UNIVERSITÉ PARIS 13 NORD

IMACS

École doctorale Galilée

Doctorat

THÈSE

pour obtenir le grade de docteur délivré par

l'Université Paris 13

Spécialité doctorale "Mathématiques Appliquées"

présentée et soutenue publiquement par

Kieran DELAMOTTE

le xx xxxx 2016

Une étude du rang du noyau de l'équation de Helmholtz : application des $\mathscr{H}$-matrices à l'EFIE

Directeur de thèse : **M. OLIVIER LAFITTE**
Co-encadrant de thèse : **M. TOUFIC ABBOUD**

Jury

NN.,	Professeur	Examinateur
NN.,	Professeur	Rapporteur
NN.,	Professeur	Examinateur
NN.,	Professeur	Examinateur

Université Paris XIII
Laboratoire Analyse, Géométrie et Applications (LAGA)
Villetaneuse, France

Résumé

La résolution de problèmes d'onde par une méthode d'éléments finis de frontière (BEM) conduit à des systèmes d'équations linéaires pleins dont la taille augmente très vite pour les applications pratiques. Il est alors impératif d'employer des méthodes de résolution dites rapides. La méthode des multipôles rapides (FMM) accèlere la résolution de ces systèmes par des algorithmes itératifs. La méthode des $\mathscr{H}$-matrices permet d'accélérer les solveurs directs nécessaires aux cas d'application massivement multi-seconds membres. Elle a été introduite et théoriquement justifiée dans le cas de l'équation de Laplace. Néanmoins elle s'avère performante au-delà de ce qui est attendu pour des problèmes d'onde relativement haute fréquence. L'objectif de cette thèse est de comprendre pourquoi la méthode fonctionne et proposer des améliorations pour des fréquences plus élevées.
Une $\mathscr{H}$-matrice est une représentation hiérarchique par arbre permettant un stockage compressé des données grâce à une séparation des interactions proches (ou *singulières*) et lointaines (dites *admissibles*). Un bloc admissble a une représentation de rang faible de type UV^T tandis que les interactions singulières sont représentées par des blocs pleins de petites tailles. Cette méthode permet une approximation rapide d'une matrice BEM par une $\mathscr{H}$-matrice ainsi qu'une méthode de factorisation rapide de type Cholesky dont les facteurs sont également de type $\mathscr{H}$-matrice.
Nous montrons la nécessité d'un critère d'admissibilité dépendant de la fréquence et introduisons un critère dit *de Fresnel* basé sur la zone de diffraction de Fresnel. Ceci permet de contrôler la croissance du rang d'un bloc et nous proposons une estimation précise de celui-ci à haute fréquence à partir de résultats sur les fonctions d'onde sphéroïdales. Nous en déduisons une methode de type HCA-II, robuste et fiable, d'assemblage rapide compressé à la précision voulue.
Nous étudions les propriétés de cet algorithme en fonction de divers paramètres et leur influence sur le contrôle et la croissance du rang en fonction de la fréquence.

Nous introduisons la notion de section efficace d'interaction entre deux clusters vérifiant le critère de Fresnel. Si celle-ci n'est pas dégénérée, le rang du bloc croît au plus linéairement avec la fréquence ; pour une interaction entre deux clusters coplanaires nous montrons une croissance comme la racine carrée de la fréquence. Ces développements sont illustrés sur des maillages représentatifs des interactions à haute fréquence.

Mots-clés BEM, Solveur Direct Rapide, $\mathscr{H}$-matrice, Noyau de Green, Approximation Directionnelle, Fonctions d'onde sphéroïdales, Électromagnétisme.

Abstract

The boundary elements method (BEM) leads to dense linear systems whose size grows rapidly in pratice ; hence the use of so-called fast methods. The fast multipole method (FMM) accelerates the resolution of BEM systems within an iterative scheme. The $\mathscr{H}$-matrix method speeds up a direct resolution which is needed in massively multiple right-hand sides problems. It has been provably introduced in the context of the Laplace equation. However, the use of $\mathscr{H}$-matrices for relatively high-frequency wave problems leads to results above expectations. This thesis main goal is to provide an explanation of these good results and thus improve the method for higher frequencies.
A $\mathscr{H}$-matrix is a compressed tree-based hierarchical representation of the data associated with an admissibility criterion to separate the near (or *singular*) and far (or *compres-*

sed) fields. An admissible block reads as a UV^T rank deficient matrix while the singular blocks are dense with small dimensions. BEM matrices are efficiently represented by $\mathscr{H}$-matrices and this method also allows for a fast Cholesky factorization whose factors are also $\mathscr{H}$-matrices.
Our work on the admissibility condition emphasizes the necessity of a frequency dependant admissibility criterion. This new criterion is based on the Fresnel diffraction area thus labelled Fresnel admissibility condition. In that case a precise estimation of the rank of a high-frequency block is proposed thanks to the spheroidal wave functions theory. Consequently, a robust and reliable HCA-II type algorithm has been developed to ensure a compressed precision-controlled assembly. The influence of various parameters on this new algorithm behaviour is discussed ; in particular their influence on the control and the growth of the rank according to the frequency. We define the interaction cross section for two Fresnel-admissible clusters and show in the non-degenerate case that the rank growth is linear according to the frequency in the high-frequency regime ; interaction of coplanar clusters results in growth like the square root of the frequency. All these results are presented on meshes adapted to high-frequency interactions.

Keywords BEM, Fast Direct Solver, H-matrix, Green kernel, Directional Approximation, Spheroidal Wave Functions, Electromagnetism.

Table des matières

Liste des figures

Liste des tableaux

Introduction

Éléments finis de frontière

De nombreuses applications industrielles relèvent de la modélisation de problèmes d'ondes. Citons par exemple en acoustique, la réduction de nuisances sonores dans l'automobile ou encore l'optimisation de l'acoustique de salles de concert. En électromagnétisme, le dimensionnement d'antennes de télécommunications ou la détection d'objets sont des autres cas d'application. Pour le problème de la diffraction d'une onde par un obstacle, le champ diffracté peut être représenté sous une forme intégrale. La condition aux limites conduit à une équation intégrale sur la surface Γ de l'obstacle dont la résolution permet de retrouver les courants équivalents sur l'obstacle. Les formules de représentation permettent ensuite de retrouver les champs en tout point de l'espace.

Nous étudions alors l'approximation des équations intégrales par la méthode des éléments finis de frontière ; par opposition à la méthode des éléments finis de volume qui s'intéresse à l'EDP. Les éléments finis de frontière (ou BEM pour *Boundary Element Method*) est une méthode de choix en électromagnétisme pour les applications civiles et militaires. Cette méthode dispose d'une base mathématique solide avec traditionnellement une forte compétence en France. Après discrétisation de la formulation variationnelle de l'équation intégrale, nous nous ramenons à la résolution d'un système linéaire avec une matrice A_h complexe pleine, souvent symétrique. En effet, les opérateurs intégraux mis en jeu ne sont pas locaux ce qui rend les coeffcients généralement non nuls. Dans la pratique, on considère des matrices de discrétisation pour des tailles allant de quelques dizaines de milliers à plusieurs millions d'inconnues.
Cette méthode est connue pour sa grande précision. En effet, la représentation intégrale basée sur le modèle discrétisé vérifie exactement l'EDP et la condition de radiation à l'infini : c'est dans le noyau de Green. L'erreur est due à la discrétisation de la géométrie et à la condition aux limites qui n'est vérifiée qu'au sens faible.
La représentation intégrale permet de simplifier l'étape de maillage (surfacique au lieu de volumique). Notons que l'étape de préparation des données de calcul est très lourde. Les coefficients de la matrice pleine A_h sont obtenus par des intégrales doubles (surface-surface) souvent coûteuses à calculer. Le temps d'assemblage et l'espace mémoire nécessaire pour une telle matrice augmentent comme $\mathcal{O}(\mathrm{N}^2)$; le temps de résolution du système linéaire par une méthode directe augmente lui comme $\mathcal{O}(\mathrm{N}^3)$. Une méthode itérative aurait un coût en $\mathcal{O}(\mathrm{N}^3 \times \mathrm{N}_{\mathrm{iter}})$ où le nombre d'itérations $\mathrm{N}_{\mathrm{iter}}$ dépend à la fois de la géométrie, du maillage, de la fréquence ainsi que du second membre (sans compter le temps de construction d'un tel préconditionneur). De plus, pour un problème d'ondes, le nombre d'inconnues croît comme le carré de la fréquence ce qui entraîne des difficultés à

traiter les cas haute fréquence d'après les complexités précédentes. Des matrices pleines si grandes sont donc évidemment délicates à manipuler dans la pratique ; l'étape dimensionnante est celle de résolution et le problème n'est plus résoluble de manière classique.

D'où une forte activité dans les communautés académiques et industrielles autour du développement de méthodes d'assemblage et de résolution rapides pour de tels systèmes. Ces méthodes consistent en la recherche d'une approximation favorable de la matrice A_h. Nous pouvons distinguer deux grandes classes de solveurs : les solveurs directs rapides et les solveurs itératifs rapides. Ces deux grandes classes de méthodes rapides ne sont pas réellement concurrentes et l'industrie a recours à chacune d'elles selon l'application visée. Par ailleurs, les développements récents tendent à souligner que les idées utilisées pour les méthodes itératives sont souvent adaptées pour les méthodes directes ; et vice versa.

Méthodes itératives rapides

La méthode de référence est la méthode des multipôles rapides (FMM pour *Fast Multipole Method*) développée à la fin des années 1980 par V. Rokhlin et L. Greengard. Cette méthode permet de réduire la complexité du produit matrice-vecteur utilisé dans les méthodes itératives de $\mathcal{O}(N^2)$ à $\mathcal{O}(N\log(N))$ opérations. Cela a immédiatement permis une montée en fréquence pour des applications en électromagnétisme comme la SER (Surface Équivalente Radar). Cette méthode est à présent très largement répandue dans l'industrie et beaucoup de variantes ont été proposées. L'augmentation des performances du matériel informatique a également été bénéfique à cette méthode rapide et des implémentations parallèles efficaces ont été développées ; ll s'agit toujours d'un sujet de recherche actif. Cependant, la FMM souffre malgré tout des défauts des méthodes itératives que sont :

— le problème de conditionnement sur des maillages raffinés ;

— la croissance linéaire du temps calcul en fonction du nombre de seconds membres traités.

Méthodes directes rapides

Pour soulever les difficultés liées aux méthodes itératives, l'emploi d'une méthode directe rapide peut être favorable. Une méthode relativement récente est la méthode des $\mathcal{H}$-matrices développées par W. Hackbusch au début des années 2000. Cette méthode repose sur une représentation hiérarchique et compressée des matrices BEM. Elle permet la construction de solveurs directs rapides présentant les avantages suivants :

— un coût en mémoire en $\mathcal{O}(N\log(N))$;

— un assemblage ainsi qu'une factorisation approchée en $\mathcal{O}(N\log(N))$ opérations dans les cas les plus favorables ;

— une excellente performance sur les problèmes d'onde moyenne fréquence (y compris pour les maillages raffinés et le grand nombre de seconds membres).

Contexte industriel

Le projet Hibox s'inscrit dans le cadre du dispositif RAPID de la DGA (Direction Générale de l'Armement) pour la période 2014-2017. IMACS est le porteur de ce projet auquel

participent également Airbus Group Innovations (AGI) et INRIA. La motivation de ce projet est de construire une bibliothèque générique de solveurs rapides pour la BEM et plus particulièrement :

— une version stable en fréquence de la FMM (*Fast Multipole Method*) ;
— une version de la méthode GMRES massivement multi-seconds membres ;
— une implémentation parallèle et haute fréquence des $\mathscr{H}$-matrices.

L'objectif de cette bibliothèque est également d'être facilement intégrable à un code de calcul client. Cette bibliothèque est constituée d'implémentations efficaces de solveurs très largement utilisés dans l'industrie. Cette thèse s'inscrit dans la partie $\mathscr{H}$-matrice de HIBOX et plus particulièrement dans les applications haute fréquence des $\mathscr{H}$-matrices. De plus, ce travail de thèse est effectué dans le cadre d'un contrat CIFRE de l'ANRT. Les thématiques étudiées ainsi que les résultats numériques obtenus par IMACS grâce à des implémentations performantes d'un code de calcul intégrale sont à la source de la problématique auquel cette thèse répond.

Motivations de cette thèse

Peu de résultats théoriques sont disponibles dans la littérature pour les cas haute fréquence. Certaines méthodes FMM sont construites en vue d'applications haute fréquence mais les solveurs directs rapides sont très peu traités, tant sur le plan théorique que sur les aspects calcul haute performance (parallélisme, *out-of-core*,...).

Pourtant, la méthode des $\mathscr{H}$-matrices usuelle avec une compression de type ACA (et ses variantes) a déjà été appliquée avec succès à des problèmes d'ondes assez généraux et relativement haute fréquence en électromagnétisme et en acoustique dans la thèse de B. Lizé. La parallélisation efficace des $\mathscr{H}$-matrices a permis de traiter des cas peu usuels mais soulève néanmoins plusieurs questions :

— Comment expliquer le bon fonctionnement de la méthode ?
— Quel est le comportement à encore plus haute fréquence ?
— Quelle est la marge d'amélioration ?

Dans le cas du noyau de Green oscillant, on s'attend à ce que le rang augmente de façon rapide et que la méthode soit rapidement inefficace. B. Lizé constate pourtant expérimentalement que la méthode demeure efficace pour des cas d'une taille de l'ordre du million d'inconnues et des fréquences allant jusqu'au cinquième de la longueur d'onde. Ces résultats et le manque d'explications disponibles dans la littérature poussent à une étude minutieuse et systématique du comportement des blocs de la matrice A_h en fonction de la fréquence : c'est la motivation principale de ce travail de thèse. Cette thèse a d'une part pour objectif d'expliquer les bons résultats pratiques observés et d'autre part d'étudier les possibilités d'amélioration. En effet, pour les applications futures il est intéressant d'anticiper les difficultés pouvant être rencontrées et jusqu'à présent inaccessibles aux implementations actuelles. Ces dernières ne permettent pas de traiter des cas à très hautes fréquences.

Plan de la thèse

Le **chapitre**1 décrit la méthode des $\mathscr{H}$-matrices dans le contexte de la résolution d'une équation intégrale par une méthode d'éléments finis de frontière. Après avoir décrit le

problème modèle de ce chapitre, on présente en détails la construction d'une approximation $\mathscr{H}$-matrice. Nous avons choisi d'effectuer une présentation didactique de la méthode en détaillant le plus possible les différentes étapes de cette dernière. Cette construction se déroule en plusieurs étapes :

— la représentation des degrés de liberté par un arbre binaire ;
— la construction d'une $\mathscr{H}$-matrice à partir du produit de deux arbres binaires ;
— l'assemblage d'une approximation de rang faible pour chaque blocs matriciel produit par l'étape précédente.

Le dernier point ci-dessus fait l'objet du **chapitre** 2. Nous avons préféré exposer cette étape à part pour plusieurs raisons. Il existe plusieurs méthodes permettant de déterminer une approximation de rang faible et certaines ne sont pas propres aux équations intégrales. En effet, d'autres méthodes rapides utilisent le même type d'approximation et les méthodes présentées ici peuvent alors être utilisées dans un autre contexte. Le **chapitre** 2 consiste en une présentation de méthodes algébriques d'une part et analytiques d'autre part.

Le **chapitre** 3 représente la contribution de cette thèse à la méthode des $\mathscr{H}$-matrices. En effet, l'approximation $\mathscr{H}$-matrice a été introduite dans le contexte de l'équation de Laplace et les estimations théoriques de la littérature se basent sur le noyau $G(x, y) = \frac{1}{|x-y|}$. Cependant, cette méthode a été appliquée avec succès à des problèmes d'onde relativement haute fréquence avec des résultats dépassant les attentes. Ce chapitre propose une explication des résultats de la littérature à partir d'une approche directionnelle du noyau. Ce chapitre contient un nouveau critère d'admissibilité dans le contexte des $\mathscr{H}$-matrices. Ce dernier permet également la construction d'une méthode d'assemblage compressée rapide adaptée au noyau oscillant.

Le **chapitre** 4 présente les expérimentations numériques que nous avons faites à partir des nouveaux développements présentés au **chapitre** 3. Plus particulièrement, nous montrons l'intégration des résultats obtenus dans un contexte $\mathscr{H}$-matrice. Cette implémentation est alors testée sur des problèmes d'ondes haute fréquence.

Les chapitres se veulent être le plus indépendants les uns des autres que possible ; c'est pourquoi nous avons choisi de présenter les références à la littérature rencontrées à la fin de chaque chapitre.
Les **chapitres** 1 et 2 permettent la construction d'une méthode $\mathscr{H}$-matrice pour le noyau de Green de l'équation de Laplace. Néanmoins, le **chapitre** 2 ne fait que très peu appel à la méthode des $\mathscr{H}$-matrices puisqu'il décrit la construction d'une approximation de rang faible d'un bloc (une matrice) quelconque. Le **chapitre** 3 est réservé à une étude du noyau de Green oscillant et n'est pas spécifique à la méthode des $\mathscr{H}$-matrices. Enfin, le **chapitre** 4 et notamment sa première partie permet de modifier un code $\mathscr{H}$-matrice existant pour y intégrer la nouvelle condition d'admissibilité développée.

Contributions

Les principales contributions de cette thèse sont :

— Une revue didactique des méthodes de compression couramment utilisées dans les méthodes rapides. Ces méthodes ne sont pas nécessairement concurrentes ; selon les besoins, nous disposons de plusieurs approches rapides, performantes et bien documentées.

— La détermination de critères d'admissibilité pour découper la zone d'interactions lointaines en fonction de la fréquence. On met en valeur un critère d'admissibilité original permettant des interactions plus proches que le critère usuel de la littérature. Ce nouveau critère permet d'expliquer en partie les bons résultats obtenus par la méthode des $\mathscr{H}$-matrices sur des cas relativement haute fréquence. Plus précisément, on constate que les interactions traitées par les $\mathscr{H}$-matrices avec le critère d'admissibilité standard ne sont pas suffisament haute fréquence pour que le critère fréquentiel puisse avoir une influence.

— Sous réserve d'admissibilité, nous présentons une approximation du noyau oscillant à l'aide d'une approche directionnelle. Cette approximation permet d'employer des résultats sur les fonctions d'onde sphéroïdales afin de fournir une estimation du rang à haute fréquence. Nous disposons alors d'un paramètre décrivant la croissance du rang du noyau dans une direction de l'espace. Par produit tensoriel nous disposons d'une estimation du rang du noyau oscillant dans la zone d'admissibilité fréquentielle déterminée.

— L'estimation du rang dont nous disposons nous permet de construire une méthode d'approximation compressée rapide, fiable et robuste du noyau oscillant. Il s'agit d'une amélioration d'une méthode de la littérature dont nous contrôlons les paramètres et la précision.

Chapitre 1

Panorama sur les $\mathscr{H}$- matrices

Sommaire

1.1 Problème modèle

1.1.1 Une équation intégrale

On considère une courbe fermée du plan $\mathbb{R}^2$ que l'on note Γ ainsi qu'un espace de Hilbert de fonctions ou de distributions sur Γ que l'on note H. On note H$'$ son dual [1] et l'on considère l'équation suivante dite équation de Fredholm du premier type

$$\mathscr{A}u = f, \tag{1.1}$$

où l'opérateur intégral $\mathscr{A}$ est défini par

$$\begin{aligned} \mathscr{A} :& \mathrm{H} \mapsto \mathrm{H}' \\ & u \mapsto \int_\Gamma \mathrm{K}(x,y)\,u(y)\,\mathrm{d}\Gamma(y), \end{aligned} \tag{1.2}$$

avec $\mathrm{K}(x,y) = -\frac{1}{2\pi}\log(\|x-y\|_2)$. Le second membre f de l'équation appartient à H$'$. L'opérateur $\mathscr{A}$ est dit opérateur de simple couche et intervient par exemple lors de la résolution de l'équation de Laplace $-\Delta u = 0$ (muni de conditions aux limites appropriées) dans $\mathbb{R}^2$.

Dans la suite de ce paragraphe, l'équation (1.1) (et plus particulièrement la définition de $\mathscr{A}$) est utilisée à titre d'exemple simple afin d'introduire les notions employées dans la suite de ce manuscrit. L'étude numérique de ce type d'équation peut être consultée dans [TA07]. Par ailleurs, le chapitre 4 se propose de résoudre le même type d'équation dans le cas de la diffraction d'une onde électromagnétique (ce qui est également traité dans [TA07]).

1.1.2 Discrétisation par éléments finis

La formulation variationnelle associée à l'équation intégrale (1.1) s'écrit

$$\begin{cases} \text{Trouver } u \in \mathrm{H} \text{ tq :} \\ a(u,v) = {}_{\mathrm{H}'}\!<f, v>_{\mathrm{H}} \quad \forall v(x) \in \mathrm{H}, \end{cases} \tag{1.3}$$

où ${}_{\mathrm{H}'}\!<.,.>_{\mathrm{H}}$ représente le crochet de dualité et,

$$\begin{aligned} a(u,v) &= {}_{\mathrm{H}'}\!<\mathscr{A}[u], v>_{\mathrm{H}} \\ &= \int_\Gamma \int_\Gamma \mathrm{K}(x,y)\,u(y)\,v(x)\,d\Gamma(y)\,d\Gamma(x) \end{aligned} \tag{1.4}$$

$${}_{\mathrm{H}'}\!<f, v>_{\mathrm{H}} = \int_\Gamma f(x)\,v(x)\,d\Gamma(x) \tag{1.5}$$

La résolution de l'équation intégrale (1.1) est équivalente à la résolution de la formulation variationnelle (1.3). On discrétise alors la formulation variationnelle à l'aide de la méthode des éléments finis.
Pour ce faire on considère une discrétisation Γ_h de Γ ainsi qu'un sous-espace de dimension finie H_h de H. On obtient alors la formulation variationnelle discrétisée suivante,

$$\begin{cases} \text{Trouver } u^h \in \mathrm{H}_h \text{ tq :} \\ a_h(u_h, v_h) = < f, v_h > \quad \forall v_h \in \mathrm{H}_h, \end{cases} \tag{1.6}$$

1. Dans l'exemple que nous allons choisir et pour les spécialistes, on prendra $\mathrm{H} = \mathrm{H}^{-\frac{1}{2}}(\Gamma)$ et $\mathrm{H}' = \mathrm{H}^{+\frac{1}{2}}(\Gamma)$, mais peu importe ici.

où,

$$a_h(u_h, v_h) = \int_{\Gamma_h}\int_{\Gamma_h} \mathrm{K}(x,y)\, u^h(y)\, v^h(x)\, \mathrm{d}\Gamma(y)\, \mathrm{d}\Gamma(x) \tag{1.7}$$

$$< f, v^h >= \int_{\Gamma_h} f(x)\, v^h(x)\, \mathrm{d}\Gamma(x) \tag{1.8}$$

L'espace H_h est de dimension finie et l'on note N sa dimension. On se donne alors une base $\{\phi_i\}_{i=1,\ldots,\mathrm{N}}$ de cet espace et l'on écrit la décomposition d'un élément v_h de H_h sur cette base,

$$v_h = \sum_{i=1}^{\mathrm{N}} v_i \phi_i(x). \tag{1.9}$$

Par linéarité, le problème variationnel discret (1.10) est équivalent à

$$\begin{cases} \text{Trouver } u^h \in \mathrm{H}_h \text{ tq :} \\ a_h(u_h, \phi_i) = (f, \phi_i) \qquad \text{pour } i = 1, \ldots, \mathrm{N}. \end{cases} \tag{1.10}$$

On écrit également u^h dans la base $\{\phi_i\}_{i=1,\ldots,\mathrm{N}}$ et la résolution du problème variationnel discret se résume alors à la résolution du système linéaire

$$\mathrm{A}\,\mathrm{U} = \mathrm{F}. \tag{1.11}$$

Les coefficients de la matrice A et du second membre $\mathrm{F} = (f_1, f_2, \ldots, f_{\mathrm{N}})^{\mathrm{T}}$ sont donnés par

$$\mathrm{A}_{ij} = \int_{\Gamma_h}\int_{\Gamma_h} \mathrm{K}(x,y)\phi_j(y)\phi_i(x)\, \mathrm{d}\Gamma_h(x)\, \mathrm{d}\Gamma_h(y) \tag{1.12}$$

$$f_i = \int_{\Gamma_h} f(x)\phi_i(x)\, \mathrm{d}\Gamma_h(x) \tag{1.13}$$

Les coefficients u_i du vecteur $\mathrm{U} = (u_1, u_2, \ldots, u_{\mathrm{N}})^{\mathrm{T}}$ représentent les inconnues du système linéaire.

1.1.3 Résolution d'un système linéaire

Dans la pratique, le traitement d'un système linéaire s'effectue en trois étapes :

1. Le calcul des coefficients A_{ij} (individuels ou par blocs) ;
2. L'assemblage : c'est la copie en mémoire des coefficients dans la matrice ;
3. La résolution du système linéaire.

L'étape (2) est une opération informatique délicate en raison des sauts en mémoire que l'on peut être amené à effectuer. Son coût n'est pas négligeable par rapport au coût de calcul de l'étape (1). Le calcul des coefficients de la matrice est de l'ordre de $c\mathrm{N}^2$ où c est le coût moyen d'un coefficient. Celui-ci varie suivant la nature du noyau $\mathrm{K}(x,y)$.
Dans le cas où les coefficients sont des réels représentés en double précision, chaque réel est décrit par 8 octets soit $8\mathrm{N}^2$ octets au total. Par exemple, pour $\mathrm{N} = 1.10^6$, la mémoire nécessaire pour stocker la matrice est de 8T*o*. Une telle taille n'est pas accessible à l'heure actuelle et cela montre qu'il est impératif d'obtenir des algorithmes permettant une forte réduction de la taille mémoire utilisée.

1.1.3.1 Méthodes de résolution

Classiquement, l'étape de résolution (3) du système (1.11) peut être effectuée de deux façons différentes :

— Par une méthode itérative ;

— Par une méthode directe.

Chaque méthode possède ses avantages et ses inconvénients et l'on retrouve les deux utilisations dans la pratique. On note $Ax = b$ le système linéaire que l'on souhaite résoudre ; A étant une matrice inversible de taille $N \times N$, x et b sont des vecteurs de taille N. Le vecteur x est la solution du système linéaire et b est le second membre.

Méthode itérative Les méthodes itératives [GL96; HS52] utilisent le produit matrice-vecteur afin de déterminer une solution approchée du système linéaire. À partir d'une approximation x_0, choisie arbitrairement, de la solution x on détermine des approximations successives $x_1, \ldots, x_{N_{iter}}$ de la solution x jusqu'à obtenir la convergence de la solution. Pour une norme choisie, par exemple la norme spectrale, on considère que l'on a atteint la convergence lorsque $\|x_{N_{iter}} - x_{N_{iter}-1}\|_2$ est plus petite qu'une tolérance fixée. Le nombre d'itéations N_{iter} requises pour obtenir la convergence de la solution dépend du conditionnement de la matrice et est d'autant plus lente que ce dernier est élevé. Le produit matrice-vecteur étant d'une complexité de l'ordre de $\mathscr{O}(N^2)$ l'obtention d'une solution approchée du système linéaire par une méthode itérative s'obtient en $\mathscr{O}(N^2\ N_{iter})$ opérations. Pour N_{sm} seconds membres, la complexité est donc de l'ordre de $\mathscr{O}(N^2\ N_{iter}\ N_{sm})$ opérations.

Les deux principaux obstacles en pratique sont le nombre de seconds membres ainsi que le conditionnement de la matrice. La complexité d'une méthode itérative devient cubique lorsque le nombre de seconds membres est de l'ordre de N et ceci rend inefficace la méthode.
Le second problème est celui du conditionnement de la matrice. La solution idéale est d'améliorer le conditionnement à l'aide d'une formulation mathématique aboutissant après discrétisation à un meilleur conditionnement. Dans la pratique, cela n'est pas toujours possible et il est plus aisé de recourir à ce que l'on appelle un préconditionneur. Il s'agit de trouver une matrice inversible C -le préconditionneur- et de résoudre le système

$$(C^{-1}A)x = C^{-1}b,$$

au lieu du système original $Ax = b$. C est construit de façon à ce que le conditionnement de $(C^{-1}A)$ soit inférieur à celui de la matrice A. Ainsi, le nombre d'itérations N_{iter} requis est plus faible. Sans donner plus de détails, soulignons toutefois qu'en général on ne construit pas directement la matrice C mais que l'on développe des algorithmes permettant le produit rapide de C^{-1} par un vecteur.
Cette approche itérative est utilisée par exemple dans le contexte de la méthode des multipôles rapides.

Méthode directe La résolution d'un système linéaire par une méthode directe est effectuée en deux étapes : la factorisation de la matrice ainsi que la résolution de systèmes linéaires plus simples.

Factorisation Pour la résolution du problème par une méthode directe, on procède à une factorisation de la matrice afin de se ramener à une structure plus favorable. Par exemple, on peut utiliser la factorisation LU d'une matrice [GL96],

$$\mathrm{A} = \mathrm{LU}, \tag{1.14}$$

où L et U sont respectivement triangulaires inférieure et supérieure. D'autres factorisations existent selon les propriétés de la matrice comme la factorisation de Cholesky $\mathrm{A} = \mathrm{LL}^{\mathrm{T}}$ dans le cas où la matrice est symétrique et définie positive ou encore la factorisation de Crout $\mathrm{A} = \mathrm{LDL}^{\mathrm{T}}$ avec D une matrice diagonale inversible [GL96].
La détermination d'une factorisation LU d'une matrice est une opération d'une complexité cubique.

Résolution À l'aide de la factorisation (1.14), on peut résoudre plus simplement le système linéaire $\mathrm{A}x = b$. En effet, on a l'écriture suivante en remplaçant A par la décomposition LU,

$$(\mathrm{LU})x = b \tag{1.15}$$

La résolution du système linéaire $\mathrm{A}x = b$ est équivalent aux deux systèmes linéaires

$$\mathrm{L}z = b, \tag{1.16}$$
$$\mathrm{U}x = z. \tag{1.17}$$

On résout dans un premier temps le système triangulaire inférieur (1.16) puis le système triangulaire (1.17) où le second membre est la solution du système inférieur. Classiquement, la résolution d'un système linéaire triangulaire requiert de l'ordre de $\frac{\mathrm{N}^2}{2}$ opérations [GL96] soit une complexité de l'ordre de N^2 opérations pour les deux étapes (1.16) et (1.17) dans le cas d'un second membre. Pour N_{sm} seconds membres, le coût de l'étape de résolution est de $\mathscr{O}(\mathrm{N}^2\ \mathrm{N}_{\mathrm{sm}})$ opérations et la résolution par une méthode directe est donc de $\mathscr{O}(\mathrm{N}^3 + \mathrm{N}^2\ \mathrm{N}_{\mathrm{sm}})$ opérations.
Contrairement à une méthode itérative, une méthode directe est particulièrement adaptée au cas « massivement seconds membres » car l'étape de factorisation n'est effectuée qu'une seule fois et est utilisée pour tous les seconds membres. La méthode des $\mathscr{H}$-matrices que l'on décrit dans ce chapitre permet de construire un solveur direct rapide.

1.1.3.2 Structure de la matrice, lien avec la complexité

La matrice de discrétisation A de l'opérateur $\mathscr{A}$ est en général pleine car l'opérateur $\mathscr{A}$ n'est pas local ; le noyau $\mathrm{K}(x, y)$ est non nul donc le terme général de la matrice est non nul également. Une méthode numérique pour résoudre le système linéaire associé à A doit être efficace pour les matrices pleines (ou matrices denses). En effet, considérons le cas d'un système d'équations linéaires, à coefficients complexes, dont la matrice pleine est d'une taille de 78000×78000. On suppose que les cefficients sont représentés en simple précision ; un réel est alors représenté par 4 bits et un complexe par 8 bits. La mémoire nécessaire pour contenir l'ensemble de la matrice est alors d'environ 25Go. Le calcul ainsi que l'assemblage (les affectations en mémoire) de la matrice sont des opérations d'une complexité quadratique. La résolution du sytème linéaire à l'aide d'une méthode directe est effectuée en environ 1h30 sur une machine dotée de six processeurs Intel(R) Core(TM) i7-3930K et 64Go de mémoire. Cette étape de résolution est cubique. En doublant la taille

de la matrice, le temps d'assemblage sera multiplié par 4 tandis que le temps de résolution sera multiplié par 8 ce qui est vite pénalisant. L'emploi d'une méthode directe est donc vite limité dans la pratique. Pour pallier cet inconvénient, nous souhaitons utiliser une approximation de la matrice de discrétisation ayant une structure particulière et qui permet un gain de temps et de calculs.

Un type de matrices très utilisées dans la pratique est le cas des matrices creuses. Il s'agit de matrices dont les coefficients sont majoritairement nuls. C'est typiquement le genre de matrice de discrétisation obtenue lors d'une discrétisation par différences finies de l'équation de Laplace. Le nombre de coefficients d'une telle matrice sont de l'ordre de $\mathscr{O}(\mathrm{N})$ et les opérations telles que le produit matrice-vecteur ou la somme de deux matrices creuses sont d'une complexité linéaire. Dans le cas d'un produit de matrices A et B, le produit AB est en général moins creux que les matrices de départ tandis qu'une factorisation LU est pleine.
Ces matrices ne sont pas adaptées à notre problème plein. Cependant, l'idée de « creuser » la matrice est envisageable dès lors qu'on partitionne la matrice en blocs matriciels dotés de « bonnes propriétés » ; ceci permet alors une réduction de la taille mémoire et du temps de calcul. L'exemple du paragraphe suivant introduit une classe de matrices dont le coût en mémoire ainsi que les opérations algébriques élémentaires sont linéaires ou quasi-linéaires.

1.1.4 Approximation dégénérée du noyau

Supposons que le noyau $\mathrm{K}(x,y)$ s'écrive sous la forme suivante,

$$\mathrm{K}(x,y) = \sum_{q=1}^{r} u_q(x)\, v_q(y), \tag{1.18}$$

où r est un entier non nul. Dans ce cas particulier, la séparation des variables et la commutativité de la somme et l'intégrale permettent d'écrire le terme général (1.12) de la matrice de discrétisation comme

$$\mathrm{A}_{ij} = \sum_{q=1}^{r} \left(\int_{\Gamma_h} u_q(x)\phi_i(x)\, \mathrm{d}\Gamma_h(x) \right) \left(\int_{\Gamma_h} v_q(y)\phi_j(y)\, \mathrm{d}\Gamma_h(y) \right). \tag{1.19}$$

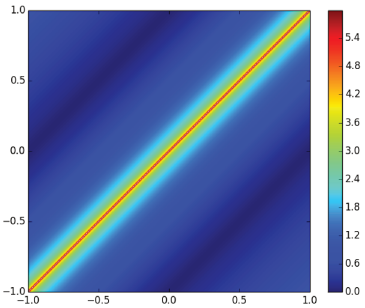
On note que cette écriture du terme général est favorable car elle découple les variables x et y et ne fait donc intervenir des intégrales simples au lieu d'une intégrale double. Pour le calcul élémentaire (1) cela représente un avantage car le coût de calcul unitaire est moins élevé.

Le noyau (1.33) est appelé **noyau dégénéré** et l'entier r est le **rang** du noyau. Pour une norme et une précision relative ϵ donnée, on peut définir le rang à la précision ϵ par

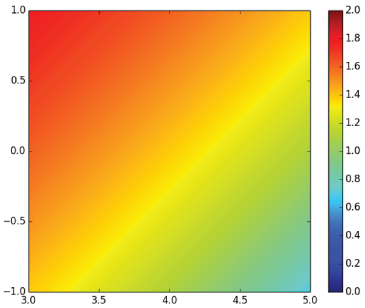
Définition 1.1 (Rang numérique d'un noyau)**.** Soit $\mathrm{K}(x,y) : \mathrm{H} \times \mathrm{H} \mapsto \mathbb{K}$ ($\mathbb{K} = \mathbb{R}$ ou $\mathbb{C}$) un noyau. On note $\|.\|_{\mathrm{H}}$ une norme sur H et ϵ un réel tel que $0 < \epsilon < 1$. On note K_r une approximation dégénérée de rang r de K (1.33). Si K_r vérifie l'inégalité

$$\|\mathrm{K} - \mathrm{K}_r\|_{\mathrm{H}} \leqslant \epsilon \|\mathrm{K}\|_{\mathrm{H}}, \tag{1.20}$$

on dit que r et le **rang numérique** à la précision ϵ du noyau. On notera r_ϵ ce rang.
On remarque que le rang numérique peut varier selon la norme employée. Par ailleurs, l'inégalité (1.35) peut être remplacé par l'inégalité $\|\mathrm{K} - \mathrm{K}_r\|_{\mathrm{H}} \leqslant \epsilon$ (avec $\epsilon > 0$) et l'on parle dans ce cas d'erreur absolue.



(a) Profil du noyau près de la diagonale ; $[-1,1]\times[-1,1]$.



(b) Profil du noyau hors diagonale ; $[3,5]\times[-1,1]$.

FIGURE 1.1 – Profil du noyau de Green 2D.

La figure 1.1a illustre la régularité du noyau en dehors de la diagonale $x = y$. Lorsqu'on est proche de la diagonale, les coefficients définis par (1.12) nécessitent un traitement particulier. En effet, même si le noyau est singulier sur la diagonale, il est intégrable et les coefficients diagonaux sont correctement définis. En dehors de la diagonale, le noyau $K(x,y)$ est régulier comme le montre la figure 1.1b.

On se place donc dans le cas où $\Gamma_i \cap \Gamma_j = \emptyset$. Le cas singulier $\Gamma_i = \Gamma_j$ correspond à l'auto-réaction (*self-interaction* dans la littérature) et nécessite un traitement particulier du fait de la singularité. Il s'agit d'un point crucial et la qualité d'un code de calcul intégral dépend de ces intégrales proches. Ce point ne fait pas parti des développements de cette thèse.

1.1.4.1 Rang du noyau $K(x,y)$

L'exemple suivant présenté plus en détails dans [GGMR09] présente une approximation à variables séparées et introduit la notion de matrice de rang faible.

Exemple 1.2. On considère les domaines $\Gamma_i = [3,5]$ et $\Gamma_j = [-1,1]$ et on considère $x \in \Gamma_i$ et $y \in \Gamma_j$. Dans ce cas, on a $x > y$ et le noyau s'écrit simplement

$$K(x,y) = -\frac{1}{2\pi}\log(x-y).$$

Admissibilité de l'interaction La configuration que l'on s'impose permet d'introduire la notion de **séparation**. En effet, les intervalles $[3,5]$ et $[-1,1]$ sont séparés en ce sens où la distance entre ces intervalles est supérieure à leurs diamètres. Plus précisément, on a les définitions suivantes

Définition 1.3 (Diamètre d'un intervalle)**.** Pour deux réels a et b tels que $a < b$, on note $\mathrm{Diam}([a,b])$ le diamètre de l'intervalle défini par

$$\mathrm{Diam}([a,b]) = b - a. \tag{1.21}$$

Définition 1.4 (Distance entre deux intervalles)**.** Pour deux intervalles non vides $[a,b]$ et

$[c,d]$, on note $\mathrm{Dist}([a,b],[c,d])$ la distance entre les deux intervalles définie par

$$\mathrm{Dist}([a,b],[c,d]) = \begin{cases} c-b & \text{si } c > b \\ a-d & \text{si } a > d \\ 0 & \text{sinon} \end{cases} \tag{1.22}$$

Ces deux notions interviennent à travers la condition suivante

Définition 1.5 (Condition d'admissibilité)**.** Pour deux intervalles $[a,b]$ et $[c,d]$, on appelle **condition d'admissibilité** l'inégalité suivante

$$\min(\mathrm{Diam}([a,b]),\mathrm{Diam}([c,d])) \leqslant \eta \mathrm{Dist}([a,b],[c,d]), \tag{1.23}$$

où η est un petit paramètre positif.

Pour l'exemple 1.2, on vérifie la condition

$$\min(\mathrm{Diam}([3,5]),\mathrm{Diam}([-1,1])) \leqslant \mathrm{Dist}([3,5],[-1,1]). \tag{1.24}$$

On parle alors de **condition d'admissibilité**. Cette dernière est une condition nécessaire pour la construction d'une bonne approximation du noyau sur le domaine considéré. En effet, la fonction $\mathrm{K}(x,y)$ est analytique sur le domaine $[3,5]\times[-1,1]$ et cette fonction coïncide avec sa série de Taylor dans un voisinage de y pour $y \in [-1,1]$. Pour l'interaction entre les domaines $[3,5]$ et $[-1,1]$ on parle d'interaction admissible car elle vérifie une relation de la forme (1.24). On parle d'**interactions lointaines** (ou en **champ lointain**) pour les interactions admissibles et d'**interactions proches** (ou en **champ proche**) pour celles ne vérifiant pas la condition (1.24).

1.1.4.2 Approximation d'une interaction admissible

On verra dans un prochain paragraphe que la condition d'admissibilité (1.24) correspond à la condition nécessaire pour que la série de Taylor du noyau converge. Pour cet exemple, cette condition est respectée et le développement de Taylor du noyau au voisinage de $y=0$, x étant fixé dans $[3,5]$, s'écrit

$$-\frac{1}{2\pi}\log(x-y) = -\frac{1}{2\pi}\left(\log(x) + \sum_{q=2}^{\infty} \frac{(-1)^q}{q.x^q}\, y^q\right). \tag{1.25}$$

Afin d'obtenir une approximation dégénérée, on souhaite tronquer la série (1.25) précédente après r termes.

La troncature du développement aux r premiers termes fournit une erreur de l'ordre de 3^{-r} [GGMR09]. C'est la majoration du terme de reste dans le développement de Taylor qui fournit ce résultat.

Ainsi pour une précision relative ϵ donnée, le nombre r de termes nécessaires pour approcher le noyau dans $[3,5]\times[-1,1]$ à la précision ϵ vérifie

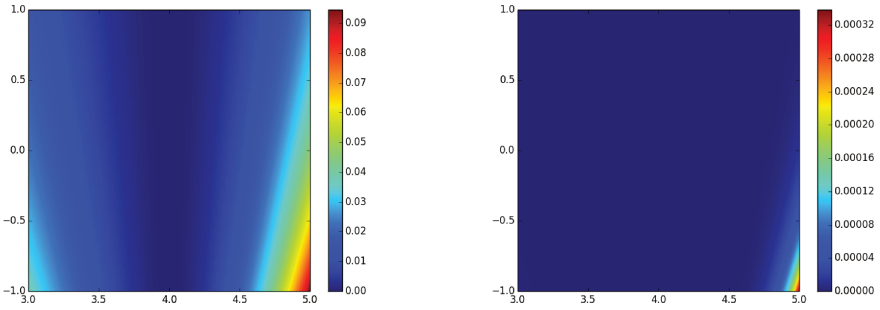
$$r = \mathscr{O}(|\log_3(\epsilon)|). \tag{1.26}$$

La décomposition (1.25) est une **approximation à variables séparées** $\tilde{\mathrm{K}}$ du noyau que l'on écrit sous la forme

$$\tilde{\mathrm{K}}(x,y) = \sum_{q=0}^{r-1} u_q(x)\, v_q(y). \tag{1.27}$$

Erreur discrète en norme l^∞ La figure suivante représente le maximum de l'erreur entre le noyau $\mathrm{K}(x,y)$ et l'approximation $\tilde{\mathrm{K}}(x,y)$ obtenue par le développement de Taylor (1.25) sur une grille cartésienne composée de 200×200 nœuds représentant le domaine $[3,5]\times[-1,1]$. Pour tous les couples (x_p,y_q) de la discrétisation de $[3,5]\times[-1,1]$, on représente l'erreur absolue E_{pq} définie par

$$\mathrm{E}_{pq} = |\mathrm{K}(x_p,y_q) - \tilde{\mathrm{K}}(x_p,y_q)|. \tag{1.28}$$



(a) Approximation de rang $r=2$. (b) Approximation de rang $r=8$.

FIGURE 1.2 – Erreur absolue en norme l^∞ de la troncature de la série de Taylor.

Matrice de rang faible On peut alors utiliser cette expression dans l'expression des coefficients de la matrice de rigidité et former une approximation $\tilde{\mathrm{A}}_{ij}$ de A_{ij} définie par

$$\begin{aligned}
\tilde{\mathrm{A}}_{ij} &= \int_\Gamma\int_\Gamma \tilde{\mathrm{K}}(x,y)\,\phi_i(x)\phi_j(y)\,\mathrm{d}\Gamma(x)\,\mathrm{d}\Gamma(y),\\
&= \sum_{q=0}^{r-1}\int_\Gamma\int_\Gamma u_q(x)\,v_q(y)\,\phi_i(x)\phi_j(y)\,\mathrm{d}\Gamma(x)\,\mathrm{d}\Gamma(y),\\
&= \sum_{q=0}^{r-1}\int_\Gamma u_q(x)\phi_i(x)\,\mathrm{d}\Gamma(x)\int_\Gamma v_q(y)\phi_j(y)\,\mathrm{d}\Gamma(y),
\end{aligned}$$

et telle que $|\mathrm{A}_{ij}-\tilde{\mathrm{A}}_{ij}| = \mathscr{O}(\epsilon)$.
Cette écriture correspond au produit scalaire de deux vecteurs U_i et V_j respectivement de tailles r définis pour $q\in\{1,\dots,r\}$ par

$$\begin{aligned}
(\mathrm{U}_i)_q &= \int_\Gamma u_q(x)\phi_i(x)\,\mathrm{d}\Gamma(x),\\
(\mathrm{V}_j)_q &= \int_\Gamma v_q(y)\phi_j(y)\,\mathrm{d}\Gamma(y),\\
\mathrm{A}_{ij} &= \mathrm{U}_i\mathrm{V}_j^{\mathrm{T}}.
\end{aligned} \tag{1.29}$$

Cet exemple est plus simple que les cas que l'on souhaite traiter dans la pratique mais il illustre parfaitement l'idée d'exprimer le noyau comme une somme de produits à variables séparées. C'est ce qui motive la méthode des $\mathscr{H}$-matrices ainsi que les approximations étudiées dans ce manuscrit. Nous souhaitons construire de telles approximations pour des noyaux plus complexes comme le noyau de Helmholtz $\mathrm{G}(x,y) = \frac{e^{ik\|x-y\|}}{\|x-y\|}$. Ceci sera l'objet du chapitre 4.

Lorsque l'on regroupe plusieurs degrés de liberté $\{i_1,\dots,i_m\}$ d'une part et $\{j_1,\dots,j_n\}$ d'autre part, leurs interactions définissent un sous-bloc matriciel $\mathrm{A}^{m\times n}$ de la matrice de rigidité que l'on cherche de façon similaire à (1.29) à écrire sous la forme

$$\mathrm{A}^{m\times n} = \mathrm{U}\mathrm{V}^{\mathrm{T}}, \tag{1.30}$$

où U et V sont des matrices de tailles respectives $m\times r$ et $n\times r$ avec r le rang de l'approxmation du noyau. Dans la suite, on désigne par s et t respectivement les ensembles $\{i_1,\dots,i_m\}$ et $\{j_1,\dots,j_n\}$ de degrés de liberté. On considère les sous-domaines de Γ correspondants

$$\Gamma_s = \bigcup_{i\in s=\{i_1,\dots,i_m\}} \Gamma_i, \tag{1.31}$$

$$\Gamma_t = \bigcup_{j\in t=\{j_1,\dots,j_n\}} \Gamma_j, \tag{1.32}$$

et l'on notera $\mathrm{A}_{s\times t}$ la matrice correspondant à l'interaction des degrés de liberté s et t.

Supposons que le noyau $\mathrm{K}(x,y)$ s'écrive sous la forme suivante,

$$\mathrm{K}(x,y) = \sum_{q=1}^{r} u_q(x)\,v_q(y), \tag{1.33}$$

où r est un entier non nul. Dans ce cas particulier, la séparation des variables et la commutativité de la somme et l'intégrale permettent d'écrire le terme général (1.12) de la matrice de discrétisation comme

$$\mathrm{A}_{ij} = \sum_{q=1}^{r} \left(\int_{\Gamma_h} u_q(x)\phi_i(x)\,\mathrm{d}\Gamma_h(x)\right)\left(\int_{\Gamma_h} v_q(y)\phi_j(y)\,\mathrm{d}\Gamma_h(y)\right). \tag{1.34}$$

On note que cette écriture du terme général est favorable car elle découple les variables x et y et ne fait donc intervenir des intégrales simples au lieu d'une intégrale double. Pour le calcul élémentaire (1) cela représente un avantage car le coût de calcul unitaire est moins élevé.

Le noyau (1.33) est appelé **noyau dégénéré** et l'entier r est le **rang** du noyau. Pour une norme et une précision relative ϵ donnée, on peut définir le rang à la précision ϵ par

Définition 1.6 (Rang numérique d'un noyau)**.** Soit $\mathrm{K}(x,y) : \mathrm{H}\times\mathrm{H}\mapsto\mathbb{K}$ ($\mathbb{K}=\mathbb{R}$ ou $\mathbb{C}$) un noyau. On note $\|.\|_{\mathrm{H}}$ une norme sur H et ϵ un réel tel que $0<\epsilon<1$. On note K_r une approximation dégénérée de rang r de K (1.33). Si K_r vérifie l'inégalité

$$\|\mathrm{K}-\mathrm{K}_r\|_{\mathrm{H}} \leqslant \epsilon\|\mathrm{K}\|_{\mathrm{H}}, \tag{1.35}$$

on dit que r et le **rang numérique** à la précision ϵ du noyau. On notera r_ϵ ce rang.
On remarque que le rang numérique peut varier selon la norme employée. Par ailleurs, l'inégalité (1.35) peut être remplacé par l'inégalité $\|\mathrm{K}-\mathrm{K}_r\|_{\mathrm{H}}\leqslant\epsilon$ (avec $\epsilon>0$) et l'on parle dans ce cas d'erreur absolue.

1.1.4.3 Obtention d'une interaction admissible

Le paragraphe précédent montre qu'il est possible d'approcher correctement le noyau $\mathrm{K}(x,y)$ pourvu qu'il soit régulier dans le domaine connsidéré. Considérons le cas suivant où les deux domaines Γ_s et Γ_t sont voisins. On illustre ce cas par les deux intervalles $\Gamma_s=[0,1]$ et $\Gamma_t=[-1,0]$ dont le seul point commun est $\{0\}$.

Exemple 1.7. On considère la matrice de discrétisation $A_n, (n > 0)$ de taille $n \times n$ définie par

$$A_n(i,j) = \log(x_i - y_j), \tag{1.36}$$

avec pour $i \in \{1,\dots,n\}$ et $j \in \{1,\dots,n\}$,

$$x_i = \frac{i-1}{n}, \tag{1.37}$$

$$y_j = -1 + \frac{j-1}{n}. \tag{1.38}$$

Les points $\{x_i\}_{i=1,\dots,n}$ et $\{y_j\}_{j=1,\dots,n}$ sont respectivement dans les intervalles $[0,1]$ et $[-1,0]$. Dans ce cas, on note s et t les ensembles d'indices de lignes et de colonnes respectivement, soit

$$s = \{1,\dots,n\}, \tag{1.39}$$

$$t = \{1,\dots,n\}. \tag{1.40}$$

À un indice $i \in s$ est alors associé le point x_i tandis que y_j est associé à l'indice $j \in t$.

À une constante multiplicative près, on obtient la matrice définie par 1.7 lorsque l'on considère une partition en n sous-domaines de Γ_s et Γ_t et que l'on approche les intégrales à l'aide d'un seul point de Gauss. Dans le cas de l'exemple 1.7, on ne peut utiliser directement le développement de Taylor car les deux domaines ne sont pas séparés.
Une approche naturelle est alors de considérer des subdivisions des domaines Γ_s et Γ_t. Pour illustrer ce processus de subdivision, on suppose que l'on ne procède qu'à une subdivision de l'intervalle $\Gamma_s = [0,1]$ mais l'on pourrait (et on le fait dans la suite) effectuer également une partition de l'intervalle $\Gamma_t = [-1,0]$.

On considère la subdivison suivante de l'intervalle $[0,1]$,

$$[0,1] = [0,1/2] \cup [1/2,1]. \tag{1.41}$$

— Le sous-intervalle $[0,1/2]$ possède toujours un point en commun avec l'intervalle $[-1,0]$ mais le second intervalle $[1/2,1]$ est à présent séparé de $[-1,0]$ et vérifie de plus la relation (1.24). Il s'agit de l'exemple précédent où le noyau peut être approché par son développement de Taylor avec un nombre de termes de l'ordre de $\mathscr{O}(|\log(\epsilon)|)$ pour une précision de l'ordre de ϵ. Les indices i des points x_i appartenant à $[1/2,1]$ sont regroupés dans le sous-ensemble s_0. (représenté en bleu sur la figure 1.3)

— On découpe l'intervalle $[0,1/2]$ de la façon suivante,

$$[0,1/2] = [0,1/4] \cup [1/4,1/2]. \tag{1.42}$$

Comme précédemment, l'intervalle $[1/4,1/2]$ est séparé de l'intervalle $[-1,0]$ et vérifie la condition (1.24). On approche alors le noyau à l'aide de $\mathscr{O}(|\log(\epsilon)|)$ pour une précision de l'ordre de ϵ. Les indices i des points x_i appartenant à $[1/4,1/2]$ sont regroupés dans le sous-ensemble s_1. (représenté en rouge sur la figure 1.3)

— De même,

$$[0,1/4] = [0,1/8] \cup [1/8,1/4], \tag{1.43}$$

et l'intervalle $[1/8, 1/4]$ est bien séparé de l'intervalle $[-1, 0]$ et on peut approcher le noyau à l'aide de $\mathscr{O}(|\log(\epsilon)|)$ termes. Les indices i des points x_i appartenant à $[1/8, 1/4]$ sont regroupés dans le sous-ensemble s_2. (représenté en vert sur la figure 1.3)

— On peut itérer le processus jusqu'à obtenir $P = |\log_2(n)|$ intervalles disjoints de $[0, 1]$.

La partition de l'intervalle $[0, 1]$ est équivalente à la partition de l'ensemble $s = \mathscr{I}$ suivante

$$s = \bigcup_{k=0}^{P} s_k. \tag{1.44}$$

La figure ci-dessous représente les sous-blocs matriciels construits par le schéma de subdivision ci-dessus pour s_0, s_1 et s_2. Les sous-matrices $(A_n)_{s_2 \times t}$,$(A_n)_{s_2 \times t}$ et $(A_n)_{s_2 \times t}$ sont de tailles respectives $n \times r$, $\frac{n}{2} \times r$ et $\frac{n}{4} \times r$ où r est le rang des approximations du noyau dans les domaines correspondants et vérifiant l'approximation (1.24).

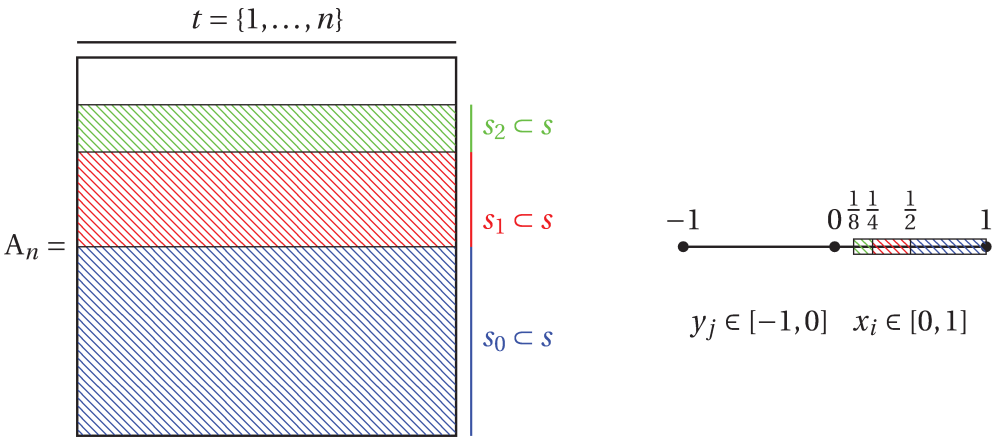


FIGURE 1.3 – Séparation des degrés de liberté en 1D.

La matrice A_n écrite sous sa forme par blocs,

$$A_n = \begin{bmatrix} \vdots \\ (A_n)_{s_2 \times t} \\ (A_n)_{s_1 \times t} \\ (A_n)_{s_0 \times t} \end{bmatrix}, \tag{1.45}$$

peut également être considérée comme une somme de P matrices,

$$A_n = \sum_{k=1}^{P} \begin{bmatrix} 0 \\ \vdots \\ (A_n)_{s_k \times t} \\ \vdots \\ 0 \end{bmatrix}, \tag{1.46}$$

où chaque matrice est de rang inférieur à $r = \mathscr{O}(|\log_3(\epsilon)|)$. La majoration du rang d'une somme est donnée par la somme des rangs des termes et l'on ne dispose pas de borne plus précise dans le cas général. Ainsi le rang de la matrice A_n vérifie

$$\begin{aligned} \mathrm{rg}(A_n) =& \mathscr{O}(Pr) \\ =& \mathscr{O}(\log_2(n) \log_3(\epsilon)). \end{aligned} \tag{1.47}$$

On retrouve cette estimation et la remarque suivante dans [GGMR09]. À l'aide d'une décomposition SVD ([GL96]), on peut déterminer le rang numérique pour une précision ϵ donnée en formant une décomposition de la matrice de la forme

$$A_n = U_n \Sigma V_n^T + \mathscr{O}(\epsilon), \tag{1.48}$$

où U_n et V_n sont de tailles $n \times r'$ avec $r \leqslant r'$ et Σ une matrice diagonale strictement positive de taille $r' \times r'$. r' est appelé le rang numérique de A_n à la précision ϵ. Le chapitre 2 contient une exposition plus détaillée des méthodes permettant des décompositions de la forme (1.48). On remarquera qu'en écrivant $\Sigma = \Sigma^{1/2}\Sigma^{1/2}$ (les coefficients de Σ sont strictement positifs), on peut se ramener à la forme (1.30) en effectuant les produits $U_n\Sigma^{1/2}$ et $V_n\Sigma^{1/2}$. Pour l'exemple considéré, on obtient les résultats suivants,

ϵ	$n = 10$		$n = 100$		$n = 1000$	
	r (1.47)	r_{SVD} (1.48)	r	r_{SVD}	r	r_{SVD}
10^{-3}	21	4	41	5	62	5
10^{-4}	27	4	55	6	83	7
10^{-5}	34	5	69	7	104	9
10^{-6}	41	6	83	8	125	11

TABLEAU 1.1 – Comparaison entre la borne (1.47) et le rang numérique déterminé par la décomposition SVD (1.48) pour $\epsilon \in \{10^{-6}, 10^{-5}, 10^{-4}, 10^{-3}\}$ et $n \in \{10, 100, 1000\}$ pour l'exemple 1.7

Le rang numérique obtenu est toujours inférieur à la borne (1.47) que l'on a déterminée. Cependant, le point important que l'on peut retenir est la faible dépendance avec la taille n de la matrice (au pire $\log(n)$ d'après l'estimation (1.47) et bien mieux d'après les résultats contenus dans le tableau 1.1. Il s'agit d'un point important pour une méthode d'approximation car dans le cas de deux géométries Γ_s et Γ_t non admissibles et non confondues, la matrice de discrétisation $A_{s\times t}$ peut être partitionnée en des blocs matriciels dont le rang est faible et ce indépendemment de la taille des blocs.
La seule condition est une **condition géométrique** sur les domaines Γ_s et Γ_t. C'est la condition d'admissibilité (1.24).

1.1.4.4 Résumé

Les deux exemples précédents ont permis de mettre en avant les points principaux d'une méthode d'approximation d'une matrice dont les coefficients sont de la forme (1.12). Le premier exemple a montré que dès que l'on est suffisament loin de la singularité $x = y$ le noyau $K(x, y)$ admet une approximation dégénérée (*i.e* de rang fini) à variables séparées. Le second exemple montre que des domaines voisins peuvent être récursivement découpés de sorte que le noyau puisse être approché dans les sous-domaines construits. Ces propriétés sont directement liées à celles de l'opérateur $\mathscr{A}$ (notamment sa compacité) ce que l'on ne détaille pas dans ce chapitre.

Une méthode d'approximation efficace de la matrice de rigidité A décrivant tout le domaine Γ repose sur trois points principaux :

— Une procédure pour subdiviser efficacement Γ de façon à produire des domaines pour lesquels leurs interactions mutuelles conduisent à des matrices de rang faible (c'est une partition des indices de lignes et de colonnes de la matrice) ;

— Un critère pour arrêter la subdivision : le critère d'admissibilité décrit par la formule (1.24) (c'est une partition de la matrice en sous-blocs) ;

— Un algorithme d'approximation pour les sous-matrices admissibles, il peut être analytique (1.27) ou algébrique (1.48).

