let us begin by case (C.10), which has only one block. If the mean of the coefficients is calculated separately for each kind of connectivity relation between coarse basis functions *i.e.* the means of, respectively, central, vertical and horizontal, and diagonal interactions, we obtain the matrix:

$$\begin{pmatrix} 4.8 & -0.92 & -0.92 & -0.34 \\ -0.92 & 4.8 & -0.34 & -0.92 \\ -0.92 & -0.34 & 4.8 & -0.92 \\ -0.34 & -0.92 & -0.92 & 4.8 \end{pmatrix} \tag{C.11}$$

that corresponds to the nine-point stencil given by:

$$\begin{matrix} -0.34 & -0.92 & -0.34 \\ -0.92 & 4.8 & -0.92 \\ -0.34 & -0.92 & -0.34 \end{matrix} \tag{C.12}$$

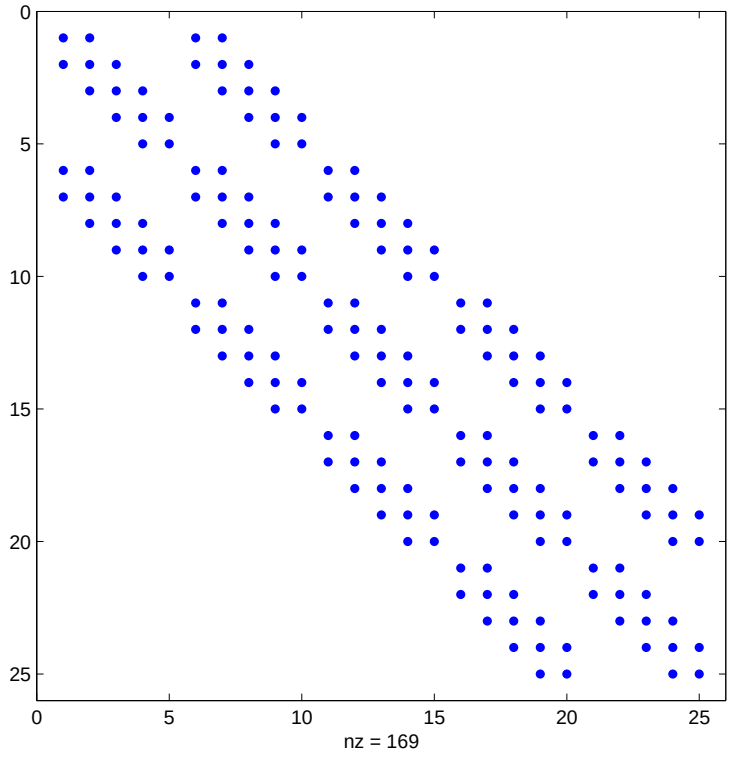


Figure C.7: Pattern of the non-zero entries in the coarse matrix associated with mesh C.2(b).

More generally, studying Fig. C.7 indicates a block-Toeplitz-Toeplitz-block (BTTB) pattern but the matrix itself is not BTTB. Then, we compute the means of the coefficients corresponding to the similar connectivity relations and we obtain the nine-point stencil given by (C.13):

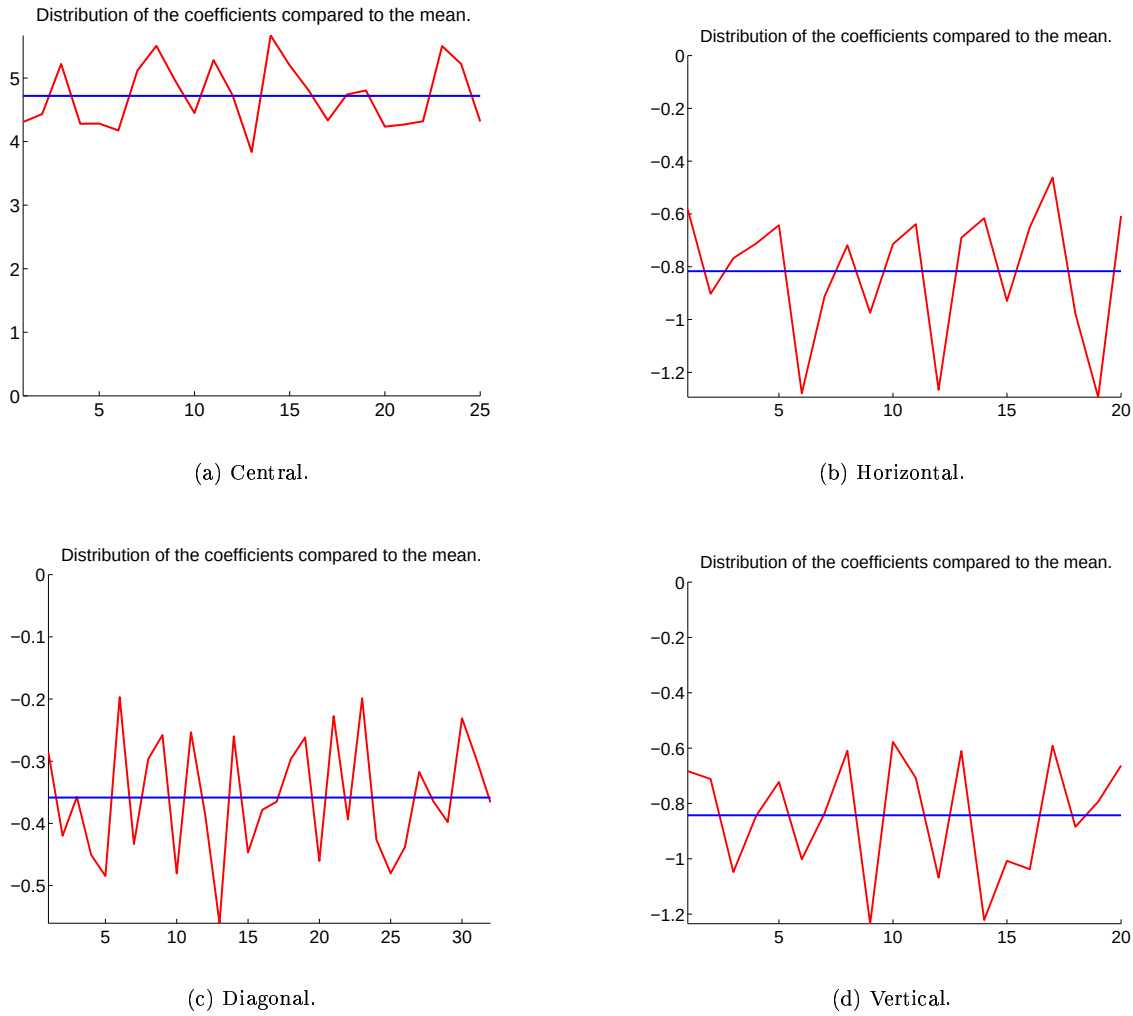
$$\begin{array}{ccc} -0.36 & -0.83 & -0.36 \\ -0.83 & 4.72 & -0.83 \\ -0.36 & -0.83 & -0.36 \end{array} \tag{C.13}$$

By using this mean stencil, the matrix which is obtained is BTTB. This property allows us to consider using fast direct solvers.

In order to evaluate the accuracy of this process and more precisely the distance to the original matrix, Fig. C.8 and Table C.2 give a statistical view of the results. The stencil obtained by using the means is called the mean stencil.

Coefficient	Central	Horizontal	Vertical	Diagonal
Mean	4.72	-0.82	-0.84	-0.36
Standard deviation	0.49	0.24	0.20	0.09
Maximal deviation	0.95	0.35	0.27	0.16
L^1 deviation	0.43	0.20	0.18	0.08

Table C.2: Statistics comparing the mean stencil with the original coarse basis matrix.



(a) Central.

(b) Horizontal.

(c) Diagonal.

(d) Vertical.

Figure C.8: Distribution of the coefficients compared to the mean coefficient: central, horizontal, diagonal and vertical.

C.5 Comparison of the number of iterations and of the conditioning on different meshes

For solving the linear systems coming from the discretization by finite elements, the stopping criterion is $\|r\|_2 \leq 10^{-10}\|r_0\|_2$, with r being the current residual and r_0 being the initial residual.

Table C.3 compares for different cases the number of unknowns on the coarse grid.

	DFDC	NFDC
Mesh C.2(a)	9	4
Mesh C.2(b)	36	25
Mesh C.2(c)	144	121

Table C.3: Dimensions of the coarse space.

Table C.4 compares the condition number based on the 2-norm of the matrix BA where A is the matrix of the system and B the preconditioning matrix. Five different preconditioning methods are used:

- symmetric Gauss-Seidel (GSsym);
- two-level method with one presmoothing and one postsmoothing, using symmetric Gauss-Seidel and 4 different kinds of coarse matrix, DFDC or NFDC, with or without using the mean stencil.

	Mesh C.2(a)	Mesh C.2(b)	Mesh C.2(c)
Without preconditioning	56.7	273.5	1343.5
Symmetric Gauss-Seidel	20	81	304.8
V-cycle$(1,1)$ GSsym smoothing — DFDC	2.5	4.1	8.9
Idem + mean stencil	3.4	7.6	19.7
V-cycle$(1,1)$ GSsym smoothing — NFDC	2.3	3.85	7
Idem + mean stencil	2.3	4	7.8

Table C.4: Condition number based on the 2-norm of the matrix BA, B being the preconditioning matrix.

We observe that the two-level methods are comparable in the case when we use the original matrix on the coarse space. Moreover, if the mean stencil is used, the best case is with the NFDC method (see Section C.3). In this case, we obtain a similar behaviour for the original matrix and for the BTTB matrix deduced from the mean stencil.

Let us add that the behaviour of two-level methods does not seem optimal, although it is much more effective than the one-level method.

In Table C.5, the number of iterations for solving the linear system with the same preconditioning techniques is given.

We can also benefit from the factorisations accomplished during the computation of the coarse basis functions, in order to implement a preconditioner or a smoother using the multiplicative Schwarz method (GSsymblc). The number of iterations needed then to solve the system are reported in Table C.6.

C.6 Comparison for other boundary conditions

We infer from the numerical experiments reported in Table C.4, C.5 and C.6 that the NFDC method is the best. In this section, we explore the applicability of this method to problem involving boundary conditions which are not everywhere Dirichlet conditions. Thus, we replaced the Dirichlet conditions on

	Mesh C.2(a)	Mesh C.2(b)	Mesh C.2(c)
Without preconditioning	45	98	X
Symmetric Gauss-Seidel	23	45	88
V-cycle$(1,1)$ GSsym smoothing — DFDC	12	15	18
Idem + mean stencil	12	19	25
V-cycle$(1,1)$ GSsym smoothing — NFDC	13	15	17
Idem + mean stencil	13	15	17

Table C.5: Number of iterations for solving the linear system for different preconditioners. X means that the algorithm did not converge after 100 iterations.

	Mesh C.2(a)	Mesh C.2(b)	Mesh C.2(c)
Symmetric block Gauss-Seidel DFDC	8	12	22
Symmetric block Gauss-Seidel NFDC	7	12	22
V-cycle$(1,1)$ GSsymblc smoothing — DFDC	5	7	10
Idem + mean stencil DFDC	5	8	13
V-cycle$(1,1)$ GSsymblc smoothing — NFDC	5	7	9
Idem + mean stencil NFDC	5	7	9

Table C.6: Number of iterations for solving the linear system for different preconditioners.

some parts of the boundary by periodicity conditions. The problem to solve can then be written:

$$\begin{cases} -\triangle u = f \text{ on } \Omega =]0;1[\times]0;1[, \\ u(0,y) = u(1,y) = 0 \ \forall y \in]0;1[, \\ u(x,0) = u(x,1) \ \forall x \in]0;1[, \\ f = 5\pi^2 \sin(\pi x)\sin(2\pi y). \end{cases} \tag{C.14}$$

We solved this linear system and we obtained the results gathered in Table C.7.

	Mesh C.2(a)	Mesh C.2(b)	Mesh C.2(c)
Without preconditioning	53	X	X
Symmetric Gauss-Seidel	26	51	99
V-cycle$(1,1)$ GSsym smoothing — NFDC	13	16	17
Idem + mean stencil	14	16	17

Table C.7: Number of iterations for solving the linear system for different preconditioners. X means that the algorithm did not converge after 100 iterations.

Then, we also added Neumann conditions on the part of the boundary where $x = 1$ and changed the right-hand side ($f = 4.25\sin(\frac{\pi}{2}x)\sin(2\pi y)$). The results obtained are gathered in Table C.8. There is now a difference between the results with the mean stencil or the initial matrix. This should be due to the rough way used to compute the mean stencil or to the disagreement between the boundary conditions implicitly chosen by the mean stencil and the true boundary conditions.

	Mesh C.2(a)	Mesh C.2(b)	Mesh C.2(c)
Without preconditioning	62	X	X
Symmetric Gauss-Seidel	29	58	X
V-cycle$(1,1)$ GSsym smoothing — NFDC	14	15	17
Idem + mean stencil	16	20	25

Table C.8: Number of iterations for solving the linear system for different preconditioners. X means that the algorithm did not converge after 100 iterations.

C.7 Some results about the Helmholtz equation

In order to test the robustness of the NFDC algorithm and in view of the applications, Laplace's equation is replaced by Helmholtz' equation. The formulation of the problem then becomes:

$$\begin{cases} -\triangle u - k^2 u = f \text{ on } \Omega =]0;1[\times]0;1[, \\ u = 0 \text{ on } \partial\Omega, \\ f = (2\pi^2 - k^2)\sin(\pi x)\sin(\pi y). \end{cases} \tag{C.15}$$

The wave number is denoted by k. It can also be written $k = 2\pi/\lambda$ where λ is the wavelength. Thus, a study can be performed with varying wavelength. An interesting comparison can be made by following the evolution of the number of iterations as a function of the ratio H/λ where H is the width of the initial subdomains. The results given in Table C.9 are obtained with mesh C.2(b).

In the case $H/\lambda = 1/2$, all the diagonal coefficients became negative, and we observed a serious deterioration of the results.

The mean of the diagonal coefficients for the coarse matrix as a function of the ratio H/λ is given in Table C.10.

H/λ	1/12	1/6	1/3	5/12	1/2
Without preconditioning	X	X	X	X	X
Symmetric Gauss-Seidel	48	57	92	X	X
V-cycle$(1,1)$ GSsym smoothing NFDC	16	20	42	48	87
Idem + mean stencil	16	19	42	47	87

Table C.9: Number of iterations for solving the linear system for different preconditioners. X means that the algorithm did not converge after 100 iterations.

H/λ	1/12	1/6	1/3	5/12	1/2
Mean of the diag. coeff.	4.56	4.07	2.08	0.9	-1.3

Table C.10: Mean of the diagonal coefficients for the coarse matrix with the ratio H/λ.

C.8 Conclusion

These different tests allowed us to understand the principle of the coarse basis construction by energy minimisation. They also demonstrated the possibility, in very simple cases, to use well-structured matrices close to the original matrix and this gave satisfactory results. However, we also observe the limits of this approach for the Helmholtz problem.

Appendix D

Bases nodale et d'arête grossières compatibles et fonctionnelles d'énergie

Compatible coarse nodal and edge elements through energy functionals[1]

François Musy, Laurent Nicolas and Ronan Perrussel

Abstract. *We propose new algorithms for the setup phase of algebraic multigrid (AMG) solvers for linear systems coming from edge element discretization. The construction of coarse levels is performed by solving an optimization problem with a Lagrange multiplier method: we minimize the energy of coarse bases under a constraint linking coarse nodal and edge element bases. On structured meshes, the resulting AMG method and the geometric multigrid method behave similarly as preconditioner. On unstructured meshes, our method compares favorably with the AMG method of Reitzinger and Schöberl.*

D.1 Introduction

Edge element discretization plays a key-rôle in computational electromagnetism; it can be implemented for evaluating the electric or magnetic field using the vector wave equation or for evaluating the magnetic potential using the eddy current formulation. As this finite element discretization leads to sparse but generally large linear systems, efficient methods are required for the resolution. Multilevel techniques have been introduced by Hiptmair [41] and Arnold et al. [4]; they are shown to be optimal on a hierarchy of nested grids. However, the inherent need of a hierarchical finite element mesh in geometric multigrid method is very restricting for industrial applications. It is more convenient to make use of a grey-box multilevel algorithm which needs a single grid but takes into account additional information on the initial problem.

For this purpose, algebraic multilevel methods have already been developed. Beyond the use of specific smoothers as Hiptmair or Arnold et al. proposed, Reitzinger and Schöberl [66] have highlighted an essential geometric compatibility relation which means that the gradient of a coarse nodal element function must belong to the coarse edge element space. Together with this relation, Reitzinger and Schöberl achieved an efficient method which is now widespread in the computational electromagnetism community [48, 53, 55, 81, 80]. However, the combinatorial approach used there does not lead to an optimal convergence rate. Some improvements have been proposed by Bochev et al. [13, 43] based on smoothed aggregation techniques [76] and a compatibility with a larger class of nodal prolongation operators.

Here, we propose a somewhat different approach. Following ideas from [79] and [84], we construct coupled nodal and edge coarse bases by energy minimization and we enforce the compatibility relation as a constraint. A connection can be made with the smoothed aggregation method [51], which also involves energy minimization. In order to simplify the notations, we present our construction in a two-level framework.

[1] Article soumis à SIAM Journal on Scientific Computing; voir [61].

In Section D.2, the problem and the finite element space with properties are reviewed, especially the essential geometric relation. The constrained minimization problem is presented in Section D.3. Then we introduce the resulting linear system with Lagrange multipliers and the edge prolongation matrix as unknowns. Section D.4 details the conditions to obtain a well-posed minimization problem. First, we clarify the construction of the edge function supports from a given graph and the coarse nodal function supports and we point out the satisfied constraints. Then, we prove that the coarse graph has to satisfy a connectivity condition if the implicit constraints are not enforced and we propose a construction of the coarse nodal function supports together with a suitable definition of the coarse graph. Finally in Section D.5, we evaluate the efficacy of various versions of the algorithm corresponding to different choices of energy norm. We give numerical results relative to 2D and 3D problems on structured and unstructured meshes. On a hierarchy of nested meshes, the edge bases being computed from the geometric nodal bases at each level, the resulting AMG method and the geometric multigrid method behave similarly as preconditioner. On unstructured meshes, our method is somewhat better than the method of Reitzinger and Schöberl [66]. However, we pay this improvement with a large increase of the computing cost for the edge bases.

D.2 Definition of the continuous problem and its discretization

D.2.1 Formulation

The following problem has to be solved on a domain Ω:

$$\begin{cases} \text{To find } \mathbf{E} \in V \text{ such that: } a(\mathbf{E}, \mathbf{E}') = F(\mathbf{E}'), \ \forall \mathbf{E}' \in V_0, \\ \text{with } a(\mathbf{E}, \mathbf{E}') = \displaystyle\int_\Omega \delta \operatorname{rot} \mathbf{E} \cdot \operatorname{rot} \mathbf{E}' + \int_\Omega \gamma \mathbf{E} \cdot \mathbf{E}'. \end{cases} \tag{D.1}$$

V is an affine subspace of $\mathbb{H}(\mathrm{rot}, \Omega)$ [57] taking into account essential boundary conditions, V_0 is the vector subspace parallel to V, F is a linear form on V_0 which describes the source term, δ and γ are strictly positive functions, so that a is coercive; F, δ and γ depend on the applications under consideration.

This formulation includes many static and transient electromagnetic models: potential vector formulation for magnetostatic or eddy currents, electric or magnetic field formulation in the transient case.

D.2.2 Finite element space and properties

Problem (D.1) is discretized by using the lowest order edge elements introduced by Nédélec [62], which are conforming in $\mathbb{H}(\mathrm{rot}, \Omega)$. We consider triangular and tetrahedral meshes. For a tetrahedron K, the local polynomial space V_K is defined by:

$$V_K = \{x \mapsto \mathbf{p} \times \mathbf{x} + \mathbf{q}, \ \mathbf{x} \in K \text{ and } \mathbf{p}, \mathbf{q} \in \mathbb{R}^3\}. \tag{D.2}$$

The symbol $\times$ denotes the cross-product.

The local degrees of freedom which permit to ensure the conformity in $\mathbb{H}(\mathrm{rot}, \Omega)$ are given by path integrals along the edges of the element:

$$\mathbf{E}_h \mapsto \int_e \mathbf{E}_h \cdot \mathbf{t} \ \mathrm{ds} \tag{D.3}$$

where e is an edge of K, $\mathbf{t}$ is a tangential vector to e and $\mathbf{E}_h$ belongs to V_K. Based on this local description, an edge finite element space can be defined on a mesh of the domain. V_h will denote the edge finite element space, defined on a mesh T_h of the domain Ω, and taking into account essential boundary conditions.

Therefore using this finite element space, we are led to solve a linear system:

$$Ax = f. \tag{D.4}$$

A fundamental property links the edge elements to the nodal P_1-Lagrange elements: the gradient of a nodal function belongs to the space of edge functions. If we denote the finite element bases as follows:

1. $(w_p^{0,h})_{p=1,\dots,N^h}$ is the nodal basis,
2. $(w_i^{1,h})_{i=1,\dots,E^h}$ is the edge basis,

this property can be recast as:

$$\operatorname{grad} w_p^{0,h} = \sum_{i=1}^{E^h} G_{ip}^h w_i^{1,h}, \ \forall p \in \{1,\dots,N^h\}, \tag{D.5}$$

where G^h is known and is the edge-node incidence matrix of the mesh T_h; in what follows, the superscript h for the fine level will be opposed to H used for the coarse level. G^h is a discrete analogue of the gradient operator on this mesh. Moreover, if the domain is contractible, the subspace spanned by the $\operatorname{grad} w_p^{0,h}$'s coincides exactly with the kernel of the rot operator in the edge element space.

Our aim, as highlighted in [66], is to preserve the analogous compatibility relation for coarse bases, which is natural with a hierarchy of nested grids. Thus, a coherent representation of the kernel of the rot operator is preserved, which permits the efficient use of the smoothers proposed by Hiptmair [41] and by Arnold et al. [4]. Therefore the gradient of a coarse nodal function must be a combination of coarse edge functions.

Let us introduce the following notations:

1. $(w_n^{0,H})_{n=1,\dots,N^H}$ is the coarse nodal basis,
2. $(w_e^{1,H})_{e=1,\dots,E^H}$ is the coarse edge basis,

the compatibility condition is written algebraically as:

$$\operatorname{grad} w_n^{0,H} = \sum_{e=1}^{E^H} G_{en}^H w_e^{1,H}, \ \forall n \in \{1,\dots,N^H\}. \tag{D.6}$$

where G^H is the discrete analogue of the gradient operator and is the edge-node incidence matrix of a coarse graph. This operator G^H or equivalently the associated coarse graph has to be constructed before the computation of the coarse edge basis. Some complements will be given in Subsections D.4.1 and D.4.4.

These coarse bases are constructed so as to satisfy the inclusion of finite element spaces, the "coarse" being included in the "fine", which is expressed by the following algebraic relations:

$$w_n^{0,H} = \sum_{p=1}^{N^h} \alpha_{pn} w_p^{0,h}, \ \forall n \in \{1,\dots,N^H\}, \tag{D.7a}$$

$$w_e^{1,H} = \sum_{i=1}^{E^h} \beta_{ie} w_i^{1,h}, \ \forall e \in \{1,\dots,E^H\}. \tag{D.7b}$$

According to relations (D.5), (D.6) and (D.7), the components α_{pn} and β_{ie} of the coarse bases must satisfy:

$$\sum_{e=1}^{E^H}\sum_{i=1}^{E^h} \beta_{ie} G_{en}^H w_i^{1,h} = \sum_{p=1}^{N^h}\sum_{i=1}^{E^h} G_{ip}^h \alpha_{pn} w_i^{1,h}, \ \forall n \in \{1,\dots,N^H\}, \tag{D.8a}$$

i.e. we would like to ensure:

$$\sum_{e=1}^{E^H} \beta_{ie} G_{en}^H = \sum_{p=1}^{N^h} G_{ip}^h \alpha_{pn}, \ \forall i \in \{1,\dots,E^h\}, \ \forall n \in \{1,\dots,N^H\}. \tag{D.8b}$$

Finally, if α denotes the matrix of components α_{in} and β the matrix of components β_{ie}, the compatibility relation (D.6) writes also as:

$$\beta G^H = G^h \alpha. \tag{D.9}$$

D.3 Overview of the coarse bases construction

D.3.1 Energy minimization problems

The domain Ω is decomposed into overlapping subdomains Ω_n^H, for n in $\{1,\ldots,N^H\}$; this decomposition enables us to restrict the support of the coarse nodal basis functions.

In the same way, Ω is decomposed into overlapping subdomains $\mathcal{U}_e$, for e in $\{1,\ldots,E^H\}$ in order to localize the supports of the coarse edge basis functions. The subdomain $\mathcal{U}_e$ will be the intersection of two subdomains Ω_l^H and Ω_m^H.

Given the subdomains Ω_n^H, the choice of the matrix G^H determines the definition of the subdomains $\mathcal{U}_e$; we will clarify this point in Subsection D.4.1.

Two minimization problems under constraints are solved successively. In analogy with the nodal element case [79], we first solve:

$$\begin{cases} \text{To find } (w_n^{0,H})_{n=1..N^H} \text{ minimizing } \displaystyle\sum_{n=1}^{N^H} c(w_n^{0,H}, w_n^{0,H}) \text{ under the constraints:} \\ \displaystyle\sum_{n=1}^{N^H} w_n^{0,H}(x) = 1,\ \forall x \in \overline{\Omega} \text{ and } \mathrm{supp}(w_n^{0,H}) \subset \overline{\Omega_n^H},\ \forall n \in \{1,\ldots,N^H\}. \end{cases} \tag{D.10}$$

The bilinear form c can be for instance $c(\phi,\psi) = \int_\Omega \gamma \,\mathrm{grad}\,\phi \cdot \mathrm{grad}\,\psi$ or more simply the bilinear form associated with the Laplacian of the graph *i.e.* $(G^h)^t G^h$. Here the superscript t denotes transposition.

Next, from the coarse nodal basis $(w_n^{0,H})_{n=1,\ldots,N^H}$ we compute the coarse edge basis $(w_e^{1,H})_{e=1,\ldots,E^H}$ by solving the problem:

$$\begin{cases} \text{To find } (w_e^{1,H})_{e=1,\ldots,E^H} \text{ minimizing } \displaystyle\sum_{e=1}^{E^H} b(w_e^{1,H}, w_e^{1,H}) \text{ under the constraints:} \\ \mathrm{grad}\, w_n^{0,H} = \displaystyle\sum_{e=1}^{E^H} w_e^{1,H} G_{en}^H,\ \forall n \in \{1,\ldots,N^H\} \text{ and } \mathrm{supp}(w_e^{1,H}) \subset \overline{\mathcal{U}_e},\ \forall e \in \{1,\ldots,E^H\}. \end{cases} \tag{D.11}$$

The letter b denotes a scalar product on V_h which can be a from (D.1) or variants, which will be introduced in Subsection D.5.2 and implemented in numerical examples in Section D.5.

Problem (D.10) can be solved by using the method described in [79]. This is the reason why we mainly concentrate our attention on Problem (D.11).

We introduce algebraic notations which encode the support constraints. First for all e in $\{1,\ldots,E^H\}$, I_e is a subset of the set of indices of the fine edge basis functions whose support is included in $\overline{\mathcal{U}_e}$. We assume that for all e I_e is non empty and we define :

$$\tilde{M} = \sum_{e=1}^{E^h} |I_e|. \tag{D.12}$$

The term $|Z|$ denotes the number of elements in a finite set Z. Then for e in $\{1,\ldots,E^H\}$ with $I_e = \{i_1,\ldots,i_{|I_e|}\}$, we introduce the projection operator which keeps only the components indexed by I_e:

$$\begin{aligned} &Q_e : \mathbb{R}^{E^h} \to \mathbb{R}^{|I_e|}, \\ &\forall x \in \mathbb{R}^{E^h},\ (Q_e x)_k = x_{i_k},\ \forall k \in \{1,\ldots,|I_e|\}. \end{aligned} \tag{D.13}$$

It is straightforward that the transposed operation Q_e^t is defined by :

$$\begin{aligned} &Q_e^t : \mathbb{R}^{|I_e|} \to \mathbb{R}^{E^h}, \\ &\forall y \in \mathbb{R}^{|I_e|},\ (Q_e^t y)_i = 0 \text{ if } i \notin I_e \text{ and } y_k \text{ if } i = i_k. \end{aligned} \tag{D.14}$$

Secondly, in order to restrict the number of constraints in Problem (D.11), we introduce sets J_n for all n in $\{1,\dots,N^H\}$. The sets J_n as the sets I_e are subsets of indices of the fine edge basis functions.

The sets J_n will be carefully defined with the help of support conditions, so as to decrease as much as possible the number of constraints coming from (D.8b) and to obtain a well-posed problem; the construction of the J_n's is described in Subsection D.4.2. We introduce an integer for denoting the number of constraints :

$$M = \sum_{n=1}^{N^H} |J_n|. \tag{D.15}$$

Since some J_n's can be empty, we introduce the set:

$$F = \{n \in \{1,\dots,N^H\} \ / \ J_n \neq \emptyset\}. \tag{D.16}$$

Similarly to the operator Q_e from set I_e, we introduce a canonical projection operator associated to J_n for n in F:

$$R_n : \mathbb{R}^{E^h} \to \mathbb{R}^{|J_n|}. \tag{D.17}$$

Let K be the matrix whose coefficients are $b(w_j^{1,h}, w_i^{1,h})$. In order to obtain a matrix form of the minimization problem, we define matrices which are restrictions to the subdomains under consideration of the matrix K:

$$K_e = Q_e K Q_e^t. \tag{D.18}$$

We also define:

$$\xi_n = R_n G^h \alpha_{\bullet n}, \ \forall n \in F \text{ and } \beta_e = Q_e \beta_{\bullet e}, \ \forall e \in \{1,\dots,E^H\}. \tag{D.19}$$

where $\alpha_{\bullet n}$ denotes the n-th column of α and $\beta_{\bullet e}$ the e-th column of β.

After solving Problem (D.10), which gives the matrix α, we compute the β_e's by solving the problem:

$$\begin{cases} \text{To minimize } \displaystyle\sum_{e=1}^{E^H} \beta_e^t K_e \beta_e \text{ under the constraints:} \\ \\ R_n\Big(\displaystyle\sum_{e=1}^{E^H} G_{en}^H Q_e^t \beta_e\Big) = \xi_n, \ \forall n \in F. \end{cases} \tag{D.20}$$

D.3.2 Solution of Problem D.20

The constrained problem (D.20) can be solved by a Lagrange multiplier method. Let us introduce column vectors ξ and ρ of $\mathbb{R}^M$ and matrices D and T.