Une méthode populaire dans la pratique est la méthode des $\mathscr{H}$-matrices qui est constituée des trois points précédents. Cette méthode, comme d'autres méthodes rapides, utilisent des approximations de blocs matriciels de la forme (1.30) afin de déterminer des approximations avec une complexité inférieure à $\mathscr{O}(N^2)$.
Le paragraphe suivant décrit brièvement les opérations disponibles sur l'ensemble des matrices de la forme (1.30). On trouve dans la littérature, la dénomination de matrice de rang faible ([BGH12; Beb00]) pour de telles matrices.

1.2 Matrices de rang faible

On introduit dans ce chapitre les notions et définitions utilisées tout au long de ce manuscrit. Les méthodes d'approximation développées par la suite sont basées sur l'emploi de matrices de rang faible. Par définition, ces matrices contiennent beaucoup moins d'information à garder en mémoire car on peut exhiber de nombreuses combinaisons linéaires entre les colonnes et/ou les lignes de la matrice.
Nous verrons dans la suite que de telles matrices sont très utiles en pratique afin de réduire la quantité d'espace disque utilisée ainsi que le nombre de calculs effectués lors de la résolution d'un problème d'équation intégrale. L'obtention d'approximations de cette forme est l'objet du chapitre prochain et l'on décrit ici l'utilisation que l'on peut faire de ces matrices. On suppose alors dans cette partie que de telles matrices peuvent être obtenues pour approcher un sous-bloc de la matrice de rigidité.

1.2.1 Représentation efficace

Puisque parmi les n colonnes d'une matrice $A \in \mathbb{C}_k^{m\times n}$ seulement k sont nécessaires pour représenter la totalité des coefficients de la matrice, on peut stocker la matrice A autrement que de manière totale, les coefficients superflus pouvant être omis.

Définition 1.8 (Rang d'une matrice)**.** Pour une matrice A quelconque de taille $m \times n$, on note Im(A) l'espace image de A. On définit le rang d'une matrice noté rg(A) comme la dimension de cet espace,

$$\mathrm{rg}(A) = \dim \mathrm{Im}(A). \tag{1.49}$$

De plus, on a $\mathrm{rg}(A) \leqslant \min(m,n)$. Dans la suite, on note également $r = \mathrm{rg}(A)$.

Définition 1.9 (Matrice de rang faible)**.** On note $\mathbb{R}_r^{m\times n}$ (respectivement $\mathbb{C}_r^{m\times n}$) l'ensemble des matrices réelles (respectivement complexes) de taille $m \times n$ et de rang inférieur à r.

Théorème 1.10. *Une matrice* $A \in \mathbb{R}^{m\times n}$ *(respectivement* $\mathbb{C}^{m\times n}$*) appartient à* $\mathbb{R}_r^{m\times n}$ *(respectivement* $\mathbb{C}_r^{m\times n}$*)si et seulement si il existe deux matrices* $U_A \in \mathbb{R}^{m\times r}$ *(resp.* $\mathbb{C}^{m\times r}$*) et* $V_A \in \mathbb{R}^{n\times r}$ *(resp.* $\mathbb{C}^{n\times r}$*) telles que*

$$A = U_A V_A^{\mathrm{T}} \tag{1.50}$$

Cette représentation (1.50) est appelée représentation compressée ou somme de produits tensoriels. En effet, si l'on note u_q et v_q pour $q = 1,\dots,r$ les colonnes respectives de U_A et V_A, alors la représentation (1.50) est équivalente à

$$A = \sum_{q=1}^{r} u_q v_q^{\mathrm{T}}. \tag{1.51}$$

Ainsi, au lieu de stocker les mn coeffcients de la matrice A, on se contente des $r(m+n)$ coefficients des vecteurs u_q et v_q.

Outre le fait de réduire l'espace mémoire utilisé, cette représentation facilite aussi le produit matrice-vecteur

$$\begin{aligned} Ax &= (U_A V_A^T)x \\ Ax &= U_A(V_A^T)x. \end{aligned} \tag{1.52}$$

La mise à jour $y := y + Ax$ n'est pas effectuée élément par élément, on effectue le produit Ax en deux étapes

1. $z := V_A^T x$, z est un vecteur de taille n ;

2. $y := U_A z$;

ce qui ne nécessite que $2r(m+n) - r$ opérations au lieu des $2mn$ opérations usuelles. On note cependant que cette représentation peut ne pas être efficace ! Supposons que $m = n$ et que la matrice considérée est de rang maximal,alors la représentation tensorielle amène à manipuler $2n^2$ coefficients soit deux fois plus que sous sa forme usuelle. On caractérise alors les matrices de faible rang à l'aide de la définition suivante.

Définition 1.11 (Matrice de rang faible)**.** Une matrice $A \in \mathbb{C}_r^{m\times n}$ est appelée matrice de rang faible si et seulement si

$$r(m+n) << mn. \tag{1.53}$$

Par abus de langage, on parlera aussi de matrice compressée. On notera toujours une telle matrice sous sa forme tensorielle(ou compressée) $U_A V_A^T$ plutôt que sous la forme A.

Calcul de normes Typiquement calculées en tant que critère d'arrêt d'un algorithme, les normes matricielles sont d'autant plus faciles d'accès avec des matrices compressées. Ainsi, la norme de Fobénius d'une matrice compressée de rang r peut être calculée en $2r^2(m+n)$ opérations en remarquant que

$$\|U_A V_A^T\|_F^2 = \sum_{i,j=1}^{r} (u_i^T u_j)(v_i^T v_j), \tag{1.54}$$

tandis que la norme spectrale, donnée par

$$\|A\|_2 = \sqrt{\rho(V_A U_A^T U_A V_A^T)}, \tag{1.55}$$

$$= \sqrt{\rho(U_A^T U_A V_A^T V_A)}, \tag{1.56}$$

$\rho(A)$ désignant le rayon spectral de la matrice A, peut être calculée en $\mathcal{O}(r^2(m+n))$ opérations en calculant les matrices $U_A^T U_A$ et $V_A^T V_A$ de tailles $r \times r$ puis la plus grande valeur propre du produit. On note que cela est avantageux si l'on vérifie l'inégalité $r^2(m+n) << mn$ ce qui est une condition bien plus restrictive que la définition de faible rang précédente.

Décomposition orthonormale Il peut parfois être utile que les colonnes de U_A et V_A soient orthonormales. Dans ce cas, on doit introduire une matrice $X \in \mathbb{R}^{r\times r}$ de coefficients et remplacer la décomposition sous forme tensorielle par une décomposition de la forme

$$A = U_A X V_A^T. \tag{1.57}$$

Dans ce cas, le calcul de la norme de Frobénius se simplifie de la façon suivante

$$\|U_A X V_A^T\|_F = \|X\|_F \tag{1.58}$$

$$= \sqrt{\sum_{i,j=1}^{r} |x_{ij}|^2}. \tag{1.59}$$

Ceci ne requiert que $\mathscr{O}(r^2)$ opérations. La norme spectrale vérifie quant à elle la relation suivante

$$\|U_A X V_A^T\|_2 = \|X\|_2, \tag{1.60}$$

menant au calcul de la plus grande valeur propre d'une matrice de taille $r \times r$.

1.2.2 Opérations élémentaires sur les matrices compressées

Le stockage d'une matrice A de rang faible sous sa forme compressée $U_A V_A^T$ permet de gagner en espace mémoire. Cependant, l'utilisation de cette forme permet aussi d'économiser des calculs lors d'opérations matricielles élémentaires comme le produit matrice-vecteur. Les paragraphes suivants décrivent les opérations courantes effectuées sur les matrices à savoir la multiplication et l'addition. On présente également une décomposition SVD approchée pour les matrices de rang faible ce qui permet d'affiner l'algorithme d'addition. On trouve le descriptif de ces opérations dans [Beb00] ainsi que dans [BGH12]. On présente les résultats pour des matrices réelles, les résultats s'adaptant aisément au cas complexe.

1.2.2.1 Multiplication de matrices compressées

On rappelle que lorsqu'une matrice A de taille $m \times n$ peut être représentée sous la forme d'une somme de produits tensoriels $U_A V_A^T$ avec U_A et V_A respectivement de taille $m \times r$ et $n \times r$ où $r << m, n$ est le rang de l'approximation. On dit alors que A est *compressée*.
On veut obtenir sous forme compressée le produit matriciel C = AB où A est compressée. On distingue alors deux cas suivant la nature de la matrice B en comparant dans chaque cas le nombre d'opérations effectuées avec la multiplication de matrices usuelles.

B est non-compressée On effectue le produit C = AB et l'on note $C = U_C V_C^T$ la représentation compressée de la matrice C. Comme A est une matrice compressée, on a $AB = (U_A V_A^T)B$. Par substitution, on obtient l'égalité suivante,

$$U_C V_C^T = (U_A V_A^T)B. \tag{1.61}$$

On cherche à déterminer les matrices U_C et V_C en fonction de U_A, V_A et B. On peut définir U_C et V_C de la façon suivante,

$$U_C = U_A, \tag{1.62}$$

$$V_C^T = V_A^T B. \tag{1.63}$$

Dans ce cas U_C est déterminé sans aucun calcul (c'est une copie en mémoire de U_A) et ainsi le nombre d'opérations pour le produit C = AB repose uniquement sur le calcul de $V_C^T = V_A^T$.B soit (rnp) opérations par rapport à (mnp) dans le cas usuel. Le rapport entre ces deux quantités est

$$\frac{knp}{mnp} = \frac{k}{m} << 1 \qquad \text{si } k << m. \tag{1.64}$$

Le gain de calcul est donc du premier ordre en $\frac{k}{m}$.

B est compressée On suppose à présent que les deux matrices A et B sont données sous forme compressée, $B = U_B V_B^t$ où $U_B \in \mathbb{R}^{n\times r}$ et $V_B \in \mathbb{R}^{p\times r}$ avec $r << n, p$. Avec les mêmes notations que précédemment pour les matrices A et C, on a

$$\begin{aligned} U_C V_C^T &= (U_A V_A^T)(U_B V_B^T) && (1.65)\\ &= U_A (V_A^T U_B) V_B^T && (1.66)\\ &= U_A (X) V_B^T. && (1.67) \end{aligned}$$

La matrice intermédiaire $X = V_A^T U_B$ est un produit scalaire avec un nombre réduit d'opérations (knr). On doit alors effectuer le produit de X avec U_A ou V_B^T ce qui respectivement mène à (mkr) ou (rkn) opérations. Finalement, le coût total du produit C = AB est de $\{knr + mkr\}$ ou $\{knr + rkn\}$. Le rapport entre les coûts des méthodes compressée et usuelle est donné par

$$\frac{knr + mkr}{mnp} = \frac{kr(m+n)}{mnp} \ll 1 \qquad \text{si } k, r << m, n, p. \tag{1.68}$$

Dans le cas où les matrices sont carrées $(m = n = p)$ le rapport est du second ordre

$$\frac{knr + mkr}{mnp} = 2\frac{kr}{m^2} \ll 1 \qquad \text{si } k, r \ll m.$$

1.2.2.2 Décomposition SVD d'une matrice compressée

L'accès à une décomposition en valeurs singulières d'une matrice peut avoir un intérêt lorsque l'on souhaite en extraire l'information principale portée par ses plus grands éléments singuliers. Un exposé plus détaillé de cette décomposition et ses propriétés est présent au chapitre 2. On dispose dans le cas de matrices compressées d'une méthode peu coûteuse afin d'en construire une décomposition SVD.

On considère une matrice compressée donnée par sa représentation efficace $A = U_A V_A^T$ avec $U_A \in \mathbb{R}^{m\times k}, V_A \in \mathbb{R}^{n\times k}$ et $k << \min(m, n)$. On souhaite obtenir la décomposition suivante

$$U_A V_A^T = \mathscr{U} \Sigma \mathscr{V}^T, \tag{1.69}$$

avec $\mathscr{U} \in \mathbb{R}^{m\times k}, \mathscr{V} \in \mathbb{R}^{n\times k}$ des matrices dont les colonnes sont orthogonales et $\Sigma \in \mathbb{R}^{k\times k}$ une matrice diagonale sont les coefficients sont positifs.
On obtient cette décomposition en commençant par orthonormaliser les colonnes des matrices U_A et V_A à l'aide de la factorisation QR :

$$\begin{aligned} U_A &= Q_U R_U, && (1.70)\\ V_A &= Q_V R_V. && (1.71) \end{aligned}$$

On forme alors le produit $R = R_U.R_V^T$ en utilisant le fait que $k << \min(m,n)$. En effet, dans les décompositions QR ci-dessus, cette inégalité implique que les $m_A - k$ et les $n_A - l$ dernières colonnes respectives de R_U et R_V sont nulles. Ainsi le calcul de R ne nécessite que la connaissance des k premières colonnes de R_U et R_V.

On effectue une décomposition en valeurs singulières de la matrice réduite R,

$$R = \hat{\mathscr{U}} \Sigma \hat{\mathscr{V}}^T, \tag{1.72}$$

où $\hat{\mathscr{U}} \in \mathbb{R}^{k\times k}, \hat{\mathscr{V}} \in \mathbb{R}^{k\times k}$ et $\Sigma \in \mathbb{R}^{k\times k}$.
Puisque les matrices Q_U et Q_V sont orthogonales, les matrices $\mathscr{U}$ et $\mathscr{V}$ définies par

$$\mathscr{U} = Q_U \hat{\mathscr{U}}, \tag{1.73}$$
$$\mathscr{V} = Q_V \hat{\mathscr{V}}, \tag{1.74}$$

le sont également et l'on obtient ainsi une décomposition SVD approchée de la matrice A. Pour la différencier de la décomposition SVD « usuelle » d'une matrice on appellera cette décomposition SVD approchée rSVD.

Coût de la décomposition Le détail des opérations effectuées à chaque étape de la construction précédente est résumé dans le tableau (1.2).

Décomposition QR de U_A et V_A	$4r^2(m+n) - \frac{8}{3}r^3$
Calcul de R	$\frac{2}{3}r^3 + \frac{11}{6}r^2 - \frac{1}{3}r$
Décomposition SVD de $R_U R_V^T$	$22r^3$
Calcul de $\mathscr{U}$ et $\mathscr{V}$	$r(2r-1)(m+n)$
Total	$\simeq 6r^2(m+n) + 20r^3$

TABLEAU 1.2 – Coût de la décomposition SVD approchée d'une matrice de rang faible $A = UV^T$.

1.2.2.3 Addition de matrices compressées

On note pour toute cette partie,

$$A = U_A V_A^T \tag{1.75}$$
$$B = U_B V_B^T \tag{1.76}$$
$$C = U_C V_C^T \tag{1.77}$$

avec $U_A \in \mathbb{R}^{m\times r_A}, V_A \in \mathbb{R}^{n\times r_A}, U_B \in \mathbb{R}^{m\times r_B}, V_B \in \mathbb{R}^{n\times r_B}, U_C \in \mathbb{R}^{m\times r}, V_C \in \mathbb{R}^{n\times r}$ et $r_A, r_B \leq \min(m,n), r \leq r_A + r_B$.

Concaténation On effectue l'addition $C = A + B$ en construisant les concaténations suivantes

$$U_C = [U_A U_B],$$
$$V_C = [V_A V_B], \tag{1.78}$$

U_C et V_C étant des matrices de tailles respectives $m \times (r_A + r_B)$ et $n \times (r_A + r_B)$. Le produit $U_C V_C^t$ correspond à la somme $A + B$ et la forme compressée du produit est obtenue sans effectuer aucun calcul (hormis les copies en mémoire éventuelles).

Cependant, on ne possède pas de borne inférieure sur le rang d'une somme de matrices. Ainsi, il est possible (on verra dans les applications numériques que cela est bien le cas) que le rang r de la somme soit nettement inférieur à $r_A + r_B$ et c'est précisement ce rang inférieur qui nous intéresse dans la pratique.

On dispose fort heureusement d'un moyen économique pour parvenir à déterminer le rang (et éventuellement le réduire) de la somme à partir des concaténations. Cette diminution du rang permet donc de réduire l'espace mémoire nécessaire au stockage de la somme et par là même de réduire le nombre d'opérations élémentaires lors de calculs impliquant cette somme.

Recompression Étant données les matrices concaténées $U_C \in \mathbb{R}^{n\times(k_A+k_B)}$ et $V_C \in \mathbb{R}^{n\times(k_A+k_B)}$ du paragraphe précédent, on peut construire une approximation de rang plus faible de la somme $C = A + B$ à l'aide de la décomposition SVD vue auparavant en $\mathscr{O}((k_A + k_B)^3)$ opérations.

Proposition 1.12. *Soient* $A \in \mathbb{R}^{m\times n}_{r_A}$ *et* $B \in \mathbb{R}^{m\times n}_{r_B}$ *deux matrices de rang faible* r_A *et* r_B *respectivement. Soit* $r \in \mathbb{N}$ *tel que* $r \leq r_A + r_B$. *On note* $C = A + B$. *Ainsi, on peut déterminer une matrice de rang faible* S *de taille* $m \times n$ *et de rang* r *réalisant le problème de minimisation suivant*

$$\|A + B - S\|_2 = \min_{M\in\mathbb{R}^{m\times n}_r} \|A + B - M\|_2. \tag{1.79}$$

Cette matrice S *peut être obtenue en* $6(r_A + r_B)^2(m + n) + 20(r_A + r_B)^3$ *opérations.*

Il s'agit là d'une application directe de la décomposition SVD d'une matrice compressée vue à la section précédente à la représentation efficace obtenue par les concaténations $U_C = [U_A\, U_B], V_C = [V_A\, V_B]$ en tronquant la SVD aux r premières valeurs singulières.

Remarque 1.13. On peut être amené lors de certaines applications à calculer une somme $A = A_1 + A_2 + \ldots + A_N, A_i \in \mathbb{R}^{m\times n}_{k_i}, i = 1\ldots N$ de matrices compressées pour lesquelles la concaténation amènerait un coefficient $(\sum_{i=1}^{N} r_i)^2$ devant $(m + n)$. Pour éviter cette situation, on effectuera les additions deux à deux ce qui réduit le nombre d'opérations à

$$6\sum_{i=1}^{N-1}(r_i + r_{i+1})^2(m + n) + 20\sum_{i=1}^{N-1}(r_i + r_{i+1})^3, \tag{1.80}$$

au détriment de la meilleure approximation.

Remarque 1.14. On considère une somme de matrices compressées du type $A = A_1 + A_2 + \ldots + A_N, A_i \in \mathbb{R}^{m\times n}_{r_i}, i = 1\ldots N$. Nous avons constaté dans la pratique qu'il est préférable de les trier par rang croissant au préalable. Ainsi, on commence par sommer (au sens compressé) les matrices de plus petit rang et l'on s'attend à ce que la recompression lors de la somme permette au rang de ne pas trop croître. Les blocs dont les rangs sont les plus élevés ne "polluent" alors pas les sommes puisqu'ils sont sommés à la fin. Généralement, cette manipulation permet de réduire le temps de calcul car les rangs restent petits tout au long des sommations. La précision demeure la même.

1.2.3 Agglomération de matrices de rang faible

On décrit dans cette partie une autre opération possible sur les matrices de rang faible. Cette étape est décrite dans [Beb00]. Il s'agit d'une opération cruciale de la méthode des

$\mathscr{H}$-matrices que l'on va décrire plus loin. Cette opération permet de regrouper des sous-blocs de rang faible constituant une partition de la matrice en une matrice plus grande mais toujours de rang faible.
Sans perdre de généralité, on considère une matrice A de taille $2m \times 2n$ définie en tant que matrice bloc 2×2 de la sorte

$$\mathrm{A} = \begin{bmatrix} \mathrm{A}_1 & \mathrm{A}_2 \\ \mathrm{A}_3 & \mathrm{A}_4 \end{bmatrix}, \tag{1.81}$$

où chacune des matrices A_i est de taille $m \times n$ et représentée sous une forme compressée de rang r par

$$\mathrm{A}_i = \mathrm{U}_i \mathrm{V}_i^{\mathrm{T}}, \tag{1.82}$$

avec pour $i = 1,\ldots,4$, U_i et V_i respectivement de taille $m \times r$ et $n \times r$. Dans la pratique, on est amené à traiter des approximations de rangs différents sans que cela ne change la méthode. On se restreint pour l'exposition de la méthode au cas constant.
On cherche une décomposition de la matrice A sous la même forme

$$\mathrm{A} = \mathrm{U}\mathrm{V}^{\mathrm{T}}, \tag{1.83}$$

avec U et V respectivement de tailles $2m \times r'$ et $2n \times r'$ et r' tel que $r' << \min(m,n)$.

Première approche naïve Comme dans le cas d'une somme compressée, on peut remarquer que A s'exprime comme une somme de quatre matrices

$$\begin{bmatrix} \mathrm{A}_1 & \mathrm{A}_2 \\ \mathrm{A}_3 & \mathrm{A}_4 \end{bmatrix} = \begin{bmatrix} \mathrm{A}_1 & 0 \\ 0 & 0 \end{bmatrix} + \begin{bmatrix} 0 & \mathrm{A}_2 \\ 0 & 0 \end{bmatrix} + \begin{bmatrix} 0 & 0 \\ \mathrm{A}_3 & 0 \end{bmatrix} + \begin{bmatrix} 0 & 0 \\ 0 & \mathrm{A}_4 \end{bmatrix}. \tag{1.84}$$

On peut alors construire une somme compressée de ces matrices en utilisant la décomposition SVD approchée du paragraphe 1.2.2.2. En effet, on a la décomposition suivante,

$$\mathrm{A} = \hat{\mathrm{U}}\hat{\mathrm{V}}^{\mathrm{T}}, \tag{1.85}$$

avec

$$\hat{\mathrm{U}} = \begin{bmatrix} \mathrm{U}_1 & \mathrm{U}_2 & 0 & 0 \\ 0 & 0 & \mathrm{U}_3 & \mathrm{U}_4 \end{bmatrix} \tag{1.86}$$

$$\hat{\mathrm{V}} = \begin{bmatrix} \mathrm{V}_1 & 0 & \mathrm{V}_3 & 0 \\ 0 & \mathrm{V}_2 & 0 & \mathrm{V}_4 \end{bmatrix} \tag{1.87}$$

Dans ce cas, la présence des blocs nuls augmente le nombre de coefficients à garder en mémoire. Il ne s'agit pas de la façon la plus rapide de procéder.

Une meilleure approche : orthogonaliser les concaténations On décrit à prséent une façon rapide d'obtenir l'agglomération des blocs en suivant la méthode décrite dans [Beb00].
On forme les décompositions QR des concaténations suivantes

$$[\mathrm{U}_1\,\mathrm{U}_2] = \mathrm{Q}_1\mathrm{R}_1, \tag{1.88}$$

$$[\mathrm{U}_3\,\mathrm{U}_4] = \mathrm{Q}_2\mathrm{R}_2, \tag{1.89}$$

$$[\mathrm{V}_1\,\mathrm{V}_3] = \mathrm{Q}_3\mathrm{R}_3, \tag{1.90}$$

$$[\mathrm{V}_2\,\mathrm{V}_4] = \mathrm{Q}_4\mathrm{R}_4, \tag{1.91}$$

où les matrices R_i sont de taille $r \times 2r$. On les partitionne de la façon suivante, avec R'_i et R''_i toutes deux de taille $r \times r$,

$$R_i = [R'_i \, R''_i]. \tag{1.92}$$

La matrice A peut alors s'écrire sous la forme

$$\hat{U} = \begin{bmatrix} Q_1 & 0 \\ 0 & Q_2 \end{bmatrix} \tag{1.93}$$

$$\hat{V} = \begin{bmatrix} Q_3 & 0 \\ 0 & Q_4 \end{bmatrix} \tag{1.94}$$

$$\hat{X} = \begin{bmatrix} R'_1 R'^{T}_3 & R''_1 R'^{T}_4 \\ R'_2 R''^{T}_3 & R''_2 R''^{T}_4 \end{bmatrix} \tag{1.95}$$

Comme précédemment pour la méthode SVD compressée, la matrice intermédiaire $\hat{X}$ est de dimension plus faible que celle de A ($2m \times 2n$). On peut alors écrire une décomposition SVD de la matrice $\hat{X}$,

$$\hat{X} = \mathscr{U} \Sigma \mathscr{V}^T, \tag{1.96}$$

et on peut alors effectuer les produits

$$U = \hat{U}\mathscr{U}, \tag{1.97}$$

$$V = \hat{V}\mathscr{V}, \tag{1.98}$$

et ainsi former une approximation de la forme

$$A = U\hat{X}V. \tag{1.99}$$

Coût de la méthode Les étapes de cette méthode sont similaires à celles développées au paragraphe 1.2.2.2. De la même façon, on regroupe le détail des opérations dans le tableau 1.3.

Décompositions QR de	$16r^2(m+n) - 5r^3$
Calcul de $\hat{X}1$,$\hat{X}2$,$\hat{X}3$,$\hat{X}4$	$2r^3$,$4r^3$,$4r^3$,$6r^3$
Décomposition SVD de $R_U R_V^T$	$22(4r)^3$
calculs unitary $\mathscr{U}$ et $\mathscr{V}$	$8r^2(m+n)$
Total	$\simeq 24r^2(m+n) + 1408\frac{2}{3}r^3$

TABLEAU 1.3 – Coût de l'agglomération d'une matrice bloc 2×2 dont les blocs sont des matrices de rang faible.

Le coût total est alors linéaire par rapport à la taille des blocs et la partie dominante est celle en $\mathscr{O}(r^2(m+n))$. Cette opération d'agglomération est efficace d'un point de vue de la mémoire si et seulement si

$$r'(2m+2n) < \sum_{i=1}^{4} r_i(m+n), \tag{1.100}$$

où r' est le rang de l'approximation de rang faible de l'agglomération. Le cas échéant, nous sommes amenés à garder plus de coefficients en mémoire ce qui dégrade les performances de la méthode.

Remarque 1.15. On pourrait agglomérer les sous-blocs $[A_1\, A_2]$ et $[A_3\, A_4]$ avant d'agglomérer les résultats. Cepdendant, on ne peut garantir que le résultat obtenu est la meilleure approximation de rang fini possible ([Beb00]).

Ce paragraphe montre que les matrices de rang faible peuvent être manipulées avec un nombre réduit d'opérations. Par exemple, pour des matrices de taille n, la multiplication de matrices est une opération qui nécessite de l'ordre de n^3 opérations pour le cas général. Dans le cas où les matrices possèdent une structure particulière comme (1.30), il est également possible de garder en mémoire le produit sous la forme (1.30) en $\mathscr{O}(rr'n)$ opérations.
Ces matrices que l'on a exhibées à partir des exemples 1.2 et 1.7 sont au cœur de la méthode des $\mathscr{H}$-matrices. Dans les deux exemples développés, nous avons considéré des sous-domaines de la courbe Γ dont la courbure est nulle afin de se ramener à des intervalles réels. Dans la pratique,on doit cependant traiter des sous-ensembles de Γ qui ne sont pas des intervalles. Dans le cas d'une courbe dans le plan ou d'une surface dans $\mathbb{R}^3$, il est nécessaire de disposer d'un outil pratique pour mesurer l'éloignement des domaines Γ_t et Γ_s. Une façon suffisante et maniable est d'utiliser des boîtes englobantes contenant les domaines.

1.3 Approximation $\mathscr{H}$matrice

1.3.1 Motivations

Dans toute la suite de ce manuscrit et sauf mention contraire, Γ fait référence à une surface régulière de $\mathbb{R}^3$ et l'on considère à présent l'équation intégrale du type (1.1) sur Γ. La méthode de résolution conduit également à considérer un nuage de N degrés de liberté (on parle également de particules dans un contexte de problème à N corps) $\{x_{i_k}\}_{k=1,\dots,N}$ de $\mathbb{R}^3$ à partir desquels on construit une matrice d'interaction A de taille $N \times N$ dont les coefficients sont définis par

$$A_{ij} = \int_\Gamma \int_\Gamma G(x,y)\phi_j(y)\phi_i(x)\, d\Gamma(y)\, d\Gamma(x), \tag{1.101}$$

où comme précédemment les fonctions $\phi_j(y)$ et $\phi_(x)$ sont des fonctions de base localisées aux degrés de liberté $\{x_{i_k}\}_{k=1,\dots,N}$. En dimension trois, le noyau $G(x,y)$ est défini par

$$G(x,y = \frac{1}{4\pi}\frac{1}{\|x-y\|}. \tag{1.102}$$

Le noyau $G(x,y)$ est le noyau de l'équation dans Laplace dans l'espace libre et est défini comme la solution élémentaire de l'équation de Laplace munie de la condition de radiation à l'infini. Comme le noyau $K(x,y)$, il possède une singularité en $x = y$ et on parle également de singularité sur la diagonale. Cette solution élémentaire intervient par exemple en électrostatique lorsque l'on travaille avec une représentation intégrale du champ électrique où les solutions de l'équation de Poisson sont déterminées en effectuant un produit de convolution avec G.

L'assemblage de la matrice de discrétisation s'effectue en $\mathscr{O}(N^2)$ opérations tandis que sa factorisation s'effectue en $\mathscr{O}(N^3)$ opérations. Comme pour le cas en dimension deux, l'approche usuelle consiste à exploiter les propriétés analytiques du noyau G. En effet,

pour deux degrés de liberté distincts et de support disjoints, leur interaction est principalement déterminée par le comportement du noyau $G(x,y)$.

Sous certaines hypothèses, on peut regrouper les particules suivant leurs proximité géométrique. La méthode des $\mathscr{H}$-matrices que nous nous proposons de décrire ici est basée sur une représentation par arbre du nuage de points $\{x_{i_k}\}_{i_k}$. Une méthode basée sur une structure d'arbre utilisant les coordonnées des particules permet de rapidement partitionner ces dernières de façon à scinder les interactions en interactions proches et lointaines. En « croisant » les groupes d'inconnues associés aux lignes et aux colonnes, on réalise bien une découpe hiérarchique des blocs de la matrice originale A en blocs proches et lointains.

Les blocs décrivant les interactions proches contiennent la singularité du noyau et sont de rang maximal. Ce sont des matrices pleines et elles sont assemblées avec une complexité quadratique. Les blocs représentant les interactions lointaines possèdent une propriété de rang faible que l'on peut exploiter efficacement par la suite. Afin de minimiser le nombre d'opérations d'interactions proches à effectuer, on souhaite alors obtenir des blocs d'interactions lointaines les plus gros possibles.

On décrit brièvement dans ce paragraphe la construction d'une approximation compressée de la matrice de discrétisation A : c'est une approximationpar une $\mathscr{H}$-matrice $A_{\mathscr{H}}$. Dans un premier temps, on introduit la notion de boîte englobante ce qui facilite la manipulation des degrés de liberté (1.3.2.1). Le paragraphe 1.3.2.2 décrit un point crucial de la construction, la partition des degrés de libertés en groupe ou paquets de degrés de liberté. Enfin, on peut écrire la matrice de discrétisation comme une partition de blocs décrivant l'interaction de groupes de degrés de libertés suffisament éloignés. Pour ce faire, on a besoin d'un critère de séparation appelé critère d'admissibilité. On fera référence aux exemples et notions introduites dans le cas en dimension deux car il s'agit de notions communes et indépendantes de la dimension. Cependant, les applications visées et développées au chapitre 4 sont, elles, en dimension 3.

1.3.2 Construction d'une $\mathscr{H}$-matrice

1.3.2.1 Boites englobantes

La manipulation des degrés de liberté dans le plan ou dans l'espace est facilitée par l'emploi de conteneurs tels que les boîtes englobantes. Il s'agit d'un outil de base utilisé par la plupart des méthodes rapides [Syl02; BGH12].

Définition et propriétés Pour faciliter la construction et la gestion de groupes d'inconnues on utilise des boîtes englobantes dont les côtés sont parallèles aux axes. Elles présentent l'avantage de travailler directement avec les coordonnées des points du nuage.

Définition 1.16 (Boîte englobante)**.** Considérons N points $\{x_{i_k}\}_{k=1,\dots,N}$ de $\mathbb{R}^3$. On définit une boîte englobante B par le produit cartésien suivant,

$$B = [a_1, b_1] \times [a_2, b_2] \times [a_3, b_3], \tag{1.103}$$

de telle façon à ce que pour tout point $x_{i_k} = \left((x_{i_k})_1, (x_{i_k})_2, (x_{i_k})_3\right)^T$ de $\mathbb{R}^3$, on a

$$a_\nu \leqslant (x_{i_k})_\nu \leqslant b_\nu \qquad \text{pour } \nu = 1,2,3. \tag{1.104}$$

Dans la suite, on indexe la boîte par le sous-ensemble d'indices qu'elle contient. Ainsi, pour $t = \{i_1, \ldots, i_N\}$, on notera B_t la boîte englobante contenant les points d'indices $i_k \in t$.

Dans le cas général, les côtés de la boîte ne sont pas nécessairement de la même dimension. De plus, la détermination des dimensions de la boite se fait en $\mathscr{O}(N)$ opérations (boucle simple sur les points). La figure (1.4) illustre ces boîtes englobantes dans le cas d'une courbe fermée de $\mathbb{R}^2$.

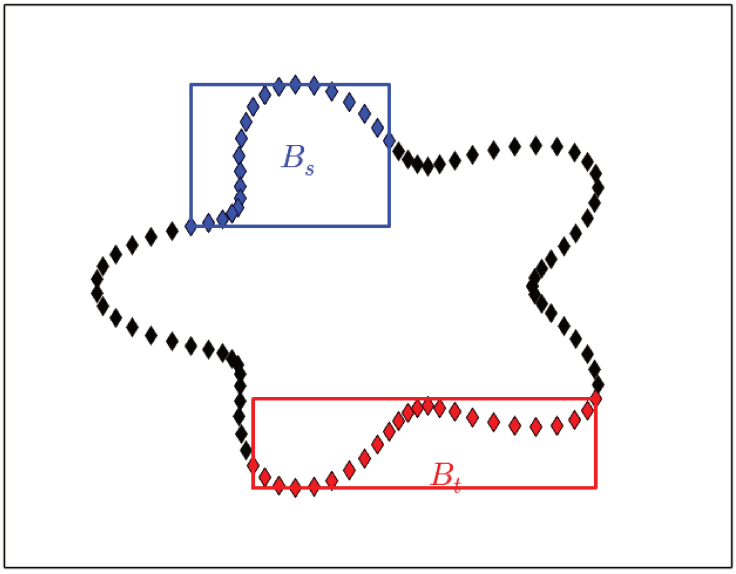


FIGURE 1.4 – Boîtes englobantes B_t et B_s pour deux sous-ensembles de degrés de liberté. En bleu les degrés de liberté associés au sous-ensemble s, en rouge ceux associés à l'ensemble t et en noir les autres degrés de liberté du domaine Γ. Les degrés de liberté de s et t définissent une interaction entre les domaines Γ_s et Γ_t.

Pour les applications, on souhaite pouvoir calculer le diamètre d'une boîte et la distance entre deux boîtes. Ceci est fait en $\mathscr{O}(1)$ opérations à l'aide des formules suivantes.

Proposition 1.17 (Diamètre d'une boîte)**.** *On note* $B = [a_1, b_1] \times [a_2, b_2] \times [a_3, b_3]$*, une boîte englobante. Son diamètre* diam(B) *est défini par la relation suivante*

$$\mathrm{diam}(B) = \sqrt{\sum_{q=1}^{3} (b_q - a_q)^2}. \tag{1.105}$$

Proposition 1.18 (Distance entre deux boîtes)**.** *On définit la distance* $\mathrm{dist}(B_t, B_s)$ *entre deux boîtes* $B_t = [a_1^t, b_1^t] \times [a_2^t, b_2^t] \times [a_3^t, b_3^t]$ *et* $B_s = [a_1^s, b_1^s] \times [a_2^s, b_2^s] \times [a_3^s, b_3^s]$ *de la façon suivante*

$$\mathrm{dist}(B_t, B_s) = \sqrt{\sum_{q=1}^{3} \mathrm{dist}([a_q^t, b_q^t], [a_q^s, b_q^s])^2}. \tag{1.106}$$

L'utilisation de ces boites est utile dans la pratique lorsque l'on cherche à regrouper des points par proximité géométrique. Lorsque les points sont distribués suivant une géométrie particulière (par exemple sur une surface comme dans le cas du domaine Γ de la

figure (1.4)), ces boites ne sont pas optimales en ce sens où l'on peut trouver une orientation dans laquelle la boîte englobante est plus petite. On peut construire une boîte orientée à l'aide de l'analyse en composantes principales ([GL96; Mes11]).

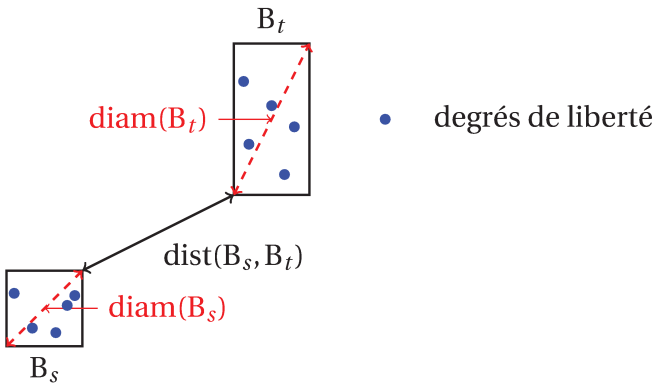


FIGURE 1.5 – Illustration graphique des diamètres des boîtes englobantes ainsi que de la distance entre ces boîtes.

Boîte englobante orientée On considère les mêmes N points $\{x_{i_k}\}_{k=1,\dots,N}$ de $\mathbb{R}^3$. On cherche à construire une boîte englobante orientée qui serait plus adaptée à la distribution des points $\{x_{i_k}\}_{k=1,\dots,N}$). On note x_g le centre de gravité de ce nuage de points et l'on note X la matrice des coordonnés centrées en x_g (*i.e* x_g est l'origine du système) et de taille $3 \times N$.
La méthode consiste à déterminer les directions principales du nuage de points et correspond à ce que l'on trouve dans la littérature sous l'appellation d'analyse en composantes principales. On note C_X la matrice symétrique dite **matrice de covariance**, de taille 3×3 définie par

$$C_X = XX^T. \tag{1.107}$$

La diagonalisation de la matrice C_X fournit alors une nouvelle base orthonormée U telle que

$$C_X = UDU^T, \tag{1.108}$$

où $D := \mathrm{diag}(\lambda_1, \lambda_2, \lambda_3)$ est une matrice diagonale dont les valeurs propres sont positives. On note alors u_1, u_2, u_3 les colonnes de cette matrice U. Quitte à exprimer les coordonnées des points du nuage de degrés de liberté original dans la nouvelle base, on peut obtenir une boite orientée B_t dont les dimensions sont adaptées à la distribution de points. En effet, en notant $\tilde{X}$ les coordonnées du nuage dans le système de coordonnées (x_g, U), la relation de changement de base s'écrit pour chaque degré de liberté $x_{i_k}, k \in \{1, \dots, N\}$ avec $\tilde{x}_{i_k} = ((\tilde{x}_{i_k})_1, (\tilde{x}_{i_k})_2, (\tilde{x}_{i_k})_3)^T$,

$$x_{i_k} = x_g + U\tilde{x}_{i_k}. \tag{1.109}$$

Comme $U^{-1} = U^T$, on obtient

$$\tilde{x}_{i_k} = U^T(x_{i_k} - x_g). \tag{1.110}$$

Les dimensions de la nouvelle boite $B_t = [\tilde{a}_1^s, \tilde{b}_1^s] \times \dots \times [\tilde{a}_d^s, \tilde{b}_d^s]$ centrée en x_g sont

$$\tilde{a}_\nu = \min_{i=1,\dots,N} \tilde{x}_{i,\nu}, \tag{1.111}$$

$$\tilde{b}_\nu = \max_{i=1,\dots,N} \tilde{x}_{i,\nu}, \tag{1.112}$$

pour $\nu = 1, \ldots, d$. La figure suivante illustre les deux boîtes englobantes B_t possibles évoquées sur l'exemple de du domaine Γ_t de l'exemple .

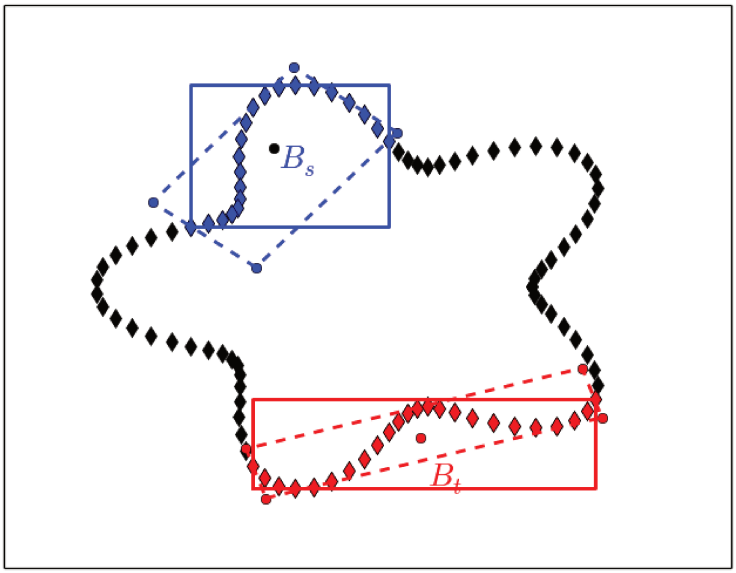


FIGURE 1.6 – Boîtes englobantes B_t et B_s pour deux sous-ensembles de degrés de liberté. En trait plein, les boîtes englobantes dans le système d'axe cartésien. En trait pointillé, la boîte englobante adaptée à la direction principale du groupe d'inconnues considéré.

On constate sur l'exemple décrit par la figure 1.6 que la boîte orientée B_t (resp. B_s) est plus adaptées à la géométrie Γ_t (resp. Γ_s). On utilisera au chapitre 4 une construction similaire pour l'approximation du noyau de Helmholtz.

1.3.2.2 Regroupement des degrés de liberté

On considère N points $\{x_{i_k}\}_{k=1,\ldots,N}$ de $\mathbb{R}^3$ appelés degrés de liberté. On souhaite construire des groupes de points de façon hiérarchique suivant en cela l'exemple 1.7. Dans la suite, on note $I = \{1, \ldots, N\}$. Un élément $k \in \mathscr{I}$ est associé au degré de liberté x_{i_k}.

Différentes numérotations Dans la suite de ce chapitre, plusieurs numérotations coexistent et selon les objets considérés nous n'utiliserons pas la même numérotation.

Degrés de liberté La numérotation des degrés de liberté est la première numérotation que l'on manipule. Les degrés de liberté x_{i_k} sont indexés par les indices i_k pour $k \in \{1, \ldots, N\}$. Les indices i_k ne sont pas nécessairement contigus. Dans la pratique, cette numérotation provient d'un logiciel de CAO ou d'un mailleur et suivant les applications visées cette numérotation a un lien avec la physique du problème. Les indices k forment eux un ensemble d'indices contigus.

Matrice de discrétisation À partir de N degrés de liberté, on assemble une matrice de discrétisation de taille $N \times N$. Ainsi, l'ensemble $\mathscr{I} = \{1, \ldots, N\}$ est un ensemble d'indices

naturellement adapté à la description de la matrice. La k^{e} ligne de la matrice de discrétisation correspond à l'interaction du degré de liberté x_{i_k} avec l'ensemble des degrés de liberté.

Groupe d'inconnues, arbre et stratégies de partition Comme nous l'avons déjà mentionné, la méthode des $\mathscr{H}$-matrices repose sur une construction hiérarchique en scindant les inconnues à l'aide d'un arbre. Cette idée est très répandue dans la pratique ([Cip00]) pour construire des méthodes rapides. On rappelle les notions sur les arbres utiles pour la compréhension de la méthode des $\mathscr{H}$-matrices [BGH12; Liz14].

Définition 1.19 (Groupe d'indices [BGH12])**.** On considère un ensemble d'indices non vide I. On définit un groupe (ou *cluster* dans la littérature en anglais) d'indices t comme un sous-ensemble d'indices de I. Par exemple, $t = \{23, 17, 89\}$ est un groupe d'indices inclus dans $\mathrm{I} = \{1, \ldots, 100\}$. Dans la suite, on parle simplement du groupe t. On notera $|t|$ le cardinal de l'ensemble t.

Pour chaque élément k d'un groupe $t \subset \mathrm{I}$, on peut considérer le nuage de points $\mathrm{X}_t := \{x_{i_k}\}_{k \in t}$ correspondant. De la même façon que l'on associe à l'indice i_k le degré de liberté x_{i_k}, on associe au groupe t le nuage X_t. Dans un contexte matriciel, t désignera ainsi un sous-ensemble de lignes (ou colonnes) tandis que dans un contexte d'approximation du noyau de Green, t désignera le nuage X_t ou son support.

Pour une quantité géométrique ω_{i_k} associée au degré de liberté x_{i_k}, on peut associer au groupe t la quantité

$$\omega_t = \bigcup_{k \in t} \omega_{i_k}. \tag{1.113}$$

Par exemple, le support $\mathrm{Supp}(t)$ d'un groupe t est l'union des supports des degrés de liberté le constituant et ainsi

$$\begin{aligned} \mathrm{Supp}(t) &= \bigcup_{k \in t} \mathrm{Supp}(x_{i_k}) && (1.114)\\ &= \mathrm{Supp}(\mathrm{X}_t) && (1.115) \end{aligned}$$

Les exemples des précédents paragraphes montrent que le noyau de Green $\mathrm{G}(x, y)$ peut s'approcher de manière efficace lorsque les degrés de liberté sont séparés géométriquement. D'un point de vue matriciel, cela revient dans un premier temps à considérer des sous-ensembles des lignes et des colonnes.

En tant qu'ensemble, il est possible de partitionner un groupe suivant plusieurs critères afin d'obtenir des sous-groupes (« les enfants » en terme d'arbre).

Définition 1.20 (Partition d'un groupe)**.** Si un groupe t est partitionné en n_p sous-groupes, on les notera $t_1, \ldots, t_{n_p}$. Ces n_p sous-ensembles respectent les conditions suivantes :

$$\begin{aligned} t_i \subset t \quad &\text{pour} \quad i = 1, \ldots, n_p \\ t \quad &= \quad \bigcup_{i=1}^{n_p} t_i \\ t_i \cap t_j = \emptyset \quad &\text{si} \quad i \neq j. \end{aligned}$$

Le groupe t est dit **père** des **fils** $t_1, \ldots, t_{n_p}$.

Arbre de groupes La partition répétée de manière hiérarchique (*i.e* appliquée récursivement aux fils) aboutit à la construction d'un **arbre**, noté $T_{\mathscr{I}}$. À chaque étape, le nombre maximal de fils est appelé l'**arité** de l'arbre. On fournit les points de vocabulaire [Liz14] relatifs aux arbres et utilisés par la suite :

— On appelle **nœuds** de l'arbre les différents groupes construits à chaque étape de la partition hiérarchique. On parle également de sommets d'un arbre ;

— L'unique nœud n'ayant pas de père (de prédécesseur) est appelé la **racine** ;

— Pour un **nœud** t, on note $S(t)$ l'ensemble de ses fils ;

— Un **nœud** n'ayant pas de fils est appelé une **feuille**. On notera $\mathscr{L}(\mathscr{T}_I)$ l'ensemble des feuilles de l'arbre

$$\mathscr{L}(\mathscr{T}_I) = \{t \in \mathscr{T}_I \ : \ S(t) = \emptyset\}. \tag{1.116}$$

— Un arbre est dit **complet** si toutes ses feuilles sont à la même distance de la racine.

— Le **niveau** d'un nœud t est la distance entre ce nœud et la racine. Par exemple, les fils de la racine sont au niveau 1 de l'arbre tandis que leurs fils sont eux au niveau 2. On note niveau(t) le niveau du groupe t.

— On note $L(\mathscr{T}_I)$ la **profondeur** de l'arbre. Cette dernière est définie par

$$L(\mathscr{T}_I) = \max_{t \in \mathscr{T}_I} \text{niveau}(t). \tag{1.117}$$

La profondeur de l'arbre correspond au niveau maximal des feuilles de l'arbre.

Dans la pratique, la méthode des $\mathscr{H}$-matrices consiste à effectuer une partition des degrés de liberté à l'aide d'un arbre binaire (*i.e* d'arité 2) tandis que la méthode des multipôles rapides utilise elle un arbre d'arité 8 (un *octree*). Il existe aussi dans la littérature ([Beb00],[BGH12]) des arbres ternaires (arité 3). Un arbre est de plus dit **entier** si chaque nœud est soit le père de deux fils, soit une feuille.

Dans la pratique, on se fixe un paramètre entier $n_{\max} > 0$ que l'on appelle **taille de feuille maximale**. Cet entier correspond à la taille des groupes de plus haut niveau dans l'arbre. Les feuilles ayant un cardinal inférieur à n_{max} ne sont pas partitionnnées.

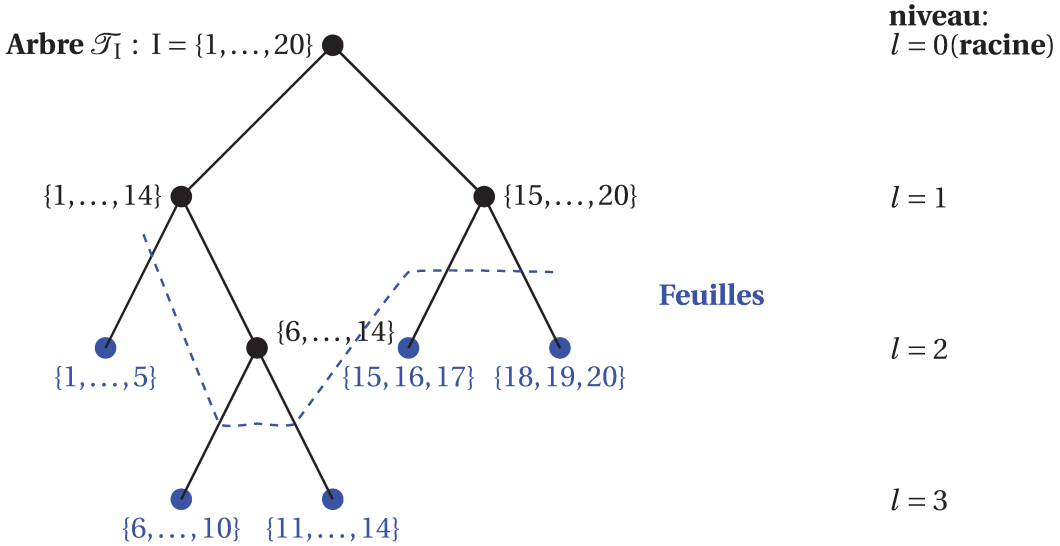


FIGURE 1.7 – Exemple d'un arbre de groupe $\mathscr{T}_I$ basé sur l'ensemble $I = \{1,\dots,20\}$ avec $n_{\max} = 5$.

La figure (1.7) illustre les définitions précédentes. À partir de l'ensemble $\mathrm{I} = \{1,\dots,20\}$, on construit un **arbre de groupes**. Chaque point (noir ou bleu) est un **nœud** de l'**arbre** $\mathscr{T}_\mathrm{I}$. Si un nœud possède des **fils**, on fait figurer une arête (trait noir plein) reliant ces nœuds. Par exemple, le nœud $t = \{1,\dots,14\}$ possède deux fils $t' = \{1,\dots,5\}$ et $t'' = \{6,\dots,14\}$. Ainsi, on a

$$\mathrm{S}(t) = \big\{\{1,\dots,5\},\ \{6,\dots,14\}\big\}.$$

Le fils d'un nœud possède un **niveau** supérieur à celui de son père (*cf* côté droit de la figure). Les nœuds n'ayant pas d'**enfant** sont les **feuilles** de l'arbre (points bleus sur la figure). Sur la figure, les feuilles sont tous les nœuds sous la ligne pointillée bleue et l'ensemble $\mathscr{L}(\mathscr{T}_\mathrm{I})$ des feuilles de cet arbre $\mathscr{T}_\mathrm{I}$ est donné par

$$\mathscr{L}(\mathscr{T}_\mathrm{I}) = \big\{\{1,\dots,5\},\{6,\dots,10\},\{11,\dots,14\},\{15,16,17\},\{18,19,20\}\big\}.$$

L'arbre construit ici est **entier** mais n'est pas **complet** car les feuilles ne sont pas toutes au même niveau (ici, 2 et 3). La profondeur $\mathrm{L}(\mathscr{T}_\mathrm{I})$ de l'arbre est donnée par

$$\begin{aligned}\mathrm{L}(\mathscr{T}_\mathrm{I}) &:= \max_{t\in\mathscr{T}_\mathrm{I}}\{\text{niveau}(t)\}\\ &=3.\end{aligned}$$

Enfin, la taille de feuille maximale sur cet exemple est fixée à $n_{\max} = 5$ *i.e* les feuilles de $\mathscr{T}_\mathrm{I}$ ne contiennent pas plus de 5 éléments.
Pour cet exemple, nous avons choisi de représenter les éléments d'un sous-ensemble de manière consécutive. On peut aisément se ramener à ce choix dans la pratique en effectuant une renumérotation à chaque bissection.

Algorithme de partition L'algorithme 1 suivant réalise la partition de l'ensemble I à partir de la racine $t = \mathrm{I}$. Le paramètre n_{max} est choisi par l'utilisateur.

Algorithme 1 Construction d'un arbre de groupes

```
1: Fonction CreationArbre(t) :
2: Si |t| < n_max Alors
3:     S(t) = ∅
4: Sinon
5:     t_1, t_2 = Partition(t)
6:     CreationArbre(t_1)
7:     CreationArbre(t_2)
8:     S(t) = {t_1, t_2}
9: Fin Si
10: finFonction
```

Cet algorithme nécessite une fonction réalisant la bissection (partition en deux sous-ensembles) d'un groupe t donné. Plusieurs stratégies sont envisageables et l'on en détaille dans la suite trois différentes. Le point commun de ces méthodes est de construire des sous-ensembles de degrés de liberté qui sont géométriquement proches. Pour ce faire, on utilise la boîte englobante B_t du groupe définie par 1.16. La partition du groupe t consiste alors à couper la boîte B_t en deux par un plan. Tous les indices associés aux degrés de liberté situés d'un côté du plan forment le premier fils tandis que les autres constituent le second fils. C'est cette opération que la fonction **Partition** de l'algorithme 1 réalise.

Position du plan de séparation Avec les notations de la définition (1.16), les vecteurs $(a_1, a_2, a_3)^{\mathrm{T}}$ et $(b_1, b_2, b_3)^{\mathrm{T}}$ définissent les points extrémaux de B_t. La plus grande dimension $i^\star$ est alors déterminée par

$$i^\star = \mathrm{argmax}_{i=1,2,3} |b_i - a_i|. \tag{1.118}$$

Le plan de séparation est alors le plan orthogonal à la direction $i^\star$. On peut à présent déplacer ce plan le long de la direction $i^\star$. Deux choix classiques sont possibles pour réaliser la découpe souhaitée :

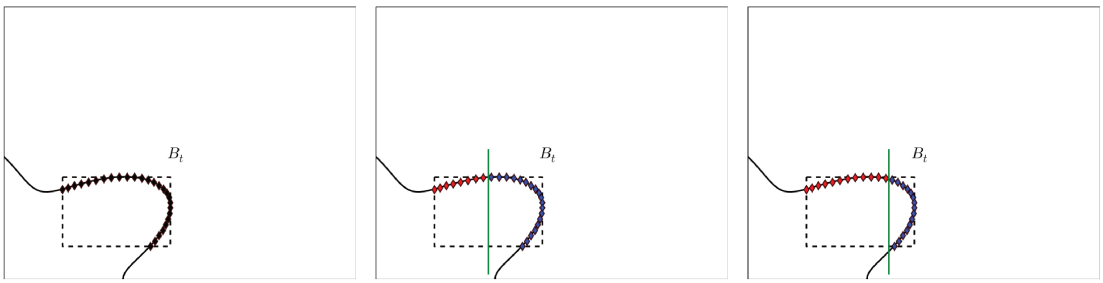
Découpe médiane : on positionne le plan de séparation de sorte à obtenir approximativement le même nombre de degrés de liberté de chaque côté du plan ;

Découpe géométrique : on positionne le plan de séparation de façon à découper la boîte B_t en deux boîtes $\mathrm{B}_t^{(1)}$ et $\mathrm{B}_t^{(2)}$ de même taille. Les points présents dans la boîte $\mathrm{B}_t^{(1)}$ forment le fils t_1 et les autres le fils t_2. En d'autres termes,

$$t_1 := \left\{ k \in t \,:\, (x_{i_k})_{i^\star} < \frac{1}{2}(b_{i^\star} + a_{i^\star}) \right\}, \tag{1.119}$$

$$t_2 := \left\{ k \in t \,:\, (x_{i_k})_{i^\star} \geqslant \frac{1}{2}(b_{i^\star} + a_{i^\star}) \right\}. \tag{1.120}$$

$$\tag{1.121}$$



(a) Boîte originale (b) Découpe géométrique (c) Découpe médiane

FIGURE 1.8 – Illustration du positionnement du plan de séparation en 2D. En vert, le plan de séparation et en pointillés noirs la boîte englobante du groupe à partitionner. En rouge et en bleu, les éléments associés aux fils du groupe t.

Dans les deux cas, on remarque que la partition d'une boîte englobante (donc d'un groupe) ne produit pas de boîte vide ne contenant pas de degrés de liberté. L'arbre ainsi créé est alors entier.

Ces deux choix obéissent à des objectifs distincts. La découpe médiane permet de diviser par deux le cardinal des groupes à chaque niveau de l'arbre et fournit ainsi un arbre de profondeur minimale. La découpe géométrique permet quant à elle de diviser par deux le volume de la boîte englobante à chaque niveau sans se préoccuper du nombre de degrés de liberté contenus dans chaque boîte.

La création de l'arbre de groupes est un point important de la méthode des $\mathscr{H}$-matrices et doit être créé avec minutie. D'un point de vue informatique, la découpe médiane produit un arbre convenable mais les boîtes englobantes associées aux nœuds de l'arbre possèdent des tailles très hétérogènes. Il s'agit du cas typique où la distribution de degrés de liberté est concentrée en une zone de l'espace (cas d'un raffinement local). L'emploi d'une

condition du type (1.24) faisant intervenir les diamètres des boîtes montre que cela n'est pas un avantage de travailler avec de grandes boîtes. La découpe géométrique permet de manipuler des boîtes de dimensions homogènes mais le nombre de degrés de liberté dans chaque boîte n'est pas homogène. On est alors amené à découper plusieurs fois les boîtes afin d'atteindre la taille de feuille maximale. Ainsi, la profondeur de l'arbre $\mathscr{T}_{\mathrm{I}}$ peut être plus élevée que lors de la découpe médiane, ce qui représente également un inconvénient.

La figure (1.9) montre différentes partitions dans le cas d'un cône-sphère. Il s'agit d'une demi-sphère attachée à la base d'un cône de révolution. Cet objet est un cas test usuel pour valider les résultats d'un code de diffraction en électromagnétisme et possède la particularité d'avoir une plus forte densité de degrés de liberté dans la partie sphérique. La figure (1.9) souligne les différences entre les deux stratégies de découpe mentionnées.

On remarque qu'à chaque étape, les boîtes englobantes épousent le plus possible la géométrie. En effet, la boîte englobante est recalculée à chaque niveau de l'arbre afin de travailler avec des boîtes les plus petites possibles. Un critère d'éloignement sera dans la suite basé sur ces diamètres et il serait néfaste de travailler avec des boites trop grandes. Sans surprise la découpe géométrique fournit des boîtes plus adaptée à la géométrie. Nous n'avons pas connaissance d'une découpe « optimale » pour nos applications mais l'on préférera une découpe géométrique pour les applications visées. En effet, nous développerons des approximations basées sur les dimensions des boîtes et non sur le nombre de degrés de liberté qu'elles contiennent. Par ailleurs nos approximations seront orientées dans $\mathbb{R}^3$ et se rapprochent d'une méthode de bissection par la direction principale.

Partition équilibrée par direction principale Le précédent paragraphe a montré qu'il est également possible de déterminer une boîte orientée d'un groupe t. La diagonalisation de la matrice de covariance en (1.108) fournit une base orientée de $\mathbb{R}^3$. On note $\vec{w}$ la direction associée à la plus grande valeur propre, c'est la direction principale du groupe. On peut alors considérer un plan de séparation normal à cette direction et passant par le centre de gravité x_g du groupe. On possède alors un moyen de partitionner le groupe t en deux sous-groupes t_1 et t_2,

$$t_1 := \{k \in t \ : \ (x_{i_k} - x_g) \cdot \vec{w} < 0\}, \tag{1.122}$$
$$t_2 := \{k \in t \ : \ (x_{i_k} - x_g) \cdot \vec{w} \geqslant 0\}. \tag{1.123}$$

On illustre cette construction en dimension deux par la figure 1.10.