ρ will be the vector whose components are Lagrange multipliers relative to the constraints in (D.20) and ξ the vector defined from (D.19). If we suppose that $F = \{i_1,\dots,i_{|F|}\}$, ξ and ρ are defined block-wise as:

$$\xi = \begin{pmatrix} \xi_{i_1} \\ \vdots \\ \xi_{i_{|F|}} \end{pmatrix}, \quad \rho = \begin{pmatrix} \rho_{i_1} \\ \vdots \\ \rho_{i_{|F|}} \end{pmatrix} \text{ with } \xi_{i_n}, \rho_{i_n} \in \mathbb{R}^{|J_n|}. \tag{D.21}$$

The matrix D of dimension (M,M) is the block-diagonal matrix whose diagonal blocks are the matrices K_e, $e \in \{1,\dots,E^H\}$. Finally, the matrix T maps a vector from $\mathbb{R}^M$ to $\mathbb{R}^{\tilde{M}}$ in the following way:

$$T : \rho \mapsto \begin{pmatrix} Q_1\Big(\displaystyle\sum_{n\in F} G_{1n}^H R_n^t \rho_n\Big) \\ \vdots \\ Q_{E^H}\Big(\displaystyle\sum_{n\in F} G_{E^H n}^H R_n^t \rho_n\Big) \end{pmatrix}. \tag{D.22}$$

The minimization problem (D.20) can now be written:

$$\text{To find } \bar{\beta}_c \in \mathbb{R}^{\tilde{M}} \text{ minimizing } \bar{\beta}^t D \bar{\beta} \text{ in } \mathbb{R}^{\tilde{M}} \text{ under the constraint } T^t \bar{\beta} = \xi, \tag{D.23}$$

or by introducing Lagrange multipliers:

$$\begin{cases} \text{To find a critical point } (\bar{\beta}_c, \rho_c) \in \mathbb{R}^{\tilde{M}} \times \mathbb{R}^M \text{ of the Lagrangian } \mathcal{L} \text{ defined by:} \\ \mathcal{L}(\bar{\beta}, \rho) = \dfrac{1}{2}\bar{\beta}^t D \bar{\beta} + \rho^t(\xi - T^t\bar{\beta}). \end{cases} \tag{D.24}$$

This critical point must verify the system of equations:

$$\begin{pmatrix} D & -T \\ T^t & 0 \end{pmatrix} \begin{pmatrix} \bar{\beta} \\ \rho \end{pmatrix} = \begin{pmatrix} 0 \\ \xi \end{pmatrix} \tag{D.25}$$

This linear system can be solved in the following way:

— first, the vector ρ_c of Lagrange multipliers is determined by applying an iterative method to the system:

$$T^t D^{-1} T \rho = \xi. \tag{D.26}$$

At this point, we see that $T^t D^{-1} T$ in (D.26) is symmetric, we do not know yet that it is positive definite; the construction of the R_n's will precisely ensure this property (see Subsection D.4.2).

— then, we return to the computation of $\bar{\beta}_c$ by solving:

$$D\bar{\beta} = T\rho_c. \tag{D.27}$$

The matrix $T^t D^{-1} T$ will not be assembled and its multiplication by a vector can be made efficiently; details will be given in Subsection D.5.1.

D.4 Elements required by the construction

D.4.1 Algebraic decomposition into subdomains

Associated to the fine nodal basis $(w_p^{0,h})_{p=1,\dots,N^h}$, we give us an oriented simple connected graph by its set of vertices indexed by $\{1,\dots,N^h\}$ and its set of edges $\mathcal{S}^h \subset \{1,\dots,N^h\}^2$. The orientations of the edges are arbitrary; there are no loops *i.e.* no elements (p,p) in $\mathcal{S}^h$ and if (p,q) belongs to $\mathcal{S}^h$, (q,p) does not belong to $\mathcal{S}^h$. By convention, p is the origin and q is the end of the edge (p,q). The edges could be the geometric edges of the mesh, but we do not restrict ourselves to this situation: starting from the second finest grid, we will have no more geometric edges.

The edges are numbered by $i \in \{1,\dots,E^h\}$, where $E^h = |\mathcal{S}^h|$; the i-th edge is denoted by $(p(i), q(i))$ and conversely if $p = p(i)$ and $q = q(i)$, we will write $i = \overline{pq}^h$.

Let us now give the precise definition of the edge-node incidence matrix G^h of size $E^h \times N^h$:

$$G^h_{il} = \begin{cases} -1 \text{ if } l = p(i), \\ +1 \text{ if } l = q(i), \\ \ \ 0 \text{ otherwise.} \end{cases} \tag{D.28}$$

Two assumptions on the fine bases are supposed to be satisfied at the beginning. We suppose that the fine nodal basis $(w_p^{0,h})_{p=1,\dots,N^h}$ satisfies the property:

$$\sum_{p=1}^{N^h} w_p^{0,h}(x) = 1, \ \forall x \in \overline{\Omega}. \tag{D.29}$$

We assume that there exists a set of subdomains $(\Omega_p^h)_{p=1,\dots,N^H}$ of Ω such that:

$$\Omega = \bigcup_{p=1}^{N^h} \Omega_p^h \text{ and } \forall p \in \{1,\dots,N^h\}, \ \overline{\Omega_p^h} \cap \left(\bigcup_{q \neq p} \Omega_q^h \right)^c \neq \emptyset, \tag{D.30}$$

where the superscript c is the standard set-complement, and satisfying :

$$\begin{aligned} &\text{supp}(w_p^{0,h}) \subset \overline{\Omega_p^h}, \ \forall p \in \{1,\dots,N^H\} \\ &\text{and } \text{supp}(w_i^{1,h}) \subset \overline{\Omega_p^h \cap \Omega_q^h} \text{ if } i = \overline{pq}^h. \end{aligned} \tag{D.31}$$

Notations and principle of the construction for the coarse nodal basis

Several steps have to be considered:

1. In order to localize the supports of the coarse nodal functions $(w_n^{0,H})_{n=1,\dots,N^H}$, we introduce sets $(L_n)_{n=1,\dots,N^H}$ of indices in $\{1,\dots,N^h\}$ such that:

$$\bigcup_{n=1}^{N^H} L_n = \{1,\dots,N^h\}, \tag{D.32}$$

and we will write:

$$\Omega_n^H = \bigcup_{p\in L_n} \Omega_p^h. \tag{D.33}$$

Then it is equivalent to state $\mathrm{supp}(w_n^{0,H}) \subset \overline{\Omega_n^H}$, and to require the unknowns (α_{pn}) in (D.7a) to satisfy:

$$\forall n \in \{1,\dots,N^H\},\ \forall p \in \{1,\dots,N^h\} \setminus L_n,\ \alpha_{pn} = 0. \tag{D.34}$$

We introduce a reciprocal set-valued function $\tilde{L}$ defined by:

$$\forall p \in \{1,\dots,N^h\},\ \tilde{L}_p = \big\{n \in \{1,\dots,N^H\} \ / \ p \in L_n\big\}. \tag{D.35}$$

Because of (D.32), $\tilde{L}_p$ is non empty. Observe that (D.34) can be rewritten:

$$\forall p \in \{1,\dots,N^h\},\ \forall n \in \{1,\dots,N^H\} \setminus \tilde{L}_p,\ \alpha_{pn} = 0. \tag{D.36}$$

An illustration of these definitions is given in Figure D.1. For the fine graph in Figure D.1(a), we set $L_1 = \{1,2,3,4,5,6,7\}$, $L_2 = \{5,6,8,9,13,14\}$ and $L_3 = \{7,8,10,11,12\}$. One obtains, for instance, the sets $\tilde{L}_7 = \{1,3\}$ and $\tilde{L}_4 = \{1\}$.

2. The constant preservation constraint of (D.10) is obtained by enforcing:

$$\sum_{n=1}^{N^H} \alpha_{pn} = 1,\ \forall p \in \{1,\dots,N^h\}. \tag{D.37}$$

Indeed, we substitute (D.7a) into the constant preservation constraint and we use (D.29):

$$\begin{aligned}\sum_{n=1}^{N^H} w_n^{0,H}(x) &= \sum_{n=1}^{N^H}\Big(\sum_{p=1}^{N^h} \alpha_{pn} w_p^{0,h}(x)\Big)\\ &= \sum_{p=1}^{N^h} w_p^{0,h}(x)\Big(\sum_{n=1}^{N^H} \alpha_{pn}\Big) = \sum_{p=1}^{N^h} w_p^{0,h}(x) = 1.\end{aligned}$$

3. We compute the prolongation matrix α with the encoded support constraint (D.34) by minimizing the energy functional from (D.10) under the constant preservation constraint (D.37). A method is given in [79].

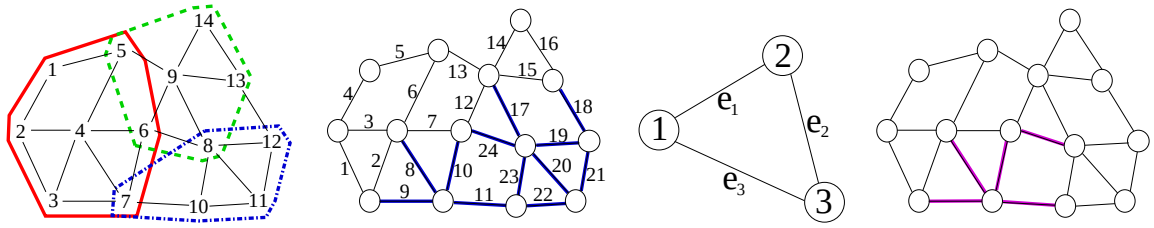
More notations and definitions: how to relate coarse and fine, edge and nodal elements

For all n in $\{1,\dots,N^H\}$, we let C_n be a set of fine edge indices; these fine edges start or end in a node whose index is in L_n:

$$C_n = \big\{i \in \{1,\dots,E^h\} \ / \ i = \overline{pq}^h \text{ with } p \text{ or } q \in L_n\big\}. \tag{D.38}$$

By (D.31) and (D.33), for i in C_n, the support of the fine edge basis function $w_i^{1,h}$ is included in the n-th coarse nodal domain:

$$i \in C_n \implies \mathrm{supp}(w_i^{1,h}) \in \overline{\Omega_n^H}. \tag{D.39}$$



(a) Fine graph with node indices. Sets $(L_n)_{n=1,\dots,3}$ are surrounded.

(b) Fine graph with edge indices. Set C_3 is represented by bold edges.

(c) Coarse graph.

(d) Set I_{e_3} is represented by bold edges.

Figure D.1: Representation of the fine and coarse graphs, sets $(L_n)_{n=1,\dots,3}$, C_3 and I_{e_3}.

We say that the fine edge function $w_i^{1,h}$ contributes to the gradient of the coarse nodal function $w_n^{0,H}$ if i belongs to C_n. Indeed, from (D.5) and (D.7a), we get:

$$\operatorname{grad} w_n^{0,H} = \sum_{p=1}^{N^h} \alpha_{pn} \left(\sum_{i=1}^{E^h} G_{ip}^h w_i^{1,h} \right) = \sum_{i=1}^{E^h} (G^h \alpha_{\bullet n})_i w_i^{1,h},$$

and from (D.28), (D.34) and (D.38):

$$\forall n \in \{1,\dots,N^H\},\ \forall i \in \{1,\dots,E^h\} \setminus C_n,\ (G^h \alpha_{\bullet n})_i = 0. \tag{D.40}$$

We introduce the reciprocal set-valued function $\tilde{C}$:

$$\forall i \in \{1,\dots,E_h\},\ \tilde{C}_i = \big\{ n \in \{1,\dots,N^H\} \ / \ i \in C_n \big\}. \tag{D.41}$$

In Figure D.1(b), the fine edges are numbered, set C_3 is highlighted and we can draw out, for instance, the set $\tilde{C}_8 = \{1,3\}$.

According to (D.35), (D.38) and (D.41) we infer:

$$\tilde{C}_i = \tilde{L}_p \cup \tilde{L}_q, \text{ if } i = \overline{pq}^h. \tag{D.42}$$

As a consequence, $\tilde{C}_i$ cannot be empty and the constant preservation constraint (D.37) with (D.36) becomes:

$$\sum_{n \in \tilde{C}_i} \alpha_{pn} = \sum_{n \in \tilde{C}_i} \alpha_{qn} = 1, \text{ if } i = \overline{pq}^h. \tag{D.43}$$

By analogy with the beginning of Subsection D.4.1, we let $\mathcal{S}^H$ be the set of edges of a simple oriented connected graph whose vertices are indexed by $\{1,\dots,N^H\}$; the elements of $\mathcal{S}^H$ are indexed by $e \in \{1,\dots,E^H\}$, and the inverse of the bijection $e \mapsto (n(e), m(e))$ is $(n,m) \mapsto \overline{nm}^H$. Similarly to the matrix G^h, we introduce the coarse edge-node incidence matrix G^H of dimension $E^H \times N^H$ with $E^H = |\mathcal{S}^H|$.

The set $\mathcal{S}^H$ has to satisfy a compatibility relation with the sets C_n:

$$\forall (n,m) \in \mathcal{S}^H,\ C_n \cap C_m \neq \emptyset. \tag{D.44}$$

For e in $\{1,\dots,E^H\}$, we define the non empty index set I_e of the coarse edge corresponding to $e = \overline{nm}^H$:

$$I_e = C_n \cap C_m \text{ if } e = \overline{nm}^H. \tag{D.45}$$

Then from (D.39), if i belongs to I_e, the fine edge function λ_i^h has its support included in $\overline{\mathcal{U}_e}$:

$$i \in I_e \implies \operatorname{supp}(w_i^{1,h}) \in \overline{\mathcal{U}_e} \text{ with } \mathcal{U}_e = \Omega_n^H \cap \Omega_m^H.$$

The coarse graph in Figure D.1(c) is related to the fine in Figure D.1(a). Set I_{e_3} is represented in Figure D.1(d).

In order to satisfy the support constraint $\mathrm{supp}(w_e^{1,H}) \subset \overline{\mathcal{U}_e}$, we impose on the unknowns (β_{ie}) in (D.7b) the constraints:

$$\forall e \in \{1,\dots,E^H\},\ \forall i \in \{1,\dots,E^h\} \setminus I_e,\ \beta_{ie} = 0. \tag{D.46}$$

By introducing the reciprocal set-valued function $\tilde{I}$ defined by:

$$\forall i \in \{1,\dots,E^h\},\ \tilde{I}_i = \big\{e \in \{1,\dots,E^H\} \ / \ i \in I_e\big\}, \tag{D.47}$$

the relation (D.46) can be rewritten as:

$$\forall i \in \{1,\dots,E^h\},\ \forall e \in \{1,\dots,E^H\} \setminus \tilde{I}_i,\ \beta_{ie} = 0. \tag{D.48}$$

The set $\tilde{I}_i$ might be empty for some $i \in \{1,\dots,E^h\}$ *i.e.* some fine edge function might not contribute to any coarse edge function but, as the sets I_e are never empty, we get:

$$\bigcup_{i=1}^{E^h} \tilde{I}_i = \{1,\dots,E^H\}.$$

Graph- and set-theoretical properties of $\tilde{I}$ and $\tilde{C}$

We will show that it is not necessary to enforce the constraints (D.8b) for all (i,n). We need the following lemma which is a direct consequence of (D.41), (D.45) and (D.47):

Lemma D.1. *If $\tilde{I}_i$ is non empty, the following equivalence holds:*

$$\overline{mn}^H \in \tilde{I}_i \Leftrightarrow (m,n) \in \mathcal{S}^H \text{ and } \{m,n\} \subset \tilde{C}_i.$$

Thus the edges indexed by $\tilde{I}_i$ are the edges of $\mathcal{S}^H$ whose extremities have their index in $\tilde{C}_i$. From this result, we will first deduce that the relations in (D.8b) reduce to $0 = 0$ in the case $n \notin \tilde{C}_i$.

Proposition 4.1. *$\forall i \in \{1,\dots,E^h\}$, $\forall n \in \{1,\dots,N^H\} \setminus \tilde{C}_i$, the constraints* (D.8b) *of index (i,n) are implicitly satisfied.*

Proof. According to (D.40), the (i,n) coefficient of the right-hand side of (D.8b) vanishes.

Conversely, according to (D.48) the left-hand side of (D.8b) is given by:

$$\sum_{e \in \tilde{I}_i} \beta_{ie} G_{en}^H.$$

It vanishes if $\tilde{I}_i$ is empty. If $e = \overline{lm}^H$ belongs to $\tilde{I}_i$, Lemma D.1 implies that l and m belongs to $\tilde{C}_i$. However for G_{en}^H not to vanish, m and l must be equal to n and this contradicts the assumption $n \notin \tilde{C}_i$. □

Secondly, the following proposition points out that some relations in (D.8b) are linearly dependent.

Proposition 4.2. *Let i be a fixed index in $\{1,\dots,E^h\}$ and m some index in $\tilde{C}_i$. If the constraints* (D.8b) *of index (i,n) are satisfied for all n in $\tilde{C}_i \setminus \{m\}$, the constraint of index (i,m) is also satisfied.*

Proof. It is sufficient to prove the equality:

$$\sum_{n \in \tilde{C}_i} \left(\sum_{e=1}^{E^H} \beta_{ie} G_{en}^H \right) = \sum_{n \in \tilde{C}_i} \left(\sum_{r=1}^{N^h} G_{ir}^h \alpha_{rn} \right).$$

For $i = \overline{pq}^h$, according to (D.28) and (D.43), the right-hand side vanishes since:

$$\sum_{n \in \tilde{C}_i} \left(\sum_{r=1}^{N^h} G_{ir}^h \alpha_{rn} \right) = \sum_{n \in \tilde{C}_i} \alpha_{qn} - \sum_{n \in \tilde{C}_i} \alpha_{pn}.$$

For the left-hand side, according to (D.48), it comes:

$$\sum_{n\in\tilde{C}_i}\Big(\sum_{e=1}^{E^H}\beta_{ie}G^H_{en}\Big)=\sum_{e\in\tilde{I}_i}\beta_{ie}\Big(\sum_{n\in\tilde{C}_i}G^H_{en}\Big).$$

It vanishes if $\tilde{I}_i=\emptyset$ and also if $\tilde{I}_i\neq\emptyset$ since for all e in $\tilde{I}_i$, by Lemma D.1 and the definition of G^H:

$$\sum_{n\in\tilde{C}_i}G^H_{en}=\sum_{n=1}^{N^H}G^H_{en}=0.$$

□

D.4.2 How to choose the R_n's in order to simplify the computational process and to have a unique coarse edge basis

The choice of the index set J_n and therefore the definition of the projections R_n can now be explicitly defined.

All the matrices $(K_e)_{e=1,\dots,E^H}$ are assumed to be symmetric positive definite; from the definition of the matrix D (Subsection D.3.2), it is clear that D^{-1} is symmetric positive definite. The matrix $T^tD^{-1}T$ is also positive definite if and only if T is one-to-one. Therefore, our aim is to choose the R_n's so that the kernel of T will be then reduced to the set $\{0\}$.

We define the reciprocal set-valued function $\tilde{J}$:

$$\forall i\in\{1,\dots,E^h\},\ \tilde{J}_i=\big\{n\in\{1,\dots,N^H\}\ /\ i\in J_n\big\}. \tag{D.49}$$

Since the sets $\tilde{I}_i$ and $\tilde{J}_i$ may be empty, we introduce the set $\tilde{F}$:

$$\tilde{F}=\big\{i\in\{1,\dots,E^h\}\ /\ \tilde{I}_i\neq\emptyset\text{ and }\tilde{J}_i\neq\emptyset\big\}. \tag{D.50}$$

Then, for i in $\tilde{F}$, we can define $\tilde{G}^{H,i}$ the matrix of dimensions $|\tilde{I}_i|\times|\tilde{J}_i|$ extracted from the matrix G^H by keeping only the rows $e\in\tilde{I}_i$ and the columns $n\in\tilde{J}_i$.

Proposition 4.3. *The following conditions are necessary and sufficient for T to be one-to-one:*

$$\forall i\in\tilde{F},\ \ker(\tilde{G}^{H,i})=\{0\}, \tag{D.51a}$$

$$\forall i\in\{1,\dots,E^h\},\ \tilde{I}_i=\emptyset\Rightarrow\tilde{J}_i=\emptyset. \tag{D.51b}$$

Proof. Let ρ be in $\mathbb{R}^M$. For n in F defined in (D.16), we observe that:

$$R^t_n\rho_n=\sum_{i\in J_n}\rho_{n,i}u_i, \tag{D.52}$$

where u_i is the i-th vector of the canonical basis of $\mathbb{R}^{E^h}$. We let T operate on ρ, and for that purpose we look at the block-wise result. Let e be in $\{1,\dots,E^H\}$, by the definition of T in (D.22):

$$(T\rho)_e=Q_e\Big(\sum_{n\in F}G^H_{en}\Big(\sum_{i\in J_n}\rho_{n,i}u_i\Big)\Big), \tag{D.53a}$$

$$=\sum_{n=1}^{N^H}\sum_{i\in J_n}G^H_{en}\rho_{n,i}Q_e(u_i). \tag{D.53b}$$

From (D.49), the components of ρ indexed by the set of couples (n,i) with n in $\{1,\dots,N^H\}$ and i in J_n can equivalently be indexed by i in $\{1,\dots,E^h\}$ and n in $\tilde{J}_i$. Thus $(T\rho)_e$ can be rewritten as:

$$(T\rho)_e = \sum_{i=1}^{E^h} \Big(\sum_{n\in\tilde{J}_i} G^H_{en}\rho_{n,i}\Big) Q_e(u_i), \tag{D.53c}$$

but $Q_e(u_i)$ is equal to 0 if i is not in I_e, then:

$$(T\rho)_e = \sum_{i\in I_e} \Big(\sum_{n\in\tilde{J}_i} G^H_{e,n}\rho_{n,i}\Big) Q_e(u_i). \tag{D.53d}$$

Since the vectors $(Q_e(u_i))_{i\in I_e}$ are independent in $\mathbb{R}^{|I_e|}$, relation (D.53d) means that $T\rho$ vanishes iff:

$$\forall e \in \{1,\dots,E^H\},\ \forall i \in I_e,\ \sum_{n\in\tilde{J}_i} G^H_{e,n}\rho_{n,i} = 0, \tag{D.54a}$$

Remarking that $i \in I_e$ corresponds to $e \in \tilde{I}_i$, (D.54a) is equivalent to:

$$\forall i \in \{1,\dots,E^h\},\ \forall e \in \tilde{I}_i,\ \sum_{n\in\tilde{J}_i} G^H_{e,n}\rho_{n,i} = 0, \tag{D.54b}$$

which can be written, according to the definition of $\tilde{G}^{H,i}$, as:

$$\forall i \in \tilde{F},\ \sum_{n\in\tilde{J}_i} \tilde{G}^{H,i}\rho^i = 0, \tag{D.54c}$$

where $\rho^i \in \mathbb{R}^{|\tilde{J}_i|}$ is of components $\rho_{n,i}$ for n in $\tilde{J}_i$. We infer from the above equivalences:

$$T\rho = 0 \Leftrightarrow \forall i \in \tilde{F},\ \rho^i \in \ker(\tilde{G}^{H,i}). \tag{D.55}$$

We prove now the sufficiency of (D.51). If we assume condition (D.51b) is satisfied then:

$$\tilde{F} = \big\{ i \in \{1,\dots,E^h\} \ / \tilde{J}_i \neq \emptyset \big\},$$

and ρ, vector of $\mathbb{R}^M$, is defined by the knowledge of its blocks ρ^i for all i in $\tilde{F}$. Condition (D.51a) associated with relation (D.55) enables us to conclude:

$$T\rho = 0 \Rightarrow \rho = 0.$$

Then, conditions (D.51) are sufficient for T to be one-to-one.

Conversely, assume that:

$$\exists i_0 \in \tilde{F},\ \ker(\tilde{G}^H_{i_0}) \neq \{0\},$$

Let ρ be in $\mathbb{R}^M \setminus \{0\}$ such that:

$$\rho^j = \{0\},\ \forall j \neq i_0 \text{ and } \rho^{i_0} \in \ker(\tilde{G}^H_{i_0}) \setminus \{0\}.$$

By relation (D.55), ρ is in $\ker(T)$. Hence, condition (D.51a) is necessary for T to be one-to-one.

In the same way, assume that:

$$\exists i_0 \in \{1,\dots,E^h\},\ \tilde{I}_{i_0} = \emptyset \text{ and } \tilde{J}_{i_0} \neq \emptyset,$$

Let ρ in $\mathbb{R}^M \setminus \{0\}$ be such that:

$$\rho^j = \{0\},\ \forall j \neq i_0.$$

Since i_0 is not in $\tilde{F}$, ρ is in $\ker(T)$ from (D.55). Condition (D.51b) is necessary for T to be one-to-one. □

In order to reduce the number of enforced constraints, from Propositions 4.1 and 4.2, a suitable choice of sets $\tilde{J}_i$ is

$$\forall i \in \{1, \ldots, E^h\},\ \tilde{J}_i = \tilde{C}_i \setminus \{m\},\ \text{with } m \in \tilde{C}_i.$$

For such a choice, the condition of Proposition 4.3 can be given in terms of connectivity of induced subgraphs of the graph of edges in $\mathcal{S}^H$. More precisely, for any fine edge i of $\mathcal{S}^h$, let $\mathcal{S}^{H,i}$ be the induced subgraph of the graph of edges in $\mathcal{S}^H$, whose vertices are indexed by $\tilde{C}_i$ and edges by $\tilde{I}_i$.

Corollary D.2. *For $\tilde{J}_i = \tilde{C}_i \setminus \{m\}$, $i = 1, \ldots, E^h$, T is one-to-one iff for all i in $\{1, \ldots, E^h\}$ the induced subgraph $\mathcal{S}^{H,i}$ is connected.*

Proof. Observe first that $\tilde{J}_i$ is empty if and only if the subgraph $\mathcal{S}^{H,i}$ reduces to one vertex which happens by Lemma D.1 if $\tilde{I}_i$ is empty.

Assume now $\tilde{I}_i$ is non empty. By Lemma D.1 the edge-node incidence matrix of the subgraph is deduced from G^H by keeping only the rows $e \in \tilde{I}_i$ and the columns $n \in \tilde{C}_i$. By deleting one column of index $m \in \tilde{C}_i$, the obtained matrix $\tilde{G}^{H,i}$ satisfies the condition (D.51a) of Proposition 4.3 iff the subgraph $\mathcal{S}^{H,i}$ is connected. □

The following subsection is devoted to the construction of index sets L_n and an incidence matrix G^H which satisfy the connectivity condition on the induced subgraph $\mathcal{S}^{H,i}$.

D.4.3 Construction of the index sets L_n

We start from an $N^h \times N^h$ nodal symmetric matrix B^h.When passing from a fine level to a coarser level, we just require:

$$B^h_{pp} \neq 0,\ \forall p \in \{1, \ldots, N^h\} \tag{D.56}$$

and a compatibility condition with the set $\mathcal{S}^h$:

$$(p, q) \in \mathcal{S}^h \implies B^h_{pq} \neq 0. \tag{D.57}$$

This condition means that B^h_{pq} must not vanish on the couples (p, q) corresponding to the edges belonging to $\mathcal{S}^h$, but it might also be different from zero on other couples. In other words, there may be more edges (p, q) determined by the condition $B^h_{pq} \neq 0$ than elements (p, q) or (q, p) in $\mathcal{S}^h$.

As Reitzinger and Schöberl do in [66], we define a map:

$$\begin{aligned} \text{ind} :\ & \mathbb{R}^{N^h} \to \mathbb{R}^{N^H} \\ & p \mapsto \text{ind}(p). \end{aligned} \tag{D.58}$$

The inverse images of n by the mapping ind make up a partition of $\{1, \ldots, N^h\}$ into sets. The set H_n is the aggregate of fine indices indexed by n:

$$H_n = \big\{p \in \{1, \ldots, N^h\} \ /\ \text{ind}(p) = n\big\}, \tag{D.59}$$

An example of a fine graph with the partition of the nodes is given in Figure D.2(a). The H_n's make up an arbitrary partition of $\{1, \ldots, N^h\}$, but there are algorithms for choosing good partitions, for instance the aggregation algorithm proposed in [76, Section 5].

For all fine index p in $\{1, \ldots, N^h\}$, we define:

$$\mathcal{S}_p = \big\{q \in \{1, \ldots, N^h\} \ /\ B^h_{pq} \neq 0\big\}. \tag{D.60}$$

This is the set of all the B^h-neighbors of the nodal variable p and the variable itself. From the compatibility relation (D.57), we remark that:

$$(p, q) \in \mathcal{S}^h \implies q \in \mathcal{S}_p. \tag{D.61}$$

For all n in $\{1, \ldots, N^H\}$, we will write:

$$L_n = \bigcup_{p \in H_n} \mathcal{S}_p \tag{D.62}$$

Observe that the set $\tilde{L}_p$ defined in (D.35) is related to $\mathcal{S}_p$ as follows:

$$\tilde{L}_p = \mathrm{ind}(\mathcal{S}_p). \tag{D.63}$$

We also infer from (D.60) and the symmetry of B^h:

$$p \in \mathcal{S}_q \Leftrightarrow q \in \mathcal{S}_p.$$

Some sets L_n corresponding to the partition of Figure D.2(a) are represented in Figure D.2(b).

Remark D.1. *On the initial mesh, some choices for B are proposed by Reitizinger and Kaltenbacher in [48]. Observe that the simplest way to take into account the connectivity of the initial mesh and to satisfy conditions* (D.56) *and* (D.57) *is to put $B^h = (G^h)^t G^h$.*

For the coarser level, the matrix can be constructed from the Galerkin product $\alpha^t B^h \alpha$.

D.4.4 Definition of a compatible coarse edge incidence matrix G^H

By analogy with Reitzinger and Schöberl in [66], we choose $\mathcal{S}^H$ such that (n,m) or (m,n) is an edge if there exists $p \in H_n$ and $q \in H_m$ defined in (D.59) such that p and q are B^h-neighbors, *i.e.*:

$$\begin{aligned}\mathcal{S}^H = \big\{(n,m) \in \{1,\dots,N^H\}^2 \ / \ n \neq m,\ \exists p,q \in \{1,\dots,N^h\} \\ \text{with } \mathrm{ind}(p) = n,\ \mathrm{ind}(q) = m,\ p < q \text{ and } B^h_{pq} \neq 0\big\}.\end{aligned} \tag{D.64}$$

From this definition, we easily verify that the connectivity property of the fine graph $\mathcal{S}^h$ remains for the coarse graph $\mathcal{S}^H$. Observe also that $B^h_{pq} \neq 0$ implies that p is in $L_{\mathrm{ind}(p)} \cap L_{\mathrm{ind}(q)}$. Then $C_{\mathrm{ind}(p)} \cap C_{\mathrm{ind}(q)}$ is non empty and $\mathcal{S}^H$ satisfies condition (D.44).

The coarse graph coming from the graph partition of Figure D.2(a) is represented in Figure D.2(c). Some sets $\tilde{L}_p$ are also represented in Figure D.2(d).

For proving the connectivity of the induced graph $\mathcal{S}^{H,i}$, we need an intermediate result.