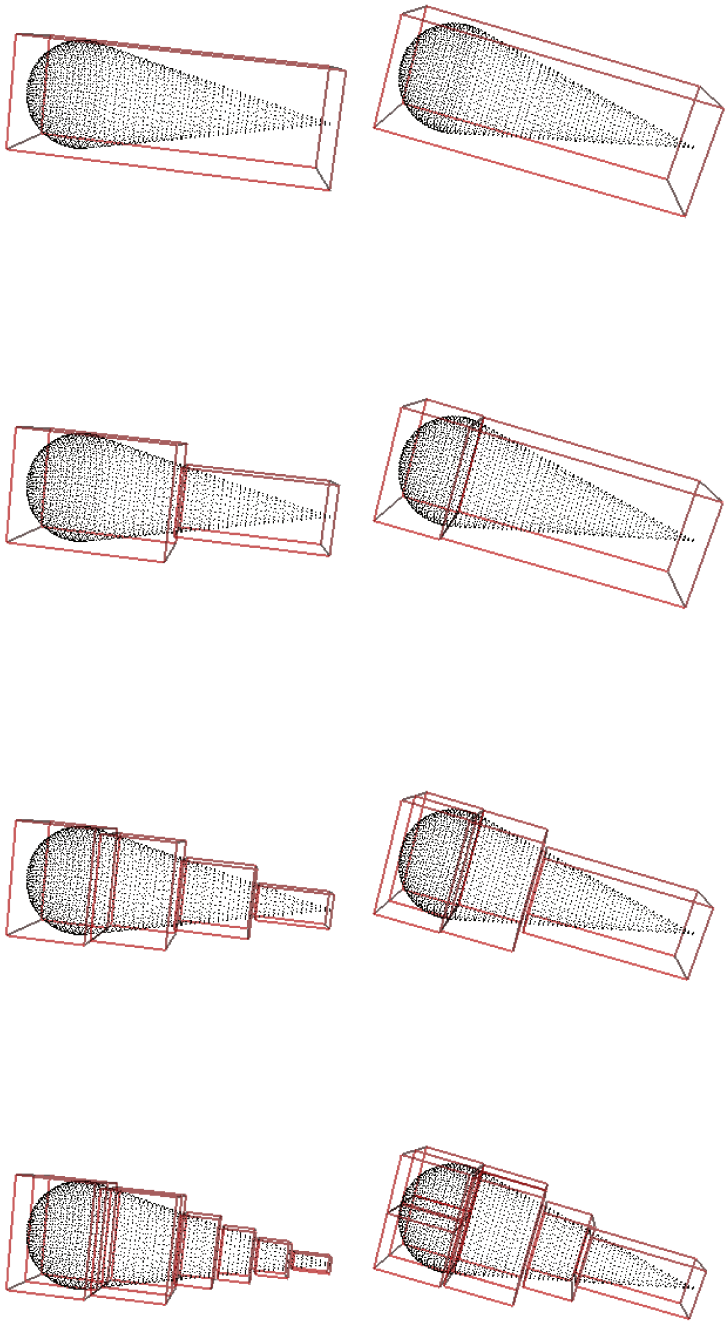
Pour des géométries ne présentant pas de raffinement local très fin, on constate de plus dans la pratique que l'analyse en composantes principales fournit des fils contenant un nombre d'éléments similaire.
Cette méthode nécessite d'effectuer des calculs supplémentaires pour déterminer la direction principale des élements lors de chaque bissection. Pour la partition de N indices, l'étape onéreuse est la détermination de la matrice de covariance dont chaque coefficient est un produit scalaire de vecteurs de taille N (soit une complexité linéaire). La diagonalisation de la matrice de covariance s'effectue en $\mathscr{O}(1)$ opérations. La complexité de la partition est donc bien linéaire.

Remarque 1.21 (Autres méthodes de regroupement). Il existe beaucoup de méthodes permettant de répartir des points dans l'espace suivant des critères variés (par exemple géométrique ou ordinal). Dans la pratique, on peut citer la méthode de bissection spectrale, des méthodes algébriques lors de l'utilisation de matrices creuses ou la décomposition de domaines ([BGH12; Beb00; Mes11]).

Complexité La proposition suivante permet de relier les dimensions des groupes de $\mathscr{T}_\mathrm{I}$ à la dimension de l'ensemble de départ I. Ce type d'estimation sera utilisé par la suite pour obtenir des estimations sur les $\mathscr{H}$-matrices.

Proposition 1.22 ([Beb00], Lemme 1.21)**.** *Soit $\mathscr{T}_\mathrm{I}$ un arbre de groupe pour l'ensemble* I. *Alors,*

$$\sum_{t \in \mathscr{T}_\mathrm{I}} |t| \leqslant \mathrm{L}(\mathscr{T}_\mathrm{I})\,|\mathrm{I}|, \tag{1.124}$$
$$\sum_{t \in \mathscr{T}_\mathrm{I}} |t| \log(|t|) \leqslant \mathrm{L}(\mathscr{T}_\mathrm{I})\,|\mathrm{I}| \log(|\mathrm{I}|). \tag{1.125}$$



(a) Découpage géométrique (b) Découpage médian

FIGURE 1.9 – Illustration graphique de la méthode de regroupement inconnues sur le cas d'un cône-sphère. La colonne de gauche correspond à un découpage géométrique tandis que la colonne de droite représente une découpe équilibrée basée sur le nombre d'inconnues.

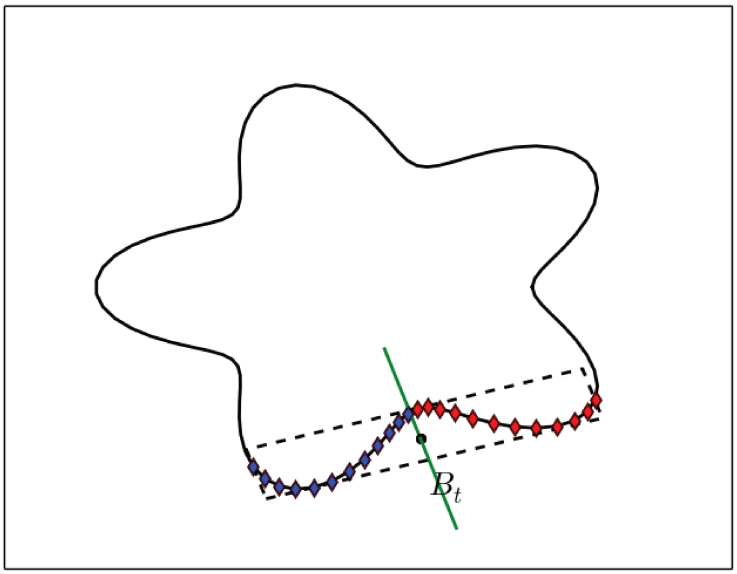


FIGURE 1.10 – Boîte englobante orientée B_t. En vert, le plan de séparation normal à la direction principale. En rouge ; les degrés de libertés appartenant à t_2 et en bleu ceux appartenant à t_1.

De plus, pour un réel $c > 0$, on a

$$\sum_{t\in\mathscr{T}_I} \min\{c, |t|^2\} \leqslant 3\sqrt{c}|I|. \tag{1.126}$$

(1.127)

L'étape de regroupement des degrés de liberté est une étape économique. Dans le cas d'un découpage équilibré, la profondeur de l'arbre est au plus celle d'un arbre binaire complet plus un. On a ainsi

$$L(\mathscr{T}_I) \leqslant \lfloor \log_2(N) \rfloor + 2. \tag{1.128}$$

Chaque niveau de l'arbre nécessite un nombre linéaire d'opérations pour déterminer la position du plan de séparation des boîtes englobantes. L'arbre comportant $L(\mathscr{T}_I)$ niveau, l'étape de regroupement des degrés de liberté est donc en $\mathscr{O}(N\log(N))$ opérations. Pour un nuage de N points.
Le cas d'une découpe géométrique peut engendrer un arbre déséquilibré. Dans ce cas, il est impossible de fournir une borne meilleure que $\mathscr{O}(N^2)$ à cause de cas particuliers (voir par exemple [Liz14]). Dans la pratique, les degrés de liberté sont très souvent associés à des éléments géométriques tels des arêtes, des triangles ou des sommets et dans la plupart des applications, on observe un coût bien meilleur que la borne pessismiste.
Dans la pratique, le comportement moyen constaté de la partition des degrés de liberté à l'aide d'un arbre binaire est de $\mathscr{O}(N\log(N))$ opérations.

La partition des indices en groupes réalisée correspond à une partition d'indices de lignes et/ou de colonnes de la matrice de discrétisation. Chaque groupe créé représente une petit « paquet » de degrés de liberté localisé dans l'espace. En suivant l'exemple 1.7, on s'intéresse à la possibilité de découper la matrice de discrétisation en blocs représentant des interactions distantes de paquets de degrés de liberté. C'est l'étape de la création d'un arbre de blocs.

1.3.3 Découpe hiérarchique d'une matrice

Dans la suite, on considère une matrice carrée de taille N × N. Les indices de lignes et de colonnes sont partitionnés et cette partition est représentée par l'arbre de groupes $\mathscr{T}_\mathrm{I}$. La méthode que l'on va décrire demeure valable pour des matrices quelconques à condition de considérer les arbres d'indices de lignes $\mathscr{T}_\mathrm{I}$ et de colonnes $\mathscr{T}_\mathrm{J}$ adaptés [BGH12; Beb00]. La répartition des indices en clusters revient à créer des petits paquets imbriqués de points dans $\mathbb{R}^3$. La motivation première de ce travail est l'assemblage de la matrice d'interaction **G**. Ainsi un bloc matriciel extrait de **G** donné correspond à l'interaction de deux paquets de points. Intuitivement, la création d'un arbre de clusters fournit une partition des indices des lignes et/ou des colonnes. On considère alors deux ensembles d'indices I et J à partir desquels on a respectivement construit les arbres $\mathscr{T}_\mathrm{I}$ et $\mathscr{T}_\mathrm{J}$.

1.3.3.1 Arbre de blocs

De la même façon que l'on a représenté le nuage de degrés de liberté sous la forme d'un arbre binaire, on va donner une représentation sous forme d'arbre des blocs extraits de la matrice de discrétisation. On définit un arbre de blocs de la façon suivante

Définition 1.23 (Arbre de blocs)**.** On note $\mathrm{I} = \{1, \ldots, \mathrm{N}\}$ et l'on considère un arbre de groupes $\mathscr{T}_\mathrm{I}$. Un arbre $\mathscr{T}$ est appelé arbre de blocs pour le produit $\mathscr{T}_\mathrm{I} \times \mathscr{T}_\mathrm{I}$ si et seulement si les conditions suivantes sont vérifiées :

1. La racine de $\mathscr{T}$ est le produit de la racine de $\mathscr{T}_\mathrm{I}$ avec elle-même :
$$\text{racine}(\mathscr{T}) = \text{racine}(\mathscr{T}_\mathrm{I}) \times \text{racine}(\mathscr{T}_\mathrm{I}). \tag{1.129}$$
2. Chaque nœud b de $\mathscr{T}$ est de la forme
$$b = t \times s, \tag{1.130}$$
avec $t \in \mathscr{T}_\mathrm{I}$ et $s \in \mathscr{T}_\mathrm{I}$.
3. Comme dans le cas des arbres de groupes, on note $\mathrm{S}(b)$ l'ensemble des fils d'un nœud b. Pour chaque nœud b tel que $\mathrm{S}(b) \neq \emptyset$, on a
$$\mathrm{S}(b) = \begin{cases} \{t \times s' : s' \in \mathrm{S}(s)\} & \text{si } \mathrm{S}(t) = \emptyset, \mathrm{S}(s) \neq \emptyset \\ \{t' \times s : t' \in \mathrm{S}(t)\} & \text{si } \mathrm{S}(t) \neq \emptyset, \mathrm{S}(s) = \emptyset \\ \{t' \times s' : t' \in \mathrm{S}(t),\, s' \in \mathrm{S}(s)\} & \text{sinon} \end{cases} \tag{1.131}$$

On associe à un élément de $\mathscr{T}_\mathrm{I}$ les degrés de libertés correspondant et de mêmes, on associe aux nœuds de $\mathscr{T}$ des blocs matriciels et l'on note $\mathscr{T}_{\mathrm{I}\times\mathrm{I}}$ cet arbre de blocs. L'ensemble I×I vérifie le premier point de la liste ci-dessus et correspond à la matrice de discrétisation A tandis que les feuilles de $\mathscr{T}_{\mathrm{I}\times\mathrm{I}}$ forment une partition disjointe de la matrice de départ.

Pour un nœud (matrice) de l'arbre $\mathscr{T}_{\mathrm{I}\times\mathrm{I}}$, le troisième point de la définition 1.23 précise que l'on peut obtenir trois types de fils (sous-matrices).
Pour deux groupes t et s, le nœud $b = t \times s$ correspond à la matrice $\mathrm{A}_{t\times s}$. Dans le cas où $\mathrm{S}(t) = \emptyset$ et $\mathrm{S}(s) \neq \emptyset$, on effectue la partition 1×2 suivante
$$\mathrm{A}_{t\times s} = \begin{bmatrix} \mathrm{A}_{t\times s_1} & \mathrm{A}_{t\times s_2} \end{bmatrix}. \tag{1.132}$$

Le cas $\mathrm{S}(t) \neq \emptyset$ et $\mathrm{S}(s) = \emptyset$ correspond à la partition en 2×1 suivante

$$\mathrm{A}_{t\times s} = \begin{bmatrix} \mathrm{A}_{t_1\times s} \\ \mathrm{A}_{t_2\times s} \end{bmatrix}. \tag{1.133}$$

Enfin, le cas $\mathrm{S}(t) \neq \emptyset$ et $\mathrm{S}(s) \neq \emptyset$ correspond à la découpe en une matrice bloc 2×2,

$$\mathrm{A}_{t\times s} = \begin{bmatrix} \mathrm{A}_{t_1\times s_1} & \mathrm{A}_{t_1\times s_2} \\ \mathrm{A}_{t_2\times s_1} & \mathrm{A}_{t_2\times s_2} \end{bmatrix}. \tag{1.134}$$

Ainsi, chaque nœud de l'arbre de blocs peut avoir trois types de fils. À ce stade, cet arbre possède des feuilles qui sont toutes des matrices de taille au plus $n_{\max} \times n_{\max}$ par définition des arbres de groupes $\mathscr{T}_{\mathrm{I}}$. L'exemple (1.7) suggère que sous réserve que les groupes t et s soient suffisament éloignés, la matrice de l'interaction $t \times s$ est de rang faible et peut être approchée de manière rapide et efficace.
Par ailleurs, cet exemple ainsi que l'exemple (1.2) suggèrent que le rang d'une interaction $t \times s$ ne dépend pas de la taille des groupes mais de la précision ϵ de l'approximation souhaitée. On va donc chercher à « élaguer » l'arbre des blocs de façon à obtenir des feuilles de tailles hétérogènes et de plus grande taille possible. Pour cela, on utilise deux ingrédients distincts.

Définition 1.24 (Interactions de niveau constant)**.** On considère un arbre de blocs $\mathscr{T}_{\mathrm{I}\times\mathrm{I}}$ associé à un arbre de groupe $\mathscr{T}_{\mathrm{I}}$. On dit que l'arbre de blocs est à niveau constant si et seulement si pour tout nœud $b = t \times s$, on vérifie la propriété suivante,

$$\text{niveau}(b) = \text{niveau}(t), \tag{1.135}$$
$$\text{niveau}(b) = \text{niveau}(s). \tag{1.136}$$

Avec cette définition, les partitions (1.132) et (1.133) ne sont pas tolérées et chaque bloc matriciel est soit conservé intact soit subdivisé en une matrice bloc 2×2. On rappelle que les arbres de groupes sont entiers. Chaque nœud possèdant des fils en possède exactement deux. Dans le cas de l'arbre de blocs, l'arité de l'arbre est de quatre, soit le carré de celle de l'arbre de groupes.
Cette limitation de l'arbre est un **choix** et est en partie motivée par l'obtention d'une structure de matrice blocs 2×2 lors des subdivisions de matrices. Ceci facilite les opérations algébriques récursives que l'on effectue par la suite.

Pour obtenir des blocs matriciels de grande taille, il est nécessaire d'utilser un autre ingrédient afin de modifier l'arbre $\mathscr{T}_{\mathrm{I}\times\mathrm{I}}$ pour obtenir des feuilles de plus grandes dimensions : c'est la condition d'admissibilité.

1.3.3.2 Conditon d'admissibilité

L'exemple (1.2) a montré que si les groupes t et s sont séparés alors le noyau est régulier et l'on peut l'approcher de manière efficace. La relation (1.24) était la condition choisie pour le noyau logarithmique $\mathrm{K}(x, y)$ et des intervalles de $\mathbb{R}$. Dans le cas d'une surface de $\mathbb{R}^3$, on peut généraliser cette condition à l'aide des boîtes englobantes B_t et B_s des groupes t et s :

$$\min\left(\mathrm{diam}(\mathrm{B}_s), \mathrm{diam}(\mathrm{B}_t)\right) \leqslant \eta\, \mathrm{dist}(\mathrm{B}_s, \mathrm{B}_t), \tag{1.137}$$

où $\mathrm{diam}(\mathrm{B}_s)$ (resp. $\mathrm{diam}(\mathrm{B}_t)$) est le diamètre de la boîte englobante B_s (resp .B_t) et $\mathrm{dist}(\mathrm{B}_s, \mathrm{B}_t)$ est la distance entre les boîtes B_s et B_t.

Il s'agit de la condition d'admissibilité utilisée pour le noyau de Green de l'équation de Laplace dans la littérature [BGH12; Beb00; Liz14; BG05]. Par ailleurs la condition de séparation usuelle dans la FMM est également une condition de ce type liant le rayon $\frac{a\sqrt{3}}{2}$ d'une boîte cubique de côté a à la distance R des centres des boîtes [Syl02]. La condition de séparation requiert que les boîtes ne se touchent pas ce qui est vérifié dès que $\mathrm{R} \geqslant 2a\sqrt{3}$.
Dans le cadre des $\mathscr{H}$-matrices, on parle dans ce cas d'admissibilité faible ; l'admissibilité forte correspondant à la condition similaire en prenant $\max(\mathrm{dist}(\mathrm{B}_s), \mathrm{diam}(\mathrm{B}_t))$. Par opposition au cas du noyau oscillant $\mathrm{G}(x,y) = \frac{e^{ik\|x-y\|}}{4\pi\|x-y\|}$, on qualifiera par la suite cette condition de statique.
La condition (1.137) réalise la séparation des interactions proches et lointaines et sous réserve de satisfaire à cette condition, on verra au chapitre prochain que le noyau de Green $\mathrm{G}(x,y) = \frac{1}{4\pi}\frac{1}{\|x-y\|}$ peut être approché par une somme dégénérée à variables séparées. On présentera au chapitre 3 une condition adaptée au noyau oscillant garantissant également une approximation dégénérée à variables séparées.

La condition d'admissibilité est un critère binaire. On désigne par $f(s,t)$ la fonction suivante, à valeurs dans l'ensemble $\{0,1\}$,

$$f : \mathscr{T}_\mathrm{I} \times \mathscr{T}_\mathrm{I} \mapsto \{0,1\} \tag{1.138}$$
$$(s,t) \mapsto f(s,t) = \begin{cases} 1 & \text{si s et t sont satisfont (1.137)} \\ 0 & \text{sinon} \end{cases}$$

1.3.3.3 Création de l'arbre des blocs

Grâce à la propriété de niveau constant et à la condition d'admissibilité (1.137), on peut construire un arbre de blocs avec des feuilles de grandes tailles représentant des interactions lointaines de rang faible. La procédure est récursive et conduit à une partition hiérarchique de la matrice de discrétisation.
Partant de la matrice complète décrite par $t \times s = \mathrm{I} \times \mathrm{I}$, on teste l'admissibilité (1.137) de tous les fils $t' \times s'$ de $t \times s$. En cas de succès, un fils admissible devient une feuille de l'arbre de blocs et le cas échéant, on applique la même procédure à ses fils de manière récursive. Les feuilles inadmissibles sont par construction de petites tailles, au plus de taille $n_{\max} \times n_{\max}$. L'algorithme (2) permet la création d'un arbre de blocs.

Algorithme 2 Construction d'un arbre de blocs

```
1: Fonction CreationArbreBlocs(t ∈ T_I, s ∈ T_I) :
2: Si f(t,s) = 1 ou S(t) = ∅ ou S(s) = ∅ Alors
3:    S(t × s) = ∅
4: Sinon
5:    S(t × s) = {t' × s' : t' ∈ S(t), s' ∈ S(s)}
6:    Pour t' × s' ∈ S(t × s) faire
7:       CreationArbreBlocs(t', s')
8:    Fin Pour
9: Fin Si
10: finFonction
```

Remarques On peut effectuer les remarques suivantes à propos de l'algorithme (2) :

— L'étape de récursion montre que l'arité de $\mathcal{T}_{\mathrm{I}\times\mathrm{I}}$ est le carré de celle de $\mathcal{T}_{\mathrm{I}}$;
— L'arbre de blocs $\mathcal{T}_{\mathrm{I}\times\mathrm{I}}$ n'est pas complet grâce à la condition d'admissibilité. C'est une caractéristique voulue pour obtenir des feuilles à un bas niveau de l'arbre (*i.e* de grandes dimensions). L'arbre est néanmoins entier ;
— Les feuilles non-admissibles sont de petite taille car $\min(|t|,|s|) \leqslant n_{\max}$;
— Selon la condition d'admissibilité, les feuilles peuvent être arbitrairement grandes.

1.3.4 L'ensemble des $\mathscr{H}$-matrices

L'arbre de blocs $\mathcal{T}_{\mathrm{I}\times\mathrm{I}}$ décrit une partition possible de la matrice de discrétisation A suivant des critères choisis par l'utilisateur d'une part et liés au noyau de Green (condition d'admissibilité). Il s'agit du squelette d'une matrice et l'on peut définir une $\mathscr{H}$-matrice de la sorte

Définition 1.25 ($\mathscr{H}$-matrice)**.** Soit A une matrice carrée de taille N×N. On note $\mathrm{I} = \{1,\dots,\mathrm{N}\}$ et $\mathcal{T}_{\mathrm{I}}$ l'arbre de groupes correspondant. On note également $\mathcal{T}_{\mathrm{I}\times\mathrm{I}}$ l'arbre des blocs sur $\mathcal{T}_{\mathrm{I}}$. Pour un entier r donné, l'ensemble

$$\mathscr{H}(\mathcal{T}_{\mathrm{I}\times\mathrm{I}},r) := \left\{\mathrm{A} \text{ de taille } \mathrm{N}\times\mathrm{N} \ : \mathrm{rg}(\mathrm{A}_{t\times s}) \leqslant r, \forall\ t\times s \in \mathscr{L}(\mathcal{T}_{\mathrm{I}\times\mathrm{I}})\right\}, \tag{1.139}$$

est l'ensemble des $\mathscr{H}$-matrices définies sur la partition $\mathcal{T}_{\mathrm{I}\times\mathrm{I}}$. Un élément $\mathrm{A}_{\mathscr{H}}$ de $\mathscr{H}(\mathcal{T}_{\mathrm{I}\times\mathrm{I}},r)$ est appelé une $\mathscr{H}$-matrice. C'est une approximation de A représentée hiérarchiquement. On parle également d'approximation compressée car les blocs $\mathrm{A}_{t\times s}$ admissibles de $\mathrm{A}_{\mathscr{H}}$ sont représentés par des matrices de rang faible.

Cette structure de représentation hiérarchique par blocs dépend de :
— La découpe (médiane, géométrique, par composantes principales,...) ;
— La taille de feuille maximale $n_{\max}$;
— La constante d'admissibilité η.

Par ailleurs le nombre de subdivisions effectuées sur les lignes et les colonnes sont égales au nombre de feuilles de $\mathcal{T}_{\mathrm{I}}$. On s'arrête lorsque les feuilles sont inférieures à la taille maximale précisée par $n_{\max}$. La structure de $\mathscr{H}$-matrice est composée majoritairement de blocs matriciels décrivant des interactions admissibles (lointaines).

Selon le même modèle qui a été utilisé dans les exemples en dimension deux, les feuilles admissibles sont représentées par des matrices de rang faible de la forme (1.30). Elles peuvent être approchées de deux façons différentes. Si l'on considère le bloc matriciel associé à une feuille $t\times s$, alors on peut utiliser une méthode d'approximation algébrique de la forme (1.48) pour obtenir la représentation de rang faible. L'autre possibilité est d'interpréter l'interaction $t\times s$ comme l'interaction de deux nuages de degrés de liberté et construire une approximation du noyau de la forme (1.27) et d'utiliser cette approximation pour assembler une représentation de rang faible directement à partir de l'expression des coefficients (1.12).

L'ensemble $\mathscr{H}(\mathcal{T}_{\mathrm{I}\times\mathrm{I}},r)$ n'est pas un espace vectoriel. En effet, le rang d'une somme de deux matrices étant borné par la somme des rangs des matrices, on a pour $\mathrm{A},\mathrm{B} \in \mathscr{H}(\mathcal{T}_{\mathrm{I}\times\mathrm{I}},r)$, $\mathrm{A}+\mathrm{B} \in \mathscr{H}(\mathcal{T}_{\mathrm{I}\times\mathrm{I}},2r)$. Cela n'est pas un obstacle à l'utilisation de la méthode car dans la pratique la borne n'est pas nécessairement atteinte. De plus, au lieu de fixer un rang maximal r, on préfère se donner une erreur ϵ et construire des approximations

satisfaisant l'erreur prescrite. Cela revient à considérer l'ensemble $\mathscr{H}(\mathscr{T}_{I\times I},\epsilon)$. Sauf mention contraire, lorsque l'on considère une approximation $\mathscr{H}$-matrice, on considère qu'il s'agit d'un élément de $\mathscr{H}(\mathscr{T}_{I\times I},\epsilon)$.

L'avantage majeur de cette méthode d'approximation est que l'on dispose d'opérations algébriques sur les $\mathscr{H}$-matrices. Sous réserve de travailler avec des partitions de blocs compatibles, on peut les sommer et les multiplier entre elles.

Dans le cas où la matrice est inversible, la résolution d'un système linéaire de la forme

$$A_{\mathscr{H}}\cdot x = b, \tag{1.140}$$

peut s'effectuer en construisant une approximation $\mathscr{H}$-matrice de l'inverse A^{-1} ou à l'aide d'une factorisation A = LU ce qui permet d'écrire le système sous la forme

$$L_{\mathscr{H}}U_{\mathscr{H}}\cdot x = b, \tag{1.141}$$

où les matrices $L_{\mathscr{H}}$ et $U_{\mathscr{H}}$ sont respectivement triangulaires inférieure et supérieure. On est alors amené à résoudre successivement deux systèmes triangulaires. Cela est particulièrement adapté à la résolution de systèmes linéaires pour un grand nombre de seconds membres.

1.3.5 Estimations liées à l'assemblage d'une $\mathscr{H}$-matrice

Les points précédemment développés ont montré qu'il est possible d'assembler une approximation hiérarchique et compressée d'une matrice à partir de matrices de rang faible. À l'échelle d'un bloc admissible, l'utilisation de matrices de rang faible permet de réduire la quantité de mémoire requise pour stocker la matrice $A_{\mathscr{H}}$. On s'intéresse dans ce paragraphe à des estimations *a priori* sur la mémoire totale nécessaire à ce stockage.
On se limitera cependant au cas du noyau de Laplace dans ce paragraphe et l'on soulignera le point de l'analyse où cette limitation devient nécessaire. Dans ce cas du noyau de Laplace, ces estimations ont été vérifiées dans la pratique à plusieurs reprises sur des cas de nature diverses ([Beb00], [BGH12],[Liz14]).

1.3.5.1 Constante de rareté

Une mesure de la complexité de la partition en blocs décrite par l'arbre de blocs $\mathscr{T}_{I\times J}$ est donnée par la constante dite de rareté (*sparsity constant* en anglais dans la littérature, [GH03]). Cette quantité est l'analogue de la constante que l'on retrouve lors de la manipulation de matrices creuses pour lesquelles on s'intéresse au nombre maximal de termes non nuls par ligne et par colonne. Dans le contexte d'une découpe hiérarchique par blocs, cette rareté décrit le nombre maximum de blocs dans la partition associés à un groupe donné. Les estimations *a priori* sur les $\mathscr{H}$-matrices développées dans la littérature l'ont été sans cette notion ([Hac99; HK00]). Cependant, la constante de rareté permet de démontrer des résultats plus généraux et nécessitant moins d'hypothèses. On renvoie à [Beb00] pour un exposé détaillé de cette notion.

Définition 1.26 (Constante de rareté)**.** Soient $\mathscr{T}_I$ et $\mathscr{T}_J$ des arbres de groupes associés aux indices I et J respectivement. On note $\mathscr{T}_{I\times J}$ l'arbre de blocs associé à ces arbres de groupes et $t\times s$ désigne un élément de cet arbre de blocs.
Le nombre de blocs $t\times s\in\mathscr{T}_{I\times J}$ associés à l'arbre de groupes $T_{\mathscr{J}}$ est donné par

$$c^l(\mathscr{T}_{I\times J},t) = \mathrm{Card}\left(\{s\subset J : t\times s\in\mathscr{T}_{I\times J}\}\right). \tag{1.142}$$

De même, pour un groupe $s \in \mathscr{T}_\mathrm{J}$, on définit

$$c^c(\mathscr{T}_{\mathrm{I}\times\mathrm{J}}, s) = \mathrm{Card}\left(\left\{t \subset \mathrm{I} : t \times s \in \mathscr{T}_{\mathrm{I}\times\mathrm{J}}\right\}\right), \tag{1.143}$$

qui représente le nombre de blocs $t \times s$ associés à l'arbre de groupes $\mathscr{T}_\mathrm{J}$.

La **constante de rareté**, notée c_{ra}, d'un arbre de blocs est définie par

$$c_{\mathrm{ra}} = \max\left\{\max_{t\in\mathscr{T}_\mathrm{I}} c^l(\mathscr{T}_{\mathrm{I}\times\mathrm{J}}, t), \max_{s\in\mathscr{T}_\mathrm{J}} c^c(\mathscr{T}_{\mathrm{I}\times\mathrm{J}}, s)\right\} \tag{1.144}$$

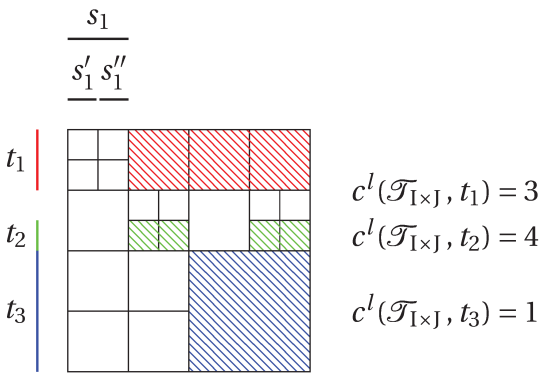


FIGURE 1.11 – Illustration graphique de la constante de rareté .

La figure (1.11) illustre la constante de rareté. Pour les trois groupes t_1, t_2 et t_3 de $\mathscr{T}_\mathrm{I}$, on détermine les quantités $c^l(\mathscr{T}_{\mathrm{I}\times\mathrm{J}}, t_1)$,$c^l(\mathscr{T}_{\mathrm{I}\times\mathrm{J}}, t_1)$ et $c^l(\mathscr{T}_{\mathrm{I}\times\mathrm{J}}, t_1)$ correspondantes à l'aide de la définition (1.26). Sur cet exemple, on constate que $\max_{t\in\mathscr{T}_\mathrm{I}} c^l(\mathscr{T}_{\mathrm{I}\times\mathrm{J}}, t) = 4$ (par exemple pour $t = t_2$). Le même travail sur l'arbre $\mathscr{T}_\mathrm{J}$ associé aux colonnes de la matrice donne $\max_{s\in\mathscr{T}_\mathrm{J}} c^c(\mathscr{T}_{\mathrm{I}\times\mathrm{J}}, s) = 3$. On remarquera dans ce cas que le nombre de blocs associés au groupe s_1 est plus élevé ($c^c(\mathscr{T}_{\mathrm{I}\times\mathrm{J}}, s) = 3$) que les nombres de blocs associés à ses fils s_1' et s_1'' ($c^c(\mathscr{T}_{\mathrm{I}\times\mathrm{J}}, s_1') = c^c(\mathscr{T}_{\mathrm{I}\times\mathrm{J}}, s_1'') = 2$). En effet, la constante de rareté c_{ra} n'est pas uniquement déterminée par les petits blocs. Pour cet exemple, on a $c_{\mathrm{ra}} = 4$.

Exemple d'utilisation de c_{ra} Beaucoup d'algorithmes sur les $\mathscr{H}$-matrices peuvent être appliqués par blocs. C'est par exemple le cas du calcul de la norme de Frobénius ou du nombre de coefficients stockés en mémoire. Dans ce cas, le coût de l'algorithme est alors la somme des coûts de chaque bloc.

Cas d'un coût borné Supposons que pour chaque bloc le coût est borné par une quantité $c > 0$. On rappelle que l'on désigne par $\mathscr{L}(\mathscr{T}_{\mathrm{I}\times\mathrm{J}})$ l'ensemble des feuilles de $\mathscr{T}_{\mathrm{I}\times\mathrm{J}}$. Le coût total est ainsi la somme des coûts des feuilles. On a alors

$$\sum_{t\times s\in\mathscr{L}(\mathscr{T}_{\mathrm{I}\times\mathrm{J}})} c = \sum_{t\in\mathscr{T}_\mathrm{I}} \sum_{s\in\mathscr{T}_\mathrm{J} \,:\, t\times s\in\mathscr{L}(\mathscr{T}_{\mathrm{I}\times\mathrm{J}})} c \tag{1.145}$$

$$= c \sum_{t\in\mathscr{T}_\mathrm{I}} \mathrm{Card}\left(\left\{s \in \mathrm{J} \,:\, t \times s \in \mathscr{T}_{\mathrm{I}\times\mathrm{J}}\right\}\right) \tag{1.146}$$

$$= c \sum_{t\in\mathscr{T}_\mathrm{I}} c^l \qquad \text{d'après (1.142)} \tag{1.147}$$

$$\leqslant c\, c^l\, |\mathrm{I}| \tag{1.148}$$

$$\leqslant c\, c_{\mathrm{ra}}\, |\mathrm{I}| \tag{1.149}$$

La majoration (1.148) se déduit du fait que le nombre d'éléments d'une partition d'un ensemble fini est inférieur au cardinal de cet ensemble.
De même, en échangeant les rôles des groupes t et s, on a

$$\sum_{t\times s\in\mathscr{L}(\mathcal{T}_{\mathrm{I}\times\mathrm{J}})} c \leqslant c\, c_{\mathrm{ra}}\, |\mathrm{J}|. \tag{1.150}$$

Cas d'un coût borné linéairement par la taille des groupes Dans le cas où le coût est borné linéairement avec la taille des blocs comme $c(|t|+|s|)$, on a

$$\begin{aligned}
\sum_{t\times s\in\mathscr{L}(\mathcal{T}_{\mathrm{I}\times\mathrm{J}})} c(|t|+|s|) &= c \sum_{t\in\mathcal{T}_{\mathrm{I}}} \sum_{\{s\in\mathcal{T}_{\mathrm{J}}\,:\, t\times s\in\mathscr{L}(\mathcal{T}_{\mathrm{I}\times\mathrm{J}})\}} |t| && (1.151)\\
&\quad + c \sum_{s\in\mathcal{T}_{\mathrm{J}}} \sum_{\{t\in\mathcal{T}_{\mathrm{I}}\,:\, t\times s\in\mathscr{L}(\mathcal{T}_{\mathrm{I}\times\mathrm{J}})\}} |s| && (1.152)\\
&\leqslant c\, c_{\mathrm{ra}} \left(\sum_{t\in\mathcal{T}_{\mathrm{I}}'} |t| + \sum_{s\in\mathcal{T}_{\mathrm{J}}'} |s| \right) && (1.153)\\
&\leqslant c\, c_{\mathrm{ra}} \mathrm{L}(\mathcal{T}_{\mathrm{I}\times\mathrm{J}}) \left(|\mathcal{T}_{\mathrm{I}}| + |\mathcal{T}_{\mathrm{J}}| \right) \qquad \text{d'après(1.22).} && (1.154)
\end{aligned}$$

$\mathcal{T}_{\mathrm{I}}'$ (resp. $\mathcal{T}_{\mathrm{J}}'$) est un sous-arbre de $\mathcal{T}_{\mathrm{I}}$ (resp. $\mathcal{T}_{\mathrm{J}}$) utilisé pour la construction de $\mathcal{T}_{\mathrm{I}\times\mathrm{J}}$ par (on obtient $\mathcal{T}_{\mathrm{J}}$' en échangeant les rôles de I et J ainsi que t et s)

$$\mathcal{T}_{\mathrm{I}}' = \left\{ t\in\mathcal{T}_{\mathrm{I}} : \exists t'\subset t \text{ et } s'\subset\mathcal{T}_{\mathrm{J}} : t'\times s'\in\mathcal{T}_{\mathrm{I}\times\mathrm{J}} \right\}. \tag{1.155}$$

1.3.5.2 Coût en mémoire d'une $\mathscr{H}$-matrice

Position du problème On note $\mathrm{A}_{t\times s}$ une feuille de la $\mathscr{H}$-matrice $\mathrm{A}_{\mathscr{H}}$. Cette feuille représente un bloc matriciel de taille $|t|\times|s|$. Ces blocs peuvent être de deux types distincts :

— Une matrice pleine, soit $|t|\times|s|$ coefficients ;

— Une matrice de rang faible de rang r, soit $r(|t|+|s|)$ coefficients.

L'estimation de la mémoire nécessaire au stockage d'une $\mathscr{H}$-matrice revient à fournir une estimation du nombre total de coefficients de la $\mathscr{H}$-matrice. Il s'agit simplement de la somme des coefficients sur les feuilles et l'on peut simplement employer l'estimation (1.154). Il est donc nécessaire de trouver une borne linéaire en $|t|+|s|$ pour chaque feuille.

Hypothèse sur le rang Dans le cas d'un bloc de rang faible, on travaille dans la pratique avec une précision donnée et le rang est un résultat de l'approximation de rang faible effectuée. Pour effectuer les estimations suivantes, il est nécessaire de se donner un rang maximal $r_{\max}$ qui n'est dépassé par aucune feuille admissible. Dans le cas du noyau de l'équation de Laplace, on peut utiliser l'estimation suivante, similaire à (1.47),

$$r = \mathscr{O}(|\log(\epsilon)|^2). \tag{1.156}$$

Ce résultat sera justifié au chapitre 2 à travers une présentation de différentes méthodes d'approximation de rang faible.
Dans le cas du noyau oscillant, le noyau de Green de l'équation de Helmholtz, cette hypothèse n'est pas correcte. En effet, une borne maximale trop grande serait inefficace pour une estimation pertinente et le rang dépend généralement de la fréquence d'étude. À haute fréquence, on s'attend à obtenir un rang plus élevé. Cette observation est justifiée

par le simple fait qu'à haute fréquence, le noyau oscille de plus en plus et qu'il est nécessaire de « capter » ces oscillations. C'est l'objet de ce manuscrit que d'étudier les variations du rang d'une approximation de rang faible en fonction de la fréquence et l'on abordera cette étude au chapitre 3. La complexité générale de la méthode des $\mathscr{H}$-matrices repose sur le comportement du noyau de Green ainsi que sur la structure de l'arbre (*c.f* BebendorfOscillant). Dans la suite de ce chapitre, on suppose pour toute feuille admissible (*i.e* de rang faible) $\mathrm{A}_{t\times s}$ que l'on vérifie l'estimation suivante,

$$\mathrm{rang}(\mathrm{A}_{t\times s}) \leqslant r_{\max}, \tag{1.157}$$

où $r_{\max}$ vérifie (1.156) pour une précision ϵ fixée. On souhaite majorer le nombre de coefficients des blocs $\mathrm{A}_{t\times s}$ pleins en fonction de $(|t|+|s|)$ comme dans le cas des blocs de rang faible. Les blocs denses sont petits et vérifient par construction des arbres de groupes,

$$\min(|t|,|s|) \leqslant n_{\max}. \tag{1.158}$$

Ainsi, on a immédiatement

$$\begin{aligned} |t|\cdot|s| &= \min(|t|,|s|)\max(|t|,|s|) && (1.159)\\ &\leqslant n_{\max}(|t|+|s|) && (1.160)\end{aligned}$$

Dans tous les cas, en notant $c_{t\times s}$ le nombre de coefficients d'un bloc $\mathrm{A}_{t\times s}$, on a

$$\begin{aligned} c_{t\times s} &\leqslant \max\{n_{\max}, r_{\max}\}(|t|+|s|),\\ &\leqslant c_0(|t|+|s|), && (1.161)\end{aligned}$$

où $c_0 := \max\{n_{\max}, r_{\max}\}$. On peut alors donner une estimation de la mémoire totale de la matrice $\mathrm{A}_{\mathscr{H}}$ grâce à l'inégalité (1.161) ainsi que l'estimation (1.154) (en prenant $c = c_0$)

Proposition 1.27 (Espace mémoire d'une $\mathscr{H}$-matrice,[Beb00])**.** *Soit $\mathrm{T}_{\mathscr{I}\times\mathscr{J}}$ un arbre de blocs associé à l'arbre de groupe $\mathrm{T}_{\mathscr{I}}$. On note $\mathscr{C}_{\mathrm{COEF}}(\mathrm{A}_{\mathscr{H}})$, le nombre de coefficients de la $\mathscr{H}$-matrice $\mathrm{A}_{\mathscr{H}}$. Alors,*

$$\mathscr{C}_{\mathrm{COEF}}(\mathrm{A}_{\mathscr{H}}) \leqslant c_{\mathrm{ra}} c_0 \left(\mathrm{L}(\mathscr{T}_{\mathrm{I}})\#\mathscr{T}_{\mathrm{I}} + \mathrm{L}(\mathscr{T}_{\mathrm{J}})\#\mathscr{T}_{\mathrm{J}}\right). \tag{1.162}$$

Pour des réels stockés en double précision, la mémoire requise pour la $\mathscr{H}$-matrice est de $8\mathscr{C}_{\mathrm{COEF}}(\mathrm{A}_{\mathscr{H}})$ octets.

1.3.5.3 Norme d'une $\mathscr{H}$-matrice

L'emploi de matrices dans une méthode numérique va souvent de pair avec le calcul de sa norme. Plus particulièrement lorsque celle-ci est une approximation d'une autre matrice, il est alors intéressant de pouvoir mesurer l'erreur commise.
La construction d'une $\mathscr{H}$-matrice s'effectue bloc par bloc et dans la pratique, on possède une bonne mesure (voire optimale si la SVD est disponible) à l'échelle d'une feuille de $\mathrm{T}_{\mathscr{I}\times\mathscr{J}}$ et l'on cherche à observer l'erreur sur la matrice entière. La constante de rareté définie au paragraphe précédent permet d'écrire des estimations globales.

Deux normes sont courament utilisées, la norme de Frobénius ainsi que la norme spectrale (aussi appelée *norme 2*). Pour une matrice quelconque A de taille $\mathrm{N}\times\mathrm{N}$, la norme de Frobénius notée $\|\mathrm{A}\|_{\mathrm{F}}$ est définie par

$$\|\mathrm{A}\|_{\mathrm{F}}^2 = \sum_{i=1}^{\mathrm{N}}\sum_{j=1}^{\mathrm{N}} |\mathrm{A}_{ij}|^2, \tag{1.163}$$

tandis que la norme spectrale $\|A\|_2$ est caractérisée par

$$\|A\|_2^2 = \max_{x\in\mathbb{C}^N, x\neq 0} \frac{\|Ax\|_2}{\|x\|_2}. \tag{1.164}$$

La norme de Frobénius en tant que somme de carrés de modules peut être décomposées en la somme sur les blocs d'une partition P donnée. La norme spectrale ne se prête pas facilement à ce genre de manipulation mais l'on peut néanmoins utiliser la norme spectrale d'un sous-bloc pour la matrice complète. Dans le cas d'une subdivision hiérarchique en matrice blocs 2×2, on peut montrer le résultat suivant issu de [Beb00],

Théorème 1.28. *On considère une partition* P *de l'ensemble* $I\times J$ *obtenue en partitionnant chaque bloc en* 2×2 *matrices au plus* L *fois. On suppose par ailleurs que pour deux matrices* A *et* B*, les blocs* A_b *et* B_b *pour* $b\in P$ *vérifient*

$$\|A_b\|_2 \leqslant \|B_b\|_2. \tag{1.165}$$

Alors, on a

$$\|A\|_2 \leqslant 2^L \|B\|_2. \tag{1.166}$$

Dans la pratique, $L = \min(|\mathcal{T}_I|, |\mathcal{T}_J|)$ est un choix fréquent et alors le coefficient 2^L est de l'ordre de $\min(|I|, |J|)$. Le théorème précédent relie l'erreur sur un bloc à l'erreur globale. La borne 2^L est par ailleurs une borne pessimiste et l'on peut s'attendre à une borne plus faible dans le cas où la partition P est plus fine. C'est typiquement le cas de la partition constituée par les feuilles d'un arbre de blocs d'une $\mathscr{H}$-matrice.
Le théorème suivant [Beb00] fournit une estimation de l'erreur en norme 2 d'une $\mathscr{H}$-matrice en fonction des erreurs commises à l'échelle des feuilles.

Théorème 1.29. *Soit* $P = \mathscr{L}(\mathcal{T}_{I\times J})$ *l'ensemble des feuilles d'un arbre de blocs* $\mathcal{T}_{I\times J}$*. Pour* A *et* B *deux* $\mathscr{H}$*-matrices définies à partir de cet arbre, on a*

— *La norme spectrale de* A *est liée au bloc de plus grande norme par*

$$\max_{b\in P}\|A_b\|_2 \leqslant \|A\|_2 \leqslant c_{ra} L(\mathcal{T}_{I\times J}) \max_{b\in P}\|A_b\|_2 \tag{1.167}$$

— *Si* $\max_{b\in P}\|A_b\|_2 \leqslant \max_{b\in P}\|A_b\|_2$,

$$\|A\|_2 \leqslant c_{ra} L(\mathcal{T}_{I\times J}) \|B\|_2 \tag{1.168}$$

Ce théorème peut permettre d'évaluer une erreur relative sur la matrice complète à partir de la connaissance des blocs en particulier grâce à (1.167).

Calcul approché de normes Le calcul de la norme de Frobénius est lui aisé car il s'agit d'une quantité additive qui peut être calculée de manière similaire à (1.154) avec au plus $c_{ra}\max\{r_{max}^2, n_{max}\}(|I|\log(|I|) + |J|\log(|J|)$.
La norme spectrale peut être calculée avec une complexité log-linéaire par une méthode de puissances itérées appliquées à $A^H A$ (A^H étant la transposée de la matrice conjuguée) à l'aide d'un produit matrice-vecteur, le produit $A^H A$ n'étant jamais effectué en tant que tel.

1.3.6 Opérations sur les $\mathscr{H}$-matrices

L'intérêt principal de la méthode des $\mathscr{H}$-matrices est la possibilité d'effectuer des opérations algébriques sur ces dernières et d'obtenir un résultat de même nature.

Principe de base Le principe de base à partir duquel on construit des méthodes sur les $\mathscr{H}$-matrices est la découpe par blocs en matrice 2×2 suivante,

$$\mathrm{A} = \begin{bmatrix} \mathrm{A}_{11} & \mathrm{A}_{12} \\ \mathrm{A}_{21} & \mathrm{A}_{22} \end{bmatrix} \tag{1.169}$$

1.3.6.1 Produit matrice-vecteur

Le produit matrice-vecteur est une opération d'une très grande importance pour des solveurs itératifs. À partir d'une approximation x_0, la résolution d'un système linéaire $\mathrm{A}x = b$ par une méthode itérative construit des approximations x_n pour $n = 1, \ldots, \mathrm{N}_{iter}$ de la solution à partir de produits matrice-vecteurs par la matrice A. Le nombre $\mathrm{N}_{\mathrm{iter}}$ d'itérations requises afin d'obtenir la convergence peut parfois être élevé et il est bon d'avoir un produit matrice-vecteur rapide. Dans le cas matrices pleines, cette opération est d'une complexité quadratique avec la taille N de la matrice. Par « rapide », on souhaite une complexité de l'ordre de $\mathscr{O}(\mathrm{N}\log(\mathrm{N}))$ (idéalement $\mathscr{O}(\mathrm{N})$).

Dans le cas des $\mathscr{H}$-matrices, une fois que l'on a construit l'arbre de blocs décrivant la découpe hiérarchique de la matrice, le produit par un veceur est une opération très simple à effectuer. L'avantage de cette opération provient du produit rapide d'une matrice de rang faible par un vecteur comme décrit par l'expression (1.52).

Remarque 1.30 (Renumérotation). On rappelle par ailleurs que les degrés de liberté ont été renumérotés lors de l'étape de construction de l'arbre de groupes. L'approximation $\mathscr{H}$-matrice $\mathrm{A}_{\mathscr{H}}$ d'une matrice A est exprimée naturellement à l'aide de cette renumérotation et il convient d'effectuer la bonne permutation du vecteur x au préalable. Ainsi, le résultat est également obtenu avec cette numérotation et l'on peut obtenir le résutat dans la numérotation originale en utilisant la permutation inverse.

Algorithme du produit matrice-vecteur L'algorithme (3) décrit le calcul récursif du produit matrice-vecteur $y = \mathrm{A}x$ d'une $\mathscr{H}$-matrice et d'un vecteur.

Algorithme 3 Produit matrice-vecteur entre une $\mathscr{H}$-matrice et un vecteur

Données: Une $\mathscr{H}$-matrice A découpée en blocs par $\mathscr{T}_{\mathrm{I}\times\mathrm{I}}$ et un vecteur x de taille N.
But: Calculer un vecteur y tel que $y = \mathrm{A}x$, y étant initialisé à 0.

1: **Fonction MatriceVecteur**($\mathrm{A}_{t\times s}$, x_s, y_t) :
2: **Si** $t \times s$ est une feuille de $\mathscr{T}_{\mathrm{I}\times\mathrm{I}}$ **Alors**
3: $\quad y_t \leftarrow y_t + \mathrm{A}_{t\times s}.x_s$
4: **Sinon**
5: $\quad$ **Pour** $t' \times s' \in \mathrm{S}(t \times s)$ **faire**
6: $\quad\quad$ **MatriceVecteur**($\mathrm{A}_{t'\times s'}$, $x_{s'}$,$y_{t'}$)
7: $\quad$ **Fin Pour**
8: **Fin Si**
9: **finFonction**

Comme pour l'assemblage d'une $\mathscr{H}$-matrice, on considère le cas d'une matrice carrée construite à l'aide de l'arbre de blocs $\mathscr{T}_{\mathrm{I}\times\mathrm{I}}$. Pour un vecteur x de taille N, on forme le produit $y = \mathrm{A}x$ à l'aide de la structure hiérarchique de la matrice A. Pour un groupe $t \in \mathscr{T}_{\mathrm{I}}$,

le vecteur y_t désigne le sous-vecteur de taille $|t|$ défini par

$$(y_t)_i = (y)_{t_i}, \tag{1.170}$$

avec pour $i = 1,\ldots,|t|$, $t_i \in t$. On définit de manière équivalente le vecteur x_s.
On n'effectue les produits qu'au niveau des feuilles de l'arbre $\mathscr{T}_{\mathrm{I}\times\mathrm{I}}$. Pour $s, t \in \mathscr{T}_{\mathrm{I}}$, on a la caractérisation suivante,

$$y_t = \sum_{t\times s\in\mathscr{L}(\mathscr{T}_{\mathrm{I}\times\mathrm{I}})} \mathrm{A}_{t\times s}.x_s \tag{1.171}$$

Dans le cas où $\mathrm{A}_{t\times s}$ est une matrice de rang faible, on effectue le produit par le vecteur x_s d'après (1.52) tandis que le cas plein est effectué de façon usuelle avec une complexité quadratique mais sur un bloc de dimensions de l'ordre de $n_{\max}\times n_{\max}$.

Complexité du produit matrice-vecteur La proposition suivante donne l'estimation de la complexité du produit matrice-vecteur. On se limite à une matrice carrée construite sur l'arbre $\mathscr{T}_{\mathrm{I}\times\mathrm{I}}$.

Proposition 1.31 (Produit matrice-vecteur)**.** *Soit $\mathscr{T}_{\mathrm{I}\times\mathrm{I}}$ un arbre de blocs associé à l'arbre de groupe $\mathscr{T}_{\mathrm{I}}$ pour l'ensemble $\mathrm{I}=\{1,\ldots,\mathrm{N}\}$. On note $\mathscr{C}_{\mathrm{MV}}(\mathrm{A}_{\mathscr{H}})$, le nombre d'opérations nécessaires pour effectuer le produit $\mathrm{A}_{\mathscr{H}}.x$. Alors, on a*

$$\mathscr{C}_{\mathrm{MV}}(\mathrm{A}_{\mathscr{H}}) \leqslant 2.\mathscr{C}_{\mathrm{COEFF}}(\mathrm{A}_{\mathscr{H}}), \tag{1.172}$$

où $\mathscr{C}_{\mathrm{COEFF}}(\mathrm{A}_{\mathscr{H}})$ est le nombre de coefficients de la $\mathscr{H}$-matrice $\mathrm{A}_{\mathscr{H}}$ défini par la proposition 1.27.

Il s'agit d'une complexité similaire à celle de l'assemblage d'une $\mathscr{H}$-matrice ce qui est attendu car le coefficient d'une matrice n'intervient qu'une seule fois lors d'un produit matrice-vecteur que ce soit au format $\mathscr{H}$-matrice ou au format usuel. On dispose alors d'un produit matrice-vecteur rapide ce qui permet de construire un solveur itératif pour lequel le produit matrice-vecteur est accéléré par l'approximation $\mathscr{H}$-matrice.

1.3.6.2 Addition de $\mathscr{H}$-matrices

La multiplication d'une matrice par un vecteur est une opération de type BLAS – 2, c'est-à-dire une opération entre une matrice et un vecteur. Les opérations d'un niveau supérieur sont celles de type BLAS – 3 qui font intervenir des oppérations entre matrices.

L'opération la plus simple entre deux $\mathscr{H}$-matrices est d'en effectuer la somme. Pour cela, il est nécessaire qu'elles soient formées d'après le même arbre de blocs. On peut trouver dans la littérature (notament [Liz14]) des algorithmes permettant la somme de matrices dont les structures sont légèrement différentes à l'aide de méthode de conversion hiérarchique basées sur l'opération d'agglomération décrite au paragraphe (1.2.3). Nous ne détaillons pas ces méthodes dans la suite et l'on renvoie à [Liz14; BGH12; Beb00].

Pour $\mathrm{A}\in\mathscr{H}(\mathscr{T}_{\mathrm{I}\times\mathrm{J}})$ et $\mathrm{B}\in\mathscr{H}(\mathscr{T}_{\mathrm{I}\times\mathrm{J}})$, on note C = A+B. Les matrices A et B étant construites sur le même arbre de blocs, la somme $\mathrm{C}_{t\times s}$ de deux blocs $\mathrm{A}_{t\times s}$ et $\mathrm{B}_{t\times s}$ peut prendre trois formes différentes :

— Le bloc $t\times s$ est une matrice de rang faible ;

— Le bloc $t \times s$ est une matrice dense ;

— Le bloc $t \times s$ n'est pas une feuille et est donc subdivisé.

Dans le cas de matrices denses, on somme les matrices élément par élément de façon standard. Notons que le fait de sommer des blocs matriciels est une opération BLAS − 3 particulièrement bien adaptée à la parallélisation et se traite simplement dans la pratique. Le cas des matrices de rang faible se traite à partir de l'opération de sommation de matrices de rang faible décrite au paragraphe 1.2.2.3 consistant à obtenir C sous sa forme compressée $U_C V_C^T$. La concaténation formée d'après les matrices de rang faible $U_A V_A^T$ et $U_B V_B^T$ est exprimée sous la forme d'une décomposition en valeurs singulières approchée

$$U_C V_C^T = \mathscr{U} \Sigma \mathscr{V}^T. \tag{1.173}$$

La troncature de la matrice Σ fournit une matrice compressée d'un rang k fixé ou à une erreur ϵ donnée. On note $A \oplus B$, la somme des ces deux matrices.

Algorithme AXPY L'algorithme (4) suivant décrit l'opération similaire

$$A \leftarrow A + \alpha B \tag{1.174}$$

où α est un scalaire. Il s'agit de l'opération AXPY de la librairie BLAS.

Algorithme 4 Algorithme d'AXPY pour des $\mathscr{H}$-matrices

```
1: Fonction AXPY (A_{t×s} , B_{t×s}, α ) :
2: Si t × s est une feuille pleine de T_{I×J} Alors
3:    A_{t×s} ← A_{t×s} + αB_{t×s}
4: Sinon Si t × s est une feuille compressée de T_{I×J} Alors
5:    A_{t×s} ← A_{t×s} ⊕ αB_{t×s}
6: Sinon
7:    Pour t' × s' ∈ S(t × s) faire
8:       AXPY (A_{t'×s'},B_{t'×s'} ,α)
9:    Fin Pour
10: Fin Si
11: finFonction
```

L'avantage de cette opération est que l'on connait facilement la structure du résultat avant le calcul car les matrices sont toutes construites sur le même arbre de groupes. De plus, le nombre limité de configurations possibles permet de fournir une estimation de la complexité avec la même méthode que celle employée pour l'analyse de l'espace mémoire d'une $\mathscr{H}$-matrice.

Complexité La complexité de l'addition de deux $\mathscr{H}$-matrices peut être obtenue de la même façon que l'assemblage. Notons $\mathscr{C}_{SM}(A,B)$ le nombre d'opérations nécessaires à la formation de la somme $A+B$. Comme pour l'assemblage, les briques élémentaires sont les opérations sur les feuilles de $\mathscr{T}_{I\times J}$.

— Le cas $A_{t\times s} + B_{t\times s}$ dense nécessite $|t| \cdot |s|$ additions (c'est le nombre de termes) et ainsi on peut écrire l'inégalité,

$$|t| \cdot |s| \leqslant n_{\max}(|t| + |s|). \tag{1.175}$$

— Pour le cas compressé, en supposant que $A_{t\times s}$ et $B_{t\times s}$ aient un rang commun fixe r, le coût de leur somme compressée est de l'ordre de $24r^2(|t|+|s|)$ d'après (1.12).

À l'aide de (1.161) on peut déterminer la complexité de la somme de deux $\mathscr{H}$-matrices,

Proposition 1.32 (Somme de $\mathscr{H}$-matrices)**.** *Soit $\mathscr{T}_{I\times I}$ un arbre de blocs associé à l'arbre de groupe $\mathscr{T}_I$. On note $\mathscr{C}_{SM}(A_{\mathscr{H}},B_{\mathscr{H}})$, le coût de la somme de deux $\mathscr{H}$-matrices. Alors,*

$$\mathscr{C}_{SM}(A_{\mathscr{H}},B_{\mathscr{H}}) \leqslant 2c_{ra}\, c_0 L(\mathscr{T}_I)|\mathscr{T}_I| \tag{1.176}$$

avec $c_0 = \max\{n_{max}, 24r_{max}^2\}$.

1.3.6.3 Multiplication de $\mathscr{H}$-matrices

La multiplication de deux $\mathscr{H}$-matrices est une autre opération de type BLAS − 3 que l'on peut réaliser avec des $\mathscr{H}$-matrices. Si la somme est une opération relativement aisée à implémenter, la multiplication de deux $\mathscr{H}$-matrices n'est pas une opération triviale et son implémentation n'est pas aussi triviale que les opérations précédentes. La référence [Beb00] contient un exposé détaillé de la multiplication de $\mathscr{H}$-matrices tandis que [Liz14] en fournit une implémentation robuste et parallèle. Dans la suite, nous nous limiterons à présenter les points difficiles et reproduire les estimations théoriques de la littérature. Cette méthode n'est par la suite pas requise pour les développements effectués sur le noyau de Helmholtz.