Lemma D.3. *Let $p \in \{1,\dots,N^h\}$ be a fine nodal index and assume that the coarse nodal set $\tilde{L}_p$ has at least two elements. Then* $\mathrm{ind}(p)$ *is connected to the other vertices in $\tilde{L}_p$ by an edge of $\mathcal{S}^H$.*

Proof. Let n be in $\tilde{L}_p \setminus \{\mathrm{ind}(p)\}$. By (D.63), there exists p_n in $\mathcal{S}_p$ such that $n = \mathrm{ind}(p_n)$, then:

$$B^h_{pp_n} \neq 0 \text{ and } \mathrm{ind}(p) \neq \mathrm{ind}(p_n).$$

From definition (D.64) of S^H, we conclude that $(n, \mathrm{ind}(p))$ or $(\mathrm{ind}(p), n)$ is in $\mathcal{S}^H$. □

Proposition 4.4. *For all i in $\{1,\dots,E^h\}$, the coarse graph $\mathcal{S}^{H,i}$ is connected.*

Proof. Let $i = \overline{pq}^h$; then from (D.42) $\tilde{C}_i = \tilde{L}_p \cup \tilde{L}_q$. From (D.61) and (D.63), we infer:

$$\{\mathrm{ind}(p), \mathrm{ind}(q)\} \subset \tilde{L}_p \cap \tilde{L}_q.$$

Assume first $\mathrm{ind}(p) \neq \mathrm{ind}(q)$. Lemma D.3 implies that $\mathrm{ind}(p)$ and $\mathrm{ind}(q)$ are connected and for all n in $\tilde{C}_i \setminus \{\mathrm{ind}(r)\}$ with $r \in \{p,q\}$, n and $\mathrm{ind}(r)$ are connected.

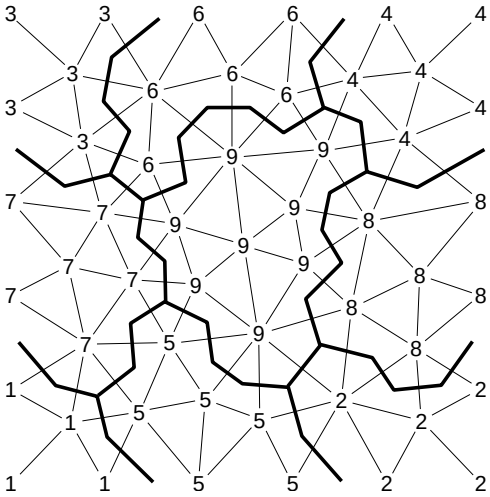
Assume now $\mathrm{ind}(p) = \mathrm{ind}(q) = n$ with $|\tilde{C}_i| > 1$. From Lemma D.3, we infer:

$$\forall m \in \tilde{C}_i \setminus \{n\},\ n \text{ and } m \text{ are connected.}$$

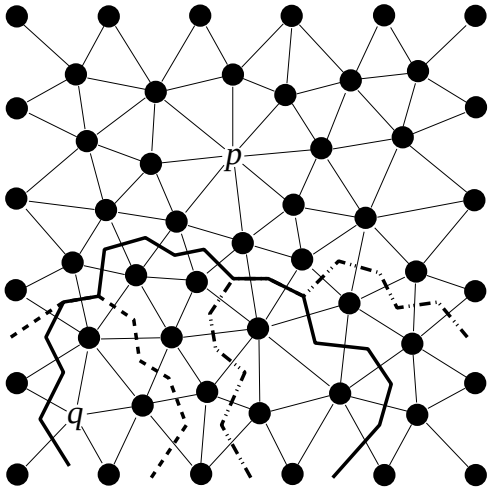
□

Remark D.2. *Reitzinger and Schöberl's approach in [66] can be described with our notations. They chose $L_n = H_n$ for all n in $\{1,\dots,N^H\}$ and the coarse graph G^H defined in* (D.64). *This leads to $|\tilde{L}_p| = 1$ for all fine nodal index p and it implies $|\tilde{C}_i| = 1$ or $|\tilde{C}_i| = 2$ with $|\tilde{I}_i| = 1$. Our construction gives larger edge aggregates, which is a nice situation because in the end nearly all fine edges will contribute to a coarse edge aggregate.*

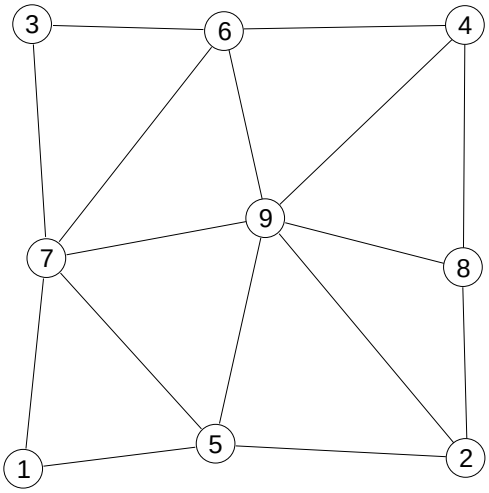
Remark D.3. *For two nested meshes τ^h and τ^H, if we take L_n as the set of indices of the τ^h-neighbors of the coarse node of index n and G^H as the matrix naturally associated with the coarse mesh τ^H, the connectivity condition for Corollary D.2 is clearly satisfied.*



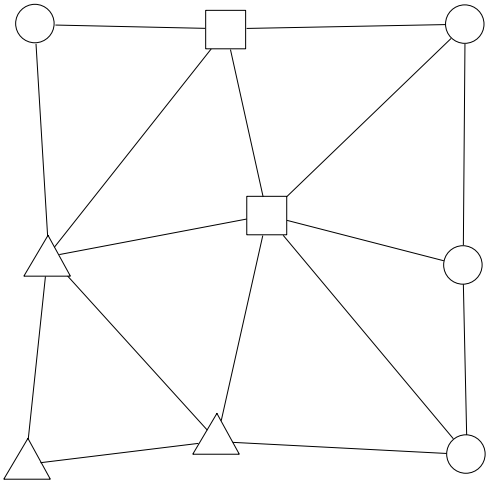
(a) Initial graph and partition



(b) Representation of L_1, L_5 and L_2.



(c) Coarse graph for partition D.2(a)



(d) Representation of $\tilde{L}_q$ (△) and $\tilde{L}_p$ (□). Nodes p and q are in Fig. D.2(b)

Figure D.2: Fig. D.2(a): Initial graph and partition of nodes in sets H_n, Fig. D.2(b): some sets L_n, Fig. D.2(c): construction of the coarse graph corresponding to the partition in Fig. D.2(c) and Fig. D.2(d): representation of the sets $\tilde{L}_q$ and $\tilde{L}_p$ referring to Fig. D.2(b).

D.5 Numerical experiments

D.5.1 Matrix multiplication algorithm

In order to solve linear system (D.26), a non-preconditioned conjugate gradient algorithm is used. Thus only the multiplication of $T^t D^{-1} T$ by a vector of $\mathbb{R}^M$ is required. We can compute $T^t D^{-1} T \rho = \tilde{\rho}$ in three steps:

- Step 1: For $e = 1, \ldots, E^H$, compute:

$$b_e = Q_e \left(\sum_{n \in F} G^H_{en} R^t_n \rho_n \right) = Q_e (R^t_m \rho_m - R^t_n \rho_n), \text{ for } e = \overline{nm}^H. \tag{D.65}$$

- Step 2: For $e = 1, \ldots, E^H$, solve:

$$K_e x_e = b_e. \tag{D.66}$$

- Step 3: For $n = 1, \ldots, N^H$, compute:

$$\tilde{\rho}_n = R_n \left(\sum_{e=1}^{E^H} G^H_{en} Q^t_e x_e \right). \tag{D.67}$$

The step 1 requires $\tilde{M}$ additions and the number of arithmetical operations in step 3 is bounded by $\max_n (\sum_{e=1}^{E^H} |G^H_{en}|) M$. The most expensive part is the resolution of all the local problems: $K_e x_e = b_e$, for all e in $\{1, \ldots, E^H\}$.

Observe that such solutions of local problems are also required for the computation of the prolongation matrix $\bar{\beta}_c$ from the Lagrange multiplier vector ρ_c by using (D.27). For an exact resolution of local problems, a factorization of each matrix K_e must be done.

D.5.2 Choice of the bilinear form b

In the following tables, different bilinear forms b are used for the minimisation. For convenience, we define a few abbreviations:

- A refers to the bilinear form a of the problem. With this choice, spatial variations of γ and δ together with mesh heterogeneity are completely taken into account for the construction of the finite element basis.

- $A + G^h M_\phi^{-1} (G^h)^t$ refers to the bilinear form defined from this matrix, M_ϕ being the mass matrix on nodal elements, with mass lumping; hence M_ϕ is diagonal. This choice enables us to improve the conditioning of local matrices K_e as it will be noted for 2D simulations.

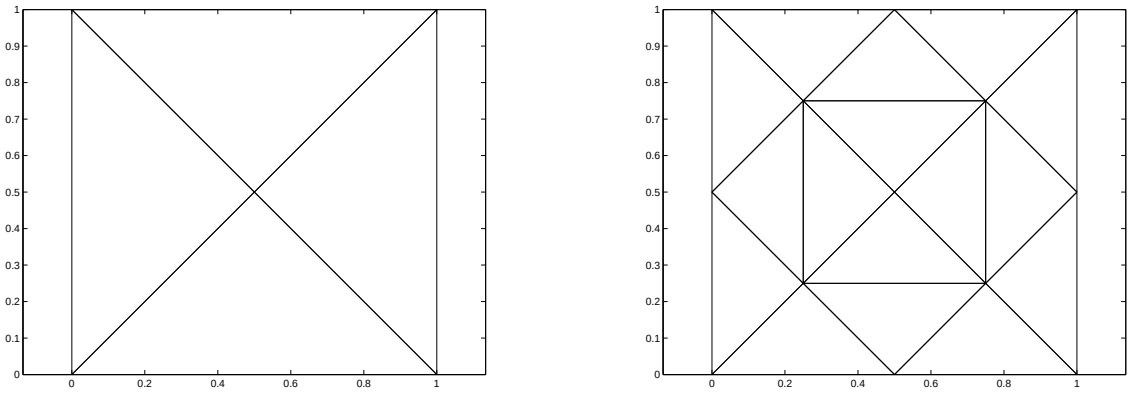
- $S + \mathrm{reg}(\eta)$ refers to the bilinear form $\int_\Omega \mathrm{rot}\, E \cdot \mathrm{rot}\, E'$ added to a local regularisation depending on the parameter η; more precisely the matrix of the local problem is $S_e + \eta \max(\mathrm{diag}(S_e))\, \mathrm{Id}$ where the matrix Id is the identity matrix and S_e is the local matrix computed from $\int_\Omega \mathrm{rot}\, E \cdot \mathrm{rot}\, E'$. The focus is to improve the conditioning of the local matrices, the essential part of the original bilinear form a being kept.

- Id refers to the use of the identity matrix for the local problems. Then no factorisation is needed but the original problem is no more taken into account in the definition of b.

The annotation GSsym means that one symmetric Gauss-Seidel iteration is performed in the local systems instead of a complete resolution in order to significantly reduce the complexity.

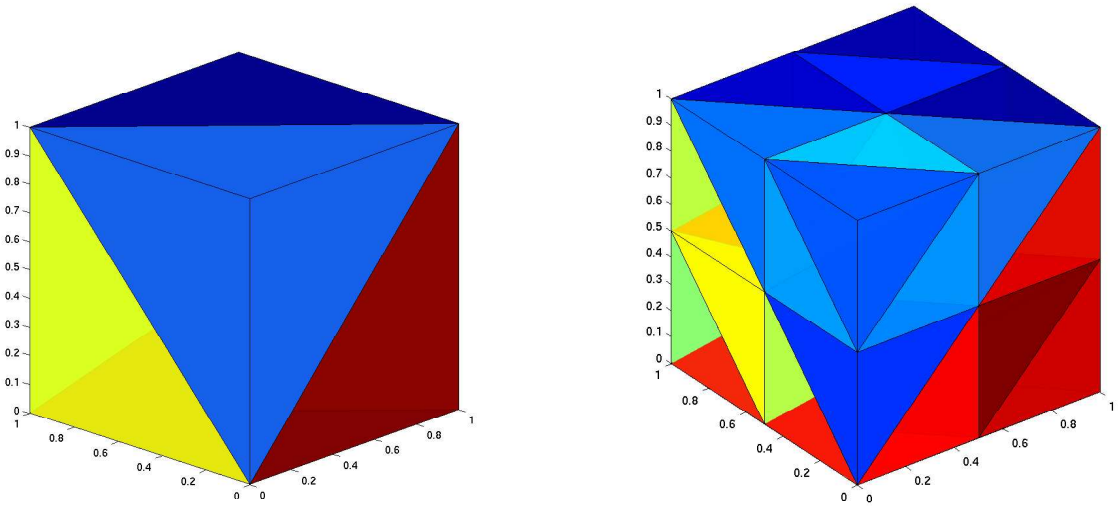
D.5.3 Structured meshes and constant coefficients

In order to validate the proposed choices of bilinear form b, we begin the numerical experiments with structured meshes and constant coefficients in the problem. The coefficients δ and γ are set equal to 1. A sequence of nested meshes $(\tau_k^h)_{k=0,\ldots,4}$ is constructed by a regular refinement starting from a simple triangular or tetrahedral mesh on the unit square and the unit cube (see Fig. D.5.3 and D.4). The source term is given by non-homogeneous boundary conditions. For each level, the edge prolongation matrix β is computed by solving system (3.28) with the right-hand side member ξ defined from the standard nodal prolongation operator [19, Chapter 6]. The sets L_n together with the edge-node incidence matrices are defined frome the meshes as described in Remark D.3.



(a) (b)

Figure D.3: Initial mesh τ_0^h (D.3(a)) and first refinement τ_1^h (D.3(b)).



(a) (b)

Figure D.4: Initial mesh τ_0^h (D.4(a)) and first refinement τ_1^h (D.4(b))

Dimensions

For the finest level, the number of Lagrange multipliers M given by the vector ρ are reported in Table D.1 for the 2D case and Table D.2 for the 3D case. The number of unknowns E^h for each problem is also recalled. Observe that the number of multipliers is less than twice the number of unknowns in the 2D case. The ratio is slightly larger in 3D but more so for problems with few unknowns.

	τ_1^h	τ_2^h	τ_3^h	τ_4^h
M = Nb of multipliers	40	152	592	2336
E^h = Nb of unknowns	20	88	368	1504

Table D.1: Number (Nb) of Lagrange multipliers and unknowns — 2D, structured meshes.

	τ_1^h	τ_2^h	τ_3^h	τ_4^h
M	164	1060	7544	56752
E^h	26	316	3032	26416

Table D.2: Number of Lagrange multipliers and unknowns — 3D, structured meshes.

Computation of Lagrange multipliers

The multipliers are initialized to zero. The number of iterations required to divide the residual by 10^3 during the iterative process is reported in Table D.3 for the 2D case and Table D.4 for the 3D case. These numbers are given for several choices of bilinear forms b. We also give the number of iterations for each level: finest level + ... + coarsest level.

Norm	τ_1^h	τ_2^h	τ_3^h	τ_4^h
A	5	42+11	64+42+14	79+64+42+14
A (GSsym)	11	12+11	12+12+11	11+11+12+11
$A + G^h M_\phi^{-1} (G^h)^t$	4	9+7	9+8+7	9+8+8+7
$A + G^h M_\phi^{-1} (G^h)^t$ (GSsym)	4	9+7	9+8+7	9+8+8+7
$S + \mathrm{reg}(0.1)$	6	19+10	19+19+9	18+19+19+10
Id	2	2+2	2+2+2	2+2+2+2

Table D.3: Number of iterations for the multiplier computation (division of the residual by 10^3) — 2D, structured meshes.

Norm	τ_1^h	τ_2^h	τ_3^h	τ_4^h
A	24	32+13	46+16+15	61+20+17+17
A (GSsym)	23	25+13	32+16+15	39+20+18+17
$A + G^h M_\phi^{-1} (G^h)^t$	41	93+21	181+27+17	381+29+21+16
$A + G^h M_\phi^{-1} (G^h)^t$ (GSsym)	26	42+18	56+21+15	62+22+16+14
$S + \mathrm{reg}(0.1)$	20	22+13	29+15+14	36+18+15+15
Id	3	3+3	3+3+3	3+3+3+3

Table D.4: Number of iterations for the multiplier computation (division of the residual by 10^3) — 3D, structured meshes.

In the 2D case, the behaviour of the algorithm with the original matrix A is not good since the number of iterations increases with the fineness of the initial mesh. The surprising thing is that it disappears when a Gauss-Seidel iteration is used. For the other cases, the behaviour is relatively homogeneous and the number of iterations is almost constant. In 3D, we notice a slow increase in the number of iterations and the behaviour for the original matrix A is not as bad as in the 2D case. The behaviour for the matrix Id is the best in both cases.

Solution of system (D.4)

System (D.4) is solved by a preconditioned conjugate gradient algorithm. The preconditioner is a multilevel method which uses one pre- and one post-smoothing step by Arnold's smoother [4]; on the coarse grid, a direct solver is used. The conjugate gradient method stops when the initial residual has been divided by 10^{10}. Table D.5 gives the results for the 2D case and Table D.6 for the 3D case. The mention "geometric" corresponds to the classical geometric multigrid.

Norm	τ_1^h	τ_2^h	τ_3^h	τ_4^h
A	5	6	7	7
A (GSsym)	5	6	7	9
$A + G^h M_\phi^{-1} (G^h)^t$	5	6	7	8
$A + G^h M_\phi^{-1} (G^h)^t$ (GSsym)	5	6	7	8
$S + \mathrm{reg}(0.1)$	5	6	7	7
Id	5	7	9	11
geometric	5	6	7	7

Table D.5: Number of iterations of the conjugate gradient preconditioned by a multilevel method (division of the residual by 10^{10}) — 2D, structured meshes.

Norm	τ_1^h	τ_2^h	τ_3^h	τ_4^h
A	4	7	10	11
A (GSsym)	4	7	10	11
$A + G^h M_\phi^{-1} (G^h)^t$	4	7	10	11
$A + G^h M_\phi^{-1} (G^h)^t$ (GSsym)	4	7	10	11
$S + \mathrm{reg}(0.1)$	4	7	10	11
Id	4	7	10	11
geometric	4	7	10	11

Table D.6: Number of iterations of the conjugate gradient preconditioned by a multilevel method (division of the residual by 10^{10}) — 3D, structured meshes.

Once the multipliers are computed, and for all choices of bilinear forms b, the efficacy of the geometric and algebraic multigrid methods are comparable.

D.5.4 Unstructured meshes and varying coefficients

We consider a 2D problem whose structure is pictured in Figure D.5; γ is set equal to 1 and δ takes the values displayed in this figure.

A mesh generator gives an initial unstructured mesh. We only control the maximal diameter $h_{\max}$ of elements, and the generator takes the interfaces into account.

The matrix B^h for the initial level is given by $(G^h)^t G^h$; see Remark D.1. The matrix α is obtained from the resolution of (D.10) with c given by $(G^h)^t G^h$. The map ind is constructed from the aggregation algorithm given in [76], on which are defined the sets L_n and the matrix G^H as in (D.62) and (D.64).

Number of unknowns

In Table D.7, the number of unknowns on each level is given for several choices of $h_{\max}$. The notation new denotes our algorithm, and RS denotes the algorithm of Reitzinger and Schöberl given in [66].

Computation of the Lagrange multipliers

Number of multipliers The number of multipliers for every mesh and every level is given in Table D.8. This number is roughly twice the number of unknowns for every level.

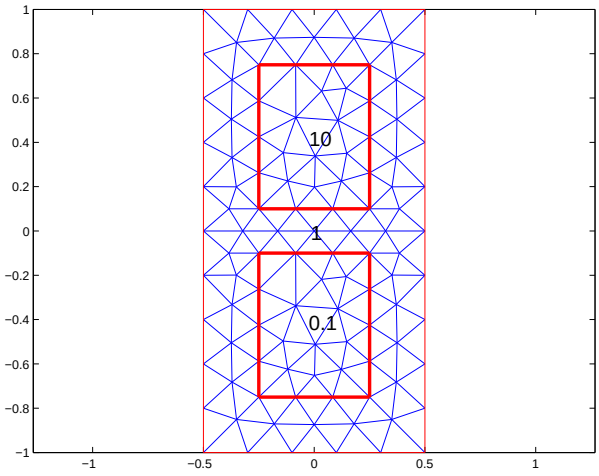


Figure D.5: Geometry of the problem with the values of δ.

Method	$h_{\max} = 0.2$	0.1	0.05	0.025
new	276+31	1005+114+6	3843+405+23	15957+1467+82+3
RS	276+31	1005+114+11	3843+405+29	15957+1467+151+21

Table D.7: Number of unknowns on each level for several choices of $h_{\max}$

$h_{\max} = 0.2$	0.1	0.05	0.025
536	2105+179	8049+759	31249+2946+118

Table D.8: Number of Lagrange multipliers

Stopping criteria on the residual The residual is divided by 10^3. The number of iterations required to compute the Lagrange multipliers with different norms are gathered in Table D.9. Compared to the

Norm	$h_{\max} = 0.2$	0.1	0.05	0.025
A	387	X+134	X+407	X+X+136
A (GSsym)	415	396+X	360+X	348+450+X
$A + G^h M_\phi^{-1} (G^h)^t$	62	103+31	117+46	120+59+26
$A + G^h M_\phi^{-1} (G^h)^t$ (GSsym)	94	128+37	112+58	113+59+27
$S + \mathrm{reg}(0.1)$	74	212+32	319+65	326+200+27
Id	8	13+7	13+12	14+15+2

Table D.9: Number of iterations for the multiplier computation (division of the residual by 10^3) — unstructured meshes. X means that the stopping criterion was not reached after 1000 iterations.

results of Table D.3, we can observe a significant increase for all methods but the results are similarly ordered.

Solution of system (D.4)

The multilevel method is used as a preconditioner; the results for the different norms are gathered in Table D.10.

	$h_{\max} = 0.2$	0.1	0.05	0.025
A	11	12	18	45
A (GSsym)	12	12	17	25
$A + G^h M_\phi^{-1} (G^h)^t$	12	13	17	26
$A + G^h M_\phi^{-1} (G^h)^t$ (GSsym)	12	13	18	27
$S + \mathrm{reg}(0.1)$	12	12	16	25
Id	13	13	18	28
R. S.	14	20	33	56

Table D.10: Number of iterations of the conjugate gradient preconditioned by a two-level method (division of the residual by 10^{10}).

The choices Id and $A + G^h M_\phi^{-1} (G^h)^t$ (GSsym) are the most efficient because they provide us with a performing preconditioner and the lowest computationnal cost for the setup phase.

The prolongation matrix β built by our method provides us with a more efficient preconditioner than the method proposed by Reitzinger and Schöberl in [66]. This validates the theoretical interest of our algebraic method.

However, the construction of the matrix β remains really more expensive than in Reitzinger and Schöberl AMG method. In order to improve the efficacy for the resolution of problem (D.20), an algorithm, which avoids the Lagrange multiplier computation, can be intended [59].

Appendix E

Commutativité entre gradient et prolongement et théorie des graphes

Gradient-prolongation commutativity and graph theory[1]

François Musy, Laurent Nicolas and Ronan Perrussel

Abstract. *This note gives conditions that must be imposed to algebraic multilevel discretizations involving at the same time nodal and edge elements so that a gradient-prolongation commutativity condition will be satisfied; this condition is very important, since it characterizes the gradients of coarse nodal functions in the coarse edge function space. They will be expressed using graph theory and they provide techniques to compute approximation bases at each level.*

E.1 Introduction

Numerical approximation of electric or magnetic field uses often edge finite elements whose relation with nodal finite elements contains important properties at discrete level [42]. In this note we restrict ourselves to lowest order approximation : P_1 for nodal elements and incomplete order 1 for edge elements. In order to solve large problems, multilevel methods are an attractive choice. While, for systems coming from edge element discretisation, Hiptmair [41] proposed multilevel methods using nested meshes, engineering applications do not usually provide structured meshes. Therefore, algebraic multilevel methods are an interesting option: we have to build coarse nodal and edge functions by using aggregates of fine nodal and edge functions. If $(w_p^{0,h})_{p=1,\dots,N^h}$ and $(w_i^{1,h})_{i=1,\dots,E^h}$ respectively denote fine nodal and edge bases, the following linear combinations define coarse nodal and edge functions:

$$w_n^{0,H} = \sum_{p=1}^{N^h} \alpha_{pn} w_p^{0,h}, \ \forall n \in \{1, \dots N^H\}, \tag{E.1a}$$

$$w_e^{1,H} = \sum_{i=1}^{E^h} \beta_{ie} w_i^{1,h}, \ \forall e \in \{1, \dots, E^H\}. \tag{E.1b}$$

By construction, the gradients of fine nodal functions belong to the space of fine edge functions:

$$\forall p \in \{1, \dots, N^h\}, \ \mathrm{grad}(w_p^{0,h}) = \sum_{i=1}^{E^h} G_{ip}^h w_i^{1,h}, \tag{E.2}$$

where G^h is the edge-node incidence matrix of the digraph naturally associated with the initial mesh. The orientation of the edges can be arbitrarily chosen.

[1] Note publiée dans Les Comptes Rendus de l'Académie des sciences Mathématiques; voir [59].

In [66], Reitzinger and Schöberl deduced their smoother from the matrix G^H involved in the relation:

$$\forall n \in \{1, \ldots, N^H\},\ \mathrm{grad}(w_n^{0,H}) = \sum_{e=1}^{E^H} G_{en}^H w_e^{1,H}, \tag{E.3}$$

which states that the gradients of the coarse nodal functions must also belong to the space of coarse edge functions. The matrix G^H is an edge-node incidence matrix as in the structured case. Relation (E.3) does not guarantee the efficacy of the algebraic multilevel method but it leads to relevant strategies.

Gathering Equations (E.1), (E.2) and (E.3), we obtain the matrix relation:

$$G^h \alpha = \beta G^H. \tag{E.4}$$

The matrix α is constructed following for instance the methods defined in [51], which provides a family of coarse nodal functions, making up a partition of unity, whose supports satisfy appropriate conditions.

Knowing the left-hand side of (E.4), we want to choose G^H as an edge-node incidence matrix of a digraph $\mathcal{S}^H$, and we will give conditions on the coarse graph $\mathcal{S}^H$, which ensure the existence of a matrix β satisfying (E.4). Moreover, the proof of the proposition indicates how to choose the degrees of freedom which enables us to define the coarse edge functions. It also helps us to construct β.

E.2 Notation and statement of the problem

Let $(L_n)_{n=1,\ldots,N^H}$ be sets of indices in $\{1, \ldots, N^h\}$ such that:

$$\bigcup_{n=1}^{N^H} L_n = \{1, \ldots, N^h\}. \tag{E.5}$$

The matrix α describes the coarse nodal basis; we assume that it is has been previously computed and it has the following properties:

- the coarse nodal functions make up a partition of unity, which can be algebraically stated as:

$$\forall p \in \{1, \ldots, N^h\},\ \sum_{n=1}^{N^H} \alpha_{pn} = 1, \tag{E.6}$$

- in order to restrict the support of each coarse basis function $w_n^{0,H}$, the indices of the non-zero components of $w_n^{0,H}$ are included in the set L_n, *i.e.*:

$$p \in \{1, \ldots, N^h\} \setminus L_n \implies \alpha_{pn} = 0. \tag{E.7}$$

  The fine nodal function $w_p^{0,h}$ contributes to the coarse nodal function $w_n^{0,H}$ if p belongs to L_n.

We have a reciprocal set-valued function $\tilde{L}$: the set $\tilde{L}_p$ is the set of coarse nodal function indices to which the fine nodal function $w_p^{0,h}$ contributes. For the fine graph in Figure E.1(a), we set $L_1 = \{1, 2, 3, 4, 5, 6, 7\}$, $L_2 = \{5, 6, 8, 9, 13, 14\}$ and $L_3 = \{7, 8, 10, 11, 12\}$. One obtains, for instance, the set $\tilde{L}_7 = \{1, 3\}$.

We define two families of sets of fine edge function indices. We will denote a directed fine edge i by $\overline{pq}^h$ where p and q are respectively the starting and ending nodes of the edge i. A similar notation is used for a directed coarse edge $e = \overline{mn}^H$.

The set C_n is the set of indices of fine edges which have an extremity in L_n:

$$C_n = \left\{ i \in \{1, \ldots, E^h\} : i = \overline{pq}^h,\ p \in L_n \text{ or } q \in L_n \right\}. \tag{E.8}$$

The fine edge function $w_i^{1,h}$ contributes to the gradient of the coarse nodal function $w_n^{0,H}$ if i belongs to C_n. Indeed, for the directed fine edge $i = \overline{pq}^h$, G_{ir}^h is equal to -1 if $r = p$ and $+1$ if $r = q$. Moreover, if p and q are not in L_n, the components α_{pn} and α_{qn} vanish according to (E.7); therefore:

$$i \in \{1, \ldots, E^h\} \setminus C_n \implies (G^h \alpha_{\bullet n})_i = 0, \tag{E.9}$$

where $\alpha_{\bullet n}$ denotes the n-th column of α. The reciprocal set-valued function $\tilde{C}$ is such that $\tilde{C}_i$ is the set of coarse nodal function indices to whose gradient the fine edge function $w_i^{1,h}$ contributes. On Figure E.1(b), the fine edges are numbered, set C_3 is highlighted and we can note, for instance, the set $\tilde{C}_8 = \{1, 3\}$.

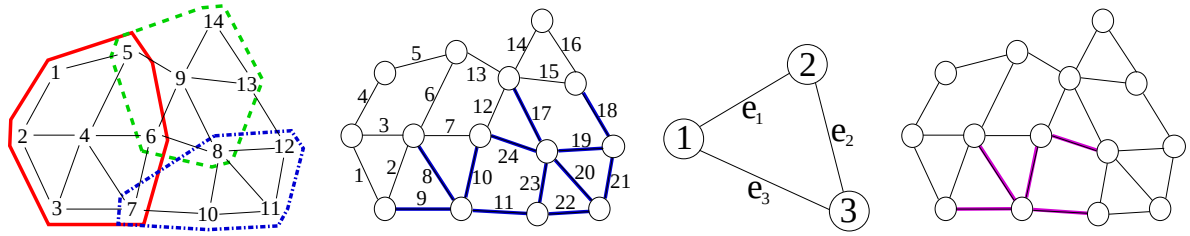
Let $e = \overline{mn}^H$ be an edge of the coarse graph $\mathcal{S}^H$; we define:

$$I_e = C_n \cap C_m. \tag{E.10}$$

By analogy with the structured case and for restricting the support of $w_e^{1,H}$, we enforce:

$$i \in \{1, \ldots, E^h\} \setminus I_e \implies \beta_{ie} = 0. \tag{E.11}$$

The fine edge function $w_i^{1,h}$ contributes to the coarse edge function $w_e^{1,H}$ if i belongs to I_e. The set-valued function $\tilde{I}$ is such that $\tilde{I}_i$ is the set of coarse edge function indices to which the fine edge function $w_i^{1,h}$ contributes. The coarse graph in Figure E.1(c) is related to the fine in Figure E.1(a). Set I_{e_3} is represented in Figure E.1(d).