Structure du résultat Contrairement à la somme de deux $\mathscr{H}$-matrices, le produit de deux $\mathscr{H}$-matrices nécessite d'introduire un nouvel arbre de blocs. En effet, dans le cas où l'on considère $A \in \mathscr{H}(\mathscr{T}_{I\times J})$ et $B \in \mathscr{H}(\mathscr{T}_{J\times K})$, la compatibilité entre les colonnes de A et les lignes de B implique que le produit de ces deux matrices est une matrice associée au produit des deux arbres de groupes $\mathscr{T}_I$ et $\mathscr{T}_K$. Il est donc nécessaire de disposer d'un arbre de blocs dont la racine est $\mathscr{T}_I \times \mathscr{T}_K$. On définit alors l'**arbre produit** $\mathscr{T}_{IJK}$ de la façon suivante [Beb00],

Définition 1.33 (Arbre produit)**.** Soient $\mathscr{T}_{I\times J}$ et $\mathscr{T}_{J\times K}$ deux arbres de blocs associés respectivement aux arbres de groupes $\mathscr{T}_I$ et $\mathscr{T}_J$ d'une part et $\mathscr{T}_J$ et $\mathscr{T}_K$ d'autre part. On définit l'arbre produit de $\mathscr{T}_{I\times J}$ et $\mathscr{T}_{J\times K}$, noté $\mathscr{T}_{IJK}$, par

1. racine$(\mathscr{T}_{IJK}) = I \times K$.
2. Au niveau $(l \geqslant 0)$ de $\mathscr{T}_{IJK}$, l'ensemble $S(t\times s)$ des fils de $t\times s$ est défini par

$$S(t\times s) = \left\{ t'\times s' \,:\, \exists r \in \mathscr{T}_J{}^{(l)}, r' \in \mathscr{T}_J{}^{(l+1)}, t'\times r' \in S_{I\times J}(t\times r) \text{ et } r'\times s' \in S_{J\times K}(r\times s) \right\} \tag{1.177}$$

On illustre -partiellement- cette définition à l'aide d'un schéma. On considère pour plus de simplicité, le cas suivant souvent présenté dans la littérature.

Trois arbres de groupes sont mis en jeu lors de cette opération, $\mathscr{T}_I$, $\mathscr{T}_J$ et $\mathscr{T}_K$.

Exemple de produit Pour les deux matrices de la figure, on détermine l'arbre de blocs produit $\mathscr{T}_{IJK}$. On commence par appliquer la définition pour le niveau $l = 0$ et l'on détermine les fils de $t\times s$. D'après le schéma des arbres de groupes, on a les partitions suivantes $t = t'\cup t'', r = r'\cup r''$ et $s = s'\cup s''$. L'ensemble des fils de $t\times s$ est constitué de

— $t'\times s'$ en considérant le fils r' dans (1.177).

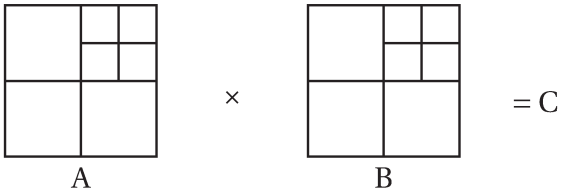


FIGURE 1.12 – Exemple de produit de matrices C = AB.

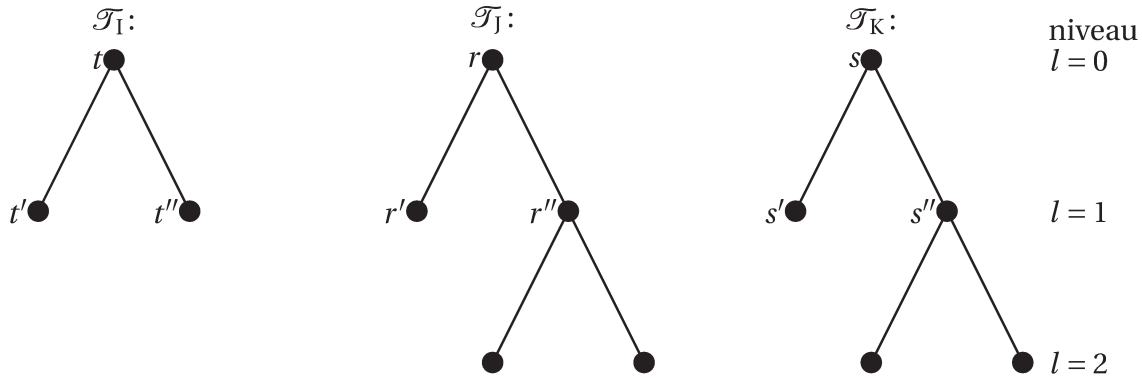


FIGURE 1.13 – Arbres de groupes mis en jeu dans le produit C = AB.

— $t' \times s''$ en considérant le fils r' dans (1.177).
— $t'' \times s'$ en considérant le fils r' dans (1.177).
— $t'' \times s''$ en considérant le fils r'' dans (1.177).

Le premier niveau de l'arbre de blocs $\mathcal{T}_{IJK}$ est représenté par la figure (1.14).

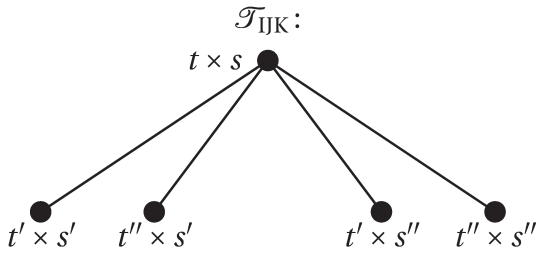


FIGURE 1.14 – Arbre produit $\mathcal{T}_{IJK}$.

Pour le niveau $l = 1$, on détermine alors les fils de chacun des nœuds de l'arbre. Dans ce cas précis, t' et t'' n'ayant pas de fils, la définition (1.177) donne $S(t' \times s') = \emptyset$. Dans ce cas, aucun enfant de $t \times s$ n'a de descendance et ces enfants sont donc des feuilles de l'arbre de bloc $\mathcal{T}_{IJK}$. L'arbre produit $\mathcal{T}_{IJK}$ est donc représenté par la figure (1.15) ci-desssus et la matrice C a donc la structure suivante, On peut constater sur cet exemple que la détermination de l'arbre produit $\mathcal{T}_{IJK}$ n'est pas une chose aisée à la main. Dans la pratique, on peut utiliser un parcours en largeur des arbres et tester la condition (1.177) pour chaque candidat.
Même si les matrices A et B sont identiques, il n'y a pas de règle quant à la structure du produit et l'arbre produit peut être plus grossier, plus fin voir complètement différent [Liz14] comme le montre d'ores et déjà l'exemple précédent.

Estimation de la constante de rareté La constante de rareté de l'arbre produit vérifie l'estimation suivante issue de [Beb00].

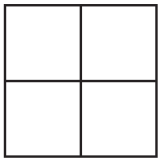


FIGURE 1.15 – Structure de la matrice produit C = AB.

Proposition 1.34 (Constante de rareté de $\mathscr{T}_{\mathrm{IJK}}$)**.** *L'arbre produit $\mathscr{T}_{\mathrm{IJK}}$ est un arbre de blocs basé sur les groupes $\mathscr{T}_{\mathrm{I}}$ et $\mathscr{T}_{\mathrm{K}}$. Sa profondeur, notée $\mathrm{L}(\mathscr{T}_{\mathrm{IJK}})$ vérifie l'inégalité*

$$\mathrm{L}(\mathscr{T}_{\mathrm{IJK}}) \leqslant \min\{\mathrm{L}(\mathscr{T}_{\mathrm{I\times J}}), \mathrm{L}(\mathscr{T}_{\mathrm{J\times K}})\}. \tag{1.178}$$

La constante de rareté de $\mathscr{T}_{\mathrm{IJK}}$ satisfait l'inégalité suivante,

$$c_{\mathrm{ra}}(\mathscr{T}_{\mathrm{IJK}}) \leqslant c_{\mathrm{ra}}(\mathscr{T}_{\mathrm{I\times J}})\, c_{ra}(\mathscr{T}_{\mathrm{J\times K}}). \tag{1.179}$$

L'estimation (1.179) est une borne maximale de la constante de rareté du produit $\mathscr{T}_{\mathrm{IJK}}$ et l'on peut obtenir une constante plus petite. L'exemple illustratif précédent met en jeu des matrices dont la constante de rareté est 2 tandis que le résultat possède également une constante de rareté de 2 ce qui est inférieur (4 dans ce cas) à la borne maximale donnée par (1.179).

Algorithme de multiplication La construction de l'arbre produit $\mathscr{T}_{\mathrm{IJK}}$ représente la première étape du produit de matrice. Une fois que l'on a créé cette structure, on effectue le calcul effectif des produits. L'algorithme de multiplication doit tenir compte de la structure de l'arbre produit. En effet, le calcul du produit dépend de la structure du résultat.

Dans le cas de la multiplication, on a A et B respectivement construites sur les arbres $\mathscr{T}_{\mathrm{I\times J}}$ et $\mathscr{T}_{\mathrm{J\times K}}$. La matrice C est alors construite à partir de l'arbre $\mathscr{T}_{\mathrm{I\times K}}$ et ces matrices sont compatibles et l'on peut effectuer le procuit C = AB. Chacune de ces trois matrices peut intervenir sous trois formes différentes :

— Une matrice de rang faible ;

— Une matrice pleine ;

— Une matrice subdivisée en une matrice-bloc 2×2 (une $\mathscr{H}$-matrice).

Au total, on a donc $3^3 = 27$ combinaisons possibles et seul le cas où toutes les matrices mises en jeu sont des $\mathscr{H}$-matrices est un cas de récursion. [Liz14] fournit une description exhaustive de cet algorithme que l'on ne reproduit pas ici. Signalons tout de même qu'il peut être nécessaire d'agglomérer des matrices de rang faible à l'aide de la méthode décrite au paragraphe (1.2.3). *A contrario*, il est possible de créer une partition d'une matrice de rang faible à partir de sa représentation compressée. En effet, pour un bloc compressé $\mathrm{A}_{t\times s}$ s'écrivant sous sa forme réduite UV^{T}, on écrit les partitions suivantes,

$$\mathrm{U} = \begin{bmatrix} \mathrm{U}_1 \\ \mathrm{U}_2 \end{bmatrix}, \tag{1.180}$$

$$\mathrm{V} = \begin{bmatrix} \mathrm{V}_1 \\ \mathrm{V}_2 \end{bmatrix}, \tag{1.181}$$

et l'on peut obtenir la partition en matrice bloc 2×2 de la matrice $A_{t\times s}$ de la façon suivante,

$$A_{t\times s}=\begin{bmatrix} U_1V_1^T & U_1V_2^T \\ U_2V_1^T & U_2V_2^T \end{bmatrix} \tag{1.182}$$

Il s'agit donc d'une matrice bloc 2×2 dont les sous-matrices sont des matrices de rang faible. Cette opération fait apparaître quatres matrices de rang faibles virtuelles qui ne sont pas présentes dans l'arbre original. Elles représentent simplément une étape intermédiaire dans le calcul et d'ailleurs sans traitement particulier, cette opération double la mémoire nécessaire pour stocker l'interaction $t\times s$. Ceci montre que cette étape doit être traitée avec précaution sous peine de dégrader les performances de la méthode. Ce sont les étapes de conversions hiérarchiques développées dans [Liz14].

Complexité du produit matriciel Contrairement aux opérations précédentes, le fait que la structure de $\mathscr{T}_{IJK}$ soit différente rend l'analyse de la complexité plus délicate. Cela est dû au fait que la structure du résultat est totalement différente de la structure originale ce qui requiert de nombreuses conversions hiérarchiques. On introduit alors une constante dont le rôle est de mesurer la qualité de l'arbre produit $\mathscr{T}_{IJK}$. C'est la constante d'idempotence.

Définition 1.35 (Constante d'idempotence)**.** Soit $\mathscr{T}_{I\times I}$ un arbre de blocs basé sur l'arbre de groupes $\mathscr{T}_I$. Pour un bloc $t\times s$ de $\mathscr{T}_{I\times I}$, on définit la **constante d'idempotence** $c_{id}(t\times s)$ pour ce bloc par

$$c_{id}(t\times s)=|\{t'\times s' \ : \ t'\in S(t), s'\in S(s') \text{ et } \exists r'\in\mathscr{T}_I \ : \ t'\times r'\in\mathscr{T}_{I\times I}, r'\times s'\in\mathscr{T}_{I\times I}\}|. \tag{1.183}$$

À l'échelle de la matrice, on définit la constante d'idempotence $c_{id}(\mathscr{T}_{I\times I})$ par

$$c_{id}(\mathscr{T}_{I\times I})=\max_{t\times s\in\mathscr{L}(\mathscr{T}_{I\times I})} c_{id}(t\times s) \tag{1.184}$$

Cette constante s'interprète comme le nombre de termes composant un terme du produit matriciel. Ceci établit le lien avec l'estimation de complexité suivante,

Proposition 1.36 (Multiplication d'$\mathscr{H}$-matrices,[Liz14])**.** *Soit $\mathscr{T}_{I\times I}$ un arbre de blocs associé à l'arbre de groupe $\mathscr{T}_I$. On note p la profondeur de l'arbre $\mathscr{T}_{I\times I}$ et l'on rappelle que c_{ra} désigne la constante de rareté de la matrice. Pour deux $\mathscr{H}$-matrices $A\in\mathscr{H}(\mathscr{T}_{I\times I},r)$ et $B\in\mathscr{H}(\mathscr{T}_{I\times I},r)$, leur produit AB peut être représenté par une matrice de $\mathscr{H}(\mathscr{T}_{I\times I},r)$ et déterminé en $\mathscr{C}_{MM}$ opérations avec $\mathscr{C}_{MM}$ vérifiant l'estimation suivante,*

$$\mathscr{C}_{MM}\leqslant K.c_{ra}^3c_{id}^3r^3(p+1)^3\max\{|I|,|\mathscr{L}(\mathscr{T}_{I\times I})|\}, \tag{1.185}$$

K étant une constante indépendante de la dimension des matrices.

1.3.6.4 Autres opérations

Les paragraphes précédents ont montré que l'on peut additionner et multiplier des $\mathscr{H}$-matrices entre elles et que le résultat est encore une $\mathscr{H}$-matrice. C'est la nature hiérarchique des $\mathscr{H}$-matrices qui est exploitée par ces opérations et les rend efficaces. De plus, en effectuant l'hypothèse que le rang n'explose pas au cours des opérations, on a exhibé des estimations *a priori* de la complexité de ces opérations et ces dernières sont plus rapides que leurs équivalents classiques respectifs.
Le produit matrice-vecteur développé permet d'utiliser un solveur itératif rapide mais

dans le cas où le nombre de seconds membres est important, on préfère utiliser un solveur direct afin de résoudre un système linéaire $Ax = b$. Pour cela, on peut envisager plusieurs méthodes selon la matrice A.

Les opérations effectuées ne le sont pas toujours au plus bas niveau de l'arbre et peuvent être effectuées sur un nœud de l'arbre de blocs. Ce nœud peut être :

— Une matrice dense ;

— Une matrice de rang faible ;

— Une $\mathscr{H}$-matrice.

Les deux premiers cas correspondent à des feuilles et l'on peut effectuer des opérations sur ces dernières de manière efficace. Les matrices denses sont par construction de petite taille tandis que les matrices de rang faible peuvent être manipulées efficacement grâce aux opérations décrites au paragraphe 1.2.2. Dans le cas où le nœud est une $\mathscr{H}$-matrice (*i.e* lui-même subdivisé en blocs) on peut continuer de manière hiérarchique l'algorithme jusqu'à atteindre une feuille (c'est le cas de l'addition) ou effectuer une opération sur ce nœud en tant que $\mathscr{H}$-matrice dans le cas des factorisations que l'on décrit à présent.

Nous nous sommes restreints à des arbres de groupes binaires et ainsi, les nœuds de l'arbre de blocs sont des matrices 2×2 par bloc. Si le nœud $t \times s$ de l'arbre $\mathscr{T}_{I\times J}$ n'est pas une feuille, le bloc matriciel associé à ce dernier est de la forme

$$A_{t\times s} = \begin{bmatrix} A_{t'\times s'} & A_{t'\times s''} \\ A_{t''\times s'} & A_{t''\times s''} \end{bmatrix}, \tag{1.186}$$

où t', t'' et s', s'' sont les fils respectifs des groupes t et s. L'idée principale des opérations suivantes est d'exploiter cette structure de matrice bloc 2×2 en écrivant de façon hiérarchique les algorithmes écrits pour les matrices 2×2. Notons que dans le cas d'un nœud $t \in \mathscr{T}_I$ quelconque, ses fils ne sont pas nécessairement de même cardinal et les blocs de la matrice 2×2 par blocs ne sont pas de la même dimension. Ceci ne pose aucun problème pour écrire les algorithmes suivants.

Dans toute la suite de ce paragraphe, on considère un arbre de blocs $\mathscr{T}_{I\times I}$ associé à un arbre de groupes $\mathscr{T}_I$. Cette restriction correspond à considérer une matrice de discrétisation du nuage de degrés de liberté sur lui-même soit « l'objet » en entier.

Inversion d'une matrice À l'aide de la somme et du produit de $\mathscr{H}$-matrices, on peut construire une inverse d'une matrice $A_{\mathscr{H}}$ sous la forme d'une $\mathscr{H}$-matrice également. On considère la matrice inversible A suivante

$$A = \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix}, \tag{1.187}$$

où le bloc A_{11} est inversible. Le complément de Schur S [GL96] du bloc A_{22} dans A est donné par

$$S = A_{22} - A_{21}A_{11}^{-1}A_{12}. \tag{1.188}$$

S est inversible et l'on peut relier A^{-1} à S^{-1} par

$$A^{-1} = \begin{bmatrix} A_{11}^{-1} + A_{11}^{-1}A_{12}S^{-1}A_{21}A_{11}^{-1} & -A_{11}^{-1}A_{12}S^{-1} \\ -S^{-1}A_{21}A_{11}^{-1} & S^{-1} \end{bmatrix}, \tag{1.189}$$

Dans le cas où A est une $\mathscr{H}$-matrice, le calcul de l'inverse de A_{11}^{-1} ainsi que du complément de Schur S^{-1} est effectué de manière hiérarchique. Le calcul du cas de base (au plus haut niveau de l'arbre de blocs) s'effectue nécessairement sur une matrice dense. En effet, une matrice de rang faible n'est par définition pas inversible et la quantité A^{-1} n'est pas définie. On n'explicite pas ici le détail de cette récursion et l'on renvoie à [Liz14; BGH12] pour un exposé plus approfondi. [Liz14] fournit également des tests numériques de l'algorithme sur des cas pratiques en électromagnétisme et en acoustique.
Ce calcul d'une matrice inverse approchée $A_{\mathscr{H}}^{-1}$ de $A_{\mathscr{H}}$ permet de résoudre le système linéaire $A_{\mathscr{H}}x = b$ en calculant le produit matrice-vecteur

$$x = A_{\mathscr{H}}^{-1} b. \tag{1.190}$$

De plus, on peut également se servir de cet inverse approché comme préconditionneur dans un solveur itératif [Beb00]. Cependant, on préfère dans la pratique utiliser une factorisation de la matrice $A_{\mathscr{H}}$.

Factorisation LU La méthode usuelle pour résoudre un système linéaire $Ax = b$ consiste en deux étapes. Dans un premier temps on cherche une décomposition particulière de la matrice puis l'on résout un système linéaire équivalent mais plus simple [GL96]. Les raisons de cette décomposition sont à la fois la complexité de la méthode et la stabilité numérique de la méthode. En effet, le coût de la résolution d'un système linéaire avec une factorisation LU est deux fois plus faible que si l'on utilise une méthode d'inversion de la matrice [LT86; GL96].

On sait que toute matrice A de taille $N \times N$ peut s'écrire sous la forme

$$A = P^{-1} LU, \tag{1.191}$$

où P est une matrice de permutation, L une matrice triangulaire inférieure dont les éléments diagonaux sont unitaires et U une matrice triangulaire supérieure (on peut choisir la diagonale de U comme unitaire à la place de celle de L). Le système $Ax = b$ devient alors $P^{-1}LUx = b$, ce qui est équivalent à résoudre successivement deux systèmes triangulaires.

Pour des matrices par blocs, la factorisation LU sans pivotage (*i.e* $P = I_N$) est stable et ne présente pas d'instabilité dans la pratique si la matrice est symétrique et définie positive ou à diagonale dominante [Liz14] contrairement au cas général où la factorisation LU sans pivotage d'une matrice A peut présenter un comportement instable [GL96; TB97]. C'est cette version par blocs et sans pivotage que l'on va décrire ici.

Comme pour l'inversion, nous nous contentons de décrire le motif de récursion et renvoyons à [Liz14] pour un exposé détaillé des algorithmes. L'objectif fixé est d'écrire une factorisation LU d'une matrice bloc 2×2,

$$A = \begin{bmatrix} L_{11} & 0 \\ L_{21} & L_{22} \end{bmatrix} \cdot \begin{bmatrix} U_{11} & U_{12} \\ 0 & U_{22} \end{bmatrix}. \tag{1.192}$$

En effectuant le produit et en identifiant terme à terme, on obtient les quatre équations

suivantes

$$A_{11} = L_{11}U_{11}, \tag{1.193}$$
$$A_{12} = L_{11}U_{12}, \tag{1.194}$$
$$A_{21} = L_{21}U_{11}, \tag{1.195}$$
$$A_{22} = L_{21}U_{12} + L_{22}U_{22}. \tag{1.196}$$

L'équation (1.193) est une factorisation LU d'un sous-bloc de A et représente un cas de récursion. Dans un contexte de $\mathscr{H}$-matrices, on effectue l'opération de factorisation de manière hiérarchique sur A_{11}. Les équations (1.194) et (1.195) sont respectivement la résolution d'un système triangulaire inférieur et supérieur où les seconds membres sont respectivement A_{12} et A_{21}. Ces opérations sont également effectuées de façon récursive dans un contexte de $\mathscr{H}$-matrices.
Enfin, l'équation (1.196) peut s'écrire sous la forme

$$A_{22} - L_{21}U_{12} = L_{22}U_{22},$$

ce qui correspond à la factorisation LU de la matrice $A_{22} - L_{21}U_{12}$. Cette dernière correspond à un produit et une somme de $\mathscr{H}$-matrices tel que décrit précédemment. Toutes les opérations ci-dessus sont donc effectuées hiérarchiquement et l'on peut alors former la factorisation LU d'une $\mathscr{H}$-matrice $A_{\mathscr{H}}$ et l'on peut alors résoudre le système $A_{\mathscr{H}}x = b$ à l'aide de cette factorisation [BGH12; Beb00; Liz14].

Complexités Dans le cas de l'inversion hiérarchique, [BGH12] fournit une estimation de la complexité en fonction de la complexité de la multiplication de matrices car il s'agit en effet de l'opération la plus utilisée lors de l'inversion.
Dans le cas de la factorisation, l'analyse de la complexité n'est pas aisée à obtenir. En particulier, l'hypothèse effectuée sur le rang paraît trop optimiste et on constate une croissance du rang au cours de la factorisation. De plus, les opérations ne sont pas effectuées avec des opérandes de même type (feuilles entre elles) et on a recours à de multiples conversions hiérarchiques. On constate une complexité de l'ordre de $\mathscr{O}(N\log^2 N)$ pour des problèmes de taille N et pour N allant jusqu'à plusieurs millions de degrés de liberté. [Liz14] fournit une implémentation parallèle et efficace de la factorisation LU d'une $\mathscr{H}$-matrice et constate également cette croissance asymptotique sur des matrices issues de la discrétisation du noyau de Helmholtz.

Remarque 1.37 (Autres décompositions). D'autres décompositions sont étudiées dans la littérature, la factorisation QR dans [Beb00] ou la factorisation de Crout LDL^T dans [Liz14].

1.4 Conclusion

La discrétisation par éléments finis de l'équation modèle considérée en début de chapitre conduit à un système linéaire plein. Cette équation modèle est une équation intégrale fréquemment utilisée dans la résolution de problèmes physiques comme la résolution des équations de l'électrostatique ou en acoustique et le besoin de méthodes approchées rapides et robustes résolvant ce type d'équation est réel dans l'industrie.
Sans méthode adaptée, la complexité d'assemblage de la matrice de discrétsation de taille $N \times N$ est en $\mathscr{O}(N^2)$ tandis que la résolution du système linéaire associé est en $\mathscr{O}(N^3)$. Ces complexités s'avèrent être vite pénalisantes et l'on cherche à construire des méthodes

d'assemblage et de résolution rapides *i.e* d'une complexité de l'ordre de $\mathscr{O}(N\log(N))$ opérations.

L'idée principale d'une méthode rapide est d'utiliser au maximum les propriétés analytiques du noyau de l'équation. Dans notre exposé, nous avons considéré un noyau singulier sur la diagonale et montré par des exemples simples mais représentatifs que l'éloignement de la diagonale est un point clé d'une méthode rapide car cela permet de construire des approximations efficaces du noyau. Il est donc nécessaire d'éloigner les degrés de liberté du problème afin d'obtenir de telles approximations.

La méthode des $\mathscr{H}$-matrices que l'on a développée dans ce chapitre réalise exactement cette séparation des degrés de liberté. Le point initial de cette méthode consiste à représenter l'ensemble des degrés de liberté à l'aide d'une structure d'arbre binaire. Cette structure permet la séparation rapide et efficace des inconnues du problème et représente une partition des lignes et des colonnes de la matrices. L'interaction entre deux nœuds de cet arbre binaire définit un sous-bloc matriciel de la matrice de discrétisation et le choix de ces nœuds s'effectue grâce à un critère de séparation géométrique appelé critère d'admissibilité. Ce critère permet d'obtenir à partir de l'arbre binaire représentant les degrés de liberté (arbre de groupe) un nouvel arbre représentant une partition hiérarchique par blocs de la matrice : c'est une $\mathscr{H}$-matrice.
Ces blocs correspondant ainsi à des interactions lointaines, on peut les représenter efficacement à l'aide de matrices de rang faible, c'est-à-dire des matrices dont le rang est très inférieur à ses dimensions. Des matrices de rang faible de taille $m \times n$ et de rang r peuvent être sommées et multipliées avec une complexité de l'ordre de $\mathscr{O}(r^2(m+n))$ ce qui est une complexité linéaire à l'échelle d'un bloc. Ces propriétés sont exploitées hiérarchiquement afin de construire les opérations d'addition et de multiplication sur les $\mathscr{H}$-matrices.
On a également présenté un produit matrice-vecteur pour les $\mathscr{H}$-matrices en $\mathscr{O}(N\log(N))$ opérations ce qui permet l'utilisation d'un solveur itératif rapide à partir des $\mathscr{H}$-matrices. Par ailleurs, un avantage majeur des $\mathscr{H}$-matrices est la possibilité d'effectuer le calcul d'un inverse approché en conservant le même format ainsi que de construire une factorisation LU d'une matrice. Ces opérations rendent possible la construction d'un solveur direct rapide à partir des $\mathscr{H}$-matrices. Il s'agit d'une méthode actuellement utilisée dans l'industrie pour la résolution de grands systèmes linéaires pleins avec un temps de calcul et une précision maîtrisés. Par ailleurs, le développement d'implémentations parallèles efficaces de cette méthode est un champ de recherche dynamique.

Les matrices de rang faible apparaissent suite à l'approximation du noyau de l'équation intégrale par une décomposition analytique ou lorsque l'on effectue une approximation algébrique d'un bloc matriciel déjà assemblé. Le chapitre suivant porte sur les méthodes -analytiques et algébriques- existantes pour obtenir une matrice de rang faible : c'est la compression d'une matrice.

1.5 Références

[Beb00] M. Bebendorf. *Hierarchical Matrices.* Springer, 2000. 20, 22, 25, 26, 28, 34, 38, 41, 43, 45, 48, 49, 51, 53, 54, 58, 59

[BG05] S. Börm and L. Grasedyck. Hybrid cross approximation for integral operators. *Numerische Mathematik*, 2005. 43

[BGH12] S. Börm, L. Grasedyck, and W. Hackbusch. Hierarchical matrices. Technical report, 2012. 20, 22, 29, 33, 34, 38, 41, 43, 45, 51, 58, 59

[Cip00] B. A. Cipra. The best of the 20th century : Editors name top 10 algorithms. *SIAM News*, 33(4), 2000. 33

[GGMR09] L. Greengard, D. Gueyffier, P.-G. Martinsson, and V. Rokhlin. Fast direct solver for integral equations in complex three-dimensional domains. *Acta Numerica*, 18 :1–33, 2009. 13, 14, 19

[GH03] L. Grasedyck and W. Hackbusch. Constructionand arithmetics of h-matrices. *Computing*, 70(4) :295–334, 2003. 45

[GL96] G.H. Golub and C.F. Van Loan. *Matrix Computations (3e éd.)*. Johns Hopkins Studies in the Mathematical Sciences, 1996. 10, 11, 19, 31, 57, 58

[Hac99] W. Hackbusch. A sparse matrix arithmetic based on h-matrices part i. *Computing*, 62(2) :89–108, 1999. 45

[HK00] W. Hackbusch and B.N. Khoromskij. A sparse h-matrix arithmetic : general complexity estimates. *Journal of Computational and Applied Mathematics*, 125 :479–501, 2000. 45

[HS52] M. R. Hestenes and E. Stiefel. Methods of conjugate gradients for solving linear systems. *Journal of Research of the National Bureau of Standards*, 49(6), 1952. 10

[Liz14] B. Lizé. *Résolution Directe Rapide pour les Éléments Finis de Frontière en Électromagnétisme et Acoustique : H-Matrices. Parallélisme et Applications Industrielles*. PhD thesis, Université Paris 13, 2014. 33, 34, 40, 43, 45, 51, 53, 54, 55, 56, 58, 59

[LT86] P. Lascaux and R. Théodor. *Analyse numérique matricielle appliquée à l'art de l'ingénieur*. Masson, 1986. 58

[Mes11] M. Messner. *Fast Boundary Element Methods in Acoustics*. PhD thesis, Graz University of Technology,Institute of Applied Mechanics, 2011. 31, 38

[Syl02] G. Sylvand. *La Méthode Multipôle Rapide en Électromagnétisme. Per- formances, Parallélisation, Applications*. PhD thesis, ENPC/CERMICS, 2002. 29, 43

[TA07] I. Terrasse and T. Abboud. *Modélisation des phénomènes de propagation d'ondes*. École Polytechnique, 2007. 8

[TB97] L. N. Trefethen and D. Bau. *Numerical Linear AlgebraAnalyse numérique matricielle appliquée à l'art de l'ingénieur*. SIAM, 1997. 58

Chapitre 2

Compression d'une matrice

Sommaire

2.1 Introduction

L'objectif de ce chapitre est de présenter des méthodes permettant de compresser une matrice (un sous-bloc dans le cas d'une méthode hiérarchique). La compression d'une matrice est l'un des principaux ingrédients d'une méthode rapide telle que la FMM ou les $\mathscr{H}$-matrices. Pour une matrice M de taille $m \times n$ et une tolérance ϵ données, on cherche une représentation approchée de la forme $M \simeq AB^T$ où A et B sont de tailles respectives $m \times r$ et $n \times r$ avec la condition $r << m, n$ et telles que

$$\|M - AB^T\|_2 \leqslant \epsilon \|M\|_2. \tag{2.1}$$

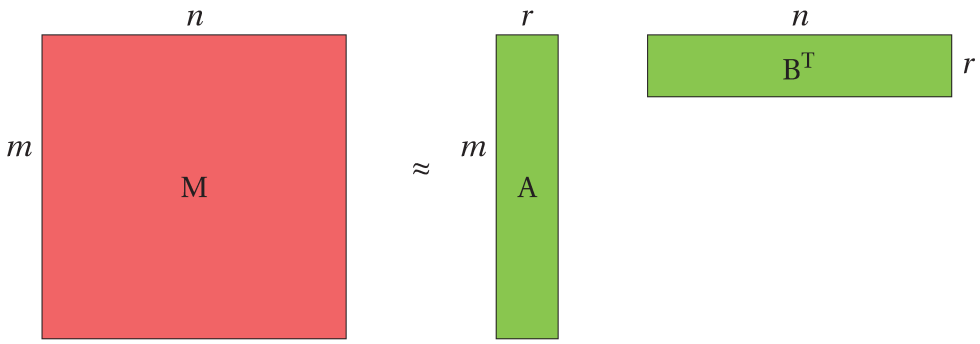


FIGURE 2.1 – Représentation compressée d'une matrice. La matrice est exprimée en tant qu'une somme de produit tensoriels.

Dans un contexte d'éléments finis de frontière, on s'intéresse à des matrices dont le terme général est donné par une intégrale double de la forme suivante,

$$M_{ij} = \int_\Gamma \int_\Gamma G(x,y)\Phi_j(y)\Phi_i^t(x)d\Gamma(x)d\Gamma(y). \tag{2.2}$$

On distinguera deux classes de méthodes selon que l'on approche le noyau de Green $G(x,y)$ (méthodes analytiques) ou la matrice de discrétisation (méthodes algébriques). Ces deux approches peuvent être complémentaire et généralement sont toutes les deux utilisées dans un code de calcul.

Méthodes algébriques Dans le cas des méthodes algébriques, on ne réalise pas réellement la séparation des variables. Par exemple, pour une précision ϵ donnée, la décomposition en valeurs singulières (SVD) fournit l'approximation de rang minimal satisfaisant 2.1. Pourtant, cette décomposition ne sépare pas les variables x et y. Elle traite simplement la matrice en totalité pour trouver un nombre minimal de vecteurs (donc un rang) pour représenter la matrice à la précision voulue.

Ces méthodes supposent peu d'hypothèses sur la matrice et ont en général un comportement de type "boîte noire". L'algorithme de référence pour effectuer la compression d'une matrice générale est la décomposition en valeurs singulières. Cette décomposition est très largement utilisée, au moins de manière théorique, dès que l'on souhaite obtenir des estimations d'erreur. La factorisation QR et ses dérivés permettent de trouver une base de l'image de M (notée Im(M)). La notion de compression intervient dans la réduction de cette base afin de ne conserver que les éléments qui contribuent le plus à décrire l'espace Im(M). Il est également possible d'obtenir une base de l'espace constitué des

lignes de M en appliquant la factorisation QR à la transposée M^T de M. On introduit à la section 2.2.3 une version probabiliste de la méthode QR afin d'accélérer les résultats. En effet, il suffit de remarquer que si Q_Y est une base de l'image de $Y = M\Omega$ alors Q_Y est également une base de l'image de M. C'est le produit par une bonne matrice test Ω qui permet la réduction rapide de l'image de M. Pour toutes ces méthodes, l'intégralité de la matrice est utilisée afin de déterminer l'approximation de rang fixe voulue et cela impose une complexité d'au moins $\mathcal{O}(mn)$ en l'absence d'hypothèse particulière (matrice creuse, matrice structurées, ...). On améliore ce point à l'aide d'une version heuristique de l'élimination de Gauss, l'algorithme ACA. Cet algorithme est détaillé en 2.2.6 et on constate qu'il fournit de très bons résultats dans la pratique dès que les valeurs singulières de la matrice affichent un profil à décroissance rapide.

Méthodes analytiques Contrairement aux méthodes algébriques, ces dernières sont plus intrusives et nécessitent en général une bonne connaissance du noyau $G(x, y)$. Leur fonctionnement est malgré tout commun en ce sens où elles cherchent à obtenir une approximation du noyau $G(x, y)$ à variables séparées et de rang fini de la forme

$$\tilde{G}(x, y) \simeq \sum_{q=0}^{r-1} u_q(x)\, v_q(y). \tag{2.3}$$

En injectant l'aproximation 2.3 dans 2.2, on se ramène à une somme de produits de deux intégrales simples au lieu d'une intégrale double,

$$\begin{aligned}
M_{ij} &= \int_{\Gamma_x}\int_{\Gamma_y} G(x,y)\Phi_j(y)\Phi_i^t(x)d\Gamma_x(x)d\Gamma_y(y) \\
&\simeq \int_{\Gamma_x}\int_{\Gamma_y}\left(\sum_{q=0}^{r-1} u_q(x)v_q(y)\right)\Phi_j(y)\Phi_i^t(x)d\Gamma_x(x)d\Gamma_y(y) \\
&\simeq \sum_{q=0}^{r-1}\left(\int_{\Gamma_x} u_q(x)\Phi_i^t(x)d\Gamma_x(x)\right)\left(\int_{\Gamma_y} v_q(y)\Phi_j(y)d\Gamma_y(y)\right).
\end{aligned}$$

En posant

$$\begin{aligned}
A_{iq} &= \left(\int_{\Gamma_x} u_q(x)\Phi_i^t(x)d\Gamma_x(x)\right), \\
B_{jq} &= \left(\int_{\Gamma_y} v_q(y)\Phi_j(y)d\Gamma_y(y)\right),
\end{aligned}$$

on retrouve bien l'approximation matricielle 2.1.

Dans le cas des équations de Laplace et de Helmholtz, on dispose d'une décomposition adéquate à travers le développement multipolaire pourvu que x et y soient séparées d'une distance suffisante dite condition d'admissibilité. Ce développement est à la base de la méthode des multipôles rapides (FMM) développée par Rokhlin et Greengard dans les années 80. Il est néanmoins très spécifique et présente certains inconvénients dans la pratique. Pour des applications particulières que nous ne développerons pas ici, d'autres méthodes utilisant une famille de fonctions orthogonales ont été utilisées. Sous certaines hypothèses de régularité sur le noyau, on peut utiliser la méthode générique du développement en série de Taylor puis tronquer la série afin d'obtenir une approximation de rang fini. Une méthode plus simple à manipuler car ne nécessitant pas les dérivées successives de G est l'interpolation polynômiale. Enfin, on revient également sur la méthode

d'approximations croisées (ACA) dans le cas d'une matrice "BEM" de la forme 2.2. Avec l'hypothèse supplémentaire que G est asymptotiquement lisse, cette méthode construit une approximation dégénérée de rang r (r petit). Cette donnée d'hypothèses supplémentaires sur le noyau permet d'affiner l'algorithme notamment le choix du premier pivot. Il s'agit d'un point essentiel de la méthode d'*Hybrid Cross Approximation* (HCA) présentée ci-après qui combine approximations croisées et interpolation polynômiale. On se limitera dans ce chapitre au noyau de Green de l'équation de Laplace.

2.2 Méthodes algébriques

Notations Dans tout ce chapitre, on note par t et s les ensembles suivants, $t = \{1,\dots,m\}$ et $s = \{1,\dots,n\}$. On note également $\hat{t} = \{i_1,\dots,i_r\}$, r lignes extraites de la matrice, $\hat{t}$ est un sous-ensemble de $t = \{1,\dots,m\}$ dont les éléments ne sont pas ordonnés. De façon similaire, on note $\hat{s}$ un sous-ensemble des colonnes. On notera $\mathrm{M}|_{\hat{t}\times\hat{s}}$ la matrice extraite de taille $r \times r$ dont les coefficients sont donnés par

$$(\mathrm{M}|_{\hat{t}\times\hat{s}})_{pq} = \mathrm{M}_{i_p,j_q}.$$

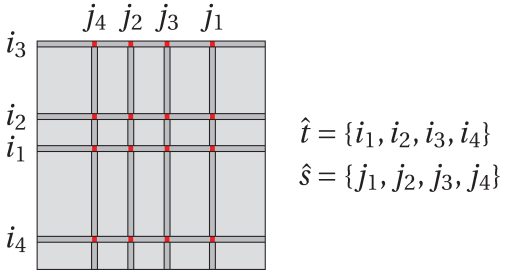


FIGURE 2.2 – Lignes et colonnes extraites d'une matrice.

Avec ces notations, $\mathrm{M}|_{\{i\}\times s}$ et $\mathrm{M}|_{t\times\{j\}}$ désignent respectivement la ligne i et la colonne j de la matrice M.

2.2.1 Décompositions en valeurs singulières (SVD)

Les résultats suivants introduisent la notion de décomposition en valeurs singulières (SVD). Ce résultat est au coeur de nombreux résultats théoriques d'algèbre linéaire. On se contente ici d'en exposer les grandes lignes nécessaires à la compression d'une matrice. On renvoie le lecteur à [GL96] pour un exposé plus précis.

Théorème 2.1 (Décomposition en valeurs singulières)**.** *Soit* $\mathrm{M} \in \mathbb{C}^{m\times n}$*, on note* $p = \min(m,n)$*. Alors il existe des matrices orthogonales* $\mathrm{U} = [u_1,\dots,u_m] \in \mathbb{C}^{m\times m}$ *et* $\mathrm{V} = [v_1,\dots,v_n] \in \mathbb{C}^{n\times n}$ *telles que*

$$\begin{aligned} \mathrm{U}^\star \mathrm{M} \mathrm{V} &= \Sigma \\ &= \mathit{diag}(\sigma_1,\dots,\sigma_p) \in \mathbb{R}^{m\times n} \\ \sigma_1 \geq \sigma_2 &\geq \dots \geq \sigma_p \geq 0. \end{aligned}$$

On appelle respectivement u_i *et* v_i *les* i^{eme} *vecteurs singuliers gauche et droit.* σ_i *est la* i^{eme} *valeurs singulières de la matrice* M *aussi notée* $\sigma_i(\mathrm{M})$*.*

On peut relier les propriétés algébriques d'une matrice quelconque M à sa décomposition en valeurs singulières $M = U\Sigma V^\star$.

Proposition 2.2 (Développement en valeurs singulières)**.** *Soit l'entier non nul r défini par*

$$\sigma_1 \geq \ldots \geq \sigma_r > \sigma_{r+1} \quad = \quad \ldots = \sigma_p = 0.$$

Alors,

$$\begin{aligned} rang(M) &= r, \\ Ker(M) &= Vec(v_{r+1}, \ldots, v_n), \\ Im(M) &= Vec(u_1, \ldots, u_r). \end{aligned}$$

On peut alors former le développement en valeurs singulières de la matrice M *par la quantité*

$$M_r \quad = \quad \sum_{i=1}^{r} \sigma_i u_i v_i^t.$$

Le développement précédent nous permet également d'établir des résultats pour différentes normes matricielles.

Proposition 2.3 (Normes matricielles et SVD)**.** *Le développement en valeurs singulières précédent et l'orthogonalité des vecteu rs singuliers impliquent les égalités suivantes,*

$$\begin{aligned} \|M\|_F^2 &= \sigma_1^2 + \ldots + \sigma_r^2, \\ \|M\|_2 &= \sigma_1, \\ \min_{x \neq 0} \frac{\|Mx\|_2}{\|x\|_2} &= \sigma_n, (m \geq n). \end{aligned}$$

Les deux propriétés précédentes montrent l'utilité de la décomposition en valeurs singulières dans la pratique. Le théorème suivant dû à Eckart et Young montre que cette dernière permet également de former une approximation d'une matrice au sens de la norme spectrale et ainsi d'en former une représentation efficace. On considère le développement en valeurs singulières de la matrice M du paragraphe précédent.

Théorème 2.4 (Meilleure approximation/Eckart-Young)**.** *Soient* $k < r = rang(M)$ *et* $M_k \in \mathbb{C}^{m \times n}$ *définie par*

$$M_k \quad = \quad \sum_{i=1}^{k} \sigma_i u_i v_i^t.$$

Alors on a,

$$\min_{rank(S)=k} \|M - S\|_2 \quad = \quad \|M - M_k\|_2 = \sigma_{k+1}.$$

Ce résultat énonce que pour un rang donné, le développement en valeurs singulières fournit la meilleure approximation au sens de la norme spectrale. Dans la pratique, on préfère se donner une tolérance ϵ et trouver le meilleur approximant possible au sens de la norme spectrale. On rappelle que le rang numérique $r_\epsilon(M)$ de M à ϵ près est défini par

$$r_\epsilon(M) = \min_{\|M-S\|_2 \leq \epsilon} rang(S)$$

Le théorème de meilleure approximation précédent entraîne que

$$\sigma_1 \geq \ldots \geq \sigma_{r_\epsilon} > \quad \epsilon \quad \geq \sigma_{r_\epsilon+1} \geq \ldots \geq \sigma_p, p = \min(m, n).$$

L'utilisation adéquate des inégalités ci-dessus permet dès lors la construction d'un approximant à précision fixe.

Application à la construction d'une approximation de rang faible Étant données une matrice $M \in \mathbb{C}^{m\times n}$ et une tolérance $\epsilon, 0 < \epsilon < 1$, on souhaite construire une approximation de la forme $M \simeq A.B^t$ telle que l'erreur relative en norme spectrale vérifie,

$$\|M - AB^t\|_2 \quad \leq \quad \epsilon\|M\|_2.$$

Puisque $\sigma_1(M) = \|M\|_2$, il suffit de normaliser les valeurs singulières, et quitte à poser $\tilde{\sigma}_i := \frac{\sigma_i}{\sigma_1}(i = 1,\ldots,p)$, on obtient

$$1 \geq \tilde{\sigma}_2 \geq \quad \ldots \quad \geq \tilde{\sigma}_p, p = \min(m,n).$$

Il suffit alors de choisir l'entier r_ϵ tel que

$$\tilde{\sigma}_{r_\epsilon} \geq \quad \epsilon \quad \geq \tilde{\sigma}_{r_\epsilon+1}.$$

La troncature M_{r_ϵ} du développement en valeurs singulières au r_ϵ premiers termes fournit alors l'approximant de rang minimal vérifiant le critère d'erreur relative.
Pour $i = 1,\ldots,r_\epsilon$, on note $\hat{u}_i := \sqrt{\sigma_i}.u_i$ et $\hat{v}_i := \sqrt{\sigma_i}.v_i$. On peut alors construire une approximation de rang faible à partir de l'approximant M_{r_ϵ}.

$$\begin{aligned} M_{r_\epsilon} &= \sum_{i=1}^{r_\epsilon} \sigma_i u_i v_i^t \\ &= \sum_{i=1}^{r_\epsilon} \hat{u}_i \hat{v}_i^t \\ &= \hat{U}\hat{V} \end{aligned}$$

où $\hat{U} = [\hat{u}_1,\ldots,\hat{u}_{r_\epsilon}] \in \mathbb{R}^{m\times r_\epsilon}$ et $V = [v_1,\ldots,v_{r_\epsilon}] \in \mathbb{R}^{n\times r_\epsilon}$. On obtient alors l'approximation voulue dont le rang est par construction r_ϵ.

Calcul numérique de la SVD Dans la pratique, nous avons utilisé la librairie d'algèbre linéaire LAPACK qui contient des implémentations performantes de plusieurs algorithmes de SVD dont l'algorithme de Golub-Reinsch.
Pour une matrice de taille $m \times n$, le coût de l'algorithme de Golub-Reinsch est $\mathscr{C}_{OP} = 4m^2n + 8mn^2 + 9n^3$. Si $m \simeq n$, ce coût est cubique ($\mathscr{O}(n^3)$). Ceci rend l'utilisation de cette méthode peu utilisable en pratique dès que les matrices sont de grandes tailles. On notera également que la construction d'une approximation de rang faible nécessite de connaître l'intégralité de la matrice.
Les autres méthodes algébriques existantes (QR,LU,...) tentent de déterminer une approximation de façon plus rapide quitte à perdre la notion de meilleure rang.

2.2.2 Factorisation QR d'une matrice

2.2.2.1 Existence de la factorisation QR

On décrit ici les principaux résultats de la factorisation QR d'une matrice. On renvoie le lecteur à la référence [GL96] pour de plus amples détails ainsi qu'un exposé complet des algorithmes. Nous ne détaillerons pas les algorithmes permettant le calcul de cette décomposition et nous suggérons de consulter -outre la référence [GL96]- la documentation de la librairie LAPACK, laquelle contient des implémentations performantes de ces algorithmes.

Proposition 2.5 (Factorisation QR)**.** *Soit* M *une matrice réelle de taille* $m \times n$*. Alors il existe une matrice orthogonale* Q *de taille* $m \times m$ *et une matrice triangulaire supérieure* R *de taille* $m \times n$ *telles que*

$$M = QR. \tag{2.4}$$

Remarque 2.6 (Processus d'orthogonalisation de Gram-Schmidt). La méthode d'orthogonalisation de Gram-Schmidt prouve par construction l'existence de la factorisation 2.5. L'existence de cette factorisation est prouvée par son calcul Notons enfin que la méthode de Householder fournit également la factorisation QR d'une matrice.

Sans perdre de généralité, on suppose que $m \geqslant n$. On va relier l'image de M avec l'image de Q grâce à la proposition suivante ce qui permet également de faire le lien avec la notion de rang puisque le rang d'une matrice est la dimension de son espace image.

Proposition 2.7. *Soit* $M = QR$ *la factorisation* QR *d'une matrice* M *de taille* $m \times n$ *de rang plein. On partitionne les matrices* M *et* Q *selon leurs colonnes et on note* $M = [M_1, \ldots, M_n]$ *et* $Q = [Q_1, \ldots, Q_m]$*. Alors :*

$$\mathrm{Vec}(M_1, \ldots, M_r) = \mathrm{Vec}(Q_1, \ldots, Q_r),\ r = 1, \ldots, n.$$

En particulier, si l'on note $Q_{11} = [Q_1, \ldots, Q_n]$ *et* $Q_{12} = [Q_{n+1}, \ldots, Q_m]$ *de telle sorte à avoir* $Q_{11} = [Q_{11}Q_{12}]$*, alors :*

$$\mathrm{Im}(M) = \mathrm{Im}(Q_{11}), \tag{2.5}$$
$$\mathrm{Im}(M)^{\perp} = \mathrm{Im}(Q_{12}), \tag{2.6}$$

et $M = Q_{11}R_{11}$ *avec* $R_{11} = R(1:n, 1:n)$*. Cette factorisation est la factorisation* QR *réduite de M. On notera de plus que cette factorisation* QR *réduite est unique.*

Grâce à la proposition précédente, on montre que le rang de M est aussi le rang de Q_{11} et ainsi, la détermination du rang d'une matrice peut être effectuée par le biais de sa factorisation QR réduite.
Dans le cas où l'on travaille à précision donnée, la proposition suivante montre qu'il est possible de relier la factorisation QR d'une matrice avec son rang numérique (cf [GL96] et [HP92]).

Proposition 2.8 (ϵ-rang)**.** *Soit une précision fixée* ϵ*. On écrit* $M = QR$ *la factorisation* QR *de la matrice* M *de taille* $m \times n$*. On décompose la matrice* R *par blocs de la façon suivante,*

$$R = \begin{bmatrix} R_{11} & R_{12} \\ 0 & R_{22} \end{bmatrix},$$

où $R_{11} \in \mathbb{R}^{r \times r}$, $R_{22} \in \mathbb{R}^{(n-r)\times(n-r)}$ *et* $R_{12} \in \mathbb{R}^{r \times (n-r)}$*. Si de plus,*

$$\sigma_{\min}(R_{11}) >> \|R_{22}\|_2 = \mathcal{O}(\epsilon),$$

alors le rang numérique de M *à la précision* ϵ *est* r*.*

Cependant, dans le cas d'une matrice de rang faible, la factorisation QR peut ne pas donner une base pour l'espace image de M (voir [GL96] p.248 par exemple). Une solution est alors d'appliquer la factorisation QR à une permutation MP des colonnes. On obtient alors :

$$MP = QR.$$

Calcul de la factorisation QR **avec pivots** On peut obtenir cette factorisation en utilisant la méthode de Housholder pour la détermination de la factorisation QR (algorithme 5.4.1 de [GL96] et voir [BG65] pour la référence originale). En effet, pour une matrice de taille $m \times n$ de rang k, l'algorithme proposé dans [GL96] a un coût $\mathscr{C}_{\mathrm{PQR}}$ donné par

$$\mathscr{C}_{\mathrm{PQR}} = 4mnr - 2r^2(m+n) + 4r^3,$$

soit $\mathscr{C}_{\mathrm{PQR}} = \mathscr{O}(mnr)$. On note que ce coût est inférieur à celui du calcul de la décomposition en valeurs propres tant que le rang est faible et est comparable au coût de la méthode de Gram-Schmidt avec pivots. Toute la difficulté réside dans la détermination d'une bonne permutation -selon des critères à définir- et ce de manière économique d'un point de vue calculs. Dans nos applications, le rang n'est pas connu *a priori* et on souhaite l'obtenir en plus d'une factorisation QR. La factorisation adéquate est liée à la notion de factorisation QR révélant le rang que nous développons dans la partie suivante.

2.2.2.2 Factorisation QR révélant le rang

On s'intéresse à une factorisation QR partielle d'une matrice complexe M de taille $m \times n$ de la forme MP = QR où Q est orthogonale et R est décomposée par blocs de la façon suivante,

$$\mathrm{R} = \begin{bmatrix} \mathrm{R}_{11} & \mathrm{R}_{12} \\ 0 & \mathrm{R}_{22} \end{bmatrix},$$

où, $\mathrm{R}_{11} \in \mathbb{R}^{r\times r}$ est bien conditionnée, $\mathrm{R}_{22} \in \mathbb{R}^{(n-r)\times(n-r)}$ est de norme spectrale petite et $\mathrm{R}_{12} \in \mathbb{R}^{r\times(n-r)}$ est linéairement indépendante de R_{11} et ses coefficients sont bornés, r étant un entier non nul. Cette forme de factorisation est particulièrement utile pour étudier le rang dans le cas où les blocs diagonaux de R sont de normes différentes.

On décrit à présent la notion de factorisation QR révélant le rang aussi appelée RRQR (*Rank Revealing QR*) dans la littérature. Dans un premier temps, on adoptera la définition fournie dans [GE94] (voir également [MG03]).

Définition 2.9. On appelle factorisation QR révélant le rang une factorisation QR par blocs de la forme précédente et dont les blocs vérifient les conditions suivantes,

$$\begin{aligned} \sigma_{\min}(\mathrm{R}_{11}) &\geqslant \frac{\sigma_r(\mathrm{M})}{p(r,n)} \\ \sigma_{\max}(\mathrm{R}_{22}) &\leqslant \sigma_{r+1}(\mathrm{M})\,p(r,n), \end{aligned}$$

où $p(r,n)$ est un polynôme en n et r de bas degré et borné.

D'autres définitions moins restrictives existent, notamment dans [HP92]. On présente une de ces définitions fortement liée à la propriété 2.8.

Définition 2.10. Supposons qu'une matrice réelle M de taille $m \times n$ ait un rang numérique r. S'il existe une permutation P de ses colonnes telle que la factorisation QR de MP s'écrive MP = QR avec,

$$\mathrm{R} = \begin{bmatrix} \mathrm{R}_{11} & \mathrm{R}_{12} \\ 0 & \mathrm{R}_{22} \end{bmatrix},$$

$\mathrm{R}_{11} \in \mathbb{R}^{r\times r}$, et

$$\sigma_{\min}(\mathrm{R}_{11}) >> \|\mathrm{R}_{22}\|_2 = \mathscr{O}(\epsilon),$$

alors la factorisation MP = QR est appelée factorisation QR révélant le rang.

Cette définition impose des conditions sur les blocs diagonaux de la matrice R qui sont liées à la notion de rang numérique. On peut utiliser une définition équivalente en travaillant avec la norme de Frobénius en utilisant le fait que pour toute matrice M, $\|M\|_2 \leqslant \|M\|_F$.

Définition 2.11. Supposons qu'une matrice M de taille $m \times n$ ait un rang numérique r. S'il existe une permutation P de ses colonnes telle que la factorisation QR de MP s'écrive MP = QR avec,

$$R = \begin{bmatrix} R_{11} & R_{12} \\ 0 & R_{22} \end{bmatrix},$$

$R_{11} \in \mathbb{R}^{r \times r}$, et

$$\frac{1}{\|R_{11}^{-1}\|_F} >> \|R_{22}\|_F = \mathcal{O}(\epsilon),$$

alors la factorisation MP = QR est appelée factorisation QR révélant le rang.

Il se peut néanmoins que la factorisation RRQR précédente ne soit pas numériquement stable si la norme de $R_{11}^{-1}R_{12}$ est grande. On introduit alors la notion de factorisation RRQR robuste abordée dans [GE94] ou [MG03]. La première définition donnait des conditions sur $\sigma_{min}(R_{11})$ et $\sigma_{max}(R_{22})$, on définit la factorisation RRQR robuste à travers une condition sur $R_{11}^{-1}R_{12}$.

Définition 2.12. En gardant les mêmes notations que dans la définition 2.9, on dit qu'une factorisation RRQR est robuste si et seulement si il existe un polynôme $q(r,n)$ borné et de bas degré en r et en n tel que

$$|(R_{11}^{-1}R_{12})_{ij}| \leqslant q(r,n), \text{pour } i \in \{1,\dots,r\} \text{ et } j \in \{1,\dots,n-r\}.$$

L'existence de telles factorisations est demontrée dans [GE94] et les auteurs fournissent par ailleurs des résultats sur ce que sont les polynômes mis en jeu. De plus, cette factorisation possède l'avantage de déterminer le rang de la matrice M comme le montre les résultats suivants, issus de [GE94] (on peut aussi consulter [CGMR05] pour une utilisation pratique de ces résultats).

Le premier théorème montre que pour toute matrice M, il est possible de trouver une permutation P de ses colonnes telle que MP puisse être décomposée sous la forme MP = QR où Q est une matrice dont les colonnes sont orthonormales et R est triangulaire supérieure. De plus, les valeurs singulières de la sous-matrice d'ordre r de R sont de bonnes approximations des r plus grandes valeurs singulières de M. Le théorème stipule également que les r premières colonnes de MP forment une base bien conditionnée de l'image de M à la précision $\sigma_{r+1}(M)$.
Le second résultat suivant montre qu'on peut obtenir une factorisation QR révélant le rang en un temps raisonnable si l'on relâche les conditions sur la matrice R.

Théorème 2.13 (Gu-Eisenstat)**.** *Soit* M *une matrice de taille* $m \times n$*, posons* $l = \min(m,n)$ *et soit* r *un entier tel que* $1 \leqslant r \leqslant l$*. Alors il existe une factorisation*

$$MP = QR$$

où P *est une matrice de permutation de taille* $n \times n$*,* Q *est une matrice ayant ses colonnes orthonormales de taille* $m \times l$ *et* R *est une matrice de taille* $l \times n$ *triangulaire supérieure. En*

décomposant Q *et* R *de la sorte :*

$$Q = \begin{bmatrix} Q_{11} & Q_{12} \\ Q_{21} & Q_{22} \end{bmatrix} \quad R = \begin{bmatrix} R_{11} & R_{12} \\ 0 & R_{22} \end{bmatrix}$$

où Q_{11} *et* R_{11} *sont de taille* $r \times r$, Q_{21} *de taille* $(m-r) \times r$, Q_{12} *de taille* $r \times (l-r)$, Q_{22} *de taille* $(m-r) \times (l-r)$, R_{12} *de taille* $r \times (n-r)$ *et* R_{22} *de taille* $(l-r) \times (n-r)$, *on obtient les inégalités suivantes*

$$\begin{aligned} \sigma_r(R_{11}) &\geqslant \frac{1}{\sqrt{1+r(n-r)}}\sigma_r(M), && (2.7) \\ \sigma_1(R_{22}) &\leqslant \sqrt{1+r(n-r)}\sigma_{r+1}(A), && (2.8) \\ \|R_{11}^{-1}R_{12}\|_F &\leqslant \sqrt{r(n-r)} && (2.9) \end{aligned}$$

On remarque que l'approximation de M par une matrice de rang r conduit naturellement à considérer que $\sigma_{r+1}(M) = \epsilon$ avec ϵ suffisamment petit.

Théorème 2.14 (Gu-Eisenstat)**.** *Étant donnée une matrice* M *de taille* $m \times n$ *décomposée sous la forme* MP = QR *précédente qui au lieu des conditions du théorème 2.13 vérifie les conditions suivantes*

$$\begin{aligned} \sigma_r(R_{11}) &\geqslant \frac{1}{\sqrt{1+nr(n-r)}}\sigma_r(M), \\ \sigma_1(R_{22}) &\leqslant \sqrt{1+nr(n-r)}\sigma_{r+1}(M), \\ \|R_{11}^{-1}\|_F &\leqslant \sqrt{nr(n-r)} \end{aligned}$$

peut être calculée en $\mathcal{O}(mn^2)$ *opérations.*

Remarque 2.15. Il s'agit là d'une borne pessimiste. Typiquement on obtient un nombre d'opérations similaire au nombre d'opérations requis pour effectuer un processus de Gram-Schmidt pivoté soit $\mathcal{O}(mnr)$.

2.2.3 Méthodes algébriques randomisées

Considérons une matrice de rang exactement r dont on cherche une base de l'espace image. Soient r vecteurs aléatoires (on ne se préoccupe pas de leurs distributions) $\omega^{(i)}, i = 1,\ldots,r$. Pour $i = 1,\ldots,r$, on forme les produits suivants,

$$y_i = M.\omega_i, i = 1,\ldots,r,$$

que l'on note matriciellement $Y = M\Omega$. Chaque $y^{(i)}$ est un élément aléatoire de l'espace image de M et de par le caractère aléatoire des $\omega^{(i)}$, l'ensemble $\{\omega^{(i)}, i = 1,\ldots,r\}$ est très certainement linéairement indépendent. Ainsi, l'ensemble $\{y^{(i)}, i = 1,\ldots,r\}$ est très certainement une base de l'image de M.
Supposons à présent que $M = B + E$ où B est une matrice de rang r et E telle que $\|E\|_2$ est petite. C'est l'information contenue par la matrice B que l'on cherche à déterminer. De manière analogue, les vecteurs $y^{(i)}, i = 1,\ldots,r$ ont à présent moins de chance de former une base de l'espace image de B à cause des perturbations dues à E. Pour pallier ce défaut, on se donne p vecteurs aléatoires supplémentaires de manière à ce que l'ensemble

enrichi $\{y^{(i)}, i = 1,\dots,r+p\}$ constitue une base de l'image de B. La pratique montre que cela fonctionne avec une probabilité d'échec faible pour un paramètre p petit.
En d'autres termes, pour une matrice M de taille $m \times n$ et de rang r, on peut chercher une base de l'espace image de M parmi les $(r+p)$ colonnes de Y au lieu des n colonnes de M ce qui diminue potentiellement le nombre de calculs, en supposant peu chère la construction de Y.

2.2.3.1 Factorisation QR randomisée à rang fixé

Algorithme randomisé Dans ce paragraphe, on fixe un entier r désignant le rang voulu de la matrice M dont on cherche une base de l'image. L'algorithme suivant décrit dans [HMT10] fournit une telle base.

Algorithme 5 Résolution de la factorisation QR à rang fixé

Données: Une matrice $M \in \mathbb{R}^{m\times n}$, des entiers $r > 0$ et $p > 0$. On note $l = r + p$.
1: Tirer aléatoirement une matrice test $\Omega \in \mathbb{R}^{n\times l}$
2: Effectuer le produit de matrices $Y = M\Omega$
3: Déterminer une base Q de Im(Y) : $Y = QR, Q \in \mathbb{R}^{m\times l}$

Coût de l'algorithme On note $\mathscr{C}_\Omega$ le coût de construire la matrice test Ω ,$\mathscr{C}_{Mx}$ le coût d'un produit d'une matrice par un vecteur et enfin $\mathscr{C}_{QR}(m,n)$ le coût d'une factorisation QR d'une matrice de taile $m \times n$.Dans la pratique, les méthodes de Gram-Schmidt, de Householder ou de Givens fournissent de bons résultats pour la détermination de la matrice Q lors de l'étape 3. Le coût de l'algorithme 5 est directement la somme des coûts de chaque étape :

$$\mathscr{C}_5 = \mathscr{C}_\Omega + l\mathscr{C}_{Mx} + \mathscr{C}_{QR}(m,l). \tag{2.10}$$

La matrice Y ayant un rang proche de l, les résultats de la section 2.2.2 fournissent $\mathscr{C}_{QR}(m,l) = ml^2$. En l'absence d'hypothèses supplémentaires sur M, le produit matrice-vecteur est en $\mathscr{O}(mn)$ opérations tandis que la construction de la matrice Ω nécessite au moins nl opérations quitte à supposer peu onéreuse la construction d'un de ses coefficients. Ainsi, le coût de cet algorithme devient

$$\mathscr{C}_5 \simeq nl + mnl + ml^2, \tag{2.11}$$

ce qui est la même complexité qu'un algorithme RRQR.

Estimation d'erreur a posteriori Dans le cas d'une matrice réelle et d'une matrice test gaussienne, on peut obtenir une estimation de l'erreur commise par l'algorithme 5 en fonction des valeurs singulières de la matrice M et des entiers r et p :

Théorème 2.16 (Erreur a posteriori,[HMT10])**.** *En supposant les calculs effectués en arithmétique exacte, si* M *est une matrice réelle de taille* $m \times n$*,* r *est le rang visé* $(r \geqslant 2)$ *et* p *est un petit paramètre de suréchantillonage* $(p \geqslant 2)$ *tel que* $r + p < \min(m,n)$*, alors*

$$\mathbb{E}\|(I - QQ^\star)M\|_2 \leqslant \left[1 + \frac{4\sqrt{r+p}}{p-1}\sqrt{\min(m,n)}\right]\sigma_{r+1}(M). \tag{2.12}$$

$\mathbb{E}$ *désigne l'espérance par rapport à la variable aléatoire* Ω *tandis que* $\sigma_{r+1}(M)$ *est la* $(r+1)^e$ *valeur singulière de* M.

Le paramètre de suréchantillonage p intervient dans l'estimation d'erreur a posteriori et l'on aimerait pouvoir quantifier le surplus qu'il représente. Dans la pratique, on constate qu'il dépend principalement de la taille de M et du comportement de ses valeurs singulières. Une matrice de grande taille nécessitera un paramètre de suréchantillonage plus élevé de même qu'une matrice dont les valeurs singulières ont une décroissance lente. On souhaiterait se donner une précision fixe et déterminer le rang de la matrice (ainsi que le paramètre p). Pour ce faire, nous devons obtenir une estimation de l'erreur d'approximation facilement calculable.

2.2.3.2 Estimation de la norme spectrale d'une matrice

Lemme 2.17 (Estimation de la norme spectrale)**.** M étant une matrice réelle de taille $m \times n$, on se donne un entier r et un réel α. On considère r vecteurs aléatoires gaussiens indépendants et identiquement distribués $\{\omega_i\}_{i=1,\dots,r}$. Alors,

$$\|M\|_2 \leqslant \alpha\sqrt{\frac{2}{\pi}} \max_{i=1,\dots,r} \|M\omega_i\|_2, \tag{2.13}$$

sauf avec une probabilité α^{-r}.

Remarque 2.18 (Estimation de la norme spectrale par une méthode de puissance). [LWPG+07] fournit un estimateur plus précis de la norme spectrale utilisant les puissances de la matrice $M^\star M$. On se donne un vecteur gaussien ω de taille n et on note $\hat{\omega} = \omega/\|\omega\|_2$. Pour un entier $q > 1$, on définit $a_q(M)$ par

$$a_q(M) = \sqrt{\frac{\|(M^\star M)^q \hat{\omega}\|_2}{\|(M^\star M)^{q-1}\hat{\omega}\|_2}}$$

On montre alors que $10.a_q(M) \geqslant \|M\|_2$ avec une probabilité supérieure à $1-4\sqrt{n/(q-1)}100^{-q}$.

2.2.3.3 Factorisation QR randomisée à précision fixée

La norme spectrale étant approchée à l'aide de produits matrice-vecteurs, son calcul devient peu onéreux et on peut l'utiliser comme estimaterur d'erreur dans l'algorithme 5. Une première solution consiste à partir d'une estimation de r et de vérifier l'erreur commise à la fin de l'algorithme. Si l'on double cette estimation en cas d'échec, on ne dégrade pas la complexité asymptotique de l'algorithme. Une autre solution consiste à construire itérativement la base Q de l'étape 3 avec une méthode de Gram-Schmidt modifiée tout en contrôlant l'erreur en cours grâce à 2.13.
ϵ étant une tolérance fixée, on cherche un entier l et une matrice orthogonale $Q^{(l)}$ tels que

$$\|(I - Q^{(l)}(Q^{(l)})^\star)M\|_2 \leqslant \epsilon. \tag{2.14}$$

On commence avec une matrice Q vide puis on construit une base en ajoutant un à un les vecteurs q_i. Pour chaque nouvel élément q_i de la base, le calcul se décompose de la sorte.

1. Tirer un vecteur gaussien ω_i et poser $y_i = M\omega_i$.
2. Calculer $q_i = (I - Q^{(i-1)}(Q^{(i-1)})^\star)y_i$.
3. Normaliser q_i et former la matrice $Q^{(i+1)} = [Q^{(i)} q_i]$.

Les q_i calculés en 3 sont ceux qui interviennent lors de l'estimation 2.13 et l'algorithme s'achève lorsque la précision est atteinte. L'algorithme suivant est le résultat de l'utilisation de l'estimateur d'erreur dans l'algorithme 5 afin de déterminer une base vérifiant 2.14.

Algorithme 6 Factorisaton QR randomisée révélant le rang

Données: Une matrice M de taille $m \times n$, une tolérance ϵ, un entier r,($r = 10$ par ex.)
But: L'algorithme suivant détermine une base orthonormale Q telle que **??** soit vérifié avec une probabilité d'au moins $1 - \min(m,n)10^{-r}$.

Tirer r vecteurs gaussiens de taille n : $\omega_1, \ldots, \omega_r$
Pour $i = 1, \ldots, r$ **faire**
  Calculer $y_i = M\omega_i$
Fin Pour
$j = 0$
$Q_0 = [\,]$ est une matrice vide de taille $m \times 0$
Tant que $\max\{\|y_{j+1}\|_2, \ldots, \|y_{j+r}\|_2\} > \epsilon/(10\sqrt{2/\pi})$ **faire**
  $j = j + 1$
  Remplacer y_j par $(I - Q_{j-1}Q_{j-1}^{\star})y_j$.
  $q_j = y_j / \|y_j\|_2$
  $Q_j = [Q_{j-1} q_j]$
  Tirer un vecteur aléatoire gaussien ω_{j+r} de taille n
  $y_{j+r} = (I - Q_j Q_j^{\star})M\omega_{j+r}$
  Pour $i = (j+1), \ldots, (j+r-1)$ **faire**
    Remplacer y_i par $y_i - < q_j, y_i > .q_j$
  Fin Pour
Fin Tant que
$Q = Q^{(j)}$.

Remarque 2.19. Le nombre d'éléments l de la base est généralement plus grand que celui de la base minimale vérifiant 2.14.

Remarque 2.20. Dans le cas où les valeurs singulières de la matrice ont une décroissance lente, il existe des variantes de ces algorithmes utilisant une méthode de puissance itérée. En effet, pour un entier p, la matrice $(QQ^{\star})^p M$ possède les mêmes vecteurs singuliers que M mais la décroissance des ses valeurs singulières est plus rapide.

2.2.3.4 Accélération du produit $Y = M\Omega$

En l'absence d'un produit matrice-vecteur $x \mapsto Mx$ rapide, l'étape la plus onéreuse de l'algorithme 5 est le produit par la matrice test Ω qui recquiert $\mathcal{O}(mnl)$ opérations pour une matrice dense. Une idée consiste à utilser une matrice test particulière. Plusieurs versions sont employées dans [HMT10] afin de réduire le coût de l'étape 2, mais nous nous contenterons de décrire le cas des matrices SRFT (*Subsampled Random Fourier Transform*) utilisées dans [LWPG+07] et [WLRT07].

Définition 2.21 (matrice SRFT)**.** Soit Ω une matrice de taille $n \times l$ de la forme suivante

$$\Omega = \sqrt{\frac{n}{l}}\mathrm{DFR} \tag{2.15}$$

— D = diag$(d_1,\dots,d_n)$, d_i étant des variables aléatoires uniformément distribuées sur le cercle unité du plan complexe.

— F est la matrice de la transformée de Fourier discrète non-normalisée de taille $n \times n$ dont les coefficients sont donnés par :

$$\mathrm{F}_{pq} = \exp\{-2i\pi(p-1)(q-1)/n\}. \tag{2.16}$$

— Pour l entiers r_j $(j = 1,\dots,l)$ aléatoires *i.i.d* uniformément dans $\{1,\dots,n\}$ sans remise, on définit la matrice R réelle de taille $n \times l$ dont les seuls éléments non nuls sont définis par :

$$\mathrm{R}_{r_j,j} = 1. \tag{2.17}$$

Cette classe de matrice peut être utilisée dans l'algorithme 5 afin de diminuer le coût de l'étape 2. En effet, la strucutre particulière d'une matrice test SRFT permet d'en calculer le produit par une matrice en $\mathcal{O}(mn\log(l))$ en utilisant un algorithme de FFT (*cf* [137]). Le nombre d'opérations est donc réduit à

$$\mathscr{C}_{\mathrm{SRFT}} = mn\log(l) + l^2 n. \tag{2.18}$$

Dans le cas où $l << \min(m,n)$ on constate numériquement (*cf* [LWPG+07]) que les tirages avec ou sans remise fournissent les mêmes résultats. Cependant, nous utilisons dans nos tests une version sans remise afin de pouvoir traiter des cas où nous ne sommes pas sûr de l'hypothèse précédente.

2.2.3.5 Obtention d'une SVD randomisée

Pour une précision donnée, l'algorithme 6 fournit le rang ainsi qu'une base de l'image de M en $\mathcal{O}(mn\log(l) + l^2 n)$ opérations vérifiant

$$\|(\mathrm{I} - \mathrm{QQ}^\star)\mathrm{M}\|_2 \leqslant \epsilon \tag{2.19}$$

L'algorithme suivant fournit une décomposition en valeurs singulières approchée de la matrice M à partir de cette base.

Algorithme 7 Décomposition SVD randomisée et approchée d'une matrice

Données: Une matrice M de taille $m \times n$ et une base Q telles que 2.19 soit satisfaite.
But: L'algorithme détermine une décomposition SVD approchée $\mathrm{M} \approx \mathrm{U}\Sigma\mathrm{V}^\star$ satisfaisant $\|\mathrm{M} - \mathrm{U}\Sigma\mathrm{V}^\star\|_2 \leqslant \epsilon$

1: Calculer $\mathrm{B} = \mathrm{Q}^\star\mathrm{M}$
2: Déterminer la SVD de la matrice B : $\mathrm{B} \approx \tilde{\mathrm{U}}\Sigma\mathrm{V}^\star$.
3: Construire la matrice U définie par $\mathrm{U} = \mathrm{Q}\tilde{\mathrm{U}}$.

Le coût de cet algorithme est dominé par la première étape consistant à former la matrice B en $\mathcal{O}(mnl)$ opérations et est intéressant dans la situation où l'on dispose d'un produit matrice vecteur rapide.

2.2.4 Décomposition interpolante

La méthode suivante exploite directement les résultats de la factorisation QR randomisée pour construire une décomposition d'une matrice M utilisant un sous-ensemble de

ses lignes(ou colonnes). On notera par ailleurs un lien avec la notion d'approximations en croix présenté à la section suivante dans [GTZ97]. On renvoie le lecteur au chapitre 7 de [MT11] pour une première approche et à [LWPG+07] et [WLRT07] pour un exposé plus approfondi. Enfin, une utilisation de cette méthode dans le cas des équations de Laplace et Helmholtz en électromagnétisme peut être consultée dans [HG11].

Dans toute la suite, M désigne une matrice complexe de taille $m \times n$ et les résultats restent valables pour une matrice réelle.

2.2.4.1 Déecomposition iD

Le théorème suivant (*cf* [MT11],Théorème 28) fournit une décomposition d'une matrice faisant intervenir un nombre réduit de colonnes extraites de cette dernière ainsi qu'une matrice dont les coefficients sont bornés en module.

Théorème 2.22. *Soient l, m et n des entiers tels que $l \leqslant m$ et $l \leqslant n$. On considère une matrice* M *de taille $m \times n$ telle que rang*(M) $\leqslant l$. *Alors il existe une matrice* B *de taille $m \times l$ dont les colonnes sont extraites de* M *et une matrice* C *de taille $l \times n$ telles que :*

$$\mathrm{M} = \mathrm{BC} \tag{2.20}$$

et

$$|\mathrm{C}_{ij}| \leqslant 1, \textit{pour } i = 1, \ldots, l \textit{ et } j = 1, \ldots, n. \tag{2.21}$$

De manière plus générale, le théorème 2 (section 3) dans [MRT06] et le théorème 3 dans [CGMR05] permettent de généraliser le résultat du théorème précédent et d'obtenir la proposition qui va suivre. Cette proposition stipule que pour toute matrice complexe M de taille $m \times n$ de rang r, il existe une matrice B de taille $m \times r$ dont les colonnes sont extraites de M ainsi qu'une matrice C de taille $r \times n$ telles que :

1. un sous-ensemble des colonnes de C forme la matrice identité d'ordre r,
2. M = BC,
3. $\|\mathrm{C}\|_2$ n'est pas trop grand.

De plus, la proposition fournit une approximation

$$\mathrm{BC} \approx \mathrm{M}, \tag{2.22}$$

dans le cas où le rang exact de M est supérieur à r mais la $(r+1)^{\text{ème}}$ valeur singulière de M est petite. On appellera décomposition interpolante ou iD, une approximation du type 2.22.

Proposition 2.23. *Soit* $\mathrm{M} \in \mathbb{C}^{m \times n}$. *Soit un entier r tel que $r \leqslant m$ et $r \leqslant n$. Il existe une matrice* $\mathrm{B} \in \mathbb{C}^{m \times r}$ *dont les colonnes sont extraites de* M *ainsi qu'une matrice* $\mathrm{C} \in \mathbb{C}^{r \times n}$ *telles que*

1. *un sous-ensemble des colonnes de* C *forme la matrice identité d'ordre r,*
2. $|\mathrm{C}_{ij}| \leqslant 1$ *pour* $i \in \{1, \ldots, r\}, j \in \{1, \ldots, n\}$
3. $\|\mathrm{C}\|_2 \leqslant \sqrt{r(n-r)+1}$,
4. *la plus petite valeur singulière de* C *est supérieure à 1.*
5. M = BC *pour $r = m$ ou $r = n$,*

6. *Pour* $r < m$ *et* $r < n$,

$$\|BC - M\|_2 \leqslant \sqrt{r(n-r)+1}\sigma_{r+1}(M), \tag{2.23}$$

où $\sigma_{r+1}(M)$ *est la* $(r+1)$ *– ème valeur singulière de* M.

Les propriétés (1) à (4) assurent que la décomposition est numériquement stable tandis que le point (6) fournit une estimation de l'erreur absolue en norme spectrale de l'approximation iD et traduit exactement le fait que BC satisfait la relation 2.22 à condition que la $(r+1)$ – ème valeur singulière de M soit petite.

Il existe des algorithmes (*cf* [GE94]) permettant de calculer la décompostion du théorème précédent mais leurs coûts sont trop élevés pour être utilisés en pratique. En effet, la condition 2.21 est particulièrement onéreuse et on préfère assouplir cette dernière.
L'algorithme décrit dans [CGMR05] (section 4), ainsi que dans [MT11] (chapitre 7), permet de construire une approximation iD satisfaisant des conditions affaiblies sur C.

La première étape consiste à écrire la matrice M sous la forme suivante :

$$M = QRP,$$

où Q est une matrice de taille $m \times r$ dont les colonnes sont orthonormales, R est une matrice triangulaire supérieure de taille $r \times n$ et P est une matrice de permutation de taille $n \times n$.
On choisit de partitionner les colonnes de R de la façon suivante,

$$R = [R_{11}|R_{12}],$$

avec R_{11} de taille $r \times r$ représentant les r premières colonnes tandis que les autres colonnes sont contenues dans R_{12} qui est de taille $r \times (n-r)$. On obtient alors une factorisation iD sous la forme 2.22, avec

$$B = QR_{11},$$

et

$$C = \left[I_r|(R_{11}^{-1})R_{12}\right]P,$$

où $(R_{11}^{-1})R_{12}$ désigne une solution du système linéaire $R_{11}X = R_{12}$ (*cf* [CGMR05]). Par construction, B est bien constituée de colonnes de M. Cependant, la matrice C ne vérifie pas la propriété 2.21 même si ses coefficients demeurent petits dans la pratique.

On peut assouplir la condition 2.21 et obtenir la proposition suivante ([MT11]).

Proposition 2.24. *Soit* $M \in \mathbb{C}^{m\times n}$. *Soit un entier* r *tel que* $r \leqslant m$ *et* $r \leqslant n$. *Pour tout* $\beta > 1$, *il existe une matrice* $B \in \mathbb{C}^{m\times r}$ *dont les colonnes sont extraites de* M *ainsi qu'une matrice* $C \in \mathbb{C}^{r\times n}$ *telles que*

1. *un sous-ensemble des colonnes de* C *forme la matrice identité d'ordre* r,
2. $|C_{ij}| \leqslant \beta$ *pour* $i \in \{1,\ldots,r\}, j \in \{1,\ldots,n\}$
3. $\|C\|_2 \leqslant \sqrt{\beta^2 r(n-r)+1}$,
4. *la plus petite valeur singulière de* C *est supérieure à 1.*
5. $M = BC$ *pour* $r = m$ *ou* $r = n$,

6. *Pour $r < m$ et $r < n$,*

$$\|BC - M\|_2 \leqslant \sqrt{\beta^2 r(n-r) + 1}\,\sigma_{r+1}(M),$$

Dans [LWPG+07], un algorithme proposé dans [CGMR05] (voir également [GE94]) pour le calcul de B et C est utilisé avec $\beta = 2$ afin d'obtenir une décomposition iD tout en gardant un coût raisonnable.
Dans la pratique, il est plus courant de se fixer une précision ϵ et de chercher le plus petit entier r tel que $\|BC - M\|_2 \approx \epsilon$. L'algorithme mentionné permet une telle modification ([WLRT07],[MT11]).

Coût de la méthode Le coût de la méthode donnée dans [CGMR05] recquiert de l'ordre de $log_\beta(n)$ fois plus de calculs qu'une factorisation QR avec pivots. Le coût total est donc de l'ordre de $\mathcal{O}\left(mnr log_\beta(n)\right)$. Cependant, dans la pratique, on observe un comportement de l'ordre de $\mathcal{O}(mnr)$ ce qui est semblable à celui d'une factorisation RRQR.

Remarque 2.25. On peut également écrire une décomposition iD sur les lignes de la matrice en considérant l'adjoint de la matrice.

2.2.4.2 Algorithmes rapides pour la décomposition iD

Un algorithme efficace pour le calcul de la décomposition iD d'une matrice est proposé dans [HMT10] et [WLRT07]. Cet algoritme utilise la factorisation QR randomisée afin d'obtenir une méthode rapide. On note Q la matrice orthogonale de rang r obtenue par l'algorithme **??** telle que

$$\|(I - QQ^\star)M\| \leqslant \epsilon. \tag{2.24}$$

On note J un sous-ensemble de $\{1,\ldots,r\}$ et $Q(J,:)$ la sous-matrice extraite de Q correspondant à ces lignes. En utilisant la proposition 2.24 et l'algorithme associé, on forme une décomposition iD des lignes de Q de la forme :

$$Q \approx XQ(J,:). \tag{2.25}$$

De 2.24, il vient que $M \approx QQ^\star M$ et en utilisant 2.25 on a

$$\begin{aligned} M &\approx QQ^\star M && (2.26)\\ &\approx XQ(J,:)Q^\star M. && (2.27)\end{aligned}$$

Puisque $X(J,:) = I_r$, $M(J,:) \approx Q(J,:)Q^\star M$ et quitte à poser $B = M(J,:)$ et $C = X$, on obtient une décomposition iD sur les lignes de M de la forme

$$M \approx CB.$$

Algorithme 8 Décomposition iD d'une matrice

Données: M de taille $m \times n$.

But: L'algorithme détermine une factorisation iD de la matrice M vérifiant à la fois la proposition 2.24 et $\|BC - M\|_2 \lesssim \sigma_{r+1}(M)$.

1: Effectuer la factorisation QR randomisée de M à l'aide de 6.

2: En utilisant la proposition 2.24 et l'algorithme associé, former une iD sur les lignes pour la base Q dterminée par QR sous la forme

$$Q \approx XQ(J,:).$$

3: Construire $B = M(J,:)$ et $C = X$ tels que

$$M \approx CB.$$

Cet algorithme probabiliste possède une probabilité d'échec faible grâce aux contrôles effectués lors de l'algorithme 6. Des tests de convergence sont proposés dans [WLRT07] pour s'assurer du succès de l'algorithme. L'estimation d'erreur 8 est prouvée dans [WLRT07] à la section 5.1 et fait intervenir la norme de la matrice test utilisée dans la factorisation QR randomisée. On précise que des implémentations de cette méthode et ses dérivées sont présentes dans la bibliothèque *Scipy*.

Coût de l'algorithme iD Pour une matrice M complexe de taille $m \times n$ de rang numérique r, r étant le rang fourni par l'algorithme 6, l'algorithme 8 a un coût de l'ordre de

$$\mathscr{C}_{iD} = \mathscr{O}\left(mn\log(r) + r^2 n\log(n)\right). \tag{2.28}$$

Dans la pratique, on observe plutôt un coût

$$\mathscr{C}'_{iD} = \mathscr{O}\left(mn\log(r) + r^2 n\right). \tag{2.29}$$

La décomposition iD construit une approximation d'une matrice à partir d'un sous-ensemble de ses lignes(ou colonnes) mais l'algorithme recquiert néanmoins l'intégralité de la matrice donc la complexité ne peut être inférieure à $\mathscr{O}(mn)$ opérations. Dans la suite, nous montrons un résultat prédisant l'existence d'une approximation à partir d'un sous ensemble de lignes et de colonnes de la matrice ainsi qu'une méthode heuristique pour obtenir une dépendance en $(m+n)$ dans la complexité.

2.2.5 Existence d'approximation extraite

2.2.5.1 Rang et sous-matrices extraites

On rappelle que $t = \{1,\dots,m\}$ et $s = \{1,\dots,n\}$. De plus, $\hat{t} \subset t$ et $\hat{s} \subset s$ désignent des sous-ensembles de t et s respectivememt.

Lemme 2.26. Soit une matrice M de taille $m \times n$ et de rang exact r. Alors il existe une sous-matrice inversible $\hat{M}$ de taille $r \times r$ de M. De plus, avec les notations précédentes, M s'écrit

$$M = C\hat{M}^{-1}R, \tag{2.30}$$

$C = M|_{t\times\hat{s}}$ et $C = M|_{\hat{t}\times s}$ $\hat{M} = M|_{\hat{t}\times\hat{s}}$ où $C = M(:,J)$ et $R = M(I,:)$ sont respectivement constituées d'un sous-ensemble des colonnes de M et d'un sous-ensemble des lignes de M.

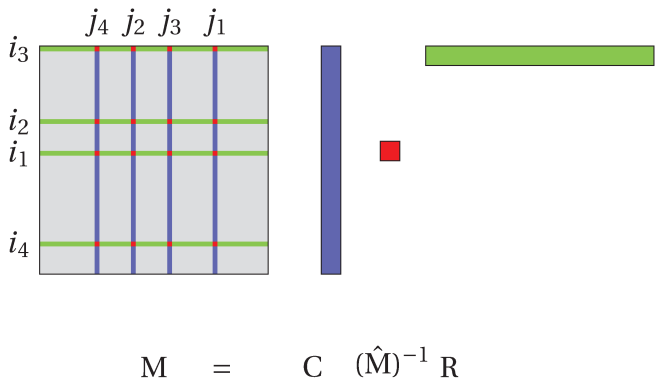


FIGURE 2.3 – Illustration d'une approximation extraite de rang $r = 4$.

La décomposition 2.30 est la décomposition squelettique de M (*skeleton decomposition*). Dans le cas où rang(M) $\approx r$, c'est-à-dire rang(M+E) $= r$ avec $\|E\|_2$ petite, on cherche à savoir si l'approximation $M \approx C\hat{M}^{-1}R$ est valide. [GTZ97] montre que l'approximation vérifie l'estimation d'erreur suivante,

$$\|M - C\hat{M}^{-1}R\|_2 = \mathcal{O}(\|M\|_2^2 \|\hat{M}^{-1}\|_2^2 \epsilon), \tag{2.31}$$

pour ϵ suffisament petit et $\hat{M}$ non singulière. Dans le cas général, on ne peut travailler directement avec $\hat{M}^{-1}$ car cette matrice est mal conditionnée. C'est pourquoi on va chercher à l'approcher par une matrice X et considérer l'approximation

$$M \approx CXR \tag{2.32}$$

qui sera la décomposition pseudo-squelettique de M. Une approche envisageable est de déterminer une sous-matrice X de taille $r \times r$ telle que $\|X^{-1}\|_2$ soit la plus petite possible. [GTZ97] cherche X comme la sous-matrice de déterminant maximal comme approximation et fournit un résultat d'existence d'une décomposition pseudo-squelettique 2.32.

2.2.5.2 Existence d'une approximation extraite

Le résultat suivant,dû à Goreinov et Tyrtyshnikov (*cf* [GTZ97]), affirme que si l'on dispose d'une approximation de faible rang raisonnablement précise, alors il existe également une approximation extraite de faible rang de qualité presque similaire.

Théorème 2.27 (Existence d'une approximation croisée)**.** *Soient* $M, N \in \mathbb{R}^{m\times n}$ *telles que* $\|M - N\|_2 \leq \epsilon$ *et* rang(N) $\leqslant r$*. Alors, il existe des sous-ensembles de* r *lignes et de* r *colonnes noté* I *et* J *ainsi qu'une matrice* $X \in \mathbb{R}^{r\times r}$ *tels que*

$$\|M - CXR\|_2 \leqslant \epsilon(1 + 2\sqrt{r}(\sqrt{m} + \sqrt{n})), \tag{2.33}$$

avec C = M(:, J),R = M(I, :) *et* X *de taille* $r \times r$*.* X *sera appelée matrice de couplage.*

En d'autres termes, si l'on dispose d'une approximation de rang r de la matrice M à la précision ϵ, il est possible d'en déduire une approximation extraite de rang r en amplifiant l'erreur en norme spectrale d'un facteur $(1+2\sqrt{r}(\sqrt{m}+\sqrt{n}))$. Notons que l'existence d'une approximation initiale N est garantie par la décomposition en valeurs singulières.

La détermination exacte de la matrice X en déterminant les pivots est théoriquement possible mais d'une complexité trop grande pour être réalisé. Pour exploiter ce théorème,

nous avons besoin d'une méthode pour construire la décomposition. Drineas *et al.* ont développé un algorithme décomposant la matrice M sous la forme A ≈ CUR où C et R sont respectivement des sous-matrices constituées de colonnes et lignes de M tandis que U est une petite matrice d'interaction. Une implémentation randomisée de l'algorithme CUR en erreur absolue apparaît dans [DKM06a; DKM06b; DKM06c]. Ces deux méthodes ont en commun d'approcher une base de l'image des colonnes et des lignes de M. La petite matrice U étant généralement la solution d'un problème de moindre carré.

Bien que non constructif, ce résultat est très intéressant en pratique car il montre qu'il est possible d'utiliser seulement un ensemble réduit de coefficients de la matrice à approcher. Toute la difficulté repose sur la détermination des pivots de lignes et de colonnes ainsi que de la matrice de couplage.

2.2.5.3 Approximation pseudo-squelettique de rang 1

On cherche une approximation de la forme 2.32 et de rang $r = 1$ d'une matrice M de taille $m \times n$. Grâce au théorème 2.27, on sait que l'on a besoin d'une seule ligne et une seule colonne de la matrice M. Il suffit de trouver les pivots (les indices de ligne et de colonne) nécessaires à la construction de la matrice de couplage X qui dans ce cas est un scalaire. Le scalaire de déterminant maximal n'est autre que le coefficient de plus grand module. On note $(i_0, j_0) = \text{argmax}_{i,j}(|\mathrm{M}_{ij}|)$ et ainsi, l'approximation 2.32 s'écrit

$$\mathrm{M} \approx \mathrm{CXR}, \tag{2.34}$$
$$\mathrm{X} = 1/\mathrm{M}|_{\{i_0\}\times\{j_0\}}, \tag{2.35}$$
$$\mathrm{C} = \mathrm{M}|_{t\times\{j_0\}}, \tag{2.36}$$
$$\mathrm{R} = \mathrm{M}|_{\{i_0\}\times s}. \tag{2.37}$$

Cette approximation est une approximation sous forme d'un produit tensoriel de rang 1 de la forme $\mathrm{M} \approx ab^{\mathrm{T}}$. On peut alors itérer le processus sur la matrice $\mathrm{M}' = \mathrm{M} - ab^{\mathrm{T}}$ pour obtenir une approximation de rang r sous forme d'une somme de produits tensoriels. En réalité, il s'agit de l'élimination de Gauss que l'on retrouve par la méthode d'approximation pseudo-squelettique.

Algorithme 9 Élimination de Gauss

1: **Tant que** $\mathrm{M} \neq 0$: **faire**
2: rang(M) ← rang(M) + 1
3: Trouver l'élément $\mathrm{M}_{i^\star j^\star}$ tel que $\mathrm{M}_{i^\star j^\star} = \max_{i,j} |\mathrm{M}_{ij}|$, $\alpha = \mathrm{M}_{i^\star j^\star}$
4: $\mathrm{M} \leftarrow \mathrm{M} - \frac{1}{\alpha}\mathrm{M}(:, j^\star)\mathrm{M}(i^\star, :)$
5: **Fin Tant que**

L'utilisation d'une heuristique pour déterminer le pivot maximal α, on peut déterminer des approximations de rang r à la volée, *i.e* sans calculer au préalable l'intégralité de la matrice que l'on souhaite approcher.

2.2.6 Approximations croisées

La détermination rapide de α est la base de tous les algorithmes rapides. ACA, ACA+ ou les autres méthodes sont des heuristiques pour cette détermination.

On présente dans cette partie l'algorithme ACA ainsi que ses dérivées. On rappelle que le but de cet algorithme est d'obtenir une décomposition d'une matrice M réelle de taille $m \times n$ sous la forme

$$\mathrm{M} = \mathrm{S}_r + \mathrm{E}_r.$$

S_r est l'approximation voulue de rang au plus r et E_r est le reste, idéalement de norme inférieure à une tolérance donnée ϵ. De plus, on souhaite que l'approximant soit donné par :

$$\mathrm{S}_r = \mathrm{AB}^{\mathrm{T}},$$

où A est de taille $m \times r$ tandis que B est de taille $n \times r$.

Le premier algorithme décrit construit une approximation de rang fixe r à l'aide d'une stratégie de pivotage total. Dans la pratique, le coût d'une telle stratégie est prohibitif et on présente une méthode alternative avec une stratégie de pivotage partiel. Enfin, une modification du critère d'arrêt des itérations de l'algorithme permet de construire un bon approximant vérifiant au choix, une précision ϵ donnée ou un rang r donné.

2.2.6.1 Approximation croisée avec pivot total

On considère une matrice M de taille $m \times n$ dont tous les coefficients sont supposés être calculés à l'avance. On cherche une approximation de la matrice sous une forme compressée et de rang approché r, de la forme :

$$\mathrm{M} \approx \mathrm{AB}^{\mathrm{T}},$$

où A et B sont respectivement de tailles $m \times r$ et $n \times r$. On notera par a_p la p^{e} colonne de A et de faon similaire, b_p la p^{e} colonne de B L'algorithme suivant, détaillé dans [BGH12] et [Beb00a], fournit une telle approximation en déterminant des approximations successives de rang 1 comme dans l'exemple du paragraphe 2.2.5.3.

Algorithme 10 Approximation en croix avec pivot total

1: **Pour** $q = 1, \dots, r$ **faire**
2: Déterminer l'élément maximal en valeur absolue :

$$(i_q, j_q) := \text{argmax}_{(i,j)} |\mathrm{M}_{ij}|, \delta := \mathrm{M}_{i_q, j_q}$$

3: **Si** $\delta = 0$ **Alors**
4: L'algorithme s'arrête avec le rang exact $q-1$ et l'approximation voulue est S_{q-1}.
5: **Sinon**
6: On calcule les éléments des vecteurs u_q et v_q :

$$\begin{aligned} (a_q)_i &:= \mathrm{M}_{i j_q} \quad , \quad i = 1, \dots, m \\ (b_q)_j &:= \mathrm{M}_{i_q j} / \delta \quad , \quad j = 1, \dots, n \end{aligned}$$

7: On soustrait l'approximation de rang 1 $a_q b_q^{\mathrm{T}}$:

$$\mathrm{M}_{ij} \leftarrow \mathrm{M}_{ij} - (a_q)_i (b_q)_j \quad , \quad i \in \{1, \dots, m\}, j \in \{1, \dots, n\}$$

8: **Fin Si**
9: **Fin Pour**

Propriétés de l'algorithme de pivot total Les trois propriétés suivantes ainsi que leurs démonstrations sont présentées dans [BGH12].

Lemme 2.28 (Décomposition exacte d'une matrice de rang r)**.** Soit M une matrice de taille $m \times n$ et de rang exactement r. Alors l'algorithme 10 fournit une matrice S_r ($\mathrm{S}_r = \sum_{p=1}^{r} u_p v_p^{\mathrm{T}}$) telle que

$$\mathrm{M} = \mathrm{S}_r.$$

Lemme 2.29 (Reproduction des lignes et colonnes pivots)**.** Soit une matrice M de taille $m \times n$ et de rang au moins r. On considère l'approximation S_r fournie par l'algorithme 10. Alors, pour tout couple de pivots $(i^\star, j^\star)$ sélectionnés par l'algorithme, on a

$$\mathrm{S}_r e_{j^\star} = \mathrm{M}|_{\mathscr{P}_{\mathrm{L}}, j^\star} \text{ et } (e_{i^\star})^{\mathrm{T}} \mathrm{S}_r = \mathrm{M}|_{j^\star, \mathscr{P}_{\mathrm{C}}}$$

Lemme 2.30. Soit une matrice M de taille $m \times n$ et de rang au moins r. On considère l'approximation S_r fournie par l'algorithme 10. Alors, cette approximation est de la forme suivante

$$\mathrm{S}_r = \sum_{p=1}^{r} a_p b_p^{\mathrm{T}} = \mathrm{M}|_{t,\hat{s}} . (\mathrm{M}|_{\hat{t},\hat{s}})^{-1} \mathrm{M}|_{\hat{t},s},$$

où $\hat{t}$ et $\hat{s}$ sont les respectivement les pivots pour les lignes et les colonnes. On remarque qu'une permutation des indices des pivots ne change pas le résultat. Ce résultat est à rapprocher de la décomposition 2.32 vu précédemment.

Remarque 2.31 (Conditionnement de la matrice de couplage). Dans la pratique, on constate que la matrice de couplage M| est très souvent mal conditionnnée. Ceci implique que l'on ne peut se servir de la fomule 2.30 pour effectuer des calculs. On peut néanmoins employer la pseudo-inverse de cette matrice plutôt que l'inverse.

Remarque 2.32 (Lien avec l'élimination de Gauss). L'algorithme 10 n'est autre que l'élimination de Gauss que l'on a écrite à la section 2.2.5.3.

Coût de l'algorithme Au cours de chaque étape $p = 1, \dots, r$ de l'algorithme 10, on doit

— déterminer les indices du pivot $(i^\star, j^\star)$ en $\mathcal{O}(mn)$ opérations,
— calculer les deux vecteurs u_p et v_p en $\mathcal{O}(m+n)$ opérations,
— mettre à jour le résidu contenu dans la matrice M de taille $m \times n$ en $\mathcal{O}(mn)$ opérations.

Ainsi, pour r étapes de cet algorithme, le coût devient en $\mathcal{O}(mnr)$ opérations ce qui est simialire au coût d'une méthode RRQR par exemple. La recherche des pivots ainsi que la mise à jour du résidu sont les points les plus onéreux de l'algorithme. Deux modifications vont nous permettre d'abaisser le coût de ces deux points. Pour contourner ce problème, on introduit une heuristique pour déterminer les pivots de façon à n'avoir besoin que de très peu de coefficients originaux de la matrice et on ne met à jour que certains éléments de la matrice.

On remarque que les lignes et les colonnes correspondantes aux pivots sélectionnés par l'algorithme 10 sont nulles après la mise à jour. On améliore ainsi cette étape en travaillant directement avec le résidu. En effet, les approximations a_p et b_p sont extraites de la matrice M otée des approximations de rang 1 successivement effectuées jusqu'à l'étape $p-1$. On peut donc se passer de mettre à jour l'intégralité de la matrice afin de réduire la complexité de l'étape de mise à jour à $\mathcal{O}(r^2(m+n))$ en modifiant la stratégie de recherche des pivots. On ne va utiliser que l'approximation a_p et/ou b_p afin de localiser un pivot.

La diminution du coût de la recherche de pivot est rendue possible grâce à une stratégie de pivot partiel. En effet, on maximise $|\mathrm{M}_{ij}|$ (M désignant soit la matrice lors de la première itération ou la mise à jour pour les itérations suivantes) seulement dans une seule direction et on garde l'autre fixée. Pour un indice de ligne quelconque et fixé $i^\star$ un indice de ligne quelconque, l'élément de plus gand module de la ligne $\mathrm{M}|_{i^\star \times s}$ peut alors être déterminée en n opérations. La paire de pivots $(i^\star, j^\star)$ ainsi construite ne correspond pas à un élément maximisant toutes les paires d'indices mais au moins une ligne de la matrice. L'algorithme poursuit en partant du pivot $j^\star$ précédemment trouvé et cherche un nouveau pivot de lignes $i^\star$ maximisant la colonne considérée. On effectue ainsi $\mathcal{O}((m+n)r)$ opérations pour la recherche des pivots au lieu de $\mathcal{O}((mn)r)$ opérations.

Les modifications décrites sont mises en place dans l'algorithme suivant lequel est présenté dans [BGH12],[ZVL05] ou encore [Liz14].

Algorithme 11 Approximation croisée avec pivot partiel

Données: M de taille $m \times n$, un rang visé r.
But: Une approximation de rang $q \leqslant r$ sous la forme $\mathrm{S}_r = \sum_{p=1}^{q} (a_p)(b_p)^{\mathrm{T}}, q \leqslant r$

$q = 1, \hat{t} = \emptyset, \hat{s} = \emptyset$
$i^\star = 1$
Tant que $q \leqslant r$ **faire**
  $\hat{t} \leftarrow \hat{t} \cup \{i^\star\}$
  $(b_q)_j = \mathrm{M}_{i^\star j} - \sum_{p=1}^{q-1} (a_p)_{i^\star} (b_j)_p$
  $j^\star = \mathrm{argmax}_{j \in s \setminus \hat{s}} |b_q|$, $\delta = (b_q)^\star_j$
  Si $\delta = 0$ **Alors**
    Si $t \setminus \hat{t} = \emptyset$ **Alors**
      retour
    Fin Si
    $i^\star = \min\{i \in \mathscr{P}_{\mathrm{L}} : i \notin \hat{t}\}$
  Sinon
    $\hat{s} \leftarrow \hat{s} \cup \{j^\star\}$
    $b_q \leftarrow b_q / \delta$
    $(a_q)_i = \mathrm{M}_{i j^\star} - \sum_{p=1}^{q-1} (b_q)^\star_j (a_q)_i$
    $i^\star = \mathrm{argmax}_{i \in t \setminus \hat{t}} |a_q|$
    $q \leftarrow q + 1$
  Fin Si
Fin Tant que
retour

Remarque 2.33 (Cas des matrices creuses). On remarque que dans le cas où $\delta = 0$ et que $\mathrm{Card}(t - \hat{t}) \neq \emptyset$, on doit déterminer une nouvelle ligne de départ $i^\star$ parmi celles restant. Ceci montre que l'algorithme d'approximation croisée avec pivot partiel est en général inefficace dans le cas d'une matrice creuse.

Coût de l'algorithme Les étapes de l'algorithme 11 ont les complexités suivantes :

— Calcul des lignes/colonnes de M en $\mathscr{O}((m+n)r)$ opérations.

— Détermination des pivots en $\mathscr{O}((m+n)r)$ opérations.

— Construction des approximations A et B en $\mathscr{O}((m+n)r^2)$ opérations.

La complexité totale de l'algorithme est donc de $\mathscr{O}((m+n)r^2)$ opérations.

Remarque 2.34 (Assemblage d'une ligne). Dans le cas de nos applications BEM, le coût d'assemblage d'une ligne est le point dominant dans les estimations précédentes. En effet, chaque coefficient d'une ligne et/ou colonne est le résultat d'un appel au calcul d'une matrice élémentaire dont on n'utilise pas tous les coefficients d'où un excès de calculs. L'algorithme se comporte comme si sa complexité était de $\mathscr{O}((m+n)r)$.

Lien avec la décomposition LU L'algorithme ACA peut être interprété comme une factorisation LU révélant le rang. Sans perdre de généralité, supposons qu'à chaque étape q de l'algorithme ACA, on a $i^\star = q$ et $j^\star = q$. Alors on a

$$\mathrm{E}_{q+1} = (\mathrm{I} - \delta_{q+1} \mathrm{E}_q e_{q+1} e_{q+1}^{\mathrm{T}}) \mathrm{E}_q = \mathrm{L}_{q+1} \mathrm{E}_q,$$

où L_q est de la forme

$$L_q = \begin{bmatrix} 1 & & & & & & \\ & \ddots & & & & & \\ & & 1 & & & & \\ & & & 0 & & & \\ & & & -\frac{e_{q+1}^T E_{q-1} e_q}{e_q^T E_{q-1} e_q} & 1 & \\ & & & \vdots & & \ddots & \\ & & & -\frac{e_m^T E_{q-1} e_q}{e_q^T E_{q-1} e_q} & & & 1 \end{bmatrix} \tag{2.38}$$

L'algorithme produit alors une décomposition LU avec un pivotage des colonnes pour approcher la matrice M. En effet, la matrice 2.38 ci-dessus ne diffère d'une matrice de Gauss qu'en le coefficient (q,q).

2.2.6.2 Approximation croisée adaptative, *ACA*

Dans la pratique, on s'intéresse au problème de trouver un rang numérique r lorsque l'on se donne une précision ϵ. Un point crucial est l'estimation de l'erreur d'approximation. Une bonne heuristique est d'obtenir une estimation du reste $M - S_p$ sous la forme d'une approximation de rang 1 car à une étape fixée, c'est le type d'approximation que l'on calcule. Ce choix d'approximation nous permet également d'approcher la norme du résidu de la façon suivante

$$\|M - S_p\| \lesssim \|M - S_{p-1}\| \approx \|S_p - S_{p-1}\| = \|u_p\| \|v_p\|. \tag{2.39}$$

Les deux vecteurs u_p et v_p étant construits durant l'approximation, le calcul de leurs normes n'entraîne pas de surcoût. Ainsi, les quantités

$$\epsilon_{abs}(p) := \|u_p\|_2 \|v_p\|_2, \tag{2.40}$$

$$\epsilon_{rel}(p) := \|u_p\|_2 \|v_p\|_2 / \|u_1\|_2 \|v_1\|_2. \tag{2.41}$$

peuvent être utilisée pour obtenir une estimation de l'erreur $\|M - S_p\|_2$ (absolue ou relative) en norme euclidienne. Il est possible de travailler en norme de Frobénius avec les quantités suivantes :

$$\epsilon_{abs}^F(p) := \|u_p\|_2 \|v_p\|_2, \tag{2.42}$$

$$\epsilon_{rel}^F(p) := \|u_p\|_2 \|v_p\|_2 / \sqrt{\sum_{q=1}^{p} \|u_q\|_2^2 \|v_q\|_2^2}. \tag{2.43}$$

Pour une précision donnée ϵ, on remplace alors le critère d'arrêt "$p \leqslant r$" de l'algorithme 11 par l'un des estimateurs suivants

$$\epsilon_{abs}^F(p-1) \leqslant \epsilon \text{ ou } \epsilon_{rel}^F(p-1) \leqslant \epsilon. \tag{2.44}$$

Cette méthode est qualifiée d'adaptative car on détermine le rang au fur et à mesure de l'approximation. Cette méthode est l' *Adaptive Cross Approximation* que l'on retrouve dans la littérature (voir par exemple [BGH12],[Beb00b],[Mes11],[Sha08]).

Coût de l'algorithme Le nombre d'opérations est de $\mathcal{O}((m+n)r^2)$ opérations. Cependant, il ne prend pas en compte le prix d'assemblage d'une ligne/colonne de M. Dans le cas d'une matrice provenant de la BEM, ce calcul est onéreux car l'assemblage est effectué à l'aide de matrice élémentaires dont on n'utilise pas tous les coefficients. C'est ce temps qui domine l'algorithme et sa complexité est donc de $\mathcal{O}((m+n)r)$ opérations.

Contre-exemple pour l'algorithme ACA Le contre-exemple suivant, décrit en détail (dans le cadre d'une approximation d'un potentiel de double couche en BEM) dans [BGH12] et [Beb00b] montre que même pour une fonction régulière, il se peut que $\epsilon(r) \to 0$ bien que l'erreur relative ou absolue soit en $\mathcal{O}(1)$. Considérons la matrice M de taille $2m \times 2n$ par blocs suivante :

$$\mathrm{M} = \begin{bmatrix} \mathrm{M}_{11} & 0 \\ 0 & \mathrm{M}_{22} \end{bmatrix},$$

où M_{11} et M_{22} sont de tailles $m \times n$.

On note que l'algorithme ACA ci-dessus commence avec un indice de ligne correspondant au bloc M_{11} et va produire un pivot non nul dans ce même bloc. Dès lors, la première approximation de rang 1 $a_1 b_1^{\mathrm{T}}$ est nulle sur les autres blocs de la matrice et ainsi le reste possède la même structure que la matrice M. Avec l'heuristique présentée pour le choix de la prochaine ligne, on observe que tous les pivots seront des éléments de M_{11} et que l'algorithme ne voit pas le bloc M_{22}. L'estimateur d'erreur sera bon pour l'approximation du premier bloc mais faux de manière globale de telle sorte que la norme du reste sera minorée par la norme du bloc non traité M_{22}.
Plusieurs variantes du ACA ont été présentées pour pallier ce défaut, notament dans [BGH12] où y est présenté une autre heuristique, nommée ACA+. ACA+ est basé sur une l'utilisation d'une ligne de référence qui aide à déterminer la nouvelle ligne pivot en plus de l'heuristique de l'algorithme ACA. Néanmoins, [Beb00b] remarque que le choix de cette ligne de référence influence l'approximation obtenue et empêche d'envisager une preuve de la convergence.

2.2.6.3 Application des algorithmes ACA sur un exemple

On effectue une comparaison des algorithmes ACA avec pivot total, ACA avec pivot partiel et la décomposition en valeurs singulières sur deux exemples où le profil de décroissance des valeurs singulières sont différents et connus.

Exemple 2.35 (Matrice de Hilbert)**.** On considère la matrice de Hilbert M_1 de taille $n \times n$ dont les coefficients sont donnés par

$$(\mathrm{M}_1)_{ij} = \frac{1}{i+j-1}. \tag{2.45}$$

Exemple 2.36 (Matrice exponentielle décroissante)**.** On considère la matrice M_2 de taille $n \times n$ dont les coefficients sont donnés par

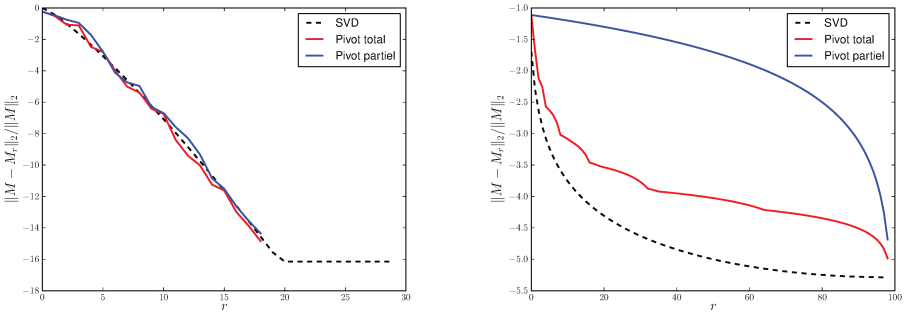
$$(\mathrm{M}_2)_{ij} = \exp\left(\frac{-\gamma|i-j|}{n}\right), \gamma = 0.01. \tag{2.46}$$

Dans le cas du premier exemple, le profil de décroissance des valeurs singulières attendu est une décroissance rapide tandis que le second exemple possède une décroissance lente. Pour les deux algorithmes étudiés, on calcul l'erreur relative d'approximation

de rang r fixé donnée par

$$\frac{\|\mathrm{M}-\mathrm{M}_r\|_2}{\|\mathrm{M}\|_2} = \frac{\|\mathrm{M}-(\mathrm{AB}^{\mathrm{T}})\|_2}{\|\mathrm{M}\|_2}$$

que l'on compare à l'erreur théorique $\sigma_r(\mathrm{M})/\sigma_1(\mathrm{M})$ fournie par la décomposition SVD.



(a) Matrice de Hilbert 2.35 (b) Matrice exponentielle 2.36

FIGURE 2.4 – Représentation graphique de l'erreur d'approximation de rang r fixe (échelle logarithmique pour les ordonnées) ;dans les deux exemples, $n = 100$

L'algorithme avec pivot total 10 fournit une bonne approximation de rang r dans les deux cas. On remarque cependant que l'algorithme 11 renvoie une approximation de qualité dans le cas d'un profil à décroissance rapide. En toute généralité, on ne peut se contenter d'effectuer r itérations de l'algorithme pour obtenir une approximation de rang r. Il est donc bien plus préférable d'avoir une estimation de l'erreur intégrée à l'algorithme quitte à sélectionner plus de pivots. Toutefois, il est possible d'éliminer les redondances en cas de surestimation du rang à l'aide d'une technique de recompression. La décomposition en valeurs singulières approchée décrite au paragraphe 1.2.2.2 le permet. On notera également la similarité avec la décomposition en valeurs singulières randomisée.Plusieurs heuristiques sont présentes dans la littérature afin de s'assurer de la convergence. [LMPLH09] propose d'effectuer quelques itérations supplémentaires de l'algorithme afin de s'assurer que la convergence a été atteinte. Le critère de convergence étant un point onéreux de l'algorithme, même sous sa forme adaptative, [HUR15] propose de déterminer l'erreur à l'aide d'une estimation statistique de la norme de Frobénius du résidu à chaque étape.

Application à des matrices issues de la BEM À ce stade, l'algorithme ACA paraît être particulièrement adapté à des matrices dont les valeurs singulières ont une décroissance rapide. En réalité, cet algorithme a été introduit dans le cadre de l'approximation d'une fonction de deux variables en tant que somme de produits de fonctions d'une variable. On détaillera ultérieurement ce lien. Les références [Beb00b](chapitre 3) et [BGH12](chapitre 4) font état de preuves de convergence de l'algorithme dans certains cas particuliers. En effet, l'algorithme ACA y est employé dans le contexte de matrices issues de la méthode des éléments finis de frontière et on y exploite la régularité du noyau de Green sous-jacent. De plus, la connaissance d'informations géométriques supplémentaires permettent d'affiner l'algorithme ACA, notamment le choix du premier pivot pour les lignes.

2.3 Méthodes analytiques

Dans toute la suite de ce chapitre, et sauf mention contraire, $G(x,y)$ désigne le noyau de Green de l'équation de Laplace défini par

$$G(x,y) = \frac{1}{|x-y|}. \tag{2.47}$$

Cette fonction possède une singularité sur la diagonale *i.e* l'ensemble des points $\{(x,y) : x = y\}$. En dehors de cet ensemble, cette fonction est analytique et on considère deux domaines X et Y de $\mathbb{R}^3$ dans lesquels vivent x et y.

Forme de l'approximation cherchée On recherche une approximation $\tilde{G}$ du noyau de la forme 2.3 :

$$\tilde{G}(x,y) = \sum_{q=0}^{r-1} u_q(x) v_q(y).$$

On introduit la notion de produit tensoriel de fonctions à l'aide de la définition suivante.

Définition 2.37 (produit tensoriel de fonctions)**.** Pour un entier $d \geqslant 1$, on considère deux fonctions $u, v : \mathbb{R}^d \mapsto \mathbb{R}$. On appelle le produit $h(x,y) = u(x)v(y)$ un produit tensoriel de fonctions. On trouve le terme de *functionnal skeleton* dans la littérature (par exemple [Beb00b]).

L'écriture 2.3 est donc une somme de produits tensoriels de fonctions.

Exemple d'application On considère pour illustrer les décompositions obtenues la matrice mentionnée par l'équation 2.2 :

$$M_{ij} = \int_\Gamma \int_\Gamma G(x,y)\Phi_j(y)\Phi_i^t(x)d\Gamma(x)d\Gamma(y). \tag{2.48}$$

où Γ est une surface fermée de $\mathbb{R}^3$ provenant de la modélisation de l'équation de Laplace par la méthode des éléments finis de frontières. Les fonctions Φ_i^t et Φ_j sont des fonctions de base telle que leurs supports respectifs sont inclus dans Γ. Il s'agit de la discrétisation de l'opérateur de simple couche,

$$\begin{aligned} \mathscr{S} \quad : \quad & H^{-1/2}(\Gamma) \mapsto H^{+1/2}(\Gamma) \\ & \mu \mapsto \mathscr{S}[\mu](x) = \int_\Gamma G(x,y)\mu(y)d\Gamma(y), x \in \Gamma. \end{aligned}$$

par la méthode des éléments finis de frontière. C'est à cette matrice à laquelle on fait référence par la dénomination de matrice de simple couche. Si requis, on effectuera l'hypothèse que $X, Y \subset \Gamma$.

Noyau asymptotiquement lisse On effectue l'hypothèse de noyau asymptotiquement lisse suivante :

Définition 2.38 (Noyau asymptotiquement lisse)**.** Soit $f(x,y)$ une fonction de $\mathbb{R}^3 \times \mathbb{R}^3$ à valeurs complexes. On dit que f est asymptotiquement lisse s'il existe des constantes C_1,C_2 et un degré de singularité $\sigma \geqslant 0$ tels que pour deux multi-indices α et β de $\mathbb{N}_0^3$ on a

$$\left|\partial_x^\alpha \partial_y^\beta f(x,y)\right| \leqslant (\alpha+\beta)! C_1 (C_2 \|x-y\|)^{-|\alpha+\beta|-\sigma}. \tag{2.49}$$

Cette condition permet d'obtenir un contrôle des dérivées d'ordres supérieures du noyau. Le noyau de l'équation de Laplace vérifie cette propriété avec $C_1 = C_2 = 1$ et $\sigma = 0$. Dans le cas du noyau oscillant, on a l'estimation suivante dont la borne dépend du nombre d'onde. On se limite à présenter les résultats connus uniquement dans le cas du noyau de Green de l'équation de Laplace même si certaines méthodes restent valables pour le noyau de Helmholtz $G(x,y) = e^{ik|x-y|}/|x-y|$. La principale difficulté liée à ce noyau est qu'une estimation du type 2.49 dépend du nombre d'onde k et l'étude de ce noyau sera effectuée séparément.

Proposition 2.39 (Noyau oscillant pour Helmholtz)**.** *On considère le noyau de Green de l'équation de Helmholtz défini par*

$$G(x,y) = \frac{e^{ik|x-y|}}{|x-y|}, \tag{2.50}$$

où k est le nombre d'onde. Alors on a l'estimation suivante sur les dérivées de G*,*

$$\left|\partial_x^\alpha \partial_y^\beta G(x,y)\right| \leqslant (\alpha+\beta)! C_1 (1+k\|x-y\|)^{|\alpha+\beta|} (C_2\|x-y\|)^{-|\alpha+\beta|-\sigma}. \tag{2.51}$$

Conditions d'admissibilité On considère deux domaines X et Y de $\mathbb{R}^3$ que l'on souhaite séparer afin de s'éloigner de la singularité du noyau.

Définition 2.40 (Diamètre et distance)**.** Pour les deux domaines X et Y, on définit leurs diamètres et leur distance par

$$\begin{aligned} \operatorname{diam}(X) &= \max_{x,x'\in X} |x-x'|, \\ \operatorname{diam}(Y) &= \max_{y,y'\in Y} |y-y'|, \\ \operatorname{dist}(X,Y) &= \min_{x\in X, y\in Y} |x-y|. \end{aligned}$$

Dans la pratique il est souvent plus aisé de travailler avec des boîtes englobantes dans $\mathbb{R}^3$. On définit une boîte englobante par

Définition 2.41 (Boîte englobante)**.** Soit $X \subset \mathbb{R}^3$. On définit la boîte englobante B_X de X par

$$B_X = [a_1,b_1] \times [a_2,b_2] \times [a_3,b_3],$$

où pour tout $x = (x_1,x_2,x_3) \in X$, les intervalles ci-dessus vérifient pour $i = 1,2,3$:

$$a_i \leqslant x_i \leqslant b_i.$$

On peut alors calculer le diamètre d'une boîte ainsi que la distance entre deux boîtes en $\mathcal{O}(1)$ opérations. Ces quantités sont des majorants de diam(X) et dist(X,Y) respectivement.

Lemme 2.42 (Diamètre d'une boîte)**.** Soit $B_X = [a_1,b_1] \times [a_2,b_2] \times [a_3,b_3]$, on définit le diamètre $\operatorname{diam}(B_X)$ d'une boîte

$$\operatorname{diam}(B_X) = \left(\sum_{i=1}^{3} (b_i-a_i)^2\right)^{1/2} \tag{2.52}$$

Lemme 2.43 (Distance entre deux boîtes)**.** Soient $B_X = [a_1, b_1] \times [a_2, b_2] \times [a_3, b_3]$ et $B_Y = [c_1, d_1] \times [c_2, d_2] \times [c_3, d_3]$, on définit la distance $\text{dist}(B_X, B_Y)$ entre les boîtes par

$$\text{dist}(B_X, B_Y) = \left(\sum_{i=1}^{3} \max(0, a_i - d_i)^2 + \max(0, c_i - b_i)^2 \right)^{1/2} \tag{2.53}$$

On supposera que les domaines vérifient la condition suivante, dite condition d'admissibilité. Cette condition réalise la séparation en champ proche et champ lointain pour des noyaux asymptotiquement lisses.

Définition 2.44 (Critères d'admissibilité)**.** Soit $\eta > 0$ un paramètre fixé. Deux domaines X et Y sont dits faiblement admissibles si

$$\min(\text{diam}(X), \text{diam}(Y)) \leqslant \eta \text{dist}(X, Y). \tag{2.54}$$

On parlera d'admissibilité forte si l'on a

$$\max(\text{diam}(X), \text{diam}(Y)) \leqslant \eta \text{dist}(X, Y). \tag{2.55}$$

Ces inégalités sont également employées pour les boîtes englobantes en substituant B_X et B_Y à X et Y respectivement. Si les boîtes englobantes B_X et B_Y sont admissibles, il s'ensuit l'admissibilité des domaines X et Y.