(a) Fine graph with node indices. Sets $(L_n)_{n=1,\ldots,3}$ are surrounded.

(b) Fine graph with edge indices. Set C_3 is represented by bold edges.

(c) Coarse graph.

(d) Set I_{e_3} is represented by bold edges.

Figure E.1: Representation of the fine and coarse graphs, sets $(L_n)_{n=1,\ldots,3}$, C_3 and I_{e_3}.

The following statement can be easily deduced from (E.8) and the definition of G^h:

Lemma E.1. *If i denotes the edge $\overline{pq}^h$, $\tilde{C}_i = \tilde{L}_p \cup \tilde{L}_q$.*

In order to simplify notations, we introduce the set $\tilde{F} = \left\{i \in \{1, \ldots, E^h\} : \tilde{I}_i \neq \emptyset\right\}$, since some fine edge functions might not contribute to any coarse edge functions.

For any fine edge i, let $\mathcal{S}^{H,i}$ be the induced subgraph defined by $\tilde{C}_i$: the vertices of $\mathcal{S}^{H,i}$ are the vertices of $\mathcal{S}^H$, which are indexed by the elements of $\tilde{C}_i$ and the edges of $\mathcal{S}^{H,i}$ are those edges of $\mathcal{S}^H$ whose extremities are vertices of $\mathcal{S}^{H,i}$.

The following lemma is a direct consequence of definition (E.10):

Lemma E.2. *For any edge $i \in \tilde{F}$, the edges of $\mathcal{S}^{H,i}$ are those edges of $\mathcal{S}^H$ which are indexed by $\tilde{I}_i$.*

We may now state precisely our main result, which gives a necessary and sufficient condition on the coarse graph $\mathcal{S}^H$ permitting the resolution of (E.4):

Proposition E.3. *For all matrices α satisfying conditions* (E.6) *and* (E.7)*, there exists a matrix β satisfying* (E.11) *and solving* (E.4) *iff for all i, the induced subgraph $\mathcal{S}^{H,i}$ is connected.*

E.3 The essential steps of the proof

First step. Many relations in (E.4) reduce to $0 = 0$: this is the case for $n \notin \tilde{C}_i$.

Indeed, according to (E.9) the (i, n) coefficient of the right-hand side of (E.4) vanishes.

Conversely, if e does not belong to $\tilde{I}_i$, according to (E.11) and the definition of $\tilde{I}_i$, β_{ie} vanishes and:

$$\sum_{e=1}^{E^H} \beta_{ie} G^H_{en} = \sum_{e \in \tilde{I}_i} \beta_{ie} G^H_{en}. \tag{E.12}$$

On the other hand if the directed coarse edge e denoted by $\overline{lm}^H$ belongs to $\tilde{I}_i$, Lemma E.2 implies that l and m belongs to $\tilde{C}_i$. However, for G^H_{en} not to vanish for all e, m or l must be equal to n, which means that n belong to $\tilde{C}_i$, and this contradicts the assumption $n \notin \tilde{C}_i$.

Second step. We look at all the other equations, *i.e.* those for which $n \in \tilde{C}_i$. We note that (E.12) remains and that the edges indexed by $\tilde{I}_i$ are precisely those of the graph $\mathcal{S}^{H,i}$ according to Lemma E.2.

We assume now $i \in \tilde{F}$ and we define $G^{H,i}$ as the edge-node incidence matrix of $\mathcal{S}^{H,i}$ and the (i,n) equation of (E.4) is rewritten:

$$\sum_{e \in \tilde{I}_i} \beta_{ie} G^{H,i}_{en} = \Theta_{i,n} \text{ where } \Theta_{i,n} = \sum_{r \in L_n} G^h_{ir} \alpha_{rn}. \tag{E.13}$$

This could be satisfied for all couples (i,n) such that $n \in \{1, \dots, N^H\}$ and $i \in C_n$ or equivalently $i \in \{1, \dots, E^h\}$ and $n \in \tilde{C}_i$. For a fixed i, we may write that $\beta_{i\bullet}$, the i-th row of β satisfies the system:

$$\sum_{e \in \tilde{I}_i} \beta_{ie} G^{H,i}_{en} = \Theta_{i,n}, \ \forall n \in \tilde{C}_i. \tag{E.14}$$

Thus, we solve line by line for β and we see that (E.14) is a flow problem whose solution is of the form:

$$\beta_{i\bullet} = \beta'_{i\bullet} + \beta''_{i\bullet}. \tag{E.15}$$

with $(\beta''_{i\bullet})^t \in \ker(G^{H,i})^t$ and $\beta'_{i\bullet}$ a particular solution.

More precisely, let $\mathcal{T}^i$ be a spanning tree for $\mathcal{S}^{H,i}$; call Γ^i the edge-node incidence matrix associated with $\mathcal{T}^i$; we know that Γ^i has $|\tilde{C}_i|-1$ rows and $|\tilde{C}_i|$ columns, and it is of rank $|\tilde{C}_i|-1$. We choose a vertex m in Γ^i and we solve the system:

$$\sum_{e \in \mathcal{E}(\mathcal{T}^i)} \beta'_{ie} \Gamma^i_{en} = \Theta_{i,n}, \ \forall n \in \tilde{C}_i \setminus \{m\}, \tag{E.16}$$

where $\mathcal{E}(\mathcal{T}^i)$ denotes the set of indices of the edges of $\mathcal{T}^i$. The system (E.16) is a regular system of $|\tilde{C}_i|-1$ equations with $|\tilde{C}_i|-1$ unknowns, and we put β'_{ie} equal to 0 if e is in $\tilde{I}_i \setminus \mathcal{E}(\mathcal{T}^i)$.

It remains to show that the forgotten equation of index m in (E.16) is automatically satisfied. Indeed, by denoting i by $\overline{pq}^h$, we sum the right-hand side of (E.14) with respect to $n \in \tilde{C}_i$:

$$\sum_{n \in \tilde{C}_i} \Theta_{i,n} = \sum_{n \in \tilde{L}_p \cup \tilde{L}_q} \alpha_{qn} - \alpha_{pn} = 0, \tag{E.17}$$

since in view of (E.6) and (E.7), $\sum_{n \in \tilde{L}_p \cup \tilde{L}_q} \alpha_{pn} = \sum_{n \in \tilde{L}_p} \alpha_{pn} = 1$ and $\sum_{n \in \tilde{L}_p \cup \tilde{L}_q} \alpha_{qn} = \sum_{n \in \tilde{L}_q} \alpha_{qn} = 1$.

On the other hand, if we sum the left-hand side of (E.14) with respect to $n \in \tilde{C}_i$, we obtain:

$$\sum_{n \in \tilde{C}_i} \sum_{e \in \tilde{I}_i} \beta_{ie} G^{H,i}_{en} = \sum_{e \in \tilde{I}_i} \beta_{ie} \sum_{n \in \tilde{C}_i} G^{H,i}_{en} = 0, \tag{E.18}$$

since each line of $G^{H,i}$ contains only two non-zero coefficients $+1$ and -1.

If $i \notin \tilde{F}$, $|\tilde{C}_i| = |\tilde{L}_p| = |\tilde{L}_q| = 1$ and the relation $\sum_{e=1}^{E^H} \beta_{ie} G^H_{en} = \Theta_{i,n}$ is satisfied from (E.12) and (E.17).

Now we assume that $\mathcal{S}^{H,i}$ is not connected and we denote by $\hat{C}_i$ the nodes of a connected component. For the same reasons as in (E.18), if β satisfies (E.11) one gets $\sum_{n \in \hat{C}_i} \sum_{e=1}^{E^H} \beta_{ie} G^H_{en} = 0$.

However we can construct a matrix α satisfying (E.6) and (E.7) such that $\sum_{n\in\hat{C}_i}\sum_{r=1}^{N^h} G^h_{ir}\alpha_{rn} \neq 0$. In fact, in view of (E.7), for $i = \overline{pq}^h$ we can write:

$$\sum_{n\in\hat{C}_i}\sum_{r=1}^{N^h} G^h_{ir}\alpha_{rn} = \sum_{n\in\hat{C}_i} \alpha_{qn} - \alpha_{pn} = \sum_{\hat{C}_i\cap\tilde{L}_q} \alpha_{qn} - \sum_{\hat{C}_i\cap\tilde{L}_p} \alpha_{pn}. \tag{E.19}$$

Since $\hat{C}_i$ is strictly included in $\tilde{L}_p \cup \tilde{L}_q$, we will have $\hat{C}_i \cap \tilde{L}_p \neq \tilde{L}_p$ or $\hat{C}_i \cap \tilde{L}_q \neq \tilde{L}_q$. Depending on the situation, we can construct a suitable matrix α such that:

$$\Big(\sum_{\hat{C}_i\cap\tilde{L}_q} \alpha_{qn} = 1 \text{ and } \sum_{\hat{C}_i\cap\tilde{L}_p} \alpha_{pn} = 0\Big) \text{ or } \Big(\sum_{\hat{C}_i\cap\tilde{L}_q} \alpha_{qn} = 0 \text{ and } \sum_{\hat{C}_i\cap\tilde{L}_p} \alpha_{pn} = 1\Big).$$

For these matrices α, the condition defined by (E.4) cannot be ensured.

E.4 Construction of the coarse edge functions

For a coarse graph satisfying the condition of Proposition E.3 and by using the decomposition (E.15), any compatible matrice can be written $\beta = \beta' + \beta''$, where the complete matrices are defined by gathering the lines of index i $\beta_{i\bullet}$, $\beta'_{i\bullet}$ and $\beta''_{i\bullet}$. The computation of each $\beta'_{i\bullet}$ can be done by solving system (E.16). As concerns $\beta''_{i\bullet}$, a basis of the kernel of $(G^{H,i})^t$ is given by a set of k_i independent cycles of $\mathcal{S}^{H,i}$. Then, $\sum_{i\in\tilde{F}} k_i$ degrees of freedom should be determined by minimising an appropriate energy functional; such a problem is introduced in [60] and can be related to explanations in [51].

We thank Michelle Schatzman for many fertile discussions.

Bibliography

[1] Yves Achdou. Décomposition de domaines pour les équations aux dérivées partielles. *Matapli*, (60):35–47, 1999.

[2] D. N. Arnold, R. S. Falk, and R. Winther. Differential complexes and stability of finite element methods. I. The de Rham complex. In *Proceedings of the IMA workshop on Compatible Spatial Discretizations for PDE*, 2005. `http://ima.umn.edu/~arnold/papers/nacomplexes.pdf`.

[3] Douglas N. Arnold. Differential complexes and numerical stability. In *Proceedings of the International Congress of Mathematicians, Vol. I (Beijing, 2002)*, pages 137–157, Beijing, 2002. Higher Ed. Press.

[4] Douglas N. Arnold, Richard S. Falk, and Ragnar Winther. Multigrid in $H(\mathrm{div})$ and $H(\mathrm{curl})$. *Numer. Math.*, 85(2):197–217, 2000.

[5] O. Axelsson and V. A. Barker. *Finite element solution of boundary value problems.* Computer Science and Applied Mathematics. Academic Press Inc., Orlando, FL, 1984. Theory and computation.

[6] Norman Balabanian and Theodore A. Bickart. *Electrical Network Theory.* John Wiley and Sons, Inc., 1969.

[7] J. Baranger, C. Brezinski, C. Carasso, J.M. Chassery, F. Chatelin, Maître J.F., J. Roux, and G. Wanner. *Analyse numérique*, volume 38 of *Enseignement des sciences.* Hermann, 1991.

[8] R. Beck. Algebraic Multigrid by Components Splitting for Edge Elements on Simplicial Triangulations. *Preprint SC 99-40, ZIB*, December 1999.

[9] R. Beck. Graph-Based Algebraic Multigrid for Lagrange-Type Finite Elements on Simplicial Meshes. *Preprint SC 99-22, ZIB*, July 1999.

[10] Rudolf Beck and Ralf Hiptmair. Multilevel solution of the time-harmonic Maxwell's equations based on edge elements. *Internat. J. Numer. Methods Engrg.*, 45(7):901–920, 1999.

[11] J. P. Bérenger. A Perfectly Matched Layer for the absorption of electromagnetic waves. *J. Comput. Phys.*, 114(2):185–200, 1994.

[12] C. Berge and A. Ghouila-Houri. *programmes, jeux et réseaux de transport.* Dunod, 1962.

[13] Pavel B. Bochev, C. Garasi, J. Hu, A. Robinson, and R. Tuminaro. An improved algebraic multigrid method for solving Maxwell's equations. *SIAM J. Sci. Comput.*, 25(2):623–642 (electronic), 2003.

[14] A. Bossavit. Whitney forms: a class of finite elements for three-dimensional computations in electromagnetism. *IEE Proc. A*, 135(8):493–500, Nov. 1988.

[15] Alain Bossavit. *Computational electromagnetism.* Electromagnetism. Academic Press Inc., San Diego, CA, 1998. Variational formulations, complementarity, edge elements.

[16] James H. Bramble, Joseph E. Pasciak, Jun Ping Wang, and Jinchao Xu. Convergence estimates for multigrid algorithms without regularity assumptions. *Math. Comp.*, 57(195):23–45, 1991.

[17] A. Brandt, S. McCormick, and J. Ruge. Algebraic multigrid (AMG) for sparse matrix equations. In *Sparsity and its applications (Loughborough, 1983)*, pages 257–284. Cambridge Univ. Press, Cambridge, 1985.

[18] Franklin H. Branin Jr. The algebraic-topological basis for network analogies and the vector calculus. In *Proceedings of the Symposium On Generalized Networks*, 1966.

[19] Susanne C. Brenner and L. Ridgway Scott. *The mathematical theory of finite element methods*, volume 15 of *Texts in Applied Mathematics*. Springer-Verlag, New York, second edition, 2002.

[20] M. Brezina, A. J. Cleary, R. D. Falgout, V. E. Henson, J. E. Jones, T. A. Manteuffel, S. F. McCormick, and J. W. Ruge. Algebraic multigrid based on element interpolation (AMGe). *SIAM J. Sci. Comput.*, 22(5):1570–1592 (electronic), 2000.

[21] Haïm Brezis. *Analyse fonctionnelle*. Collection Mathématiques Appliquées pour la Maîtrise. [Collection of Applied Mathematics for the Master's Degree]. Masson, Paris, 1983. Théorie et applications. [Theory and applications].

[22] V. Cingoski, R. Tokuda, S. Noguchi, and H. Yamashita. Fast multigrid solution method for nested edge-based finite element meshes. *IEEE Trans. on Mag.*, 36(4):1539–1542, 2000.

[23] M. Clemens and Thomas Weiland. Comparaison of Krylov-Type Methods for Complex Linear Systems Applied to High-Voltage Problems. *IEEE Trans. on Magn.*, 34(5):3335–3338, September 1998.

[24] M. Clemens and Thomas Weiland. Iterative Methods for the Solution of Very Large Complex Symmetric Linear Systems of Equations in Electrodynamics. *IEEE Trans. on Magn.*, 34(5):3335–3338, Septembre 1998.

[25] D. Colton and R. Kress. *Inverse Acoustic and Electromagnetic Scattering Theory*. Springer-Verlag, 1992.

[26] James W. Demmel, Stanley C. Eisenstat, John R. Gilbert, Xiaoye S. Li, and Joseph W. H. Liu. A supernodal approach to sparse partial pivoting. *SIAM J. Matrix Analysis and Applications*, 20(3):720–755, 1999.