Les interactions admissibles (champ lointain) sont de bonnes candidates pour l'obtention d'une approximation à variables séparées tandis que les interactions inadmissibles (champ proche) ne peuvent être approchées efficacement de la sorte.

2.3.1 Développement de Taylor du noyau

On rappelle que l'on cherche une approximation à variables séparées et de rang fini du noyau G(x,y) avec $x \in X \subset \Gamma$ et $y \in Y \subset \Gamma$, Γ une surface fermée de $\mathbb{R}^d$. En supposant le noyau suffisament régulier, on montre dans cette partie que le développement en série de Taylor du noyau fournit la décomposition voulue. En se plaçant dans le cas de l'hypothèse de noyau asymptotiquement lisse et en imposant une certaine condition entre X et Y, on peut également majorer le reste du développement de Taylor et obtenir une estimation du rang de l'approximation construite.

2.3.1.1 Formule de Taylor-Cauchy en dimension 3

On considère le noyau $G(x, y) = 1/|x - y|$ en dimension $d = 3$. La fonction G étant analytique en dehors de la diagonale $x = y$, on le développement de Taylor à l'ordre p avec reste de Cauchy autour d'un point $y_0 \in Y \subset \mathbb{R}^3$:

$$\begin{aligned} G(x, y) &= \sum_{|\alpha|=0}^{p-1} \frac{1}{\alpha!} (y - y_0)^\alpha \partial_y^\alpha G(x, y_0) + \frac{1}{p!} (y - y_0) \partial_y^p G(x, \tilde{y}) && (2.56) \\ &= s_{n_{p-1}}(x, y) + e_p(x, y) && (2.57) \end{aligned}$$

avec $\tilde{y} \in Y$ et n_{p-1} le nombre de termes dans la somme. La somme indéxée par le multi-indice α correspond à l'approximation souhaitée. On aura alors une approximation convenable si le terme de reste est contrôlé lorsqu'on augmente l'ordre dans le développement.

2.3.1.2 Majoration de l'erreur

Noyau asymptotiquement lisse On rappelle que l'hypothèse de régularité sur le noyau $G(x,y)$ implique qu'il existe des constantes c_1,c_2 et un degré de dégénérescence σ tel que pour tout multi-indice $\beta \in \mathbb{N}^3$ on a

$$|\partial_y^\beta G(x,y)| \leqslant c_1 |\beta|! c_2^{|\beta|!} |x-y|^{-\sigma-|\beta|}. \tag{2.58}$$

Pour $y_0, \tilde{y} \in Y$, on a les inégalités suivantes pour tout $x \in X$ et $y \in Y$,

$$\begin{aligned} |y-y_0| &\leqslant \operatorname{diam}(Y), \\ |x-\tilde{y}| &\geqslant \operatorname{dist}(X,Y). \end{aligned}$$

Alors le reste $e_p(x,y)$ du développement de Taylor peut être majoré de la façon suivante

$$\begin{aligned} |e_p(x,y)| &\leqslant \frac{1}{|\alpha|!} d^{|\alpha|} |y-y_0|^\alpha |\partial_y^\alpha G(x,\tilde{y})| \\ &\leqslant \frac{1}{|\alpha|!} |\alpha|! d^{|\alpha|} \frac{|y-y_0|^{|\alpha|}}{|x-\tilde{y}|^{|\alpha|}} |x-\tilde{y}|^{-\sigma} c_1 c_2^{|\alpha|} \\ &\leqslant d^{|\alpha|} \frac{\operatorname{diam}(Y)^{|\alpha|}}{\operatorname{dist}(X,Y)^{|\alpha|}} \operatorname{dist}(X,Y)^{-\sigma} c_1 c_2^{|\alpha|} \\ &\leqslant c_1 \operatorname{dist}(X,Y)^{-\sigma} (\eta d c_2)^{|\alpha|} \end{aligned}$$

Condition d'admissibilité On assure la décroissance exponentielle du reste en imposant une condition liant les domaines X et Y : c'est la condition d'admissibilité. Ainsi, on impose la condition suivante,

$$\operatorname{diam}(Y) \leqslant \eta \operatorname{dist}(X,Y), \tag{2.59}$$

avec $\eta d c_2 < 1$. Alors le reste est majoré de la sorte

$$|e_p(x,y)| \leqslant c_1 \operatorname{dist}(X,Y)^{-\sigma} (\eta d c_2)^{|\alpha|}, \tag{2.60}$$

et l'errreur commise est donc $\epsilon := c_1 \operatorname{dist}(X,Y)^{-\sigma} (\eta d c_2)^{|\alpha|}$.

Remarque 2.45 (Développement en x). On peut effectuer le même développement suivant la variable x. Dans ce cas, on impose la condition

$$\operatorname{diam}(X) \leqslant \eta \operatorname{dist}(X,Y). \tag{2.61}$$

2.3.1.3 Approximation d'une matrice BEM

On peut utiliser le développement 2.56 pour approcher la matrice donnée par 2.2 en remplaçant le noyau par son approximation, y_0 étant fixé dans Y,

$$\begin{aligned} M_{ij} &= \int_\Gamma \int_\Gamma G(x,y) \Phi_j(y) \Phi_i^t(x) d\Gamma(x) d\Gamma(y) \\ &\simeq \int_\Gamma \int_\Gamma s_{n_{p-1}}(x,y) \Phi_j(y) \Phi_i^t(x) d\Gamma(x) d\Gamma(y) \\ &= \int_\Gamma \int_\Gamma \left(\sum_{|\alpha|=0}^{p-1} \frac{1}{\alpha!} (y-y_0)^\alpha \partial_y^\alpha G(x,y_0) \right) \Phi_j(y) \Phi_i^t(x) d\Gamma(x) d\Gamma(y) \\ &= \sum_{|\alpha|=0}^{p-1} \left(\int_\Gamma \frac{1}{\alpha!} (y-y_0)^\alpha \Phi_j(y) d\Gamma(y) \right) \left(\int_\Gamma \partial_y^\alpha G(x,y_0) \Phi_i^t(x) d\Gamma(x) \right) \end{aligned}$$

En posant

$$\begin{aligned} A_{i,|\alpha|} &= \int_\Gamma \partial_y^\alpha G(x,y_0)\Phi_i^t(x)d\Gamma(x), \\ B_{j,|\alpha|} &= \int_\Gamma \frac{1}{\alpha!}(y-y_0)^\alpha \Phi_j(y)d\Gamma(y), \end{aligned}$$

on obtient la décomposition de rang faible $\tilde{M} = AB^T$ voulue.

Choix du nombre de termes En considérant $x \in X$ fixé, on remarque que la somme est un polynôme en y de degré au plus $p-1$ en dimension 3. On désigne par $\mathbb{P}^3_{p-1}$ l'ensemble de ces polynômes. Ainsi, le nombre de termes dans la somme n_{p-1} est borné par le nombre de monômes y^α linéairement indépendants dans $\mathbb{P}^3_{p-1}$ soit

$$n_{p-1} \approx c_3.p^3, \tag{2.62}$$

c_3 étant une constante.

Cas de la matrice du simple couche La matrice définie par 2.2 pour le noyau de l'équation de Laplace $G(x,y) = 1/|x-y|$ est la matrice de l'opérateur de simple couche. Dans ce cas, les constantes dans l'hypothèse de noyau asymptotiquement lisse valent $c_1 = c_2 = 1$ et $\sigma = 0$ et ainsi, le développement de Taylor du noyau à l'ordre p fournit une approximation de rang borné par $c_3 p^3$ avec une erreur de $(3\eta)^{-p}$ pour $\eta < 1/3$. Pour une précision ϵ fixée, l'ordre d'interpolation p est donné par

$$p = \mathcal{O}(-log_{10}(\epsilon)), \tag{2.63}$$

et le rang r_ϵ à la précison ϵ vérifie donc

$$r_\epsilon = \mathcal{O}(-log^3_{10}(\epsilon)). \tag{2.64}$$

Le développement de Taylor fournit une approximation dont on peut borner le rang en fonction de la précision voulue en se plaçant dans l'hypothèse de noyau assymptotiquement lisse 2.58 et en imposant une condition d'admissibilité 2.59. Cependant, cette approximation recquiert le calcul des dérivées successives du noyau et il s'agit en général d'un calcul couteux et pénible. Cependant, on est assuré de l'existence d'une approximation de rang faible avec une estimation d'erreur ainsi qu'une estimation du rang de l'approximation en fonction de la précision.

2.3.2 Approximations croisées du noyau de Green

2.3.2.1 Position du problème

Dans cette partie, $G(x,y)$ désigne le noyau de Green de l'équation de Laplace avec $x \in X$ et $y \in Y$ deux domaines de $\mathbb{R}^3$. On suppose de plus que ces deux domaines satisfont la condition d'admissibilité suivante

$$\min\{\mathrm{diam}(X), \mathrm{diam}(Y)\} \leqslant \eta \mathrm{dist}(X,Y), \tag{2.65}$$

où η est un paramètre fixé. Pour une fonction asymptotiquement lisse, cette condition suffit à garantir l'existence d'une approximation de rang faible via sa série de Taylor. On

cherche alors à obtenir l'approximation à variables séparées sans avoir à calculer des dérivées du noyau par une méthode d'approximations croisées. Pour ce faire, on écrit le noyau sous la forme suivante

$$G(x,y) = s_r(x,y) + e_r(x,y), \tag{2.66}$$

où $s_r(x,y)$ est une somme de r produits tensoriels de fonctions et e_r est un résidu tel que $|e_r(x,y)| \leqslant \epsilon$, $\epsilon \to 0$ quand $r \to \infty$.

2.3.2.2 Algorithme fonctionnel

On construit itérativement les fonctions $\{s_q\}$ et $\{e_q\}$ de la sorte

$$\begin{aligned} s_0(x,y) &= 0, \\ e_0(x,y) &= G(x,y), \end{aligned}$$

et pour $q \geqslant 0$,

$$e_{q+1}(x,y) = e_q(x,y) - \delta_{q+1} e_q(x,y_{j_{q+1}}) e_q(x_{i_{q+1}},y), \tag{2.67}$$

$$s_{q+1}(x,y) = s_q(x,y) + \delta_{q+1} e_q(x,y_{j_{q+1}}) e_q(x_{i_{q+1}},y), \tag{2.68}$$

où $\delta_{q+1} = (e_q(x_{i_{q+1}},y_{j_{q+1}}))^{-1}$ et $x_{i_{q+1}}$ et $y_{j_{q+1}}$ choisi à chaque étape q tels que $e_q(x_{i_{q+1}},y_{j_{q+1}}) \neq 0$

La première itération de l'algorithme consiste à déterminer un pivot non nul $\delta_1 = G(x_{i_1},y_{j_1})$ et on considère le *functionnal skeleton* $s_1(x,y)$ défini par

$$\begin{aligned} s_1(x,y) &= G(x,y_{j_1})(G(x_{i_1},y_{j_1}))^{-1} G(x_{i_1},y) \\ &= u_1(x) v_1(y), \end{aligned}$$

avec $u_1(x) = G(x,y_{j_1})/\sqrt{|\delta|}$ et $v_1(y) = \mathrm{sgn}(\delta).G(x_{i_1},y)/\sqrt{|\delta|}$.

La seconde itération de l'algorithme revient à effectuer la même procédure à la fonction $(G - s_1)(x,y)$ et obtenir une autre approximation associée à un pivot $(x_{i_2},y_{j_2}) \in X \times Y$. Quand le nombre de *functionnal skeleton* augmente, la fonction $s_q(x,y)$ interpole le noyau aux pivots (x_{i_l},y_{j_l}), $(l = 1,\ldots,q)$ et le résidu $e_q(x,y)$ accumule les zéros

Lemme 2.46 ([Beb00b])**.** Pour $1 \leqslant l \leqslant q$ et $x \in X$, on a $r_q(x,y_{j_l}) = 0$. On obtient le même résultat en échangeant les rôles des variables x et y.

On introduit la notation $G(x,[y]_n)$ pour désigner le vecteur de taille n suivant dont les coefficients sont des fonctions de x pour les pivots $y_{j_q}, q = 1,\ldots,n$ fixé

$$G(x,[y]_n) = \begin{bmatrix} G(x,y_{j_1}) \\ \vdots \\ G(x,y_{j_n}) \end{bmatrix}.$$

Alors, à l'étape q de l'algorithme, on a l'approximation suivante du noyau ([Beb00b],lemme 3.),

$$G(x,y) \simeq G(x,[y]_q)^{\mathrm{T}} (M_q)^{-1} G([x]_q,y), \tag{2.69}$$

où M_q est de taille $q \times q$ et définie par

$$(M_q)_{\alpha\beta} = G(x_{i_\alpha},y_{j_\beta}),$$

pour $\alpha,\beta = 1,\ldots,q$. Par analogie avec les approximations extraites traitées dans les méthodes algébriques, on appelle $(M_q)^{-1}$ une matrice de couplage.

2.3.2.3 Convergence de la méthode

On considère l'entier n_{p-1} défini par 2.62 pour le développement de Taylor et on effectue n_{p-1} étapes de l'algorithme précédent. Par analogie avec le développement de Taylor, [Beb00b] relie le résidu $e_{n_{p-1}}$ au reste d'une interpolation polynomiale dans $\mathbb{P}^3_p$. Cependant, l'interpolation polynomiale en dimension supérieure n'est pas aussi aisée que dans le cas réel ([SX95]). Son existence est garantie mais l'unicité dépend de la distribution des points d'interpolation. On obtient l'unicité de l'interpolation s'il n'existe pas de polynôme de $\mathbb{P}^3_p$ nul en chaque nœud d'interpolation.

Théorème 2.47 (Convergence du résidu $e_r(x,y)$,[Beb00b])**.** *Pour chaque étape q, un pivot x_{i_q} est choisi tel que*

$$|e_{q-1}(x_{i_q}, y_{j_q})| \geqslant |e_{q-1}(x, y_{j_q})|, x \in \mathrm{X}. \tag{2.70}$$

Alors le noyau $\mathrm{G}(x,y)$ *asymptotiquement lisse et les fonctions* $s_q(x,y)$ *et* $e_q(x,y)$ *définies par 2.67 vérifient pour* $x \in \mathrm{X}$, $y \in \mathrm{Y}$,

$$\mathrm{G}(x,y) \quad = \quad s_{n_{p-1}}(x,y) + e_{n_{p-1}}(x,y) \tag{2.71}$$

avec

$$s_{n_{p-1}}(x,y) \quad = \quad \mathrm{G}(x,[y]_{n_{p-1}})^{\mathrm{T}}(\mathrm{M}_{n_{p-1}})^{-1}\mathrm{G}([x]_{n_{p-1}},y). \tag{2.72}$$

et

$$|e_{n_{p-1}}(x,y)| \leqslant c_{p-1}\mathrm{dist}^{-\sigma}(\mathrm{X},\mathrm{Y})\eta^p, \tag{2.73}$$

et c_{p-1} *ne dépend pas de* η *mais seulement des points* $\{y_{j_q}\}_{q=1}^{n_{p-1}}$.

L'approximation $s_{n_{p-1}}(x,y)$ est une approximation à variables séparées et de rang n_{p-1} du noyau de la forme voulue. On note que la décroissance dépend du facteur d'admissibilité η associé aux domaines X et Y.

Remarque 2.48 (Choix des pivots). Le choix des pivots est crucial pour s'assurer de la convergence de l'algorithme. Les pivots $\{y_{j_q}\}_{q=1}^{n_{p-1}}$ servent de nœuds d'interpolation pour un polynôme minimisant le résidu. Plusieurs stratégies sont explicitées dans [Beb00b]. Une méthode heuristique et similaire à l'algorithme ACA+ est présentée et consiste à observer l'influence du pivots choisis sur d'autres produits tensoriels de fonctions. Une méthode plus précise consiste à s'assurer que les pivots choisis garantissent que le déterminant de Vandermonde issu de l'interpolation polynomiale du résidu ne s'annule pas.

2.3.2.4 Lien avec l'approximation croisée ACA

Pour deux nuages de points de $\mathbb{R}^3$, $\mathrm{X} = \{x_i\}_{i=1}^m$ et $\mathrm{Y} = \{y_i\}_{j=1}^n$, on note $\mathrm{B_X}$ et $\mathrm{B_Y}$ les deux boîtes englobantes contenant respectivement les nuages X et Y. On suppose que les deux boites englobantes satisfont la condition d'admissibilité suivante

$$\min\{\mathrm{diam}(\mathrm{B_X}), \mathrm{diam}(\mathrm{B_Y})\} \leqslant \eta \mathrm{dist}(\mathrm{B_X}, \mathrm{B_Y}). \tag{2.74}$$

On considère la matrice M de taille $m \times n$ définie par

$$\mathrm{M}_{ij} = \mathrm{G}(x_i, y_j), x_i \in \mathrm{X}, y_j \in \mathrm{Y}. \tag{2.75}$$

Cette matrice est l'équivalent d'une discrétisation des intégrales avec un point de Gauss dans 2.2, aussi appelée matrice de collocation. L'application de l'algorithme algébrique

ACA n'est autre que la discrétisation de l'algorithme 2.67 en utilisant comme pivots les points x_i et y_j. On rappelle que l'algorithme ACA fournit la décomposition de rang r suivante, avec les notations $t = \{1,\dots,m\}$ et $s = \{1,\dots,n\}$,

$$\mathrm{M} = \mathrm{M}|_{t\times\hat{s}}(\mathrm{M}|_{\hat{t}\times\hat{s}})^{-1}\mathrm{M}|_{\hat{t}\times s}, \tag{2.76}$$

avec $\hat{t} = \{i_q : q = 1,\dots,r\}$ et $\hat{s} = \{j_q : q = 1,\dots,r\}$ deux sous-ensembles de t et s respectivement. L'écriture 2.76 n'est autre que la discrétisation de la décomposition du noyau obtenue en 2.72 où

$$\begin{aligned}(\mathrm{M}|_{\{i\}\times\hat{s}}) &= \mathrm{G}(x_i,[y_{j_q}]_r)^{\mathrm{T}} \quad , \quad i = 1,\dots,m\\ (\mathrm{M}|_{\hat{t}\times\{j\}}) &= \mathrm{G}([x_{i_q}]_r,y_j) \quad , \quad j = 1,\dots,n\end{aligned}$$

On remarque que la matrice $\mathrm{M}|_{\hat{t}\times\hat{s}}$ est exactement la même matrice que la matrice de couplage intervenant dans 2.69 en prenant $q = r$.

Remarque 2.49 (Convergence pour l'algorithme ACA). Ce lien est à la base des preuves de convergence de l'algorithme ACA avec pivot total. En effet, Bebendorf a initialement introduit l'algorithme 2.67 dans le cadre de l'approximation de noyau asymptotiquement lisse. La discrétisation de cet algorithme n'est autre que l'algorithme de pivot total présenté usuellement dans un cadre algébrique. Bebendorf prouve ainsi la convergence de l'algorithme pour une matrice de collocation à l'aide d'une interpolation polynomiale du résidu. Rjasanow et Bebendorf ont alors étendu le résultat pour des matrices de discrétisation de type Galerkin ou Nyström.

2.3.2.5 Approximation du noyau dans $\mathrm{B_X}\times\mathrm{B_Y}$

Sous réserve d'avoir un nombre suffisant de points x_i et y_j dans X et Y, on peut obtenir une approximation du noyau dans $\mathrm{B_X}\times\mathrm{B_Y}$ en utilisant l'algorithme ACA. L'application de cet algorithme nous fournit une approximation de la forme AB^{T} de M. Quitte à renvoyer également la liste des pivots de lignes et de colonnes utilisés, on peut alors construire une approximation de rang fini à variables séparées du noyau dans $\mathrm{B_X}\times\mathrm{B_Y}$ par

$$\mathrm{G}(x,y) \simeq \sum_{l=1}^{r}\left(\sum_{q=1}^{l}\mathrm{G}(x,y_{j_q})\mathrm{C}_{lq}\right)\left(\sum_{q=1}^{l}\mathrm{G}(x_{i_q},y)\mathrm{D}_{lq}\right), \tag{2.77}$$

où les coefficients C_{lq} et D_{lq} sont donnés par l'algorithme suivant.

Algorithme 12 Calcul des coefficients C et D

Données: Une matrice de collocation M de taille $m \times n$ construite sur les nuages de points X et Y.

But: Calculer les coefficients C et D de la décomposition 2.77.

1: Compresser la matrice de collocation M par ACA pour obtenir une approximation de rang r,$M \simeq AB^T$. Renvoyer de plus la liste des pivots $\hat{t} = \{i_1, \ldots, i_r\}$ et $\hat{s} = \{j_1, \ldots, j_r\}$ utilisés par l'algorithme ACA.
2: Initialiser les matrices C et D de taille $r \times r$ par zéro. Initialiser deux vecteurs c et d de taille r par zéro.
3: **Pour** $l = 1, \ldots, r$ **faire**
4: **Pour** $i = 1, \ldots, l-1$ **faire**
5: $d_i = 0, c_i = 0$
6: **Pour** $q = 1, \ldots, i$ **faire**
7: $c_i = c_i + C_{lq} G(x_{i_l}, y_{j_q})$
8: $d_i = d_i + D_{lq} G(x_{i_q}, y_{j_l})$
9: **Fin Pour**
10: **Fin Pour**
11: $C_{ll} = 1/\sqrt{|A_{i_l l}|}$
12: $D_{ll} = \mathrm{sgn}(A_{i_l l})/\sqrt{|A_{i_l l}|}$
13: **Pour** $q = 1, \ldots, l-1$ **faire**
14: $C_{lq} = 0, D_{lq} = 0$
15: **Pour** $i = q, \ldots, l-1$ **faire**
16: $C_{lq} = C_{lq} - C_{iq} d_i C_{ll}$
17: $D_{lq} = D_{lq} - D_{iq} c_i D_{ll}$
18: **Fin Pour**
19: **Fin Pour**
20: **Fin Pour**
21: **retour** C et D

La qualité de la décomposition 2.77 du noyau dépend de l'erreur commise par l'algorithme ACA. On peut obtenir une majoration de l'erreur ponctuelle commise par l'algorithme ACA. [BG05] donne une estimation de l'erreur commise pour chaque coefficient de la matrice de collocation à partir du théorème 2.47.

Lemme 2.50. Soit M la matrice dont les coefficient sont donnés par 2.75. Alors, on a

$$|M_{ij} - (AB^T)_{ij}| = \mathcal{O}(2^r (\eta/2)^{r^{1/3}}) \tag{2.78}$$

où r est le rang déterminé par l'algorithme ACA et η est le facteur d'admissibilité.

Dans la pratique, pour le cas du simple couche, [BG05] indique que le facteur 2^r n'apparaît pas et que pour la matrice du simple couche $r = \log^2(\epsilon)$ est souvent suffisant pour obtenir une erreur de l'ordre de $\mathcal{O}(\epsilon)$.

Remarque 2.51 (Traitement de la matrice de collocation). Dans nos tests numériques, nous avons pu constater que l'emploi de l'algorithme ACA avec pivot total est souvent plus stable que celui utilisant une stratégie de pivot partiel. Dans le cas où la matrice de collocation est petite, on peut envisager d'utiliser la SVD et la pseudo-inverse de la matrice M. On introduit ce changement lorsque l'estimation d'erreur 2.78 n'est pas vérifiée.

2.3.3 Interpolation polynomiale

Contrairement au développement de Taylor qui requiert les dérivées successives du noyau, l'approximation polynomiale fournit des résultats intéressants tout en étant plus maniable dans la pratique. Cette section se veut être un bref rappel des principaux résultats sur l'interpolation polynomiale afin de pouvoir l'utiliser dans un code de calcul.
On détaille la méthode d'interpolation polynomiale par les polynômes de Lagrange et de Chebyshev dans $\mathbb{R}$ puis on généralise à $\mathbb{R}^3$ à l'aide de produits tensoriels.
On renvoie à [BGH12] pour un exposé clair et détaillé de l'interpolation par les polynômes de Lagrange dans le contexte des $\mathscr{H}$-matrices. La même méthode est employé dans le cadre de la méthode des multipôles rapides avec des polynômes de Chebyshev dans la thèse de Messner [Mes11].

2.3.3.1 Interpolation de Lagrange dans $\mathbb{R}$

On se donne une fonction f continue de $[-1,1]$ à valeurs réelles, aussi noté $f \in \mathscr{C}[-1,1]$, et notons $\mathbb{P}_m$ l'ensemble des polynômes de degré au plus m.

Définition 2.52 (Opérateur d'interpolation)**.** On définit l'opérateur d'interpolation polynomiale $\mathfrak{I}_m$ de la manière suivante

$$\begin{aligned} \mathfrak{I}_m : \mathrm{C}[-1,1] &\longrightarrow \mathbb{P}_m \\ f &\longrightarrow \sum_{\nu=0}^{m} f(\xi_\nu)\mathscr{L}_\nu \end{aligned}$$

où les points $\{\xi_\nu\}_\nu \in [-1,1]$ sont dits noeuds d'interpolation et les polynômes $\{\mathscr{L}_\nu\}_\nu$ sont les polynômes de Lagrange vérifiant $\mathscr{L}_\nu(\xi_\mu) = \delta_{\nu\mu}$ pour $\nu, \mu \in \{0,\dots,m\}$.

Lemme 2.53. On utilise la transformation affine suivante pour passer de l'intervalle de réfrence $[-1,1]$ à un intervalle $[a,b]$ quelconque.

$$\begin{aligned} \Phi_{[a,b]} : [-1,1] &\longrightarrow [a,b] \\ t &\longrightarrow \frac{b+a}{2} + \frac{b-a}{2} t \end{aligned}$$

On définit les nœuds et les polynômes d'interpolation dans $[a,b]$ grâce à l'application $\Phi_{[a,b]}$ de la sorte

$$\begin{aligned} \xi_{[a,b],\nu} &= \Phi_{[a,b]}(\xi_\nu), \\ \mathscr{L}_{[a,b],\nu}(t) &= \prod_{\mu=0,\mu\neq\nu}^{m} \frac{t-\xi_{[a,b],\mu}}{\xi_{[a,b],\nu}-\xi_{[a,b],\mu}}. \end{aligned}$$

Choix des noeuds d'interpolation Un problème typique lors de l'emploi de l'interpolation est le choix des nœuds d'interpolation. Sélectionner m points équirépartis dans l'intervalle peut conduire au phénomène de Runge pour des grandes valeurs de m et ne garantit pas la convergence uniforme. Pour atténuer ce problème, l'usage courant est de choisir comme nœuds d'interpolation les m zéros du $m^{\text{ème}}$ polynôme de Chebyshev définis dans l'intervalle de référence par

$$\xi_{[-1,1],\nu} = \cos\left(\frac{2\nu+1}{2m}\pi\right), \nu = 0,\dots,m-1. \tag{2.79}$$

2.3.3.2 Interpolation polynomiale en dimension 3

La construction d'un interpolant en dimension supérieure sur un ouvert Γ_t de $\mathbb{R}^3$ n'est pas une tâche aisée et de manière générale, on ne peut pas prouver l'unicité d'une telle approximation. Pour pallier cette difficulté, on considère une boite englobante $B = [a_1, b_1] \times [a_2, b_2] \times [a_3, b_3]$ telle que $\Gamma \subseteq B$ et on cherche une approximation dans cette boîte. Pour ce faire, on utilise l'interpolation en une dimension dans chaque direction et on construit l'opérateur d'interpolation à partir du produit tensoriel suivant.

Définition 2.54 (Interpolation par produit tensoriel)**.** Soit f une fonction de $B \subset \mathbb{R}^3$ à valeurs réelles. On définit l'opérateur d'interpolation $\mathscr{I}_m^B$ dans la boîte B par le produit tensoriel suivant,

$$\mathscr{I}_m^B = \mathscr{I}_{[a_1,b_1],m} \otimes \mathscr{I}_{[a_2,b_2],m} \otimes \mathscr{I}_{[a_3,b_3],m}. \tag{2.80}$$

Pour un triplet d'entiers inférieurs ou égaux à m, $\nu = (\nu_1, \nu_2, \nu_3)$, on définit les nœuds et les polynômes d'interpolation par

$$\begin{aligned} \xi_\nu^B &= (\xi_{[a_1,b_1],\nu_1}, \xi_{[a_2,b_2],\nu_2}, \xi_{[a_3,b_3],\nu_3}), & (2.81)\\ \mathscr{L}_\nu^B(x) &= \mathscr{L}_{[a_1,b_1],\nu_1}(x_1)\mathscr{L}_{[a_2,b_2],\nu_2}(x_2)\mathscr{L}_{[a_3,b_3],\nu_3}(x_3). & (2.82) \end{aligned}$$

Pour une fonction f de $\mathbb{R}^3$ à valeurs réelles, on notera l'interpolation de la façon suivante

$$\mathscr{I}_m^B[f] = \sum_{\nu \in \mathscr{M}} f(\xi_\nu^B)\mathscr{L}_\nu^B, \tag{2.83}$$

où $\mathscr{M} := \{0, \dots, m\}^3$.

Remarque 2.55 (Anisotropie). On notera que l'on a écrit les formules d'interpolation en prenant un ordre constant suivant chaque direction. L'analyse de l'erreur au prochain paragraphe montrera que l'on dispose de plus de souplesse dans la pratique ce qui permet de réduire le nombre de points nécessaire au calcul d'une bonne approximation.

2.3.3.3 Analyse d'erreur

On explicite dans cette partie les principaux résultats de la théorie de l'interpolation polynomiale en une dimension. Rappelons que pour une fonction continue de $[-1,1]$ à valeurs réelles, $\|f\|_{\infty,[-1,1]}$ désigne la norme sup sur l'intervalle $[-1,1]$.

Dans un premier temps, on donne une définition de la stabilité d'un schéma d'interpolation puis on utilise cette notion pour exhiber des inégalités sur l'erreur d'interpolation dans le cas de la dimension 1 puis dans le cas d'une approximation par produits tensoriels en dimension 3.Les détails des résultats suivants figurent dans [BGH12].
Notons enfin que dans toute la suite, le choix de la famille de polynômes n'intervient pas.

Résultats en une dimension d'espace Pour tout entier m, Λ_m est une constante vérifiant la condition de stabilité suivante pour toute fonction $f \in \mathscr{C}[-1,1]$,

$$\|\mathfrak{I}_m[f]\|_{\infty,[-1,1]} \leqslant \Lambda_m \|f\|_{\infty,[-1,1]}, \tag{2.84}$$

Dans le cas où l'on choisit les noeuds de Chebyshev, on a ([MH03]) :

$$\Lambda_m = \frac{2}{\pi}\ln(m+1) + 1 \leqslant m+1. \tag{2.85}$$

La stabilité et le fait que l'opérateur d'interpolation est un projecteur sur l'espace des polynômes de degré inférieur ou égal à m fournissent la propriété suivante.

Proposition 2.56 (Meilleure approximation dans $\mathbb{P}_m$)**.** *Soit $\mathfrak{I}_m$ vérifiant l'estimation de stabilité 2.84. Alors, pour toute fonction f continue de $[-1,1]$, on a*

$$\|f-\mathfrak{I}_m[f]\|_{\infty,[-1,1]} \leqslant (1+\Lambda_m)\inf\{\|f-v\|_{\infty,[-1,1]} : v\in\mathbb{P}_m\}.$$

Cette propriété traduit le fait qu'un schéma d'interpolation stable fournit la meilleure approximation possible à une constante près. En supposant la fonction f plus régulière, par exemple $f\in\mathscr{C}^{m+1}([-1,1])$, on peut améliorer l'inégalité précédente et obtenir

$$\|f-\mathfrak{I}_m[f]\|_{\infty,[-1,1]} \leqslant \frac{2^{-m}}{(m+1)!}\|f^{(m+1)}\|_{\infty,[-1,1]}. \tag{2.86}$$

Dans le cas de l'intervalle $[a,b]$, la relation 2.84 est inchangée tandis que pour une fonction $f\in\mathscr{C}^{m+1}([a,b])$, on obtient l'estimation suivante :

$$\|f-\mathfrak{I}_m[f]\|_{\infty,[a,b]} \leqslant \frac{2}{(m+1)!}\left(\frac{b-a}{4}\right)^{m+1}\|f^{(m+1)}\|_{\infty,[a,b]}.$$

Pour les applications courantes, on souhaite travailler avec les dérivées successives de la fonction f. Une solution possible serait de dériver la fonction puis de l'interpoler, ce qui nécessiterait de dériver la fonction au préalable et que cette dérivée devienne un argument du code de calcul. On préférera dériver l'interpolant car il est plus aisé de dériver un polynôme.
Le résultat suivant permet de relier la norme d'un polynôme et de sa dérivée $l^{\text{ème}}$.

Proposition 2.57 (Inégalité de Markov itérée)**.** *Soient l un entier strictement positif et $u\in\mathbb{P}_m$. On a ,*

$$\|u^{(l)}\|_{\infty,[-1,1]} \leqslant \begin{cases} \left(\frac{m!}{(m-l)!}\right)^2\|u\|_{\infty,[-1,1]} & \text{si } l\leqslant m,\\ 0 & \text{sinon.}\end{cases} \tag{2.87}$$

Cette propriété permet de définir une condition de stabilité similaire à 2.84 pour les dérivées successives.

Proposition 2.58 (Stabilité des dérivées)**.** *Soit $l\in\{0,\dots,m\}$, on considère $f\in\mathscr{C}^l([a,b])$. Si $\mathfrak{I}_m^{[a,b]}$ satisfait la condition de stabilité 2.84 alors il satisfait également la condition suivante,*

$$\|(\mathfrak{I}_m^{[a,b]}[f])^{(l)}\|_{\infty,[a,b]} \leqslant \Lambda_m^{(l)}\|f^{(l)}\|_{\infty,[a,b]}, \tag{2.88}$$

avec la constante de stabilité

$$\Lambda_m^{(l)} := \frac{\Lambda_m}{l!}\left(\frac{m!}{(m-l)!}\right)^2. \tag{2.89}$$

La condition de stabilité précédente permet de prouver une majoration de l'erreur similaire à l'estimation de meilleure approximation de la proposition 2.56.

Théorème 2.59 (Approximation des dérivées)**.** *Soient $l\in\{0,\dots,m\}$,$n\in\{0,\dots,m-l\}$ et une fonction $f\in\mathscr{C}([a,b])$. Sous réserve de stabilité (2.84) et de meilleure approximation (2.86), on a*

$$\|f^{(l)}-(\mathfrak{I}_m^{[a,b]}[f])^{(l)}\|_{\infty,[a,b]} \leqslant \frac{2(\Lambda_m^{(l)}+1)}{(n+1)!}\left(\frac{b-a}{4}\right)^{n+1}\|f^{(l+n+1)}\|_{\infty,[a,b]}. \tag{2.90}$$

À ce stade, on dispose d'une majoration de l'erreur d'interpolation dans une direction en fonction de l'ordre d'interpolation et de la dimension de l'intervalle aussi bien pour la fonction que pour ses dérivées successives. On généralise ces résultats au cas de l'interpolation par produits tensoriels en dimension 3 à l'aide de l'opérateur 2.83.

Résultats en dimension 3 Le théorème suivant prouve d'une part la stabilité de l'opérateur d'interpolation 2.83 et d'autre part fournit une majoration de l'erreur d'interpolation dans le cas de l'interpolation par produits tensoriels. La norme $\|.\|_{\infty,\mathrm{B}}$ désigne de manière similaire au cas 1D la norme de la convergence uniforme sur $\mathrm{B} \subset \mathbb{R}^3$. En l'absence de précision supplémentaire, les notations sont les mêmes que celles du paragraphe précédent.

Théorème 2.60 (Interpolation par produits tensoriels)**.** *Soit* $\mathrm{B} = [a_1, b_1] \times [a_2, b_2] \times [a_3, b_3]$. *Soit un entier* m. *On suppose pour* $i \in \{1,2,3\}$ *que les opérateurs* $\mathfrak{I}_m^{[a_i,b_i]}$ *respectent la condition de stabilité 2.84 et l'estimation d'erreur 2.86. On considère de plus* $\mathfrak{I}_m^{\mathrm{B}}$ *l'opérateur d'interpolation dans la boîte englobante* B *défini par 2.83. Pour toute fonction* $f \in \mathscr{C}(\mathrm{B})$ *cet opérateur vérifie alors l'inégalité suivante :*

$$\|\mathfrak{I}_m^{\mathrm{B}}[f]\|_{\infty,\mathrm{B}} \leqslant \Lambda_m^3 \|f\|_{\infty,\mathrm{B}}. \tag{2.91}$$

Si l'on suppose de plus que la fonction f *est* $m+1$ *fois dérivable, on a*

$$\|f - \mathfrak{I}_m^{\mathrm{B}}[f]\|_{\infty,\mathrm{B}} \leqslant 6\Lambda_m^2 \left(\frac{diam(\mathrm{B})}{4}\right)^{m+1} \frac{\max\{\|\partial_i^{m+1} f\|_{\infty,\mathrm{B}} : i \in \{1,2,3\}\}}{(m+1)!}. \tag{2.92}$$

Remarque 2.61 (Interpolation anisotrope). [BGH12] montre qu'il est possible d'utiliser un ordre moins élevé dans une direction plus petite sans modifier l'erreur commise. On construit par la suite des boites "épousant" la surface $\Gamma \subset \mathrm{B}$ à l'aide de l'analyse en composantes principales.

Enfin, de manière similaire au cas 1D, on peut montrer une inégalité sur l'approximation des dérivées successives dans une boite B.

Théorème 2.62 (Interpolation des dérivées par produit tensoriel)**.** *Soit* $\mathrm{B} = [a_1, b_1] \times [a_2, b_2] \times [a_3, b_3]$, *soient* m *un entier et* μ *un triplet tel que pour tout entier* $i \in \{1,2,3\}, \mu_i \leqslant m$. *Soit* n *un entier tel que* $n \leqslant m - \mu_i$ *pour tout entier* $i \in \{1,2,3\}$.

On suppose pour $i \in \{1,2,3\}$ *que les opérateurs* $\mathfrak{I}_m^{[a_i,b_i]}$ *respectent la condition de stabilité 2.84 et l'estimation d'erreur 2.86. On considère* $\mathfrak{I}_m^{\mathrm{B}}$ *l'opérateur d'interpolation dans la boîte* B. *Quitte à poser*

$$\Lambda_m^{(\mu)} := \prod_{i=1}^{3} (\Lambda_m^{(\mu_i)} + 1),$$

on a alors pour toute fonction $f \in \mathrm{C}^{m+1}(\mathrm{B})$,

$$\|\partial^{\mu}(f - \mathfrak{I}_m^{\mathrm{B}}[f])\|_{\infty,\mathrm{B}} \leqslant \frac{2\Lambda_m^{(\mu)}}{(n+1)!} \sum_{i=1}^{3} \left(\frac{b_i - a_i}{4}\right)^{n+1} \|\partial_i^{n+1} \partial^{\mu} f\|_{\infty,\mathrm{B}}. \tag{2.93}$$

2.3.3.4 Application au noyau de Green

On applique le schéma d'interpolation par produit tensoriel au noyau de Green de l'équation de Laplace. On se limite à ce cas dans cette partie afin de pouvoir se placer dans le cadre des noyaux asymptotiquement lisses. On effectuera l'interpolation du noyau oscillant ultérieurement.

Notations On considère deux boîtes englobantes $\mathrm{B_X}$ et $\mathrm{B_Y}$ dans $\mathbb{R}^3$ définies par

$$\begin{aligned} \mathrm{B_X} &= [a_1^x, b_1^x] \times [a_2^x, b_2^x] \times [a_3^x, b_3^x], \\ \mathrm{B_Y} &= [a_1^y, b_1^y] \times [a_2^y, b_2^y] \times [a_3^y, b_3^y]. \end{aligned}$$

On supposera que ces boîtes sont faiblement admissibles (fortement si on prend le max), *i.e* elles vérifient l'inégalité suivante, pour η réel,

$$\min(\mathrm{diam}(X), \mathrm{diam}(Y)) \leqslant \eta \mathrm{dist}(X, Y). \tag{2.94}$$

Dans toute la suite, x désigne un point de B_X et y un point de B_Y.

Contexte d'application On applique les résultats de l'interpolation polynomiale au noyau de Green $G(x, y)$. On rappelle que ce noyau vérifie l'hypothèse asymptotiquement lisse suivante pour des constantes positives c_0, C et σ :

$$|\partial_x^\alpha \partial_y^\beta G(x, y)| \leqslant C|\alpha+\beta|! c_0^{|\alpha+\beta|} |x-y|^{-\sigma-|\alpha+\beta|}.$$

En particulier, les fonctions G_x et G_y définies par

$$\begin{aligned} G_x : B_X &\mapsto \mathscr{C}^\infty(B_Y) \quad , \quad x \mapsto G(x, .), \\ G_y : B_Y &\mapsto \mathscr{C}^\infty(B_X) \quad , \quad y \mapsto G(., y), \end{aligned}$$

vérifient les estimations suivantes pour tout multi-indice α et β de $\mathbb{N}^3$

$$\begin{aligned} \|\partial^\alpha G_x(x)\|_{\infty, B_X} &\leqslant C|\alpha|! c_0^{|\alpha|} \mathrm{dist}(B_X, B_Y)^{-\sigma-|\alpha|}, \\ \|\partial^\beta G_y(y)\|_{\infty, B_Y} &\leqslant C|\beta|! c_0^{|\beta|} \mathrm{dist}(B_X, B_Y)^{-\sigma-|\beta|}. \end{aligned}$$

Interpolation du noyau dans $\mathbb{R}^3$ En utilisant l'inégalité 2.85, le théorème 2.60 nous fournit les estimations suivantes pour $x \in B_X$ et $y \in B_Y$

$$\begin{aligned} \|G_x - \mathfrak{I}_m^{B_X}[G_x]\|_{\infty, B_X} &\leqslant \frac{2Cd(m+1)^{d-1}}{\mathrm{dist}(B_X, B_Y)^\sigma} \left(\frac{c_0 \mathrm{diam}(B_X)}{4\mathrm{dist}(B_X, B_Y)} \right)^{m+1}, \\ \|G_y - \mathfrak{I}_m^{B_Y}[G_y]\|_{\infty, B_Y} &\leqslant \frac{2Cd(m+1)^{d-1}}{\mathrm{dist}(B_X, B_Y)^\sigma} \left(\frac{c_0 \mathrm{diam}(B_Y)}{4\mathrm{dist}(B_X, B_Y)} \right)^{m+1}. \end{aligned}$$

Alors l'approximation $\tilde{G}$ du noyau dans $B_X \times B_Y$ est donnée par

$$\tilde{G}(x, y) = \begin{cases} \mathfrak{I}_m^{B_X}[G_x](x)(y) & \text{si } \mathrm{diam}(B_X) \leqslant \mathrm{diam}(B_Y) \\ \mathfrak{I}_m^{B_Y}[G_y](x)(y) & \text{sinon} \end{cases} \tag{2.95}$$

Dans le cas où $\mathrm{diam}(B_X) \leqslant \mathrm{diam}(B_Y)$, on en conclut que

$$\begin{aligned} |G(x, y) - \tilde{G}(x, y)| &= |G(x, y) - \sum_{\nu \in \mathscr{M}} G(\xi_\nu^B, y) \mathscr{L}_\nu^{B_X}(x)| \\ &= |(G_x - \mathfrak{I}_m^{B_X}[G_x])(x)(y)| \\ &\leqslant \frac{2Cd(m+1)^{d-1}}{\mathrm{dist}(B_X, B_Y)^\sigma} \left(\frac{c_0 \mathrm{diam}(B_X)}{4\mathrm{dist}(B_X, B_Y)} \right)^{m+1}. \end{aligned}$$

On peut obtenir le même type d'inégalité si $\mathrm{diam}(B_X) \geqslant \mathrm{diam}(B_Y)$. La condition d'admissibilité 2.94 exprime le lien entre les diamètres et la distance de B_X et B_Y pour finalement donner

$$|G(x, y) - \tilde{G}(x, y)| \leqslant \frac{P(m)}{\mathrm{dist}(B_X, B_Y)^\sigma} \left(\frac{c_0 \eta}{4} \right)^{m+1}, \tag{2.96}$$

où $P(m) = 2Cd(m+1)^d \eta^\sigma$ est un polynôme en m. Il suffit alors d'imposer la condition $\eta < 4/c_0$ pour que le schéma d'interpolation converge exponentiellement.

Approximation de rang fini En remplaçant $G(x,y)$ par son approximation dans l'expression des coefficients de la matrice de simple couche 2.2, dans le cas où $\text{diam}(B_X) \leqslant \text{diam}(B_Y)$ on obtient

$$\begin{aligned} M_{ij} &= \int_\Gamma \int_\Gamma G(x,y)\Phi_j(y)\Phi_i^t(x)d\Gamma(x)d\Gamma(y) \\ &\simeq \int_\Gamma \int_\Gamma \left(\sum_{\nu \in \mathcal{M}} G(\xi_\nu^B, y)\mathcal{L}_\nu^{B_X}(x) \right) \Phi_j(y)\Phi_i^t(x)d\Gamma(x)d\Gamma(y) \\ &\simeq \sum_{\nu \in \mathcal{M}} \left(\int_\Gamma \mathcal{L}_\nu^{B_X}(x)\Phi_i^t(x)d\Gamma(x) \right) \left(\int_\Gamma G(\xi_\nu^B, y)\Phi_j(y)d\Gamma(y) \right) \\ &\simeq \sum_{\nu \in \mathcal{M}} A_{i,\nu}B_{j,\nu}. \end{aligned}$$

Il s'agit bien d'une approximation à variables séparées et de rang fini $\text{Card}(\mathcal{M}) = \mathcal{O}(m^3)$.

Remarque 2.63 (Critère d'admissibilité). On note que c'est le critère d'admissibilité qui implique la convergence exponentielle du schéma d'interpolation. Dans le cas de l'admissibilité forte, on peut effectuer une interpolation polynomiale dans les deux variables x et y. Dans le cas d'une admissibilité faible, seule la variable associée à la plus petit boite peut être développée.

Interpolation des dérivées du noyau dans $\mathbb{R}^3$ Dans la modélisation par les équations intégrales, on utilise fréquemment des opérateurs définis par les dérivées du noyau $G(x,y)$. Un exemple typique est l'opérateur de double couche dont les coefficients de la matrice de discrétisation sont

$$\begin{aligned} D_{ij} &= \int_\Gamma \int_\Gamma \frac{\partial G}{\partial \vec{n}_y}(x,y)\Phi_j(y)\Phi_i^t(x)d\Gamma(x)d\Gamma(y) \\ &= \int_\Gamma \int_\Gamma \frac{< x-y, \vec{n}_y >}{|x-y|^3}\Phi_j(y)\Phi_i^t(x)d\Gamma(x)d\Gamma(y) \end{aligned}$$

où $\vec{n}(y)$ est la normale extérieure à Γ au point y. En général, ce noyau n'est pas lisse et on ne peut appliquer directement les résultats précédents à ce noyau. Pour se ramener aux résultats précédents, on peut écrire la dérivée normale du noyau sous une autre forme,

$$\frac{\partial G}{\partial \vec{n}_y} = < \nabla_y G(x,y), \vec{n}(y) > . \tag{2.97}$$

$G(x,y)$ étant asymptotiquement lisse, il en va de même pour $\nabla_y G(x,y)$. On peut alors remplacer le noyau par l'approximation suivante,

$$\frac{\partial G}{\partial \vec{n}_y} \simeq < \nabla_y \tilde{G}(x,y), \vec{n}(y) > \tag{2.98}$$

Cela revient à dériver l'approximation polynomiale obtenue pour le noyau $G(x,y)$ ce qui diminue l'ordre de l'approximation de la dérivée. Ceci impose d'approcher le noyau $G(x,y)$ avec un ordre suffisament élevé sous peine de perte de précision. Sans perdre de généralité, on suppose que l'on a $\text{diam}(B_Y) \leqslant \text{diam}(B_X)$ et on effectue l'interpolation de $G(x,y)$ en la variable y, x étant fixé dans B_X. En remplaçant le noyau par son approximation il vient

$$
\begin{aligned}
\mathrm{D}_{ij} &= \int_\Gamma\int_\Gamma < \nabla_y \mathrm{G}(x,y), \vec{n}(y) > \Phi_j(y)\Phi_i^t(x)d\Gamma(x)d\Gamma(y) \\
&\simeq \int_\Gamma\int_\Gamma < \nabla_y \tilde{\mathrm{G}}(x,y), \vec{n}(y) > \Phi_j(y)\Phi_i^t(x)d\Gamma(x)d\Gamma(y) \\
&\simeq \int_\Gamma\int_\Gamma < \nabla_y \left(\sum_{\nu\in\mathscr{M}} \mathrm{G}(x,\xi_\nu^{\mathrm{B_Y}})\mathscr{L}_\nu^{\mathrm{B_Y}}(y)\right), \vec{n}(y) > \Phi_j(y)\Phi_i^t(x)d\Gamma(x)d\Gamma(y) \\
&\simeq \sum_{\nu\in\mathscr{M}} \int_\Gamma\int_\Gamma < \nabla_y \left(\mathrm{G}(x,\xi_\nu^{\mathrm{B_Y}})\mathscr{L}_\nu^{\mathrm{B_Y}}(y)\right), \vec{n}(y) > \Phi_j(y)\Phi_i^t(x)d\Gamma(x)d\Gamma(y) \\
&\simeq \sum_{\nu\in\mathscr{M}} \left(\int_\Gamma \mathrm{G}(x,\xi_\nu^{\mathrm{B_Y}})\Phi_i^t(x)d\Gamma(x)\right) \int_\Gamma \left(< \nabla_y \mathscr{L}_\nu^{\mathrm{B_Y}}(y), \vec{n}(y) > \Phi_j(y)d\Gamma(y)\right) \\
&\simeq \sum_{\nu\in\mathscr{M}} \mathrm{A}_{i,\nu}\mathrm{B}_{j,\nu}
\end{aligned}
$$

On retrouve bien l'approximation de rang faible à variables séparées de rang $\mathrm{Card}(\mathscr{M}) = \mathscr{O}(m^3)$ pour la matrice du double-couche.

Estimation d'erreur pour les dérivées Dans un cas plus général, on considère une dérivée du noyau de Green de la forme

$$\mathrm{F}(x,y) = \partial_x^\mu \partial_y^\zeta \mathrm{G}(x,y),$$

μ et ζ étant des multi-indices de $\mathbb{N}^3$.

On peut obtenir une majoration de l'erreur similaire à celle obtenu en (2.96) en se plaçant avec les mêmes hypothèses de régularité et d'admissibilité. On se limite à présenter le cas où $\mathrm{diam}(\mathrm{B_X}) \leqslant \mathrm{diam}(\mathrm{B_Y})$, l'autre cas se traitant de manière similaire. On effectue l'hypothèse supplémentaire que l'ordre d'interpolation du noyau est suffisament grand pour que $\mu_i \leqslant m$ pour $i = 1,2,3$.. On fixe un entier n tel que $n \leqslant m - \mu_i$ pour $i = 1,2,3$. D'après le théorème 2.62 on a alors la majoration de l'erreur d'interpolation pour une approximation $\tilde{\mathrm{F}}$ de F dans $\mathrm{B_X} \times \mathrm{B_Y}$:

$$\|\mathrm{F}(x,y) - \tilde{\mathrm{F}}(x,y)\|_{\infty,\mathrm{B_X}\times\mathrm{B_Y}} \leqslant \frac{\mathrm{P}^\mu(n)}{\mathrm{dist}(\mathrm{B_X},\mathrm{B_Y})^{\sigma+|\mu|}} \left(\frac{c_0\eta}{4}\right)^{n+1}, \tag{2.99}$$

où $\mathrm{P}^\mu(m)$ est un polynôme en n (*cf* [BGH12]). L'approximation $\tilde{\mathrm{F}}(x,y)$ de $\mathrm{F}(x,y)$ étant donnée par

$$\tilde{\mathrm{F}}(x,y) = \sum_{\nu\in\mathscr{M}} \partial_x^\mu \mathscr{L}_\nu^{\mathrm{B_X}}(x)\partial_y^\zeta \mathrm{G}(\xi_\nu^{\mathrm{B_X}}, y). \tag{2.100}$$

Contrairement à l'estimation d'erreur pour l'interpolation du noyau, on observe une convergence en $(c_0\eta/4)^{n+1}$ au lieu de $(c_0\eta/4)^{m+1}$. Ceci est dû à la dérivation de l'approximation du noyau et souligne encore que l'ordre d'approximation initial m doit être suffisament grand pour approcher les dérivées souhaitées.

Remarque 2.64 (Estimation de l'ordre d'interpolation). Contrairement au développement de Taylor, on ne dispose pas d'une estimation précise de l'ordre d'interpolation d'après les estimations d'erreurs. À cause des constantes qui interviennent, on trouve dans la pratique un ordre d'interpolation trop grand. En effet, le rang de la matrice approché se comporte comme m^3 et le nombre de nœuds devient vite trop important. On choisit dans la pratique de se fier à l'estimation obtenue pour le développement de Taylor et dans le cas du simple couche, on prend $m \approx \log_{10}(\epsilon)$ comme ordre d'interpolation.

2.3.4 *Hybrid Cross Approximation (HCA)*

2.3.4.1 Contexte d'utilisation

On présente la méthode d'*Hybrid Cross Apprximation*(HCA) présentée dans [BG05]. Cette méthode combine les notions d'interpolation polynomiale dans $\mathbb{R}^3$ et l'approximation croisée ACA. On vise l'obtention d'une approximation $\tilde{\mathrm{M}}$ de rang faible d'une matrice M issue de la méthode des éléments finis de frontières. On prendra comme exemple la matrice de l'opérateur de simple couche de taille $m \times n$,

$$\mathrm{M}_{ij} = \int_\Gamma \int_\Gamma \mathrm{G}(x,y)\Phi_i^t(x)\Phi_j(y)d\Gamma(x)d\Gamma(y),$$

avec $i \in t = \{i_1,\ldots,i_m\}$ et $j \in s = \{j_1,\ldots,j_n\}$.

Aux ensembles t et s, on associe les supports $\mathrm{Supp}(t)$ et $\mathrm{Supp}(s)$ définis à partir des supports des fonctions de base par

$$\begin{aligned}\mathrm{Supp}(t) &= \bigcup_{q=1}^{m} \mathrm{Supp}(\Phi_{i_q}^t)\\ \mathrm{Supp}(s) &= \bigcup_{q=1}^{n} \mathrm{Supp}(\Phi_{j_q})\end{aligned}$$

Par abus de notation, on désignera indifféremment par t et s les ensembles d'indices ou les supports suivant l'usage. Avec les notations précédentes, cela revient à considérer les domaines $\mathrm{X} = \mathrm{Supp}(t)$ et $\mathrm{Y} = \mathrm{Supp}(s)$. Dans toute la suite, on considère que $x \in t$ et $y \in s$ et on note B_t et B_s leurs boîtes englobantes respectives.

2.3.4.2 Interpolation du noyau $\mathrm{G}(x,y)$

On applique les résultats d'interpolation précédents au noyau de Green. L'opérateur d'interpolation étant appliqué successivement aux deux variables, cela recquiert l'utilisation du critère d'admissibilité forte 2.55 :

$$\max(\mathrm{diam}(\mathrm{B}_t),\mathrm{diam}(\mathrm{B}_s)) \leqslant \eta \mathrm{dist}(\mathrm{B}_t,\mathrm{B}_s).$$

En conservant les notations précédentes, l'approximation polynomiale du noyau à l'ordre p dans $\mathrm{B}_t \times \mathrm{B}_s$ est alors donnée par

$$\begin{aligned}\mathrm{G}(x,y) &\simeq \tilde{\mathrm{G}}(x,y)\\ &= \sum_{\nu\in\mathscr{M}}\sum_{\mu\in\mathscr{M}} \mathrm{G}(\xi_\nu^{\mathrm{B}_t},\xi_\mu^{\mathrm{B}_s})\mathscr{L}_\nu^{\mathrm{B}_t}(x)\mathscr{L}_\mu^{\mathrm{B}_s}(y).\end{aligned}$$

Cette écriture permet d'ores et déjà la séparation des variables x et y. Cependant, nous avons observé à la section 2.3.3 que le nombre de nœuds d'interpolation peut être élevé si l'on souhaite une approximation fine du noyau. À ce stade, notre approximation est de rang fini et bornée par $\mathrm{Card}(\mathscr{M}) = (p+1)^3 := \mathrm{K}$. En substituant $\mathrm{G}(x,y)$ par son approximation et en échangeant les signes somme et intégrale, on obtient

$$\tilde{\mathrm{M}}_{ij} = \int_\Gamma \int_\Gamma \tilde{\mathrm{G}}(x,y)\Phi_i^t(x)\Phi_j(y)d\Gamma(x)d\Gamma(y) \tag{2.101}$$

$$= \int_\Gamma \int_\Gamma \left(\sum_{\nu\in\mathscr{M}}\sum_{\mu\in\mathscr{M}} \mathrm{G}(\xi_\nu^{\mathrm{B}_t},\xi_\mu^{\mathrm{B}_s})\mathscr{L}_\nu^{\mathrm{B}_t}(x)\mathscr{L}_\mu^{\mathrm{B}_s}(y)\right)\Phi_i^t(x)\Phi_j(y)d\Gamma(x)d\Gamma(y) \tag{2.102}$$

$$= \sum_{\nu\in\mathscr{M}}\sum_{\mu\in\mathscr{M}} \left(\int_\Gamma \Phi_i^t(x)\mathscr{L}_\nu^{\mathrm{B}_t}(x)d\Gamma(x)\right)\mathrm{G}(\xi_\nu^{\mathrm{B}_t},\xi_\mu^{\mathrm{B}_s})\left(\int_\Gamma \mathscr{L}_\mu^{\mathrm{B}_s}(y)\Phi_j(y)d\Gamma(y)\right) \tag{2.103}$$

L'avantage de cette écriture est que le second terme de l'égalité peut être vu comme un produit de trois matrices.

$$\tilde{M}_{ij} = (U.S.V^T)_{ij},$$

où les matrices $U \in \mathbb{R}^{m\times K}$, $S \in \mathbb{R}^{K\times K}$ et $V \in \mathbb{R}^{n\times K}$ sont définies par

$$\begin{aligned} U_{i\nu} &= \int_\Gamma \Phi_i^t(x)\mathscr{L}_\nu^{B_t}(x)d\Gamma(x), \\ V_{j\mu} &= \int_\Gamma \Phi_j(y)\mathscr{L}_\mu^{B_s}(y)d\Gamma(y), \\ S_{\nu\mu} &= G(\xi_\nu^{B_t}, \xi_\mu^{B_s}). \end{aligned}$$

Remarque 2.65 (Calcul des intégrales). Outre la séparation des variables, on remarque que les intégrales mises en jeu ne sont à présent ques des intégrales simples et non des intégrales doubles. Ceci est également un gain de calculs. Dans la pratique, les domaines d'intégration $\Gamma \cap \mathrm{Supp}(t)$ et $\Gamma \cap \mathrm{Supp}(s)$ sont discrétisés à l'aide de n_t et n_s triangles respectivement. Chaque intégration sur un triangle est alors effectuée de manière approchée à l'aide d'une quadrature à p_g points de Gauss. Dans le cas d'une intégrale double, l'intégration sur un produit de deux triangles recquiert p_g^2 évaluations de l'intégrande ceci pour les $n_t n_s$ interactions élémentaires possibles. Les intégrales simples mises en jeu ici abaissent ce coût à $p_g(n_t + n_s)$ évaluations de l'intégrande.

La matrice de collocation S correspond à l'évaluation aux nœuds d'interpolation du noyau de Green $G(x, y)$ pour nœuds appartenant à des boîtes admissibles. On est alors exactement dans les hypothèses d'utilisation de l'algorithme ACA. Deux approches sont envisageables. La première effectue une compression algébrique de la matrice de couplage, tandis que la seconde consiste à utiliser cette matrice de couplage pour former une approximation croisée analytique du noyau. Dans [BG05], la méthode algébrique est référencée par HCA-I et l'approche analytique par HCA-II.

L'algorithme HCA est donc constitué de deux ingrédients :

— l'interpolation fournit un ensemble de nœuds d'interpolation dans un volume afin de s'assurer de la convergence du ACA. En effet, il s'agit d'une distribution en grille cartésienne des points et l'algorithme ACA est réputé robuste pour ces situations. L'interpolation fournit dans un premier temps une borne $K = (p+1)^3$ sur le rang de l'approximation.

— l'algorithme ACA extrait un sous-ensemble des nœuds d'interpolation pour fournir une représentation compressée de la matrice de couplage de rang $r < K$.