[27] Reinhard Diestel. *Graph theory*, volume 173 of *Graduate Texts in Mathematics*. Springer-Verlag, 2005.

[28] Howard C. Elman, Oliver G. Ernst, and Dianne P. O'Leary. A multigrid method enhanced by Krylov subspace iteration for discrete Helmhotz equations. *SIAM J. Sci. Comput.*, 23(4):1291–1315 (electronic), 2001.

[29] Bjorn Engquist and Andrew Majda. Absorbing boundary conditions for the numerical simulation of waves. *Math. Comp.*, 31(139):629–651, 1977.

[30] R. W. Freund and N. M. Nachtigal. A new Krylov-subspace method for symmetric indefinite linear systems. In *Proceedings of the 14th IMACS World Congress on Computational and Applied Mathematics*, pages 1253–1256, 1994.

[31] Roland W. Freund. Conjugate gradient-type methods for linear systems with complex symmetric coefficient matrices. *SIAM J. Sci. Statist. Comput.*, 13(1):425–448, 1992.

[32] Roland W. Freund, Gene H. Golub, and Noël M. Nachtigal. Iterative solution of linear systems. In *Acta numerica, 1992*, Acta Numer., pages 57–100. Cambridge Univ. Press, Cambridge, 1992.

[33] P. J. Frey and P.-L. George. *Maillages - Applications aux élḿents finis*. Hermes Science Publications, 1999.

[34] Lucas Gerardo-Giorda and Ana Alonso-Rodriguez. New non-overlapping domain decomposition methods for the time-harmonic maxwell system. Technical Report 529, CMAP, École Polytechnique, April 2004.

[35] H. De Gersem, D. Lahaye, S. Wandewalle, and K. Hameyer. Comparison of quasi minimal residual and bi-conjugate gradient iterative methods to solve complex symmetric systems arising from time-harmonic magnetics simulations. *COMPEL*, 18(3):298–310, 1999.

[36] Jayadeep Gopalakrishnan, Joseph E. Pasciak, and Leszek F. Demkowicz. Analysis of a multigrid algorithm for time harmonic Maxwell equations. *SIAM J. Numer. Anal.*, 42(1):90–108 (electronic), 2004.

[37] Paul W. Gross and P. Robert Kotiuga. *Electromagnetic theory and computation: a topological approach*, volume 48 of *Mathematical Sciences Research Institute Publications*. Cambridge University Press, Cambridge, 2004.

[38] G. Haase, U. Langer, S. Reitzinger, and J. Schöberl. A general approach to algebraic multigrid methods. Technical Report 33, SFB, 2000.

[39] G. Haase, U. Langer, S. Reitzinger, and J. Schöberl. Algebraic multigrid methods based on element preconditioning. *Int. J. Comput. Math.*, 78(4):575–598, 2001.

[40] Wolfgang Hackbusch. *Iterative solution of large sparse systems of equations*, volume 95 of *Applied Mathematical Sciences*. Springer-Verlag, New York, 1994. Translated and revised from the 1991 German original.

[41] R. Hiptmair. Multigrid method for Maxwell's equations. *SIAM J. Numer. Anal.*, 36(1):204–225 (electronic), 1999.

[42] R. Hiptmair. Finite elements in computational electromagnetism. *Acta Numer.*, 11:237–339, 2002.

[43] J. Hu, R. Tuminaro, P. Bochev, C. Garasi, and A. Robinson. Toward an h-independent algebraic multigrid method for Maxwell's equations. To appear in SIAM J. Sci. Computing, 2005.

[44] H. Igarashi. On the property of the curl-curl matrix in finite-element analysis with edge elements. *IEEE Trans. on Mag.*, 37(5):3129–3132, Sept. 2001.

[45] Frank Ihlenburg and Ivo Babuška. Finite element solution of the Helmholtz equation with high wave number. II. The *h-p* version of the FEM. *SIAM J. Numer. Anal.*, 34(1):315–358, 1997.

[46] J. Synka J.K. Kraus. An agglomeration-based multilevel-topology concept with application to 3d-fe meshes. Technical report, RICAM, 2004.

[47] Jim E. Jones and Panayot S. Vassilevski. AMGe based on element agglomeration. *SIAM J. Sci. Comput.*, 23(1):109–133 (electronic), 2001.

[48] M. Kaltenbacher and S. Reitzinger. Algebraic multigrid methods for nodal and edge based discretizations of Maxwell's equations. *International Compumag Society Newsletter*, 9(2):15–23, 2002.

[49] Ferdinand Kickinger. Algebraic multi-grid for discrete elliptic second-order problems. In *Multigrid methods V (Stuttgart, 1996)*, volume 3 of *Lect. Notes Comput. Sci. Eng.*, pages 157–172. Springer, Berlin, 1998.

[50] P. R. Kotiuga. Essential Arithmetic For Evaluating Three Dimensional Vector Finite Element Interpolation Schemes. *IEEE Trans. on Mag.*, 27(6):5208–5210, November 1991.

[51] J. Mandel, M. Brezina, and P. Vaněk. Energy optimization of algebraic multigrid bases. *Computing*, 62(3):205–228, 1999.

[52] G. Meurant. *Computer Solution of Large Linear Systems*. Elsevier Science B.V., 1999.

[53] T. Mifune, T. Iwashita, and M. Shimasaki. A fast solver for fem analyses using the parallelized algebraic multigrid method. *IEEE Trans. on Mag.*, 38(2):369–372, 2002.

[54] T. Mifune, T. Iwashita, and M. Shimasaki. Algebraic multigrid method for nonsymmetric matrices arising in electromagnetic finite-element analyses. *IEEE Trans. on Mag.*, 39(3):1670–1673, May 2003.

[55] T. Mifune, T. Iwashita, and M. Shimasaki. New algebraic multigrid preconditioning for iterative solvers in electromagnetic finite edge-element analyses. *IEEE Trans. on Mag.*, 39(3):1677–1680, 2003.

[56] T. Mifune, T. Iwashita, and M. Shimasaki. A parallel algebraic multigrid solver for fast magnetic edge-element analyses. *IEEE Trans. on Mag.*, 41(5):1660–1663, 2005.

[57] Peter Monk. *Finite element methods for Maxwell's equations*. Numerical Mathematics and Scientific Computation. Oxford University Press, New York, 2003.

[58] I. Munteanu. Tree-cotree condensation properties. *ICS Newsletter*, 9(1), March 2002.

[59] F. Musy, L. Nicolas, and R. Perrussel. Gradient-prolongation commutativity and graph theory. *C. R. Acad. Sci. Paris*, 341(11):707–712, 2005.

[60] F. Musy, L. Nicolas, R. Perrussel, and M. Schatzman. Compatible coarse nodal and edge elements through energy functionals. Technical Report 394, UMR MAPLY, 2004. `http://maply.univ-lyon1.fr/~perrussel/report.pdf`.

[61] Fran¸cois Musy, Laurent Nicolas, and Ronan Perrussel. Compatible coarse nodal and edge elements through energy functionals. Submitted to SIAM J. Sci. Computing.

[62] J.-C. Nédélec. Mixed finite elements in $\mathbf{R}^3$. *Numer. Math.*, 35(3):315–341, 1980.

[63] R. Perrussel, L. Nicolas, and F. Musy. Un préconditionneur adapté à l'opérateur rotationnel en éléments finis. 4th European Conference on Numerical Methods in Electromagnetism, NUMELEC 2003, october 2003.

[64] R. Perrussel, L. Nicolas, and F. Musy. An efficient preconditioner for linear systems issued from the finite-element method for scattering problems. *IEEE Trans. on Mag.*, 40(2):1080–1083, march 2004.

[65] R. Perrussel, L. Nicolas, F. Musy, and M. Schatzman. Preconditioners for finite element method in scattering problems. 4th European Congress on Computational Methods in Applied Sciences and Engineering, ECCOMAS 2004, july 2004.

[66] S. Reitzinger and J. Schöberl. An algebraic multigrid method for finite element discretizations with edge elements. *Numer. Linear Algebra Appl.*, 9(3):223–238, 2002.

[67] J. W. Ruge and K. Stüben. Algebraic multigrid. In *Multigrid methods*, volume 3 of *Frontiers Appl. Math.*, pages 73–130. SIAM, Philadelphia, PA, 1987.

[68] Yousef Saad. *Iterative methods for sparse linear systems*. Society for Industrial and Applied Mathematics, Philadelphia, PA, second edition, 2003.

[69] M. Schinnerl, J. Schöberl, and M. Kaltenbacher. Nested multigrid methods for the fast numerical computation of 3d magnetic fields. *IEEE Trans. on Mag.*, 36(4):1557–1560, 2000.

[70] M. Schinnerl, J. Schöberl, M. Kaltenbacher, and R. Lerch. Multigrid methods for the three-dimensional simulation of nonlinear magnetomechanical systems. *IEEE Trans. on Mag.*, 38(3):1497–1511, 2002.

[71] J. R. Shewchuk. What is a Good Linear Element? Interpolation, Conditionning, and Quality Measures. 11th International Meshing Roundtable, 2002.

[72] N. Siauve, L. Nicolas, C. Vollaire, and C. Marchal. 3D modelling of electromagnetic fields in local hyperthermia. *Eur. Phys. J. AP.*, 21:243–250, 2003.

[73] V. Spasov, S. Noguchi, and H. Yamashita. Fast 3-d edge element analysis by the geometric multigrid method using an accelerated symmetric gauss-seidel smoother. *IEEE Trans. on Mag.*, 39(3):1685–1688, 2003.

[74] U. Trottenberg, C. W. Oosterlee, and A. Schüller. *Multigrid.* Academic Press Inc., San Diego, CA, 2001. With contributions by A. Brandt, P. Oswald and K. Stüben.

[75] H. A. van der Vorst and J. B. M. Melissen. A Petrov-Galerkin type method for solving $Ax = b$, where A is symmetric complex. *IEEE Trans. Mag.*, 26(2):706–708, 1990.

[76] P. Vaněk, J. Mandel, and M. Brezina. Algebraic multigrid by smoothed aggregation for second and fourth order elliptic problems. *Computing*, 56(3):179–196, 1996. International GAMM-Workshop on Multi-level Methods (Meisdorf, 1994).

[77] C. Vollaire and L. Nicolas. Implementation of a finite element and absorbing boundary conditions package on a parallel shared memory computer. *IEEE Trans. on Mag.*, 34(5):3343–3346, September 1998.

[78] Christian Wagner. Introduction to Algebraic Multigrid. Technical report, University of Heidelberg, 1999. course notes.

[79] W. L. Wan, Tony F. Chan, and Barry Smith. An energy-minimizing interpolation for robust multigrid methods. *SIAM J. Sci. Comput.*, 21(4):1632–1649 (electronic), 1999/00.

[80] K. Watanabe and H. Igarashi. On robustness of edge-based finite-element analysis using algebraic multigrid method. *COMPEL*, 24(2):408–417, 2005.

[81] K. Watanabe, H. Igarashi, and T. Honma. Comparison of geometric and algebraic multigrid methods in edge-based finite-element analysis. *IEEE Trans. on Mag.*, 41(5):1672–1675, 2005.

[82] H. Whitney. *Geometric Integration Theory.* Princeton Univ. Press, Princeton, 1957.

[83] Jinchao Xu. Iterative methods by space decomposition and subspace correction. *SIAM Rev.*, 34(4):581–613, 1992.

[84] Jinchao Xu and Ludmil Zikatanov. On an energy minimizing basis for algebraic multigrid methods. *Comput. Vis. Sci.*, 7(3-4):121–127, 2004.

AUTORISATION DE SOUTENANCE

Vu les dispositions de l'arrêté du 25 avril 2002,

Vu la demande du Directeur de Thèse

Monsieur L. NICOLAS

et les rapports de

Monsieur A. TOSELLI
Docteur - ETH Zürich - Rämistrasse 101 - 8092 ZURICH - Allemagne

et de

Monsieur P. DULAR
Chercheur Qualifié FNRS - ELAP - Université de Liège - Institut Montefiore - B28 - 4000 LIEGE - Belgique

Monsieur PERRUSSEL Ronan

est autorisé à soutenir une thèse pour l'obtention du grade de **DOCTEUR**

Ecole doctorale Mathématique et Informatique Fondamentale

Fait à Ecully, le 17 octobre 2005

P/Le Directeur de l'E.C.L.
Le Directeur des Etudes

ECOLE CENTRALE DE LYON
SCOLARITE
3ème Cycle

J. JOSEPH

Méthodes multiniveau algébriques pour les éléments d'arête.
Application à l'électromagnétisme.

Résumé : Le calcul numérique du champ électrique ou magnétique intervient aussi bien dans la mise au point d'outils de communication performants que dans les problèmes de compatibilité électromagnétique des systèmes électriques ou de modélisation de l'interaction champ-vivant. Ce calcul est fréquemment fondé sur la discrétisation des équations de Maxwell par la méthode des éléments finis d'arête. Il conduit alors à la résolution d'un système linéaire creux mais généralement de grande taille.

L'objectif de ce travail est de proposer une méthode multiniveau algébrique pour la résolution des systèmes linéaires issus d'une discrétisation par la méthode des éléments finis d'arête. En effet, les méthodes itératives multiniveau construisent des algorithmes qui s'avèrent être les plus performants pour certaines classes d'équations aux dérivées partielles. Ces méthodes s'appuient, dans leur version géométrique, sur une hiérarchie de maillages emboîtés. Cependant pour des applications réalistes cette hiérarchie ne peut pas toujours être construite et il faut définir algébriquement les différents niveaux.

La stratégie algébrique de définition des niveaux grossiers que nous proposons repose sur la construction de fonctions grossières nodales et d'arête vérifiant une contrainte de compatibilité. En outre, les fonctions grossières d'arête doivent minimiser une fonctionnelle d'énergie. Ce problème de minimisation avec contrainte est résolu par deux techniques : l'une utilise les multiplicateurs de Lagrange, l'autre s'appuie sur la résolution d'une suite de problèmes de flot dans un graphe. Des expériences numériques illustrent les performances de différentes versions de notre méthode.

Mots-clés : électromagnétisme, éléments finis d'arête, résolution de systèmes linéaires, méthodes multigrille algébriques, minimisation sous contraintes.

Algebraic multilevel methods for edge elements.
Application to electromagnetism.

Abstract: The computation of the electric or magnetic field plays a key-role in the design of efficient communication tools, in the electromagnetic compatibility of electronic systems and also in the modelling of the interaction between the field and living tissues. This computation is frequently based on the discretisation of Maxwell's equations by the edge element method. Then, it leads to solve a linear system with a sparse but usually large matrix.

The aim of this work is to introduce an algebraic multilevel method for solving linear systems coming from the edge element method. Indeed, iterative multilevel methods generate algorithms which are the most efficient for some classes of partial differential equations. These methods are founded in their geometric version on a hierarchy of nested meshes; however, for realistic applications, this hierarchy cannot be constructed and the levels are then required to be algebraically defined.

Our algebraic strategy for defining the coarse levels is founded on the construction of nodal and edge coarse bases, which have to minimise an energy functional. This minimisation problem with constraint is solved by two techniques: for the first one Lagrange multipliers are used, for the second technique a sequence of flow problems in a graph is solved. Some numerical experiments illustrate the performances of the different versions of our method.

Keywords: electromagnetism, edge elements, resolution of linear systems, algebraic multigrid methods, constrained minimisation.

Discipline : Mathématiques appliquées.
Laboratoire : Institut Camille Jordan, UMR CNRS 5208 et Centre de Génie Électrique de Lyon, UMR CNRS 5005. Ecole Centrale de Lyon, 69134 Ecully cedex.