2.3.4.3 Traitement de la matrice de couplage

Le noyau étant supposé asymptotiquement lisse, on s'attend à observer une décroissance rapide des valeurs singulières de la matrice de couplage en tant qu'évaluation du noyau aux noeuds d'interpolation. On peut dès lors tronquer les valeurs singulières dont les contributions sont négligeables. On compresse la matrice de couplage à l'aide d'un algorithme quelconque comme une SVD, ACA, rank-revealing LU, QR, etc, ... L'avantage de compresser la matrice de couplage et non la matrice de départ est qu'elle est de dimension moindre, $K \times K$.

Approche algébrique (HCA I)

Obtention de la décomposition En utilisant une méthode algébrique telle que le ACA ou la SVD, on construit une approximation de rang faible de la matrice de couplage :

$$\mathrm{S} \simeq \hat{\mathrm{A}}.\hat{\mathrm{B}}^{\mathrm{T}},$$

avec $\hat{\mathrm{A}} \in \mathbb{R}^{\mathrm{K}\times r}$ et $\hat{\mathrm{B}} \in \mathbb{R}^{\mathrm{K}\times r}, r < \mathrm{K}$, vérifiant

$$\frac{\|\mathrm{S} - \hat{\mathrm{A}}.\hat{\mathrm{B}}^{\mathrm{T}}\|}{\|\mathrm{S}\|} \leq \epsilon.$$

On obtient directement une représentation compressée 2.1 de rang r de M

$$\mathrm{M} \simeq \mathrm{A}\mathrm{B}^{\mathrm{T}},$$

en posant

$$\begin{aligned} \mathrm{A} &= \mathrm{U}\hat{\mathrm{A}}, \\ \mathrm{B} &= \mathrm{V}\hat{\mathrm{B}}, \end{aligned}$$

où $\mathrm{A} \in \mathbb{R}^{m\times r}$ et $\mathrm{B} \in \mathbb{R}^{n\times r}$ et avec pour $i \in t = \{i_1, \ldots, i_m\}$ et $j \in s = \{j_1, \ldots, j_n\}$.

$$\mathrm{U}_{i\nu} = \int_\Gamma \Phi_i^t(x)\mathscr{L}_\nu^{\mathrm{B}_t}(x)d\Gamma(x), \tag{2.104}$$

$$\mathrm{V}_{j\mu} = \int_\Gamma \Phi_j(y)\mathscr{L}_\mu^{\mathrm{B}_s}(y)d\Gamma(y). \tag{2.105}$$

Complexité On note r le rang de l'approximation obtenue par l'algorithme HCA – I. On rappelle que le nombre de nœuds d'interpolation est K La complexité de l'algorithme HCA – I est la suivante

— Calcul de la matrice U de taille $m \times \mathrm{K}$: $\mathscr{O}(m\mathrm{K})$.

— Calcul de la matrice V de taille $n \times \mathrm{K}$: $\mathscr{O}(n\mathrm{K})$.

— Compression de la matrice de couplage : $\mathscr{O}(2\mathrm{K}r^2)$.

— Produit $\mathrm{A} = \mathrm{U}\hat{\mathrm{A}}$: $\mathscr{O}(m\mathrm{K}r)$.

— Produit $\mathrm{B} = \mathrm{V}\hat{\mathrm{B}}$: $\mathscr{O}(n\mathrm{K}r)$.

Au total, la complexité de l'algorithme est donc $\mathscr{O}((m+n)\mathrm{K}r + 2\mathrm{K}r^2)$ opérations. On note que cette méthode est intéressante à partir du moment ou le nombre initial de nœuds d'interpolation K est petit comparé aux dimensions de la matrice. C'est typiquement le cas des matrices de discrétisation des opérateurs intégraux associés au noyau de Green du Laplacien.

Approche analytique (HCA II)

Obtention de la décomposition L'application d'une méthode d'approximations croisées à la matrice de couplage revient à écrire la décomposition suivante

$$\mathrm{S}|_{t\times s} = \mathrm{S}|_{t\times\hat{s}}(\mathrm{S}|_{\hat{t}\times\hat{s}})^{-1}\mathrm{S}|_{\hat{t}\times s}, \tag{2.106}$$

où $\hat{t}$ et $\hat{s}$ sont respectivement les r lignes et les r colonnes extraites de la matrice S.

D'après les résultats décrit à la section 2.3.2, ceci nous permet de construire la décomposition $\tilde{G}$ du noyau (*c.f* 2.77) :

$$\tilde{G}(x,y) \simeq \sum_{l=1}^{r}\left(\sum_{q=1}^{l} C_{lq}.G(x,\xi_{\mu_q}^{B_s})\right)\left(\sum_{q=1}^{l} D_{lq}.G(\xi_{\nu_q}^{B_t},y)\right), \tag{2.107}$$

où C et D sont les matrices de coefficients déterminées par l'algorithme 12 et on peut utiliser l'expression 2.107 pour obtenir

$$\begin{aligned} M_{ij} &\simeq \int_\Gamma\int_\Gamma \tilde{G}(x,y)\Phi_i^t(x)\Phi_j(y)d\Gamma(x)d\Gamma(y) \\ &= \sum_{l=1}^{r}\left(\sum_{q=1}^{l} C_{lq}.\int_\Gamma G(x,\xi_{\mu_q}^{B_s})\Phi_i^t(x)d\Gamma(x)\right)\left(\sum_{q=1}^{l} D_{lq}.\int_\Gamma G(\xi_{\nu_q}^{B_t},y)\Phi_j(y)d\Gamma(y)\right) \end{aligned}$$

Cette écriture correspond à un produit de quatre matrices,

$$M \simeq (UC^T).(VD^T)^T, \tag{2.108}$$

où C et D sont les matrices de coefficients déterminées par l'algorithme 12 et,

$$U_{iq} = \int_\Gamma \Phi_i^t(x).G(x,\xi_{\mu_q}^{B_s})d\Gamma(x), \tag{2.109}$$

$$V_{jq} = \int_\Gamma \Phi_j(y).G(\xi_{\nu_q}^{B_t},y)d\Gamma(y). \tag{2.110}$$

En posant $A = UC^T$ et $B = VD^T$, on obtient l'approximation de rang r de la forme souhaitée.

On remarque que les matrices U et V dont les coefficients sont donnés par des intégrales simples sont de dimensions plus petites que celles mises en jeu dans l'algorithme HCA – I. En effet, elle ne contiennent que le nombre de nœuds d'interpolation sélectionnés par la compression ACA de la matrice de couplage. Dans le cas où le nombre initial K de points d'interpolation est élevé, ceci s'avère être un gain non négligeable.

Complexité On note r le rang de l'approximation obtenue par l'algorithme HCA – I. On rappelle que le nombre de nœuds d'interpolation est K La complexité de l'algorithme HCA – II est la suivante

- Calcul de la matrice U de taille $m \times r$: $\mathcal{O}(mr)$.
- Calcul de la matrice V de taille $n \times r$: $\mathcal{O}(nr)$.
- Compression de la matrice de couplage : $\mathcal{O}(2Kr^2)$.
- Calcul des matrices C et D : $\mathcal{O}(r^2)$.
- Produit $A = UC^T$: $\mathcal{O}(mr^2)$.
- Produit $B = VD^T$: $\mathcal{O}(nr^2)$.

Au total, la complexité de l'algorithme est donc $\mathcal{O}((m+n)r^2+2Kr^2)$ opérations. Le nombre de nœuds d'interpolation K n'intervient dans cet algorithme qu'avec un facteur r^2 contrairement à HCA – I pour lequel on a un facteur $(m+n$ qui peut être grand. Ultérieurement, on utilisera cet algorithme pour approcher le noyau oscillant. Le noyau de l'équation de Helmholtz recquiert lui plus de points d'interpolation, ce qui nous fait préférer HCA – II à l'algorithme HCA – I.

Remarque 2.66 (Intégrales simple pour les deux versions). On note, en vue d'un prochain chapitre que le coût de calcul des intégrales varie suivant la méthode utilisée, HCA – I ou HCA – II. En effet, la méthode HCA – I n'utilise que des évaluations de polynômes pour le calcul des intégrales tandis que HCA – II utilise l'évaluation du noyau. Bien que l'on ait moins d'intégrales à déterminer dans le cas d'HCA – II, le calcul unitaire peut être plus élevé dans le cas du noyau oscillant par exemple. L'évaluation du noyau en un point revient à un calcul de cosinus et de sinus. Ce type de calcul est plus élevé qu'une somme ou une mltiplication comme dans le cas du noyau de l'équation de Laplace.

2.3.4.4 Estimation d'erreur

Par construction, l'approximation de rang faible produite par la méthode HCA utilise à la fois l'interpolation polynomiale ainsi que la compression ACA. Ceci conduit naturellement à obtenir une estimation de l'erreur tenant compte de ces deux méthodes. Le résultat suivant, présentée initialement dans [BG05] (théorème 26) puis dans [BGH12] (théorème 4.11) fournit une estimation de l'erreur obtenue pour la construction de l'approximation 2.107. Une première application du schéma d'interpolation amenant une erreur $\epsilon_{\text{int}}(G)$ permet de construire la matrice de couplage S de dimension $K \times K$. On se place dans l'hypothèse où l'erreur ponctuelle ϵ_{ACA} demandée à l'algorithme ACA vérifié les inégalités 2.78. On utilise alors les pivots sélectionnés par l'algorithme ACA afin d'obtenir une seconde interpolation polynomiale donnant 2.107 avec une erreur $\epsilon_{\text{int}}(\tilde{G})$.

Théorème 2.67 (Estimation d'erreur pour HCA)**.** *L'application de la méthode* HCA – II *pour approcher le noyau de Green dans* $B_t \times B_s$ *par son approximation 2.107* $\tilde{G}$*, fournit l'erreur suivante, pour* $x \in B_t$ *et* $y \in B_s$

$$|G(x,y) - \tilde{G}(x,y)| \leqslant \epsilon_{\text{int}}(G) + \epsilon_{\text{int}}(G) + \Lambda_m^6 \epsilon_{\text{ACA}} \tag{2.111}$$

avec la constante de stabilité Λ_m *utilisée par le schéma d'interpolation polynomiale et* ϵ_{ACA} *l'erreur ponctuelle due à la compression de la matrice de couplage vérifiant*

$$|(S - \hat{A}\hat{B})_{ij}| \leqslant \epsilon_{\text{ACA}}. \tag{2.112}$$

[BGH12] établit la remarque que dans la pratique on obtient une estimation de la forme

$$|G(x,y) - \tilde{G}(x,y)\| \lesssim \max(\epsilon_{\text{int}}, \epsilon_{\text{ACA}}). \tag{2.113}$$

[Djo06],[BGH12] et [BG05] fournissent des tests numériques amenant une estimation suivante

$$|G(x,y) - \tilde{G}(x,y)\| \leqslant \alpha_1 \epsilon_{\text{int}} + \alpha_2 \epsilon_{\text{ACA}}, \tag{2.114}$$

où $\alpha_1 << 1$ et $\alpha_2 \approx 1$ pour l'opérateur de simple couche tandis que $\alpha_2 \simeq 30$ pour le potentiel de double couche. Dans la pratique, pour une précision ϵ finale visée, on augmente légérement l'ordre d'interpolation et on utilise une précision $\epsilon_{\text{ACA}} \approx \epsilon/30$ pour traiter les dérivées du noyau.

L'estimation d'erreur pour l'approximation $\tilde{M}$ de la matrice de discrétisation M suit le même comportement que l'erreur du théorème précédent (*cf* **??** corollaire 27),

$$\|M - \tilde{M}\| \leqslant \beta_1 \epsilon_{\text{int}} + \beta_2 \epsilon_{\text{ACA}}, \tag{2.115}$$

où les constantes β_1 et β_2 dépendent de la géométrie des supports de t et de s ainsi que de la discrétisation de ces supports pour le calcul des intégrales. De même, on note une faible dépendance de l'erreur d'interpolation.

2.4 Conclusion

La compression d'une matrice M de taille $m \times n$ par une représentation tensorielle de rang r est au cœur des méthodes rapides pour des solveurs directs et itératifs dans le cadre de la discrétisation par la méthode des éléments de frontières. Cette méthode consiste à exprimer la solution à l'aide d'opérateurs intégraux faisant intervenir le noyau de Green associé au problème. La discrétisation des ces opérateurs conduit à traiter des blocs matriciels dont les coefficients sont déterminés à l'aide d'une intégrale double faisant intervenir le noyau de Green. De ce caractère découle deux approches possibles.

La première approche, algébrique, consiste à approcher la matrice à l'aide de méthodes algébriques. La décomposition en valeurs singulières permet de justifier l'existence d'une représentation par produits tensoriels efficace et de rang faible (le meilleur possible !) au prix d'une complexité en $\mathcal{O}(mn\min(m,n))$ en déterminant des bases réduites de l'image et du noyau de la matrice. D'autres méthodes telle que la factorisation QR permettent la détermination de base réduite en $\mathcal{O}(mnr)$ opérations. L'ajout d'une partie probabiliste -à travers le produit par une matrice test- permet d'accélérer les méthodes algébriques afin de réduire la complexité à $\mathcal{O}\left(mn\log(l) + rln\log(n)\right)$ avec $l \sim r$. Ces méthodes sont très robustes et ont une probabilité d'échec très faible tout en conservant une estimation d'erreur convenable. Enfin, nous savons que les matrices considérées héritent des bonnes propriétés du noyau de Green et dès lors, une version heuristique de l'élimination de Gauss permet d'abaisser la complexité à $\mathcal{O}\left((m+n)r^2\right)$ en trouvant un rang proche de celui de la SVD. Cette dernière méthode est un point essentiel de la méthode des $\mathscr{H}$-matrices car sa complexité est linéaire en la dimension de la matrice.

La seconde approche consiste à obtenir une approximation analytique du noyau de Green et de l'utiliser pour le calcul rapide des coefficients de la matrice. La première étape consiste à se placer dans un cas où les variables x et y sont suffisamment éloignées pour s'éloigner de la singularité du noyau. C'est la condition d'admissibilité que l'on applique afin d'éloigner les variables. Ceci correspond à réaliser une séparation en champ proche et en champ lointain. Dans l'hypothèse supplémentaire où le noyau est régulier et asymptotiquement lisse, la troncature de son développement de Taylor fournit une approximation sous forme de produits tensoriels à variables séparées. Ce type d'approximation conduit naturellement à une représentation de rang faible pour la discrétisation de la matrice. Cependant, le développement de Taylor recquiert un nombre important de dérivées d'ordres supérieurs du noyau et cela s'avère limitant pour un code de calcul performant. Comme la décomposition en valeurs singulières, le développement de Taylor fournit néamoins des estimations du rang et on peut l'utiliser pour construire une version analytique de l'algorithme ACA. Ce nouvel outil permet alors d'obtenir une approximation du noyau correcte avec une erreur maîtrisée pourvu que l'on soit dans les hypothèse de régularité de d'admissibilité précédentes. Une autre méthode, facilement utilisable dans un code de calcul, est l'interpolation polynomiale. Plusieurs résultats existent afin d'assurer une bonne stabilité de l'approximation du noyau ainsi que de ses dérivées quitte à utiliser suffisament de nœuds d'interpolation. Ce nombre pouvant être élevé suivant les applications, la méthode HCA combine un schéma d'interpolation avec une compression algébrique d'une matrice d'évaluations du noyau aux nœuds d'interpolation. Cette compression algébrique permet une réduction importante du nombre de nœuds en ne sélectionnant que ceux dignes d'intérêts pour l'approximation. On peut alors obtenir un schéma d'approximation stable et précis en $\mathcal{O}\left((m+n)r^2 + 2\mathrm{K}r^2\right)$ opéra-

tions pour le noyau de Green. Cette complexité est alors la même que pour la méthode algébrique ACA. Par ailleurs, l'approche mixte du HCA ouvre la voie pour l'obtention d'une approximation efficace du noyau oscillant quitte à correctement déterminer le nombre intial K de nœuds d'interpolation.

2.5 Références

[Beb00a] M Bebendorf. *Hierarchical Matrices.* Springer, 2000. 82

[Beb00b] M Bebendorf. *Hierarchical Matrices.* Springer, 2000. 86, 87, 88, 89, 94, 95

[BG65] P. A. Businger and G. H. Golub. Linear least squares solutions by householder transformations. *Numer. Math.*, 7 :269–276, 1965. 69

[BG05] S. Börm and L. Grasedyck. Hybrid cross approximation for integral operators. *Numerische Mathematik*, 2005. 97, 105, 106, 109

[BGH12] Steffen Börm, Lars Grasedyck, and Wolfgang Hackbusch. Hierarchical matrices. Technical report, 2012. 82, 83, 84, 86, 87, 88, 98, 99, 101, 104, 109

[CGMR05] H. Cheng, Z. Gimbutas, P.-G. Martinsson, and V. Rokhlin. On the compression of low-rank matrices. *SIAM J. Sci. Comput.*, 2005. 70, 76, 77, 78

[Djo06] J. Djokic. *Efficient Update of Hierarchical Matrices in the case of Adaptive Discretisation Schemes.* PhD thesis, Leipzig University, 2006. 109

[DKM06a] P. Drineas, R. Kannan, and M. W. Mahoney. Fast montecarlo algorithms for matrices i : Approximating matrix multiplication. *SIAM J. Comput.*, 36(1) :132–157, 2006. 81

[DKM06b] P. Drineas, R. Kannan, and M. W. Mahoney. Fast montecarlo algorithms for matrices ii : Computing a low-rank approximation to a matrix. *SIAM J. Comput.*, 36(1) :158–183, 2006. 81

[DKM06c] P. Drineas, R. Kannan, and M. W. Mahoney. Fast montecarlo algorithms for matrices iii : Computing a compressed approximate matrix decomposition. *SIAM J. Comput.*, 36(1) :184–206, 2006. 81

[GE94] M. Gu and S. C. Eisenstat. An efficient algorithm for computing a strong rank-revealing qr factorization. *Research Report YALEU*, 1994. 69, 70, 77, 78

[GL96] G.H. Golub and C.F. Van Loan. *Matrix Computations (3e éd.).* Johns Hopkins Studies in the Mathematical Sciences, 1996. 65, 67, 68, 69

[GTZ97] S. A. Goreinov, E. E. Tyrtyshnikov, and N. L. Zamarashkin. A theory of pseudoskeleton approximations. *Linear Algebra and its Applications*, 1997. 76, 80

[HG11] K. L. Ho and L. Greengard. A fast direct solver for structured linear systems by recursive skeletonization. *Arxiv*, 2011. 76

[HMT10] N. Halko, P.-G. Martinsson, and J. A. Tropp. Finding strucutre with randomness : Probabilistic algorithms for constructing approximate matrix decompositions. *Arxiv*, 2010. 72, 74, 78

[HP92] Y.P. Hong and C.-T. Pan. Rank-revealing qr factorizations and the singular value decomposition. *Mathematics of Computations*, 58(197) :213–232, 1992. 68, 69

[HUR15] A. Heldring, E. Ubeda, and J.M. Rius. Stochastic estimation of the frobenius norm in the aca convergence criterion. *IEEE Trans. on Antennas and Propagation*, 63(3), 2015. 88

[Liz14] B. Lizé. *Résolution Directe Rapide pour les Éléments Finis de Frontière en Électromagnétisme et Acoustique : H-Matrices. Parallélisme et Applications Industrielles.* PhD thesis, Université Paris 13, 2014. 84

[LMPLH09] J. Laviada, R. Mittra, M. R. Pino, and F. Las-Heras. On the convergence of the aca. *Microwave and Optical Technology Letters*, 51(10) :2458–2460, 2009. 88

[LWPG+07] E. Liberty, F. Woolfe, Martinsson P.-G., V. Rokhlin, and M. Tygert. Randomized algorithms for the low-rank approximation of matrices. *PNAS*, 2007. 73, 74, 75, 76, 78

[Mes11] M. Messner. *Fast Boundary Element Methods in Acoustics.* PhD thesis, Graz University of Technology,Institute of Applied Mechanics, 2011. 86, 98

[MG03] L. Miranian and M. Gu. Strong rank revealing lu factorizations. *Linear Algebra and its Applications*, 367 :1–16, 2003. 69, 70

[MH03] J.C Mason and D.C. Handscomb. *Chebyshev Polynomials.* Chapman and Hall/CRC, 2003. 99

[MRT06] P.-G. Martinsson, V. Rokhlin, and M. Tygert. On interpolation and integration in finite-dimensional spaces of bounded functions. *Communication in Applied Mathematics and Computational Science*, 2006. 76

[MT11] P.-G. Martinsson and M. Tygert. Multilevel compression of linear operators : Descendants of fast multipole methods and calderòn-zygmund theory. 2011. 76, 77, 78

[Sha08] J. Shaeffer. Direct solve of electrically large integral equations for problem sizes to 1m unknowns. Technical report, 2008. 86

[SX95] T. Sauer and Y. Xu. On multivariate lagrange interpolation. *Mathematics of Computation*, 64(211) :1147–1170, 1995. 95

[WLRT07] F. Woolfe, E. Liberty, V. Rokhlin, and M. Tygert. A fast randomized algorithm for the approximation of matrices. *Technical Report YALEU*, 2007. 74, 76, 78, 79

[ZVL05] K. Zhao, M.N. Vouvakis, and J.-F. Lee. The adaptative cross approximation algorithm for accelerated method of moments computations of emc. *IEEE Trans Electromagn Compat*, 47 :763–773, 2005. 84

Chapitre 3

Approximation du noyau oscillant

Sommaire

3.1 Développement limité de la phase

3.1.1 Notations

On décrit les notations utilisées dans cette partie. La base canonique de $\mathbb{R}^3$ est notée (e_1, e_2, e_3). Pour deux points x et y distants, on considère deux sphères $\mathbb{S}_x$ et $\mathbb{S}_y$ de rayon a et de centres c_x et c_y contenant respectivement les points x et y.

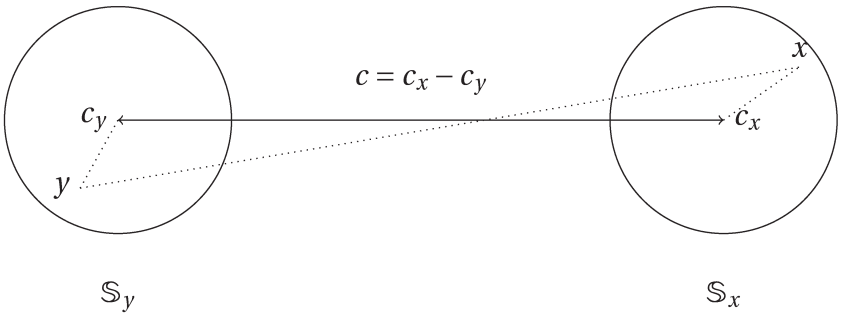


FIGURE 3.1 – Position des groupes d'inconnues et notations

On note $\mathrm{R} = \|c\|_2$ la distance entre les centres des sphères et $\vec{u}_3 = c/\|c\|_2$ est le vecteur directeur de norme 1 décrivant la direction des centres. On introduit par ailleurs $d_x = c_x - x$, $d_y = c_y - y$ ainsi que $d = d_x - d_y$. Ainsi, pour tout $x \in \mathbb{S}_x, y \in \mathbb{S}_y$,

$$|x - y| = |c + d|, \tag{3.1}$$

où c est un vecteur constant de norme R. Seul le vecteur d varie lorsque x et y changent.

On note (u_1, u_2) une base orthonormée du plan Π_{u_3} normal à u_3 avec l'origine le point $\frac{c_x+c_y}{2}$. On complète cette base par le vecteur u_3 afin de former une base (u_1, u_2, u_3) de $\mathbb{R}^3$. Dans cette nouvelle base, on a

$$x = \xi_1 u_1 + \xi_2 u_2 + \xi_3 u_3, \tag{3.2}$$
$$y = \eta_1 u_1 + \eta_2 u_2 + \eta_3 u_3, \tag{3.3}$$
$$c_x = \frac{\mathrm{R}}{2} u_3, \tag{3.4}$$
$$c_y = -\frac{\mathrm{R}}{2} u_3. \tag{3.5}$$

3.1.2 Développement limité du noyau de Green

On effectue un développement limité de la phase du noyau de Green de l'équation de Helmholtz, $\mathrm{G}(x, y) = e^{ik|x-y|}/|x - y|$.

3.1.2.1 Développement limité de la phase

Ceci revient à effectuer le développement limité de la quantité $|x - y|$ lorsque x et y sont suffisament éloignés l'un de l'autre.

Avec les notations précédentes, on a $|x-y| = |c+d|$. On se place dans le cas où $\frac{|d|}{\mathrm{R}}$ est petit et on effectue un développement limité de la quantité $|c+d|$. On a

$$\begin{aligned} |c+d|^2 &= < c+d, c+d > \\ &= \mathrm{R}^2 + |d|^2 + 2 < c, d > \\ &= \mathrm{R}^2(1+s), \end{aligned}$$

où $s = |d|^2/\mathrm{R}^2 + 2 < c, d > /\mathrm{R}^2$. On rappelle que par définition, on a $u_3 = c/\mathrm{R}$ et ainsi, $s = |d|^2/\mathrm{R}^2 + 2 < u_3, d > /\mathrm{R}$. En utilisant le développement limité de $\sqrt{1+s}$ et en négligeant les termes d'ordre 5 et plus en $|d|$, on obtient

$$\begin{aligned} |x-y| = \mathrm{R} &+ < u_3, d > + \frac{(|d|^2 - < u_3, d >^2)}{2\mathrm{R}} \\ &+ \frac{(< u_3, d >^3 - |d|^2 < u_3, d >)}{2\mathrm{R}^2} \\ &+ \frac{6|d|^2 < u_3, d >^2 - 5 < u_3, d >^4 - |d|^4}{8\mathrm{R}^3} \\ &+ \mathcal{O}(|d|^5) \end{aligned} \tag{3.6}$$

Le développement précédent fait intervenir les deux quantités $|d|$ et $< u_3, d >$. On décompose le vecteur d dans la base (u_1, u_2, u_3) afin d'isoler les comportements dans la direction u_3 et dans le plan transverse Π_{u_3}

$$\begin{aligned} d &= d_{\parallel}.u_3 + d_{\perp}, \\ d_{\parallel} &= < d, u_3 > . \end{aligned}$$

Par orthogonalité, on a

$$|d|^2 = d_{\parallel}^2 + |d_{\perp}|^2,$$

et par substitution dans (3.6) il vient,

$$|x-y| = \mathrm{R} + d_{\parallel} + \frac{|d_{\perp}|^2}{2\mathrm{R}} - \frac{|d_{\perp}|^2 d_{\parallel}}{2\mathrm{R}^2} + \frac{|d_{\perp}|^2(4d_{\parallel}^2 - |d_{\perp}|^2)}{8\mathrm{R}^3} + \mathcal{O}(|d|^5) \tag{3.7}$$

3.1.2.2 Écriture du noyau dans la base (u_1, u_2, u_3)

On réécrit le noyau G(x,y) dans la base (u_1, u_2, u_3) à l'aide du développement (3.7) et des variables $d_{\parallel}$ et $d_{\perp}$. On note

$$\mathrm{G}_0(x,y) = \frac{e^{ik\mathrm{R}}}{|x-y|}, \text{(terme d'ordre 0)} \tag{3.8}$$

$$\mathrm{G}_1(x,y) = e^{ikd_{\parallel}}, \text{(terme d'ordre 1)} \tag{3.9}$$

$$\mathrm{G}_2(x,y) = e^{+ik\frac{|d_{\perp}|^2}{2\mathrm{R}}}, \text{(terme d'ordre 2)} \tag{3.10}$$

$$\mathrm{G}_3(x,y) = e^{-ik\frac{|d_{\perp}|^2 d_{\parallel}}{2\mathrm{R}^2}}, \text{(terme d'ordre 3)} \tag{3.11}$$

$$\mathrm{G}_4(x,y) = e^{ik\frac{|d_{\perp}|^2(4d_{\parallel}^2 - |d_{\perp}|^2)}{8\mathrm{R}^3}}. \text{(terme d'ordre 4)} \tag{3.12}$$

Ainsi, le noyau se décompose sous la forme suivante

$$\mathrm{G}(x,y) = \mathrm{G}_0(x,y).\mathrm{G}_1(x,y).\mathrm{G}_2(x,y).\mathrm{G}_3(x,y).\mathrm{G}_4(x,y).\tilde{\mathrm{G}}(x,y), \tag{3.13}$$

où $\tilde{\mathrm{G}}(x,y)$ contient les termes d'ordres supérieurs ou égaux à 5. On remarque que le terme $\mathrm{G}_1(x,y)$ défini par 3.9 n'est autre qu'un produit d'ondes planes dans la direction u_3.

À l'aide de l'écriture 3.13 on peut obtenir des approximations à des ordres variables pourvu que l'on contrôle les oscillations des termes d'ordres supérieurs. Par exemple, pour décrire le noyau complet $\mathrm{G}(x,y)$ par les ondes planes, il suffit de contrôler les oscillations du terme $\mathrm{G}_2(x,y)$ car les autres termes sont plus petits que le terme d'ordre 2. La relation obtenue pour le contrôle des oscillations des termes d'ordres supérieurs correspond à une relation d'admissibilité fréquentielle, c'est-à-dire dépendant du nombre d'onde k.

3.1.3 Admissibilité fréquentielle pour le noyau $\frac{e^{ik|x-y|}}{|x-y|}$

On considère le noyau de Green de l'équation de Helmholtz pour une fréquence non nulle, soit $k>0$. L'emploi du développement limité de la phase (3.6) recquiert la condition

$$\frac{|d|}{\mathrm{R}} < 1. \tag{3.14}$$

Il s'agit de la condition d'admissibilité pour le noyau $1/|x-y|$ que l'on a déjà obtenue en utilisant le développement de Taylor.

Sous réserve de satisfaire la condition (3.14), on présente deux critères d'admissibilité fréquentiels correspondant respectivement à des approximations du premier ordre et du second ordre du noyau.

3.1.3.1 Admissibilité de type Fraunhofer

L'approximation consistant à conserver le terme d'ordre 1 On souhaite contrôler les oscillations des termes d'ordre supérieurs ou égaux à 2. Pour cela, il suffit de contrôler les oscillations de 3.10, c'est-à-dire que l'on veut borner la phase de $\mathrm{G}_2(x,y)$ de la sorte

$$k\frac{|d_\perp|^2}{2\mathrm{R}} \leqslant \alpha.2\pi,\ \alpha<1. \tag{3.15}$$

Cette condition correspond à l'approximation de Fraunhofer qui consiste à considérer que dans la zone de champ lointain, le noyau de Green est correctement décrit par les ondes planes. Il s'agit d'une approximation très largement utilisée dans la pratique.

3.1.3.2 Admissibilité de type Fresnel

La prise en compte du terme d'ordre 2 dans l'approximation du noyau recquiert de prendre en compte le comportement de $d_\parallel$. En effet, si celui-ci est nul, il convient d'étudier le terme suivant et de borner les oscillations de 3.12. On suppose dans un premier temps que $|d_\parallel|>0$ pour x et y donnés. Alors, on a la condition

$$k\frac{|d_\perp|^2|d_\parallel|}{2\mathrm{R}^2} \leqslant \alpha.2\pi,\ \alpha<1. \tag{3.16}$$

Si $|d_\parallel| = 0$, le contrôle des oscillations du terme suivant donne la condition d'admissibilité suivante

$$k\frac{|d_\perp|^4}{8R^3} \leqslant \alpha.2\pi,\ \alpha < 1. \tag{3.17}$$

Remarque 3.1 (Lien avec l'optique). Le cas où $|d_\parallel| = 0$ correspond au cas où x et y reposent respectivement sur des plans en opposition. Il s'agit d'un cas fréquent en optique lorsqu'on observe le motif de diffraction sur un écran.

Remarque 3.2 (Condition de Fresnel). Pour éviter d'avoir à distinguer les cas selon la valeur de $d_\parallel$, on peut regrouper les termes d'ordre 3 et 4 sous un seul et même critère d'admissibilité :

$$k\left(\frac{|d_\parallel||d_\perp|^2}{2R^2} + \frac{|d_\perp|^2(4|d_\parallel|^2 + |d_\perp|^2)}{8R^3}\right) \leqslant \alpha.2\pi,\ \alpha < 1. \tag{3.18}$$

3.1.4 Zones d'admissibilité fréquentielles

3.1.4.1 Expression de $d_\perp$ et $d_\parallel$ dans (u_1, u_2, u_3)

On note $(d^{(1)}, d^{(2)}, d^{(3)})$ les coordonnées du vecteur d dans la base (u_1, u_2, u_3). Avec les notations précédentes, on a

$$\begin{aligned} d^{(1)} &= \xi_1 - \eta_1, \\ d^{(2)} &= \xi_2 - \eta_2, \\ d^{(3)} &= \xi_3 - \eta_3 - R, \end{aligned}$$

d'où

$$\begin{aligned} d_\parallel &= \xi_3 - \eta_3 - R, \\ |d_\perp|^2 &= (\xi_1 - \eta_1)^2 + (\xi_2 - \eta_2)^2. \end{aligned}$$

3.1.4.2 Majoration de $d_\perp$ et $d_\parallel$ en fonction de a

Dans le cas des deux sphères de rayon a décrites en 3.1.1, on a $\xi_1 \in [-a, a]$, $\xi_2 \in [-a, a]$, $\eta_1 \in [-a, a]$, $\eta_2 \in [-a, a]$ et $\xi_3 \in [\frac{R}{2} - a, \frac{R}{2} + a]$,$\eta_3 \in [-\frac{R}{2} - a, -\frac{R}{2} + a]$. En notant $D = 2a$ le diamètre des sphères, on a

$$\begin{aligned} |d_\parallel| &\leqslant D = 2a, \\ |d_\perp|^2 &\leqslant 2D^2 = 8a^2. \end{aligned}$$

Dans le cas des deux sphères de rayon a décrites en 3.1.1, on a $|d| = |d_x - d_y| \leqslant 2a$. Le diamètre $D = 2a$ des sphères est alors une borne maximum de $|d|$ et la condition d'admissibilité (adimensionnée par la logueur d'onde λ) est alors

$$\frac{D}{\lambda} \leqslant \frac{R}{\lambda}. \tag{3.19}$$

On rappelle que le nombre d'onde k est lié à la longueur d'onde λ par $k = \frac{2\pi}{\lambda}$. Pour le critère de Fraunhofer, en remplaçant $|d_\perp|$ et $|d_\parallel|$ par leur valeurs, on obtient

$$\left(\frac{D}{\lambda}\right)^2 \leqslant \alpha\frac{R}{\lambda}. \tag{3.20}$$

Pour le critère de Fresnel, on a

$$\left(\frac{D}{\lambda}\right)^3 \leqslant \alpha\left(\frac{R}{\lambda}\right)^2. \tag{3.21}$$

On représente les diférentes zones d'admissibilités en fonction du diamètre en longueurs d'onde à partir des relations adimensionnées (3.19), (3.20) et (3.21).

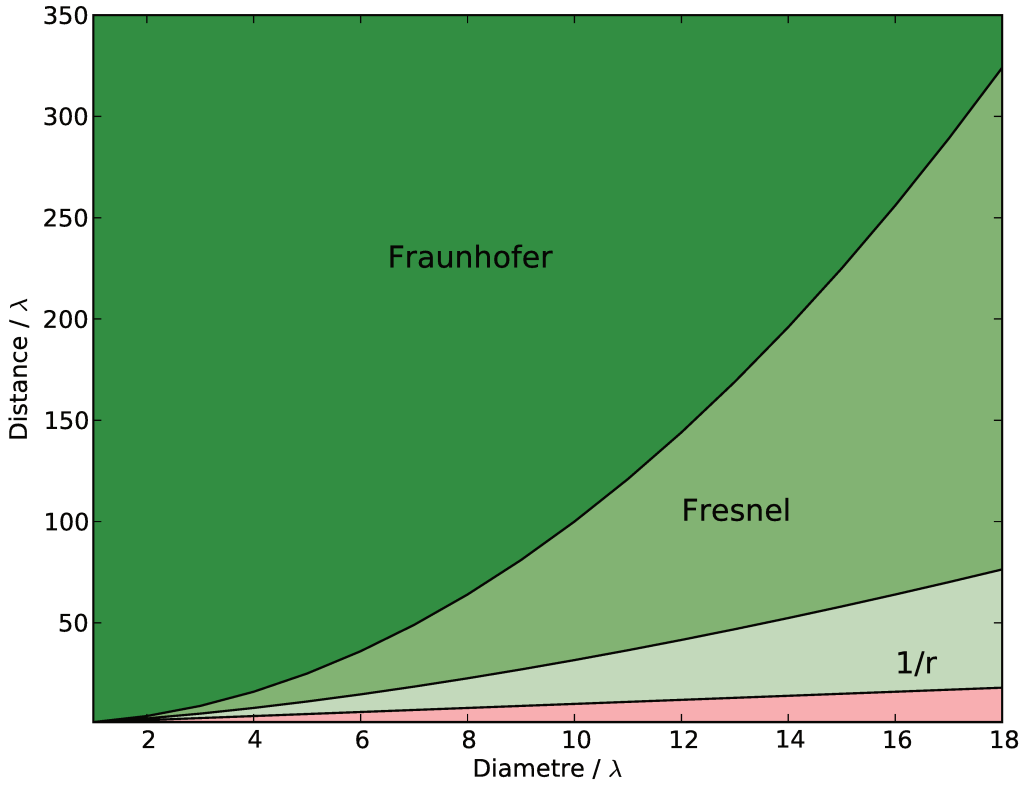


FIGURE 3.2 – Représentation graphique des zones d'admissibilités fréquentielles. En abscisse, le diamètre en longueurs d'onde. En ordonnée, la distance en longueurs d'onde.

La représentation ci-dessus permet d'établir une comparaison des différents critères d'admissibilité. Pour une sphère de diamètre D donné, une autre sphère de diamètre D est Fraunhofer-admissible si la distance les séparant vérifie le critère (3.15), on dit qu'elle est dans la zone de Fraunhofer (noté "Fraunhofer sur la figure"). Dans cette zone, les ondes planes fournissent une bonne approximation du noyau. De même, elle est Fresnel-admissible si la distance vérifie le critère (3.18) (noté "Fraunhofer sur la figure"). Si cette distance ne satisfait pas les deux critères fréquentielles mais est supérieur au diamètre, il s'agit de la zone dite en "1/r".

Pour une petite sphère, c'est-à-dire une sphère dont le diamètre ne représente que peu de longueurs d'onde (comportement basse fréquence), les zones sont plus resserées. Au contraire, pour de grandes sphères (comportement haute fréquence) elles sont plus grandes. L'introduction du critère de Fresnel possède un avantage immédiat en ce sens où la limite de champ lointain est plus proche que celle obtenue par le critère de Fraunhofer. En effet, d'après les conditions (3.15) et (3.18), la distance varie comme le carré du diamètre dans la zone de Fraunhofer tandis que dans la zone de Fresnel, la distance varie comme la puissance 3/2 du diamètre.

3.2 Application au rang du noyau

On utilise la décomposition 3.13 obtenu dans la base (u_1, u_2, u_3) afin d'étudier le rang du noyau. Plus particulièrement, on cherche à obtenir une approximation à variables séparées du noyau par une méthode de type HCA−II tout en contrôlant le nombre de points d'interpolation nécessaire.

On s'intéresse aux trois premiers termes de l'approximation 3.13 et et on écrit la décomposition suivante du noyau dans la base (u_1, u_2, u_3)

$$\mathrm{G}(x,y) = \frac{1}{|x-y|} e^{ik(\xi_3-\eta_3)} e^{\frac{ik}{2\mathrm{R}^2}(\xi_1-\eta_1)^2} e^{\frac{ik}{2\mathrm{R}^2}(\xi_2-\eta_2)^2} \tilde{\mathrm{G}}(x,y), \tag{3.22}$$

$\tilde{\mathrm{G}}(x,y)$ regroupant les termes d'ordre 3 et plus dans le développement de la phase précédent. On remarque d'ores et déjà que cette écriture permet de découpler les contributions dans le plan Π_{u_3} : le terme quadratique est un produit cartésien de deux fonctions de $\mathbb{R}\times\mathbb{R} \mapsto \mathbb{C}$.

La motivation de cette décomposition est d'obtenir une approximation de rang fini du noyau. En effet, on rappelle que l'approximation voulue est de la forme

$$\mathrm{G}(x,y) \approx \sum_{q=1}^{r} u_q(x)\, v_q(y), \tag{3.23}$$

où r est le rang approché du noyau. Ce rang dépend de la précision voulue ainsi que de la fréquence dans le cas du noyau oscillant. On peut utiliser la décomposition 3.22 afin de déterminer une borne maximum sur l'entier r en cherchant une approximation dégénérée de chacun des termes. Pour chaque terme de la décompositon 3.22, on note rg(.) l'entier désignant le rang approché par une approximation du type 3.23. Ainsi, on a

$$r \leqslant \mathrm{rg}\left(\frac{1}{|x-y|}\right).\mathrm{rg}\left(e^{ik(\xi_3-\eta_3)}\right)\mathrm{rg}\left(e^{\frac{ik}{2\mathrm{R}^2}(\xi_1-\eta_1)^2}\right).\mathrm{rg}\left(e^{\frac{ik}{2\mathrm{R}^2}(\xi_2-\eta_2)^2}\right).\mathrm{rg}\left(\tilde{\mathrm{G}}(x,y)\right). \tag{3.24}$$

Cette estimation est une borne maximale, et on peut s'attendre à pouvor construire une approximation du noyau $\mathrm{G}(x,y)$ avec un rang plus petit en exploitant l'admissibilité de Fresnel introduite plus haut et en observant chacun des termes.

— Sous l'hypothèse d'admissibilité de Fresnel, on peut supposer que la phase de $\tilde{\mathrm{G}}$ n'oscille que très peu et que ce terme ne contribue que très peu au rang du noyau. Comme le terme en $1/|x-y|$ peut être approché par un noyau dégénéré de rang faible (à l'aide d'une méthode algébrique ou analytique comme HCA − II), on fait l'hypothèse que

$$\mathrm{rg}\left(\frac{\tilde{\mathrm{G}}(x,y)}{|x-y|}\right) \approx \mathrm{rg}\left(\frac{1}{|x-y|}\right), \tag{3.25}$$

et on note $\mathrm{N}_0 = \mathrm{rg}\left(\frac{\tilde{\mathrm{G}}(x,y)}{|x-y|}\right)$.

— Le terme en $(\xi_3-\eta_3)$ est le produit de deux ondes planes chacune étant de rang 1 et ne présente pas de difficulté pour construire une approximation dégénérée de rang fini. C'est ce terme qui est utilisé dans des développements basés sur un critère de type Fraunhofer : dans le champ lointain, la phase est correctement décrite par les ondes planes.

— Les termes quadratiques en $(\xi_1-\eta_1)^2$ et $(\xi_2-\eta_2)^2$ sont des fonctions de $\mathbb{R}\times\mathbb{R} \mapsto \mathbb{C}$. On note N_1 et N_2 respectivement le rang de chacune des contributions d'ordre 2.

Finalement, l'apport de chacun des termes sous la condition de Fresnel conduit à l'estimation suivante du rang,

$$r \lesssim \mathrm{N}_0.\mathrm{N}_1.\mathrm{N}_2 \tag{3.26}$$

Pour une erreur ϵ donnée, les résultats obtenus par le développement de Taylor pour le noyau $1/|x-y|$ suggèrent que

$$\mathrm{N}_0 \sim \mathscr{O}(\log(\epsilon)). \tag{3.27}$$

Pour les rangs N_1 et N_2, on cherche à présent la dépendance vis-à-vis de la fréquence et de la précison. La contribution du terme quadratique étant un produit cartésien de deux fonctions 1D de deux variables, il suffit d'étudier le noyau $x, y \mapsto e^{\frac{ik}{2\mathrm{R}}(x-y)^2}$.

3.3 Opérateur de Fox-Li

3.3.1 Définition

On introduit l'opérateur intégral associé au noyau $e^{i\alpha(x-y)^2}, \alpha > 0$ correspondant à l'une des contributions du terme d'ordre 2 dans le développement de la phase.

Définition 3.3 (opérateur de Fox-Li/Fresnel)**.** Pour des réels $d > 0$ et $\alpha > 0$, on considère l'opérateur intégral $\mathscr{F}_\alpha$ suivant,

$$\begin{aligned} \mathscr{F}_\alpha : \mathrm{L}^2([-d,d]) &\mapsto \mathrm{L}^2([-d,d]) \\ \lambda \mapsto [\mathscr{F}_\alpha \lambda](x) &= \int_{-d}^{d} e^{i\alpha(x-y)^2} \lambda(y)\, dy, \end{aligned} \tag{3.28}$$

Cet opérateur borné est dit opérateur de Fox-Li ou encore opérateur de Fresnel dans la littérature.

Pour nos applications, on a d'une part $\alpha = k/2\mathrm{R}$ et d'autre part on suppose que d et R sont liés par la relation d'admissibilité de Fresnel. On s'intéresse aux valeurs singulières de l'opérateur $\mathscr{F}_\alpha$ afin de déterminer un rang numérique à une précision donnée une fois construit une discrétisation. L'étude des valeurs singulières revient à considérer les valeurs propres de $\mathscr{F}_\alpha^* \mathscr{F}_\alpha$. Avant de procéder au calcul explicite de $\mathscr{F}_\alpha^* \mathscr{F}_\alpha$, on montre le comportement des valeurs singulières sur un exemple numérique simple.

3.3.2 Discrétisation $\mathbb{P}_0$ à un point de Gauss

On discrétise uniformément l'intervalle $[-d,d]$ à l'aide de 5 points par longueur d'onde et on considère une distance R telle que R et d satisfont le critère de Fresnel (3.16). On considère une discrétisation $\mathbb{P}_0$ à un point de Gauss de cet opérateur. Les coefficients de la matrice M de discrétisation de $\mathscr{F}_\alpha$ sont donnés par

$$\mathrm{M}_{ij} = e^{i\alpha(x_i - y_j)^2},$$

où x_i et y_j sont des points de la discrétisation de $[-d,d]$.

Remarque 3.4 (Matrice de Töplitz)*.* On note que la matrice obtenue est une matrice de Töplitz. Ce type de matrice est très particulier et la littérature à propos des valeurs propres ou du déterminant d'une telle matrice est très riche.

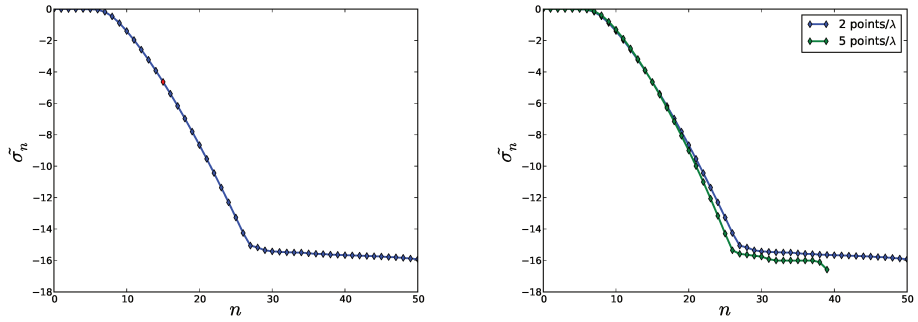
On peut obtenir les valeurs singulières de M à l'aide de la décomposition en valeurs singulières tel que l'exemple suivant l'illustre.

Exemple 3.5 (Valeurs singulières de M)**.** Les paramètres sont $d = 10\lambda$ et $R = 55\lambda$. On assemble la matrice de discrétisation M de taille $n \times n$ pour deux discrétisations uniformes différentes :

— Une discrétisation à 5 points par longueur d'onde : $n = 100$.

— Une discrétisation à 2 points par longueur d'onde : $n = 40$.

Bien que la dimension de la matrice de discrétisation varie d'un test à l'autre, ceci n'enlève aucune pertinence au test sur le comportement des valeurs singulières. Dans les deux cas, on cherche à approcher le même opérateur. On effectue alors la décomposition SVD dans chaque cas afin d'obtenir les valeurs singulières. De façon standard, on observe les valeurs normalisées afin d'observer l'erreur relative.



(a) Décroissance des valeurs singulières normalisées de M. Les 50 dernières valeurs singulières ne sont pas représentées. En rouge la valeur singulière à partir de laquelle l'erreur relative est inférieure à 1.10^{-4}.

(b) Comparaison de deux discrétisations uniformes.

FIGURE 3.3 – Décroissance des valeurs singulières normalisées de la matrice M dans le cas de l'exemple 3.5. L'échelle est logarithmique en ordonnée.

On constate que le profil de décroissance des valeurs singulières de la matrice de discrétisation de l'opérateur de Fox-Li est rapide. Les premières valeurs singulières normalisées sont proches de 1 puis tendent rapidement vers 0. Par exemple, dans le cas de la figure 3.3a, si l'on s'intéresse au rang numérique de cette matrice à la précision 1.10^{-4}, on obtient un rang numérique $r_\epsilon = 17$ pour une matrice de taille 100×100.

Dans le cas de la discrétisation à 5 points par longueur d'onde, on note que beaucoup de valeurs singulières sont inutiles et correspondent à une erreur proche de la précision machine. Ceci suggère qu'il n'est pas nécessaire de discrétiser l'intervalle avec 5 points par longueur d'onde et qu'une discrétisation plus grossière peut fournir un rang d'une aussi bonne qualité. C'est la comparaison représentée par la figure 3.3b. On compare les valeurs singulières déjà observées sur la figure 3.3a à celles obtenues par une discrétisation plus grossière. Pour cet exemple particulier, on remarque que les valeurs singulières sont à peu près les mêmes pour une erreur relative allant jusqu'à $\epsilon = 1.10^{-8}$.

Dans l'exemple suivant on procède au même test pour plusieurs valeurs du diamètre et de la distance.

Exemple 3.6 (Évolution du rang)**.** Pour la précison relative de $\epsilon = 1.10^{-4}$, on répète ce test et l'on représente le rang obtenu à l'aide d'une discrétisation à 2 points par longueur d'onde. On observe alors pour plusieurs couples (d, R) respectivement compris

entre $[\lambda, 50\lambda]$ et $[\lambda, 1000\lambda]$. On représente une cartographie des rangs obtenus par la décomposition en valeurs singulières que l'on tronque à $\epsilon = 1.10^{-4}$ en fonction du diamètre $2d$ et de la distance R, adimensionnés par la longueur d'onde.

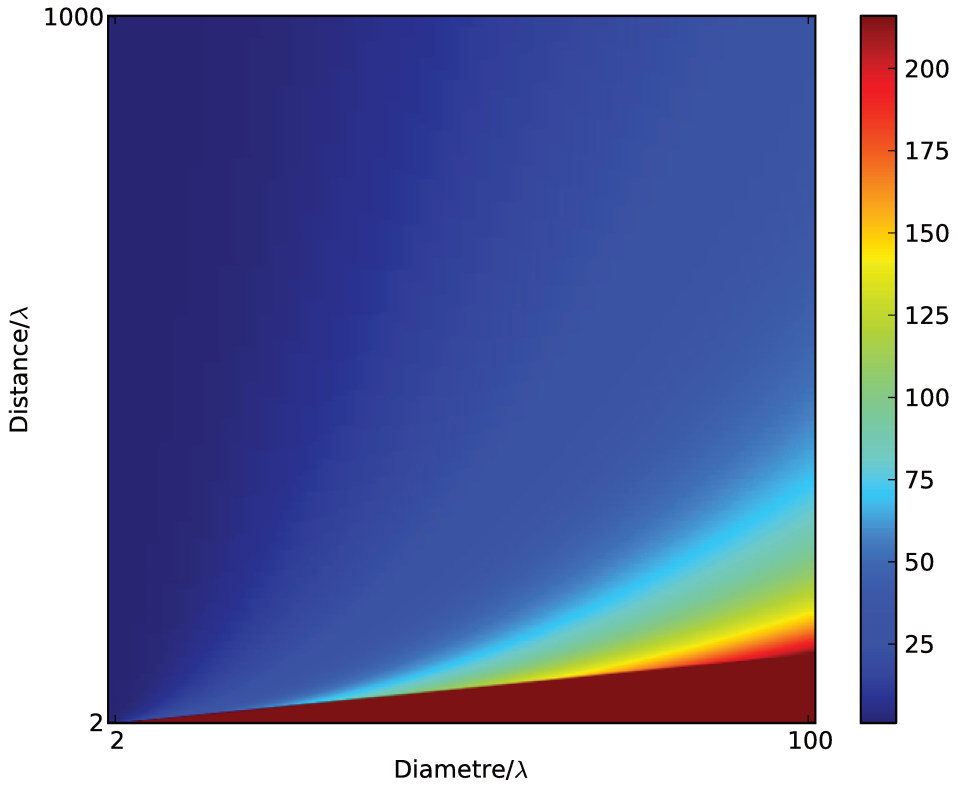


FIGURE 3.4 – Cartographie des rangs à la précision 1.10^{-4}. En rouge foncé, il s'agit de la zone proche correspondant à l'admissibilité $1/r$ où le calcul du rang n'a pas été effectué.

La partie basse (partie triangulaire rouge) de la carte correspond à des interactions proches ne satisfaisant pas le critère d'admissibilité pour le noyau $1/|x-y|$. Pour ces interactions, le calcul du rang n'a pas été effectué lors du test. Pour l'autre partie de la carte, on observe que le rang varie selon la relation entre le diamètre et la distance : à un diamètre donné, la valeur du rang décroît avec la distance. On peut comparer l'allure de cette figure avec la figure 3.2. Dans la région de Fraunhofer, le rang du noyau est principalement dû à la partie $1/|x-y|$. Ainsi, on s'attend à ce que le rang des termes quadratiques soient faibles (en bleu foncé sur la figure 3.4). Entre la partie proche et la zone de Fraunhofer, soit la zone d'admissibilité de Fresnel, le rang décroit plus faiblement (la zone dégradée rouge/bleu ciel sur la figure 3.4). C'est cette décroissance du rang qui motive l'étude de l'opérateur $\mathscr{F}_\alpha$ en fonction de la précision ϵ et de la fréquence.

3.3.3 Calcul de $\mathscr{F}_\alpha^* \mathscr{F}_\alpha$

On construit la décomposition SVD de $\mathscr{F}_\alpha$ i.e on cherche les valeurs propres de $\mathscr{F}_\alpha^* \mathscr{F}_\alpha$. Avant cela, on introduit les deux opérateurs suivants :

Définition 3.7 (Opérateurs)**.** Pour des réels $d > 0$ et $\alpha > 0$, on considère les opérateurs

suivants.

$$\begin{aligned} \mathrm{T}_\alpha : \mathrm{L}^2([-d,d]) &\mapsto \mathrm{L}^2([-d,d]) \\ \lambda &\mapsto [\mathrm{T}_\alpha \lambda](x) = e^{i\alpha x^2}\lambda(x), \end{aligned} \tag{3.29}$$

On note que $\mathrm{T}_\alpha^* \circ \mathrm{T}_\alpha = \mathrm{T}_\alpha \circ \mathrm{T}_\alpha^* = \mathrm{I}d$.

$$\begin{aligned} \mathrm{F}_{2\alpha} : \mathrm{L}^2([-d,d]) &\mapsto \mathrm{L}^2([-d,d]) \\ \lambda &\mapsto [\mathrm{F}_{2\alpha}\lambda](x) = \int_{-d}^{d} e^{i2\alpha xy}\lambda(y)\,dy, \end{aligned} \tag{3.30}$$

On a alors la relation suivante entre $\mathscr{F}_\alpha$,$\mathrm{F}_{2\alpha}$ et T_α,

$$\mathscr{F}_\alpha = \mathrm{T}_\alpha \circ \mathrm{F}_{2\alpha} \circ \mathrm{T}_\alpha \tag{3.31}$$

De par (**??**), l'étude de $\mathscr{F}_\alpha^* \mathscr{F}_\alpha$ est ramenée à l'étude de $\mathrm{F}_{2\alpha}^* \mathrm{F}_{2\alpha}$ qui est un opérateur compact de $\mathrm{L}^2([-d,d])$ dans $\mathrm{L}^2([-d,d])$. L'expression de $\mathrm{F}_{2\alpha}^* \mathrm{F}_{2\alpha}$ peut être réécrite de la sorte,

$$[\mathrm{F}_{2\alpha}^* \mathrm{F}_{2\alpha}\lambda](x) = \int_{-d}^{d}\int_{-d}^{d} e^{2i\alpha(x-z)y}\lambda(z)\,dy\,dz$$

La première intégrale en la variable y se calcule de manière exacte par

$$\int_{-d}^{d} e^{2i\alpha(x-z)y}\,dy = 2d\,\mathrm{sin}_c(2d\alpha(x-z)),$$

où sin_c désigne le sinus cardinal. Par subsitution il vient

$$[\mathrm{F}_{2\alpha}^* \mathrm{F}_{2\alpha}\lambda](x) = 2d\int_{-d}^{d} \mathrm{sin}_c(2d\alpha(x-z))\lambda(z)\,dz.$$

L'étude des valeurs singulières de $\mathrm{F}_{2\alpha}$ correspond à l'étude des valeurs propres de $\mathrm{F}_{2\alpha}^* \mathrm{F}_{2\alpha}$ et ainsi on s'intéresse aux couples (σ,λ) tels que

$$2d\int_{-d}^{d} \mathrm{sin}_c(2d\alpha(x-z))\lambda(z)\,dz = \sigma\lambda(x).$$

L'opérateur $[\mathrm{F}_{2\alpha}^* \mathrm{F}_{2\alpha}]$ est l'opérateur sinus cardinal et agit comme un filtre passe-bande. C'est ce comportement qui est illustré par le plateau dans la décroissance des valeurs singulières de l'opérateur de Fresnel.

Pour l'approximation sur l'intervalle $[-d,d]$, on utilise la fréquence f_d définie par :

$$f_d = \frac{\alpha d}{\pi} = \frac{kd}{2\pi \mathrm{R}}.$$

On discrétise alors l'intervalle avec deux points par longueur d'onde en respectant le critère de Shanon. Le pas de discrétisation h_d étant alors donné par

$$h_d = \frac{1}{2f_d}.$$

On constate numériquement sur l'exemple 3.5 que le nombre de points obtenus à l'aide du pas h_d permet de décrire le palier correspondant aux N valeurs propres proches de 1. Avec le pas h_d pour l'intervalle [-d,d], ceci correspond à

$$\mathrm{N} \simeq \frac{2d}{h_d} = \frac{2}{\pi}.\frac{kd^2}{\mathrm{R}}. \tag{3.32}$$

Le rang numérique r_ϵ de l'opérateur de Fresnel $\mathscr{F}_\alpha$ à la précision ϵ correspond au nombre de valeurs propres de $\mathrm{F}_{2\alpha}^* \mathrm{F}_{2\alpha}$ supérieures à ϵ^2. Pour $\epsilon < 1$, on a déjà

$$r_\epsilon \geqslant \mathrm{N} = \frac{2}{\pi}.\frac{kd^2}{\mathrm{R}}. \tag{3.33}$$

Dans la suite, on poursuit l'étude des valeurs singulières de l'opérateurs F_p en déterminant sa décomposition en valeurs singulières de manière explicite. Le premier résultat sur le rang r_ϵ comporte déjà une dépendance en fréquence mais la précision ϵ n'intervient pas. La dépendance en précision de l'estimation (3.33) peut être obtenue à l'aide d'un résultat analytique sur l'opérateur F_p qui est lié aux fonctions d'onde sphéroïdales d'ordre zéro.

3.4 Fonctions d'onde spheroidales

3.4.1 Introduction

Le paragraphe 3.3 a introduit l'opérateur de Fox-Li dans le cadre du développement limité de la phase du noyau de Green $\mathrm{G}(x,y) = e^{ik|x-y|}/|x-y|$. ce développement limité suggère que sous réserve de satisfaire le critère d'admissibilité de Fresnel, le rang du noyau $\mathrm{G}(x,y)$ est majoré par le rang du terme du second ordre qui s'écrit comme un produit tensoriel de deux opérateurs de Fox-Li. Dès lors, on se ramène à l'étude du rang de l'opérateur de Fox-Li. La relation (3.31) permet de lier le rang de l'opérateur de Fox-Li à celui d'un autre opérateur dont on cherche la décomposition en valeurs singulières.

Dans cette partie, on étudie le rang de l'opérateur F_p à l'aide de résultats sur les fonctions d'onde sphéroïdales d'ordre zéro. On introduit ces fonctions spécialesà travers la méthode de séparations des variables pour l'équation de Helmholtz dans le système de coordonnées sphéroidales. La séparation des variables conduit à une équation différentielle du second ordre qui commute avec l'opérateur F_p que l'on décrit plus précisément ici. Enfin, on introduit l'opérateur Q_c formé à partir de $\mathrm{F}_c^\star \mathrm{F}_c$. Cet opérateur commute également avec l'opérateur L_c ce qui permet de décrire ses fonctions propres. Finalement, on peut décrire la décomposition en valeurs singulières de l'opérateur F_c et estimer son rang numérique en fonction de c et de la précision grâce à un résultat de Widom.
La littérature sur les fonctions d'onde sphéroidale est abondante et les applications de ces dernières sont nombreuses ([Sle83; LW80], applications laser). L'approximation des valeurs propres de l'opérateur de Fox-Li est également effectuée par Trefethen avec l'emploi des Chebfun. Sans perdre de généralité, on présente les résultats dans l'intervalle de référence $[-1,1]$.

3.4.2 Coordonnées sphéroïdales

On reprend la définition adoptée dans [Fla57] ; on considère l'axe z comme étant l'axe de révolution et l'on note d la distance interfocale. Les coordonnées sphéroïdales sont

reliées aux coordonnées cartésiennes par les relations suivantes,

$$\begin{aligned} x &= \frac{d}{2}[(1-\eta^2)(\xi^2-1)]^{\frac{1}{2}}\cos\phi \\ y &= \frac{d}{2}[(1-\eta^2)(\xi^2-1)]^{\frac{1}{2}}\sin\phi \\ z &= \frac{d}{2}\eta\xi \end{aligned} \tag{3.34}$$

avec $-1 \leqslant \eta \leqslant 1$, $1 \leqslant \xi \leqslant \infty$ et $0 \leqslant \phi \leqslant 2\pi$.

La surface ξ = constante > 1 est un ellipsoïde de révolution allongé de grand axe $d\xi$ et de petit axe $d\sqrt{\xi^2-1}$. La surface dégénérée $\xi = 1$ est le segment joignant les points $z = -\frac{d}{2}$ et $z = +\frac{d}{2}$. La surface $|\eta|$ = constante < 1 est un hyperboloïde de révolution dont les asymptotes forment un angle $\cos^{-1}\eta$ avec l'axe z. La surface dégénérée $|\eta| = 1$ est l'ensemble des points z de l'axe z tels que $|z| > \frac{d}{2}$. La surface ϕ = constante est un plan contenant l'axe z et formant un angle ϕ avec le plan xOz.

3.4.3 Séparation des variables pour l'équation de Helmholtz

Dans ce système de coordonnées, l'équation de Helmholtz $-(\Delta + k^2)\psi = 0$ admet une solution à variables séparées. L'expression de l'équation de Helmholtz dans les coordonnées sphéroïdales recquiert les coefficients métriques h_η, h_ξ et h_ϕ tels que

$$dx^2 + dy^2 + dz^2 = h_\eta^2 d\eta^2 + h_\xi^2 d\xi^2 + h_\phi^2 d\phi^2. \tag{3.35}$$

Ces coefficients sont les suivants

$$\begin{aligned} h_\eta &= \frac{d}{2}\left(\frac{\xi^2-\eta^2}{1-\eta^2}\right)^{\frac{1}{2}} \\ h_\xi &= \frac{d}{2}\left(\frac{\xi^2-\eta^2}{\xi^2-1}\right)^{\frac{1}{2}} \\ h_\phi &= \frac{d}{2}\left((1-\eta^2)(\xi^2-1)\right)^{\frac{1}{2}} \end{aligned} \tag{3.36}$$

L'équation de Helmholtz devient alors

$$\left[\frac{\partial}{\partial\eta}(1-\eta^2)\frac{\partial}{\partial\eta} + \frac{\partial}{\partial\xi}(\xi^2-1)\frac{\partial}{\partial\xi} + \frac{\xi^2-\eta^2}{(\xi^2-1)(1-\eta^2)}\frac{\partial^2}{\partial\phi^2} + c^2(\xi^2-\eta^2)\right]\psi = 0, \tag{3.37}$$

où $c = \frac{1}{2}kd$.

On cherche une solution de l'équation 3.37 à variables séparées ψ_{mn} de la forme

$$\psi_{mn} = \mathrm{S}_{mn}(c,\eta)\mathrm{R}_{mn}(c,\xi)\exp(im\phi). \tag{3.38}$$

Les fonctions $\mathrm{S}_{mn}(c,\eta)$ et $\mathrm{R}_{mn}(c,\xi)$ vérifient les équations différentielles suivantes

$$\frac{d}{d\eta}\left[(1-\eta^2)\frac{d}{d\eta}\mathrm{S}_{mn}(c,\eta)\right] + \left[\lambda_{mn} - c^2\eta^2 - \frac{m^2}{1-\eta^2}\right]\mathrm{S}_{mn}(c,\eta) \tag{3.39}$$

$$\frac{d}{d\xi}\left[(\xi^2-1)\frac{d}{d\xi}\mathrm{R}_{mn}(c,\xi)\right] - \left[\lambda_{mn} - c^2\xi^2 - \frac{m^2}{\xi^2-1}\right]\mathrm{R}_{mn}(c,\xi) \tag{3.40}$$

La constante λ_{mn} est la même dans les équations ci-dessus. La fonction $S_{mn}(c,\eta)$ est dite fonction d'angle tandis que $R_{mn}(c,\eta)$ est la fonction radiale. Elles sont caractérisées par la même équation différentielle linéaire du second ordre, de la forme générale

$$\frac{d}{dx}\left[(1-x^2)\frac{du}{dx}(x)\right]+\left[\lambda-c^2x^2-\frac{\mu^2}{1-x^2}\right]u(x)=0, \tag{3.41}$$

où λ et μ sont des constantes. Chaque solution pour $c^2\neq 0$ est dite fonction sphéroidale.

Remarque 3.8 (Développement du noyau de Green). On trouve dans [Fla57] un développement du noyau de Green $G(x,y)=e^{ik|x-y|}/|x-y|$ dans le système de coordonnées sphéroïdales (ξ,η,ϕ) à partir des fonctions sphéroïdales S_{mn} et R_{mn}.

Ce développement est similaire au développement multipolaire et partagent avec lui un défaut pour les applications visées. Ce développement est vérifié dans tout l'espace ce qui est à l'encontre d'une approche directionnelle.

Dans la suite, on s'intéresse particulièrement aux propriétés des fonctions d'onde sphéroïdales d'ordre zéro.

3.4.4 Fonction d'onde sphéroïdale d'ordre zéro

3.4.4.1 Définition et premières propriétés

On considère l'opérateur différentiel L_c défini dans $[-1,1]$ de la façon suivante.

Définition 3.9 (Opérateur différentiel L_c)**.** Soit $c>0$ et ϕ une fonction de $]-1,1[$ deux fois dérivable, on définit L_c par

$$L_c[\phi](x)=-\frac{d}{dx}\left((1-x^2)\frac{d\phi}{dx}(x)\right)+c^2x^2\phi(x). \tag{3.42}$$

pour $-1<x<1$.

Le problème aux valeurs propres lié à l'opérateur L_c est donné par

Définition 3.10 (EDO d'ordre deux)**.** Soit un réel χ_n, on considère le problème aux valeurs propres dans $[-1,1]$ suivant

$$L_c\psi_n=\chi_n\psi_n. \tag{3.43}$$

Ce problème correspond à l'équation différentielle suivante,

$$(1-x^2)\psi_n''(x)-2x\psi_n'(x)+(\chi_n-c^2x^2)\psi_n(x)=0 \tag{3.44}$$

On remarque que l'équation différentielle (3.105) est de la même forme que l'équation avec $\mu=0$ intervenant lors de la séparation des variables en coordonnées sphéroidales. Les solutions de l'équation différentielle (3.105) sont appelées fonctions d'onde sphéroïdales d'ordre zéro. On notera $\psi_n^c(x)$ la nième fonction d'onde sphéroïdales d'ordre zéro afin d'expliciter la dépendance avec le paramètre c.

Le théorème suivant donne une caractérisation des zéros de ψ_n^c et plus particulièrement ceux compris dans l'intervalle $[-1,1]$.

Théorème 3.11 (Zéros de $\psi_n(x)$)**.** *Soit $c>0$ la largeur de bande et un entier $n\geqslant 0$. Les zéros de la fonction $\psi_n^c:\mathbb{C}\mapsto\mathbb{C}$ sont réels et simples. De plus, la fonction ψ_n^c s'annule une infinité de fois mais il y a exactement n zéros dans l'intervalle $[-1,1]$. Ces zéros sont symétriques par rapport à l'origine.*

Le théorème précédent stipule qu'en tant que fonction propre de l'opérateur de Sturm-Liouville L_c (*cf* (3.105)), ψ_n^c possède exactement n zéros dans l'intervalle $[-1,1]$.

3.4.4.2 Fonction à bande limitée

On définit l'opérateur intégral F_c suivant intervenant dans la définition d'une fonction à largeur de bande limitée. En effet, on définit une fonction à bande limitée de la façon suivante

Définition 3.12 (Fonction à bande limitée)**.** On appelle fonction à bande limitée de largeur de bande $c > 0$ toute fonction $f : \mathbb{R} \mapsto \mathbb{R}$ telle qu'il existe une fonction σ de $L^2([-1,1])$ vérifiant

$$f(x) = \int_{-1}^{1} \sigma(t) e^{icxt} dt. \tag{3.45}$$

En d'autres termes, la transformée de Fourier d'une fonction à bande limitée està support compact. L'opérateur intégral intervenant dans la definitiond'une fonction à bande limitée est donné par

Définition 3.13 (Opérateur F_c)**.** Soit un réel $c > 0$, on définit l'opérateur intégral F_c de la façon suivante,

$$F_c : L^2([-1,1]) \mapsto L^2([-1,1]) \tag{3.46}$$

$$\phi(x) \mapsto F_c[\phi](x) = \int_{-1}^{1} \phi(t) e^{icxt} dt. \tag{3.47}$$

F_c est un opérateur compact et en tant que tel, il possède des valeurs propres $\lambda_0, \lambda_1, \ldots$, supposées ordonnées de la façon suivante,

$$|\lambda_n| \geqslant |\lambda_{n+1}|, \forall n. \tag{3.48}$$

On note $\psi_n(x)$ les fonctions propres associées aux valeurs propres λ_n. On a alors

$$\int_{-1}^{1} \psi_n(t) e^{icxt} dt = \lambda_n \psi_n(x), \tag{3.49}$$

pour tout entier $n \geqslant 0$ et $-1 \leqslant x \leqslant 1$. On peut lier les opérateurs L_c et F_c par la proposition suivante,

Théorème 3.14 (L_c et F_c commutent)**.** *Soit une largeur de bande $c > 0$. Soit $\phi : [-1,1] \mapsto \mathbb{C}$ une fonction de $\mathscr{C}^2([-1,1])$. Alors,*

$$L_c[F_c[\phi]](x) = F_c[L_c[\phi]](x), \tag{3.50}$$

pour tout $x \in [-1,1]$. On obtient par ailleurs la même propriété entre L_c et $F_c^\star$

Comme les opérateurs commutent, ils partagent les mêmes fonctions propres à une constante multiplicative pres [ORX13]. Le théorème suivant fournit ainsi une description des valeurs et des fonctions propres de F_c.

Théorème 3.15 (Valeurs/fonctions propres de F_c)**.** *Pour une largeur de bande $c > 0$ fixée, on considère l'opérateur intégral défini par 3.13. Les fonctions propres ψ_m, $m = 0, 1, \ldots$ de F_c sont réelles, orthonormales et complètes dans $L^2([-1,1])$. Les fonctions d'indice pair sont paires et celles d'indice impair sont impaires. Chaque fonction ψ_n possède exactement n zéros dans l'intervalle $[-1,1]$*

3.4.4.3 Propriétés de l'opérateur $F_c^\star F_c$

On définit l'opérateur auto-adjoint Q_c à partir de l'opérateur F_c.

Définition 3.16 (Opérateur Q_c)**.** Soit un réel $c > 0$, on définit l'opérateur intégral Q_c de la façon suivante,

$$Q_c : L^2([-1,1]) \mapsto L^2([-1,1]) \tag{3.51}$$

$$\phi(x) \mapsto Q_c[\phi](x) = \frac{1}{\pi}\int_{-1}^{1} \frac{\sin(c(x-t))}{x-t}\phi(t)\,dt \tag{3.52}$$

D'autre part, on peut écrire la caractérisation suivante de Q_c,

$$Q_c[\phi](x) = \mathbb{1}_{[-1,1]}.\mathscr{F}^{-1}\left[\mathbb{1}_{[-c,c]}.\mathscr{F}[\phi](\xi)\right](x), \tag{3.53}$$

où $\mathscr{F} : L^2(\mathbb{R}) \mapsto L^2(\mathbb{R})$ est la transformée de Fourier et $\mathbb{1}_{[-a,a]} : \mathbb{R} \mapsto \mathbb{R}$ est la fonction indicatrice de l'intervalle $[-a,a], a > 0$. Q_c est un filtre passe-bas suivi d'une troncature en temps. Par ailleurs, Q_c et F_c sont liés par la relation suivante,

$$Q_c = \frac{c}{2\pi}.F_c^\star.F_c. \tag{3.54}$$

D'après la relation (3.54), les valeurs propres μ_n de Q_c et λ_n de F_c vérifient pour tout $n \geqslant 0$,

$$\mu_n = \frac{c}{2\pi}|\lambda_n|^2. \tag{3.55}$$

La caractérisation (3.53) donne immédiatement $\mu_n < 1$ pour tout $n \geqslant 0$.

D'après le lien entre Q_c et F_c, on peut lier l'opérateur Q_c et L_c par la propriété de commutativité suivante,

Théorème 3.17 (L_c et Q_c commutent)**.** *Soit une largeur de bande $c > 0$. Soit $\phi : [-1,1] \mapsto \mathbb{C}$ une fonction de $\mathscr{C}^2([-1,1])$. Alors,*

$$L_c[Q_c[\phi]](x) = Q_c[L_c[\phi]](x), \tag{3.56}$$

pour tout $x \in [-1,1]$.

Q_c et L_c partageant les mêmes fonctions propres, F_c et Q_c possèdent donc également les mêmes fonctions propres. ceci s'exprime par

$$\mu_n\psi_n(x) = \frac{1}{\pi}\int_{-1}^{1} \frac{\sin(c(x-t))}{x-t}\psi_n(t)\,dt, \tag{3.57}$$

Le théorème suivant résume les propriétés des fonctions propres de F_c et Q_c.

Théorème 3.18 (Fonctions propres de F_c et Q_c)**.** *Soit $c > 0$ la largeur de bande. On considère les opérateurs F_c et Q_c de $L^2([-1,1])$ dans lui-même définis par 3.13 et 3.16. Alors les fonctions d'onde sphéroïdales $\psi_0^c, \psi_1^c, \ldots$ forment une base orthonormée de $L^2([-1,1])$. Pour un indice n pair, la fonction ψ_n^c est réelle et paire ;pour n impair, la fonction ψ_n^c est réelle et impaire. De plus, pour tout entier $n \geqslant 0$, ψ_n^c est la n^e fonction propre des opérateurs intégraux F_c et Q_c associée aux valeurs propres λ_n et μ_n respectivement.*

On rappelle que les valeurs propres μ_n de Q_c sont inférieures à 1 pour tout n. Pour une tolérance $0 < \alpha < 1$, le résultat suivant dû à Widom et Landau et initialement présenté dans [LW80] permet de compter les valeurs propres de l'opérateur Q_c supérieures à α.

Théorème 3.19 (Comptage des valeurs propres de Q_c)**.** *Soit $c > 0$ une largeur de bande. Soit $0 < \alpha < 1$. On considère l'opérateur Q_c défini par la définition 3.16. On note $N(c,\alpha)$ le nombres de valeurs propres μ_n de Q_c supérieures à α. En d'autres termes, on a la caractérisation suivante*

$$N(c,\alpha) = \max\left\{k = 1,2,\ldots : \mu_{k-1} > \alpha\right\}. \tag{3.58}$$

Alors,

$$N(c,\alpha) = \frac{2c}{\pi} + \frac{1}{\pi^2}\log\left(\frac{1-\alpha}{\alpha}\right)\log(c) + o(\log(c)). \tag{3.59}$$

Selon l'expression (3.59), il y a environ $2c/\pi$ valeurs propres dont la valeur absolue est proche de 1 et de l'ordre de $\log(c)$ valeurs propres décroissant rapidement. Les valeurs propres restantes sont proches de 0. En se donnant une erreur relative ϵ telle que $\epsilon^2 = \alpha$, l'entier $N(c,\alpha)$ défini par (3.59) permet de déterminer le rang numérique à la précision ϵ de l'opérateur F_c d'après la caractérisation (3.55). L'obtention d'une décomposition en valeurs singulières de l'opérateur F_c s'obtient en calculant les valeurs propres μ_n. Cependant, on ne dispose pas de formule explicite donnant ces valeurs propres. Dans la pratique, on peut les calculer à partir des fonctions propres ψ_n^c d'après la caractérisation suivante ([ORX13]),

$$\int_{-\infty}^{\infty} \left(\psi_n^c(t)\right)^2 dt = \frac{1}{\mu_n(c)}. \tag{3.60}$$

Dans l'expression ci-dessus, on fait apparaître clairement la dépendance vis-à-vis de la largeur de bande pour la valeur propre μ_n. En effet, ce paramètre joue un rôle très important dans le comportement des fonctions propres ψ_n^c ainsi que dans la décroissance des valeurs propres μ_n.
Par exemple, [ORX13] fournit une estimation asymptotique de $1-\mu_n(c)$ lorsque $c \to \infty$ et $n << c$,

$$1 - \mu_n(c) = 4\sqrt{\pi}\cdot\frac{8^n}{n!}\cdot c^{n+1/2}\cdot e^{-c}\cdot(1+o(1)). \tag{3.61}$$

De nombreux résultats similaires existent dans la littérature ([Fla57],[ORX13]) et nous ne les détaillerons pas dans ce manuscrit. Dans la suite, on fournit une brève description du comportement des fonctions ψ_n^c en fonction de la largeur de bande c et de l'indice n.

3.4.5 Un point historique

L'étude de l'opérateur F_c précédent en lien avec l'opérateur différentiel L_c est exactement l'étude effectuée par Slepian, Landau et Pollak aux *Bell Labs* dans les années 1960 [SP61; LP61; LP62]. Les Laboratoires Bell (*Bell Labs* ou *Bell Telephone Laboratories* ou encore AT&T Bell Laboratories) furent fondés en 1925 et implantés à Murray Hill dans l'état américain du New Jersey. Ces laboratoires ont déposé de nombreux brevets dans des domaines tels que les télécommunications (réseau téléphonique, transmission télévisuelle, communications satellite) et l'informatique (Unix, C et C++). Slepian et ses collaborateurs ont étudié l'opérateur F_c à partir d'un problème d'ingénierie électrique.

Un appareil ne peut transmettre des sinusoïdes à une fréquence arbitraire sans perturbation. Ainsi, à cause de la limitation en fréquence imposée par l'appareil et la nature des signaux considérés. Par exemple, l'analyse de Fourier d'un enregistrement d'une voix d'un homme ne comporte pas de fréquences supérieures à 8000 Hz. L'ingénieur en communication est ainsi amené à considérer des signaux à bande limitée. Ce sont des signaux dont l'amplitude du spectre tend vers zéro pour des fréquences $|f| > W$ où W est une

fréquence maximale fixée. L'espace de ces signaux est noté B_W et un élément $r(t)$ de B_W s'écrit de la sorte

$$r(t) = \int_{-W}^{W} e^{i2\pi f t} R(f) df. \tag{3.62}$$

On remarque qu'il s'agit (à une normalisation près) de l'opérateur F_c introduit précédemment. De façon analogue, un signal est limité en temps s'il existe $T > 0$ tel que

$$r(t) = 0 \qquad |t| > \frac{T}{2}.$$

Outre la transmission d'un signal lisse comme la voix humaine, il est également nécessaire de pouvoir efficacement transmettre des signaux ponctuels comme une impulsion (en temps donc). Dans le domaine fréquentiel, ce type de signal comporte toutes les fréquences possibles et n'est donc pas limité en fréquence. La transmission de ces deux types de signaux de nature très différente est un problème récurrent dans le domaine des télécommunications. On souhaiterait idéalement obtenir des signaux qui sont à la fois à support compact en temps et en fréquence. Cependant, le seul signal à la fois limité en temps et en espace est le signal nul. Dès lors, on s'attache à des signaux permettant d'approcher au mieux ce comportement idéal. Pour ce faire, Slepian et ses collaborateurs se sont intéressés à une mesure de l'énergie du signal transmis : c'est le problème de concentration que Slepian décrit dans [Sle83]. Une mesure de la qualité de la transmission est donnée par les rapports suivants,

$$\alpha^2(T) = \frac{\int_{-T/2}^{T/2} r^2(t) dt}{\int_{-\infty}^{\infty} r^2(t) dt}, \tag{3.63}$$

$$\beta^2(W) = \frac{\int_{-W/2}^{W/2} |R(f)|^2 df}{\int_{-\infty}^{\infty} |R(f)|^2 df}. \tag{3.64}$$

Si le signal $r(t)$ était à support compact dans $[-T/2, T/2]$, le rapport α^2 serait maximal. Le problème étudié par Slepian, Landau et Pollak est alors de déterminer la valeur maximale de α^2 pour un signal à bande limité $r(t) \in B_W$. Après réécriture [SP61; Sle83], ils obtiennent le problème aux valeurs propres suivant,

$$\int_{-1}^{1} \frac{\sin c(x-y)}{\pi(x-y)} \psi(y) dy = \lambda \psi(x) \qquad |x| \leqslant 1, \tag{3.65}$$

où $c = \pi WT$ est un paramètre fixé. On note que le sinus cardinal est largement employé en traitement du signal afin de construire des filtres passe-bande. Ce problème (3.65) est exactement celui que l'on obtient lors de notre analyse des termes d'ordre deux dans le développement limité de la phase du noyau oscillant. Incidemment, nous avons abordé ce problème de la même façon que Slepian. Les expériences numériques précédentes montrent que le critère de Shannon fournit le nombre de valeurs propres proches de l'unité : c'est le « théorème 2WT ». Slepian décrit ce résultat heuristique utilisé en ingénierie électrique depuis longtemps par : *Si* WT *est grand, l'espace des signaux de durée approximative* T *de largeur de bande* W *a une dimension proche de* 2WT *[Sle83]*. En réalité, comme nous l'avons constaté, ce résultat fournit seulement les valeurs propres les plus grandes. À ce stade, l'analyse du problème (3.65) peut être approfondie à l'aide des résultats sur les fonctions d'onde sphéroïdales. Slepian lui-même décrit cette étude comme un travail qu'ils ont résolu complètement à sa plus grande surprise [Sle83]. La description du problème aux valeurs propres (3.65) est qualifié de fortuit (« There was a lot of serendipity

here,clearly. ») car cette résolution est rendue possible par l'opérateur différentiel L_c commutant avec l'opérateur intégral F_c. Les fonctions propres de l'opérateur différentiel L_c ne sont autres que les fonctions d'onde sphéroïdales. Les fonctions propres du problème (**??**) s'expriment ainsi à partir de ces fonctions spéciales. Par ailleurs, Slepian a généralisé ce travail dans le cas multi-dimensionnel [Sle64] et en particulier dans le cas d'un disque. Là également, l'étude est rendue possible par l'existence d'un opérateur différentiel commutant avec l'opérateur intégral.

Si le « théorème 2WT » (plutôt une heuristique) est en réalité très imprécis pour les faibles valeurs du paramètre c, il devient correct asymptotiquement. Les résultats sur les fonctions d'onde sphéroïdales permet à Slepian [Sle65] de caractériser le comportement des valeurs propres en fonction de la largeur de bande c :

$$\lim_{c\to\infty} \lambda_n = \begin{cases} 0 & n = \lfloor (1+\eta)\frac{2c}{\pi} \rfloor \\ (1+e^{\pi b})^{-1} & n = \lfloor \frac{2c}{\pi} + \frac{b}{\pi}\log(c) \rfloor \\ 1 & n = \lfloor (1-\eta)\frac{2c}{\pi} \rfloor, \end{cases} \tag{3.66}$$

où $\eta > 0$ est un petit paramètre fixé et $\lfloor x \rfloor$ désigne ici le plus grand entier inférieur à x. Pour $n \ll 2c/\pi$, la plupart des λ_n sont proches de l'unité tandis que pour $n \gg 2c/\pi$ elles sont proches de zéro. Quand $n \approx 2c/\pi$, on a $\lambda_n \approx 1/2$. L'intervalle en n entre les deux comportements extremaux croît comme $\log(c)$. Cette croissance logarithmique est déjà mentionnée par H.J. Landau et H. O. Pollak [LP62]. Ce comportement asymptotique des valeurs propres de l'opérateur sinus cardinal trouve un équivalent discret. Slepian a fourni des résultats similaires dans [Sle78]. Une fois discrétisé, les matrices considérées sont des matrices de Töplitz (les éléments sont constants sur une diagonale) et J. M. Varah fournit également des résultats dans ce cas discret [Var93]. Il note également qu'une telle matrice peut servir de cas test pour des algorithmes d'algèbre linéaire puisque le comportement des valeurs propres est connu. La suite logicielle MATLAB répertorie cette matrice dans sa gallerie de tests. De plus, L. Trefethen utilise également une discrétisation de l'opérateur F_c pour tester la convergence de l'approximation de fonction de deux variables en tant que série de Chebyshev (les Chebfun).

Le comportement asymptotique des valeurs propres dans le cas continu a été démontré par Landau et H. Widom en 1980 à partir de résultats sur les formes hermitiennes et les opérateurs de Töplitz. Avant de collaborer pour obtenir ce résultat, Landau et Widom ont débuté leur carrière par une certaine rivalité. Harold Widom a grandi à Brooklyn, New York et fréquenta le lycée Stuyvesant où il était le capitaine de l'équipe de mathématiques. Le capitaine de l'équipe rivale du lycée du Bronx n'était autre que ...Henry Landau ! Cette rivalité ne se limita néanmoins qu'aux compétitions de mathématiques [BG92] et leur amitié et collaboration donna lieu a plusieurs articles dont la démonstration rigoureuse du comportement des valeurs propres λ_n que l'on utilise ici.

3.4.6 Comportement asymptotique, série de Legendre

3.4.6.1 Polynômes de Legendre

On rappelle succintement les propriétés principales des polynômes de Legendre. Les polynômes de Legendre, notés $P_0, P_1, \ldots$ sont déterminés par la relation de récurrence suivante,

$$\begin{aligned} \mathrm{P}_0 &= 1, \\ \mathrm{P}_1 &= \mathrm{X}, \\ (k+1)\mathrm{P}_{k+1} &= (2k+1)\mathrm{X}.\mathrm{P}_k - k\mathrm{P}_{k-1} \qquad k \geqslant 1. \end{aligned} \tag{3.67}$$

D'après la définition des polynômes, il vient immédiatement que pour tout entier $n \geqslant 0$, $\mathrm{P}_n(1) = 1$ et P_n est de la même parité que n. On définit également les polynômes de Legendre comme la solution de l'équation différentielle (dite équation de Legendre) suivante

$$-\frac{d}{dx}\left[(1-x^2)\frac{d}{dx}\mathrm{P}_n(x)\right] + [n(n+1)]\mathrm{P}_n(x) = 0, x \in [-1,1]. \tag{3.68}$$

Les polynômes de Legendre constituent une base orthogonale et sont complets dans $\mathrm{L}^2([-1,1])$ muni du produit scalaire usuel mais ne sont pas orthonormés. En effet, pour tout entier $n \geqslant 0$,

$$\|\mathrm{P}_n\|^2_{\mathrm{L}^2([-1,1])} = \int_{-1}^{1} \mathrm{P}_n^2(x)dx = \frac{2}{2n+1}. \tag{3.69}$$

On construit une base orthonormale de $\mathrm{L}^2([-1,1])$ à l'aide des polynômes normalisés $\overline{\mathrm{P}_n}$ donnés par

$$\overline{\mathrm{P}_n} = \mathrm{P}_n.\sqrt{n+1/2}. \tag{3.70}$$

3.4.6.2 Approximation des fonctions d'onde sphéroïdales

La relation entre les fonctions sphéroïdales d'ordre zéro et les polynômes de Legendre permet d'obtenir des approximations des fonctions d'onde sphéroïdales d'ordre zéro et est utilisée dans la pratique depuis les années 1950. On rappelle que ces fonctions vérifient l'équation suivante

$$-\frac{d}{dx}\left[(1-x^2)\frac{d}{dx}\Phi\right] + (\chi_n - c^2x^2)\psi(x) = 0, x \in [-1,1]. \tag{3.71}$$

Par ailleurs, les polynômes de Legendre vérifient l'équation différentielle suivante pour tout $n \geqslant 0$.

$$-\frac{d}{dx}\left[(1-x^2)\frac{d}{dx}\mathrm{P}_n(x)\right] + [n(n+1)]\mathrm{P}_n(x) = 0, x \in [-1,1]. \tag{3.72}$$

Lorsque c tend vers zéro, l'équation (3.71) se résume à l'équation de Legendre (3.72). Ainsi, la solution de (3.71) doit tendre vers la fonction de Legendre solution de (3.72) lorsque c tend vers zéro. Ainsi, on a

$$\lambda_n(0) = n(n+1), n \geqslant 0. \tag{3.73}$$

Lorsque c est différent de zéro, l'équation (3.71) ne diffère de l'équation (3.72) que par l'existence d'une singularité essentielle à l'infini. Ceci suggère alors un développement de la forme suivante

$$\psi_n(x) = \sum_{k=0}^{\infty} \beta_k^{(n)} \overline{\mathrm{P}_n}(x) = \sum_{k=0}^{\infty} \alpha_k \mathrm{P}_k(x), \tag{3.74}$$

pour tout $-1 \leqslant x \leqslant 1$ avec $\beta_0^{(n)}, \beta_1^{(n)}, \ldots$, définis par

$$\beta_k^{(n)} = \int_{-1}^{1} \psi_n(x)\overline{\mathrm{P}_k}(x)dx, \tag{3.75}$$

et $\alpha_0^{(n)}, \alpha_1^{(n)}, \ldots$, définis par

$$\alpha_k^{(n)} = \beta_k^{(n)}.\sqrt{k+1/2} = (k+1/2)\int_{-1}^{1} \psi_n(x)\overline{P_k}(x)dx, \tag{3.76}$$

pour $k \geqslant 0$. Quitte à substituer $\psi_n(x)$ par son développement sur les polynômes de Legendre, on obtient les relations suivantes :

$$\begin{aligned}
A_{0,0}.\beta_0^{(n)} + A_{0,2}.\beta_2^{(n)} &= \chi_n.\beta_0^{(n)}, &(3.77)\\
A_{1,1}.\beta_1^{(n)} + A_{1,3}.\beta_3^{(n)} &= \chi_n.\beta_1^{(n)}, &(3.78)\\
A_{k,k-2}.\beta_{k-2}^{(n)} + A_{k,k}.\beta_k^{(n)} + A_{k,k+2}.\beta_{k+2}^{(n)} &= \chi_n.\beta_k^{(n)}, &(3.79)
\end{aligned}$$

pour $k = 2,3,\ldots$, avec $A_{k,k}$,$A_{k+2,k}$ et $A_{k,k+2}$ définis pour $k \geqslant 0$ par

$$\begin{aligned}
A_{k,k} &= k(k+1) + \frac{2k(k+1)-1}{(2k+3)(2k-1)}.c^2,\\
A_{k,k+2} = A_{k+2,k} &= \frac{(k+2)(k+1)}{(2k+3)\sqrt{(2k+1)(2k+5)}}.c^2. \qquad (3.80)
\end{aligned}$$

Écrit autrement, le vecteur infini $(\beta_0^{(n)}, \beta_1^{(n)}, \ldots)^{\mathrm{T}}$ vérifie le problème de valeur propre suivant

$$(A - \chi_n I).(\beta_0^{(n)}, \beta_1^{(n)}, \ldots)^{\mathrm{T}} = 0, \tag{3.81}$$

où I est la matrice infinie de l'identité tandis que les coefficients non nuls de la matrice A sont donnés par la relation 3.80.

On remarque que pour un entier k donné, seuls $k-2$, k et $k+2$ interviennent dans la définition du problème . Ceci conduit naturellement à séparer la matrice A en deux parties, l'une constituée des éléments pairs, l'autre des éléments impairs. De plus, par parité, pour tout couple d'entier (n,k),

$$\beta_k^{(n)} = 0, \text{ si } k+n \text{ est impair.} \tag{3.82}$$

Le théorème suivant présenté dans [ORX13] établit la relation entre la matrice A et la décomposition sur les polynômes de Legendre.

Théorème 3.20. *Soit une largeur de bande $c > 0$ et les matrices tridiagonales infinies $A^{(\mathrm{pair})}$ et $A^{(\mathrm{impair})}$ de la forme*

$$A^{(\mathrm{pair},n)} = \begin{bmatrix}
A_{0,0} & A_{0,2} & & & & \\
A_{2,0} & A_{2,2} & A_{2,4} & & & \\
 & A_{4,2} & A_{4,4} & A_{4,6} & & \\
 & & \ddots & \ddots & \ddots & \\
 & & & \ddots & \ddots &
\end{bmatrix}, \tag{3.83}$$

et

$$A^{(\mathrm{impair},n)} = \begin{bmatrix}
A_{1,1} & A_{1,3} & & & & \\
A_{3,1} & A_{3,3} & A_{3,5} & & & \\
 & A_{5,3} & A_{5,5} & A_{5,7} & & \\
 & & \ddots & \ddots & \ddots & \\
 & & & \ddots & \ddots &
\end{bmatrix}. \tag{3.84}$$

où les coefficients $A_{i,j}$ *sont donnés par 3.80. On note* $\beta^{(n)}_{\text{pair}} \in l^2$ *et* $\beta^{(n)}_{\text{impair}} \in l^2$ *les vecteurs définis par*

$$\beta^{(n)}_{\text{pair}} = (\beta^{(n)}_0, \beta^{(n)}_2, \ldots)^{\mathrm{T}}, \tag{3.85}$$

$$\beta^{(n)}_{\text{impair}} = (\beta^{(n)}_1, \beta^{(n)}_3, \ldots)^{\mathrm{T}}, \tag{3.86}$$

avec $\beta^{(n)}_0, \beta^{(n)}_1, \ldots$ *définis par 3.75. Si n est pair, on a*

$$A^{(\text{pair})}.\beta^{(n)}_{\text{pair}} = \chi_n \beta^{(n)}_{\text{pair}}. \tag{3.87}$$

Si n est impair, alors

$$A^{(\text{impair})}.\beta^{(n)}_{\text{impair}} = \chi_n \beta^{(n)}_{\text{impair}}. \tag{3.88}$$

Les coefficients des matrices mises en jeu ne possèdent pas de propriété de décroissance particulière lorsque l'indice de ligne et/ou de colonne augmente. Cependant, dans la on se ramène à un cas fini en considérant la partie supérieure de A. Cette opération est rendue possible car les coefficients ont une décroissance rapide.

Décroissance des coefficients Les fonctions ψ_n sont analytiques dans $\mathbb{C}$ et il s'ensuit la décroissance rapide des coefficients β dans la décomposition 3.74.

Théorème 3.21 (Décroissance des coefficients)**.** *Soit* $c > 0$. *k et m sont des entiers positifs et on note* $\overline{P_k}(x)$ *le* k^e *polynôme de Legendre normalisé. On note également* $\psi_m(x)$ *la* m^e *fonction d'onde sphéroïdale associée à la largeur de bande c. Enfin,* λ_m *est la valeur propre associée à cette fonction propre et on suppose que k vérifie l'inégalité suivante*

$$k \geqslant 2(\lfloor e.c \rfloor + 1),$$

où $\lfloor . \rfloor$ *désigne la partie entière. Alors, on a*

$$\left| \int_{-1}^{1} \overline{P_k}(x) \psi_m(x) dx \right| < \frac{1}{\lambda_m} \left(\frac{1}{2} \right)^{k-1}. \tag{3.89}$$

En particulier, si pour une tolérance $\epsilon > 0$ *k vérifie*

$$k \geqslant 2(\lfloor e.c \rfloor + 1) + \log_2\left(\frac{1}{\epsilon}\right) + \log_2\left(\frac{1}{\lambda_m}\right),$$

alors,

$$\left| \int_{-1}^{1} \overline{P_k}(x) \psi_m(x) dx \right| < \epsilon. \tag{3.90}$$

Régime c petit Dans le cas où c est petit, [Sle65] décrit le comportement asymptotique suivant

$$\psi_n(x) = \overline{P_n}(x) + \mathcal{O}(c^2), c \to 0, \tag{3.91}$$

où $\overline{P_n}$ est le n^{e} polynôme de Legendre normalisé. Plus précisément, on peut trouver dans [ORX13] au théorème 8.3 un développement avec plus de termes. Néanmoins, le développement effectué est en c^2/m

3.4.6.3 Approximation par la méthode des puissances itérées

La structure de la matrice A permet d'obtenir des résultats supplémentaires à l'aide de la méthode des puissances itérées. Pour une matrice A quelconque, si l'on dispose d'une estimation λ_0 d'une de ses valeurs propres ainsi que d'un vecteur propre approximatif x_0. On pose

$$B = A - \lambda.I \tag{3.92}$$

Alors la suite $B^{-n}x_0$ tend vers le vecteur Le vecteur $x_0 = (0,\dots,0,1,0,\dots,0)^T$ est une bonne estimation du vecteur propre.

On pose

$$a_k = \frac{2k(k+1)-1}{(2k+3)(2k-1)}.c^2, \tag{3.93}$$

$$b_k = \frac{(k+2)(k+1)}{(2k+3)\sqrt{(2k+1)(2k+5)}}.c^2. \tag{3.94}$$

Ainsi, on a

$$A_{k,k} = k(k+1) + a_k, \tag{3.95}$$

$$A_{k-2,k} = A_{k,k+2} = b_k. \tag{3.96}$$

Proposition 3.22. *Soit c une margeur de bande fixée et $0 < \mu < \nu$. La matrice $A^{\mu,\nu}$ est à diagonale dominante. De plus, si μ est suffisament grand, on a*

$$A^{\mu,\nu}_{k+1,k+1} - A^{\mu,\nu}_{k,k} > |A^{\mu,\nu}_{k+1,k}| + |A^{\mu,\nu}_{k,k+1}|, \tag{3.97}$$

pour tout $\mu < k < \nu - 1$. Ainsi, pour μ suffisament grand,

$$\lambda_k(A^{\mu,\nu}) \approx 2k.(2k+1), \tag{3.98}$$

où $\lambda_k(A^{\mu,\nu})$ représente la k^e valeur propre de la matrice $A^{\mu,\nu}$.

En considérant la matrice $A^{n-k,n-k}$ pour n grand, on peut en déduire qu'il existe une valeur propre proche de $\chi^0 = n(n+1)$ dont le vecteur propre associé est proche de $(0,...,0,1,0,...,0)^T$. En effet, pour n grand, a_n et b_n sont constants. On peut alors obtenir une approximation de meilleure qualité grâce à la méthode des puissances itérées. L'application de cette méthode fournit l'approximation suivante

Théorème 3.23 (Approximation des valeurs propres de L_c)**.** *Soit une largeur de bande $c > 0$ fixée. Alors, pour tout entier $m > 0$,*

$$\chi_m(c) = m(m+1) + \frac{c^2}{2} + \frac{c^2(4+c^2)}{32m^2} - \frac{c^2(4+c^2)}{32m^3} \tag{3.99}$$

$$\frac{c^2(28+13c^2)}{128m^4} - \frac{c^2(20+11c^2)}{64m^5} \tag{3.100}$$

$$+\frac{c^2(3904+3936c^2+160c^4+5c^6)}{8192m^6} - \tag{3.101}$$

$$\frac{c^2(5824+8416c^2+480c^4+15c^6}{8192m^7} + c^2.\mathcal{O}\left(\frac{c^8}{m^8}\right). \tag{3.102}$$

[PTVF92] pour le calcul de la solution du système symétrique tridiagonal par la méthode de factorisation QR.

3.4.7 Lien avec les polynômes d'Hermite

Les approximations précédentes ne valent que pour une largeur de bande c petite. Ces développements sont en général utilisés pour c allant de 0 à 10 ([Fla57],[ORX13]). Ceci est attendu car ce développement se base sur le comportement de l'opérateur L_c lorsque $c \to 0$. Pour de plus grandes valeurs, il est nécessaire d'observer le comportement de l'équation quand $c \to \infty$. Plus précisément, c'est la relation entre l'indice m de la fonction et la largeur de bande c qui va déterminer le comportement de la fonction $\psi_m^c(x)$. Dans le cas de l'approximation de Legendre, soit $c \to 0$, alors l'indice m vérifie $m >> c$. Par exemple, la méthode des puissances itérées appliquée à la matrice de la méthode de Bouwkamp fournit des approximations asymptotiques en c/m. Dans le cas où $c \to \infty$, on s'intéresse aux indices tels que $c >> m$ et les approximations sont en m/c. Nous ne traiterons pas ici du cas où $m, c \to \infty$ et le rapport m/c n'est pas grand. [ORX13] fournit des approxiations dans ce cas à l'aide de l'approximation BKW. On commence par donner le comoprtement asymptotique de l'opérateur différentiel L_c quand $c \to \infty$. On rappelle succintement les principaux résultats sur les polynômes d'Hermite puis l'on utilise ces derniers pour approcher les fonctions d'onde sphéroïdale d'ordre zéro.

3.4.7.1 Comportement asymptotique de L_c pour c grand

On rappelle que l'opérateur L_c est défini par

$$L_c(\phi) = -\frac{d}{dx}\left[(1-x^2)\frac{d\phi(x)}{dx}\right] + c^2x^2\phi(x), \tag{3.103}$$

et que la fonction d'onde sphéroïdale d'ordre zéro $\psi_n^c(x)$ est solution du problème de valeur propre dans $x \in [-1,1]$,

$$L_c[\psi_n^c](x) = \chi_n\psi_n^c(x), \tag{3.104}$$

soit

$$(1-x^2)\frac{d^2\psi_n^c}{dx^2}(x) - 2x.\frac{d\psi_n^c}{dx}(x) + (\chi_n - c^2x^2)\psi_n^c(x) = 0. \tag{3.105}$$

[Fla57] fournit une analyse asymptotique de l'équation **??** dans le cas général. On ne s'intéresse ici qu'aux fonctions d'ordre zéro ce qui correspond à $m = 0$ dans le développement de [Fla57].

On cherche une solution $\psi_n^c(\eta)$ de $L_c(\phi) = 0$ sous la forme

$$\psi_n^c(\eta) = (1-\eta^2)^{\frac{1}{2}}u_n^c(\eta). \tag{3.106}$$

En remplaçant dans l'équation en posant $\eta = (2c)^{-1/2}x$, on obtient

$$(2c-x^2)\frac{d^2u_n}{dx^2} - 2x\frac{du_n}{dx} + [\lambda_n - \frac{1}{2}cx^2]u_n = 0 \tag{3.107}$$

Lorsque c tend vers l' ∞, comme $x \in [-1,1]$, on a $2c >> x^2$ et ainsi

$$\frac{d^2u_n}{dx^2} + \left(\frac{1}{2}\frac{\lambda_n}{c} - \frac{1}{4}x^2\right)u_n = 0 \tag{3.108}$$

L'équation (3.108) est de la même forme que l'équation vérifiée par les fonctions paraboliques cylindriques D_r soit,

$$\frac{d^2D_r}{dx^2} + \left(r + \frac{1}{2} - \frac{1}{4}x^2\right)D_r = 0 \tag{3.109}$$

Pour un entier $r \geqslant 0$, les fonctions D_r sont liées aux fonctions d'Hermite par

$$D_r(x) = (-1)^r \exp\left(\frac{1}{4}x^2\right) \frac{d^2}{dx^2} \exp\left(\frac{1}{2}x^2\right) \tag{3.110}$$

$$= 2^{-\frac{1}{2}r} \exp\left(-\frac{1}{4}x^2\right) H_r\left(\frac{x}{\sqrt{2}}\right) \tag{3.111}$$

Lorsque c devient très grand, en considérant les zéros des fonctions, [Fla57] montre que

$$u_n(x) \propto D_n(x).$$

Par ailleurs, d'après (3.108) et (3.109) il s'ensuit que $\lambda_n(0) \to (2n+1)c$, $c \to \infty$. Ceci suggère alors un développement de la forme

$$u_n(x) \sum_{r=-\infty}^{r=+\infty} h_r^n D_{n+r}(x). \tag{3.112}$$

3.4.7.2 Polynômes d'Hermite

Polynômes d'Hermite On définit les polynômes d'Hermite comme la famille de polynômes vérifiant l'équation différentielle suivante

$$H_n''(x) - 2xH_n'(x) + 2nH_n(x) = 0 \tag{3.113}$$

Cette famille est orthogonale sur $\mathbb{R}$ avec un poids e^{-x^2}. Ainsi, pour $m\,n \geqslant 0$, on a

$$\int_{-\infty}^{\infty} e^{-x^2} H_n(x) H_m(x) dx = \sqrt{\pi} 2^n n! \delta_{n,m}. \tag{3.114}$$

De plus, ces polynômes satisfont la relation de récurrence suivante

$$H_{n+1}(x) = 2xH_n(x) - 2nH_{n-1}(x), \tag{3.115}$$

avec les conditions initiales

$$H_0(X) = 1, \tag{3.116}$$

$$H_1(X) = 2X. \tag{3.117}$$

Dans la suite, on utilise une version normalisée de ces polynômes que l'on note $H_n^a(x)$ en imposant que les polynomes $H_n^a(x)$ soient orthonormaux pour le poids $e^{-a^2x^2}, a > 0$. Autrement dit,

$$\int_{-\infty}^{\infty} e^{-a^2x^2} H_n^a(x) H_m^a(x) dx = \delta_{n,m}. \tag{3.118}$$

Par suite,

$$H_n^a(x) = \frac{\sqrt{a}}{\pi^{\frac{1}{4}}.2^{\frac{n}{2}}.(n!)^{\frac{1}{2}}} H_n(ax), \tag{3.119}$$

et ainsi,

$$\frac{1}{a^2}\frac{d^2H_n^a(x)}{dx^2} - 2x\frac{dH_n^a(x)}{dx} + 2nH_n^a(x) = 0. \tag{3.120}$$

Ils satisfont également une relation de récurrence à trouis termes :

$$H_{n+1}^a(x) = ax\sqrt{\frac{2}{n+1}} H_n^a(x) - \sqrt{\frac{n}{n+1}} H_{n-1}^a(x), \tag{3.121}$$

avec les conditions initiales

$$\mathrm{H}_0^a(\mathrm{X}) = \sqrt{a}\left(\frac{1}{\sqrt{\pi}}\right)^{\frac{1}{2}}, \tag{3.122}$$

$$\mathrm{H}_1^a(\mathrm{X}) = \sqrt{2a}\left(\frac{1}{\sqrt{\pi}}\right)^{\frac{1}{2}} ax. \tag{3.123}$$

Fonctions d'Hermite Pour un réel $a > 0$, on définit les fonctions $\phi_0^a, \phi_1^a, \phi_2^a, \ldots : \mathbb{R} \mapsto \mathbb{R}$ par la formule

$$\phi_n^a(x) = e^{-a^2x^2/2}.\mathrm{H}_n^a(x). \tag{3.124}$$

Ces fonctions satisfont la relation de récurrence à trois termes suivante pour $n \geqslant 1$

$$\phi_n^a(x) = \frac{1}{a}\sqrt{\frac{n+1}{2}}\phi_{n+1}^a(x) + \frac{1}{a}\sqrt{\frac{n}{2}}\phi_{n-1}^a(x). \tag{3.125}$$

Ces fonctions forment une base complète dans $\mathrm{L}^2(\mathbb{R})$.

Théorème 3.24 ($\mathrm{L}^2(\mathbb{R})$)**.** *Soit un réel $a > 0$. Pour $m, n \geqslant 0$,*

$$\int_{-\infty}^{+\infty} \phi_n^a(x).\phi_m^a(x)\,dx = \delta_{m,n}. \tag{3.126}$$

Pour une fonction $f \in \mathscr{C}^2(\mathbb{R})$, on peut décomposer f sur les fonctions d'Hermite normalisées :

$$f(x) = \sum_{n=0}^{\infty} \alpha_n \phi_n^a(x), \tag{3.127}$$

où les coefficients α_n sont donnés par

$$\alpha_n = \int_{-\infty}^{+\infty} f(x)\phi_n^a(x)\,dx. \tag{3.128}$$

De plus, si f est paire, alors $\alpha_{2k+1} = 0$ pour tout $k \geqslant 0$. Si f est impaire, alors $\alpha_{2k} = 0$ pour tout $k \geqslant 0$.

3.4.7.3 Approximation de $\psi_m^c(x)$ par les polynômes d'Hermite

On développe la m^{e} fonction sphéroïdale d'ordre zéro ψ_m^c suivant les polynômes d'Hermite

$$\psi_m^c(x) = \sum_{k=0}^{\infty} \alpha_k^m \phi_k^a(x), \tag{3.129}$$

où les coefficients α_k^m dépendent de $c > 0$ et $a > 0$. On note

$$\alpha_{-1} = \alpha_{-2} = \alpha_{-3} = \alpha_{-4} = 0, \tag{3.130}$$

et en substituant dans l'équation (3.105), les coefficients vérifient une relation de récurrence à cinq termes :

Théorème 3.25 (Relation de récurrence)**.** *Soit $a > 0$. Pour tout $n \geqslant 0$, les coefficients α_n^m satisfont la relation de récurrence à cinq termes suivante*

$$\begin{aligned}
&\frac{1}{4}\sqrt{-3+n}\sqrt{-2+n}\sqrt{-n+n^2}.\alpha_{n-4}^m \\
&-\frac{1}{2a^2}(a^4-c^2).\sqrt{-n+n^2}.\alpha_{n-2}^m \\
-\left(\chi_m-\frac{1}{4a^2}(-3a^2+2a^4+2c^2-2a^2n+4a^4n+4c^2n-2a^2n^2)\right)&\alpha_n^m \\
&-\frac{1}{2a^2}(a^4-c^2)\sqrt{2+3n+n^2}.\alpha_{n+2}^m \\
&+\frac{1}{4}\sqrt{3+n}\sqrt{4+n}\sqrt{2+3n+n^2}.\alpha_{n+4}^m=0
\end{aligned}$$

De la même façon que pour le développement en série de Legendre, on peut représenter cette récurrence par une matrice infinie symétrique B dont les coefficients b_{ij} sont donnés pour $n \geqslant 0$ par

$$\begin{aligned}
b_{n,n} &= \frac{1}{4a^2}(-3a^2+2a^4+2c^2-2a^2n+4a^4n+4c^2n-2a^2n^2), \\
b_{n,n+2} &= -\frac{1}{2a^2}(a^4-c^2)\sqrt{2+3n+n^2}, \\
b_{n+2,n} &= -\frac{1}{2a^2}(a^4-c^2)\sqrt{2+3n+n^2}, \\
b_{n,n+4} &= \frac{1}{4}\sqrt{(n+1)(n+2)(n+3)(n+4)}, \\
b_{n+4,n} &= \frac{1}{4}\sqrt{(n+1)(n+2)(n+3)(n+4)},
\end{aligned} \tag{3.131}$$

les autres coefficients étant nuls.

On note μ^m le vecteur de l^2 défini par

$$\mu^m = (\alpha_0^m, \alpha_1^m, \alpha_2^m, \ldots), \tag{3.132}$$

on peut alors obtenir un résultat similaire au développement de Legendre par le théorème suivant :

Théorème 3.26. *Soit $a > 0$, et on considère la matrice définie par les relations (3.131). χ_m et ψ_m désignent respectivement la valeur propre et la fonction propre de L_c. μ_m désigne le vecteur des coefficients du développement de Hermite de la fonction ψ_m^c. Alors pour tout entier $m \geqslant 0$,χ_m et μ_m sont solutions du probblème aux valeurs propres*

$$B.\mu_m = \chi_m \mu_m. \tag{3.133}$$

Les coefficients de la matrice B dépendent de a et le choix de la valeur de a permet de se ramener à une relation de récurrence à trois termes. En effet, pour

$$a = \sqrt{c},$$

les seuls éléments non nuls sont donnés pour $n \geqslant 0$ par

$$\begin{aligned}
b_{n,n} &= (2n+1)c - \frac{1}{4}(3+2n+2n^2), \\
b_{n,n+4} &= \frac{1}{4}\sqrt{(n+1)(n+2)(n+3)(n+4)}, \\
b_{n+4,n} &= \frac{1}{4}\sqrt{(n+1)(n+2)(n+3)(n+4)}.
\end{aligned}$$

On considère la sous-matrice $B^{\mu,\nu}$ définie par

$$(B^{\mu,\nu})_{ij} = B_{ij},$$

avec $\mu \leqslant i, j \leqslant \nu$. Pour $0 \leqslant \mu < \nu - 4$ et $a = \sqrt{c}$, $B^{\mu,\nu}$ est symétrique, définie positive et s'écrit comme quatre sous-matrices tridiagonales. De plus, si $\nu < 2c$, alors chacune de ces sous-matrices est à diagonale dominante.

[ORX13] cherche un développement de $\psi_m^c(x)$ de la forme :

$$\psi_m^c(x) = \sum_{i=0}^{\infty}\sum_{k=0}^{\infty} \frac{\alpha_{i,k}}{c^k}.\phi_{m+4i}^{\sqrt{c}}(x) + \sum_{i=1}^{\lfloor m/4 \rfloor}\sum_{k=1}^{\infty} \frac{\beta_{i,k}}{c^k}.\phi_{m-4i}^{\sqrt{c}}(x). \tag{3.134}$$

La série définie par 3.134 ne converge pas pour $i, k \to \infty$ mais pour un entier n fixé, la somme finie $\psi_m^{c,n}$ définie par

$$\psi_m^c(x) = \sum_{i=0}^{n}\sum_{k=0}^{n} \frac{\alpha_{i,k}}{c^k}.\phi_{m+4i}^{\sqrt{c}}(x) + \sum_{j=1}^{p}\sum_{k=1}^{n} \frac{\beta_{j,k}}{c^k}.\phi_{m-4j}^{\sqrt{c}}(x), \tag{3.135}$$

où $p = \min(n, \lfloor m/4 \rfloor)$, converge uniformément vers ψ_m^c quand $c \to \infty$.

En posant

$$\alpha_i^n = \sum_{k=0}^{n} \frac{\alpha_{i,k}}{c^k} \tag{3.136}$$

$$\beta_i^n = \sum_{k=0}^{n} \frac{\beta_{i,k}}{c^k} \tag{3.137}$$

On peut réécrire l'approximation sous la forme,

$$\psi_m^c(x) = \sum_{i=0}^{n} \alpha_i^n.\phi_{m+4i}^{\sqrt{c}}(x) + \sum_{i=1}^{p} \beta_i^n.\phi_{m-4i}^{\sqrt{c}}(x), \tag{3.138}$$

De façon analogue, les valeurs propres $\chi_m(c)$ de L_c sont approchées par

$$\chi_m^n(c) = (1+2m)c + \sum_{i=0}^{n} \frac{\gamma_i(m)}{c^i}, \tag{3.139}$$

$\gamma_i(m)$ étant un polynôme en m. Cette somme finie ne converge pas vers $\chi_m(c)$ lorsque $n \to \infty$. Cependant, à un ordre n fixé, lorsque $c \to \infty$ on a

$$\lim_{c\to\infty} (\chi_m(c) - \chi_m^n(c)) = 0. \tag{3.140}$$

[ORX13] contient des approximations à l'ordre $n = 6$ pour les valeurs propres que nous ne détaillerons pas ici. Le théorème suivant fournit le comportement asymptotique des valeurs propres χ_m ainsi que des fonctions d'onde sphéroïdale ψ_m^c :

Théorème 3.27 (Approximations asymptotiques)**.** *Pour tout entier $m \geqslant 0$ et $0 \leqslant n \leqslant 5$, on note χ_m la m^e valeur propre de L_c et ψ_m^c la fonction propre associée. On note χ_m^n l'approximation définie par 3.139 et $\psi_m^{c,n}$ l'approximation donnée par 3.138. Alors pour une largeur de bande c suffisament grande, on a*

$$|\chi_m^n - \chi_m| = \mathcal{O}\left(\frac{1}{c^n}\right) \tag{3.141}$$

et

$$\|\psi_m^{c,n} - \psi_m^c\|_{[-\infty,+\infty]} = \mathcal{O}\left(\frac{1}{c^{n+1}}\right). \tag{3.142}$$

3.5 Retour au noyau de Green

3.5.1 Approximation du terme quadratique

Sous réserve d'admissibilité de Fresnel, on approche le noyau de Green à l'aide d'un développement limité de la phase faisant intervenir un produit tensoriel de deux opérateurs de Fox-Li définis par 3.3. Dans le cas général, on a cependant $x \in [-d_1, d_1]$ et $y \in [-d_2, d_2]$. On adapte la définition 3.3 dans le cas général,

Définition 3.28 (opérateur de Fox-Li/Fresnel,cas général)**.** Pour des réels $d_1 > 0$,$d_2 > 0$ et $\alpha > 0$, on considère l'opérateur intégral $\mathscr{F}_\alpha$ suivant,

$$\begin{aligned} \mathscr{F}_\alpha : \mathrm{L}^2([-d_1, d_1]) &\mapsto \mathrm{L}^2([-d_2, d_2]) \\ \lambda \mapsto [\mathscr{F}_\alpha \lambda](x) &= \int_{-d_1}^{d_1} e^{i\alpha(x-y)^2} \lambda(y)\, dy, \end{aligned} \tag{3.143}$$

L'étude du rang de l'opérateur $\mathscr{F}_\alpha$ se ramène à l'étude du rang de l'opérateur $\mathrm{F}_{2\alpha}$ défini par

Définition 3.29 (Opérateur F_c, cas général)**.** Pour des réels $d_1 > 0$,$d_2 > 0$ et $c > 0$, on considère l'opérateur intégral F_c suivant,

$$\begin{aligned} \mathrm{F}_c : \mathrm{L}^2([-d_1, d_1]) &\mapsto \mathrm{L}^2([-d_2, d_2]) \\ \lambda \mapsto [\mathrm{F}_c \lambda](x) &= \int_{-d1}^{d_1} e^{icxt} \lambda(t)\, dt. \end{aligned}$$

L'adjoint $\mathrm{F}_c^\star$ de F_c est donné par

$$\begin{aligned} \mathrm{F}_c^\star : \mathrm{L}^2([-d_2, d_2]) &\mapsto \mathrm{L}^2([-d_1, d_1]) \\ \mu \mapsto [\mathrm{F}_c^\star \mu](x) &= \int_{-d_2}^{d_2} e^{-icxz} \mu(z)\, dz. \end{aligned}$$

F_c est un opérateur compact de $\mathrm{L}^2([-d_1, d_1])$ dans $\mathrm{L}^2([-d_2, d_2])$.

Comme dans le cas $d_1 = d_2$, on s'intéresse aux valeurs singulières de l'opérateur F_c. Pour ce faire, on regarde les valeurs propres de $\mathrm{F}_c^\star \mathrm{F}_c$. Le calcul explicite de $\mathrm{F}_c^\star \mathrm{F}_c$ donne

$$[\mathrm{F}_c^\star \mathrm{F}_c \lambda](x) = \int_{-d_1}^{d_1} \lambda(t) \left[\int_{-d_2}^{d_2} e^{icz(t-x)} dz \right] dt$$

On intègre alors l'intégrale en z de façon exacte en ayant fait au préalable le changement de variable $z = u d_2$. D'où

$$[\mathrm{F}_c^\star \mathrm{F}_c \lambda](x) = \int_{-d_1}^{d_1} \lambda(t) \frac{2}{c(t-x)} \sin\left[c d_2 (t-x)\right] dt.$$

Afin de se ramener à l'intervalle de référence $[-1, 1]$, on écrit $t = v d_1$ et $x = u d_1$ et on obtient la formulation suivante

$$\begin{aligned} [\mathrm{F}_c^\star \mathrm{F}_c \lambda](x) &= \int_{-1}^{1} \lambda(v d_1) \frac{2}{c(v-u)} \sin\left[c d_1 d_2 (v-u)\right] dv \\ &= (d_1 d_2) \frac{2\pi}{c d_1 d_2} \left(\int_{-1}^{1} \lambda(v d_1) \frac{\sin\left[c d_1 d_2 (v-u)\right]}{c(v-u)} dv \right). \end{aligned}$$

Pour $d_1 > 0$ fixé, la fonction $v \mapsto \lambda(vd_1)$ appartient à $\mathrm{L}^2([-1,1])$ et l'on peut se ramener à l'opérateur Q_c défini à la section 3.4.4.3 en posant $c_{12} = cd_1d_2$,

$$\begin{aligned}[\mathrm{F}_c^\star \mathrm{F}_c\lambda](x) &= (d_1d_2)\frac{2\pi}{cd_1d_2}\left(\int_{-1}^{1}\lambda(vd_1)\frac{\sin[cd_1d_2(v-u)]}{c(v-u)}dv\right)\\ &= \frac{2\pi}{c_{12}}(d_1d_2)\mathrm{Q}_{c_{12}}.\end{aligned}$$

On peut alors obtenir le nombre de valeurs propres (normalisées) de $\mathrm{Q}_{c_{12}}$ supérieures à une tolérance α, que l'on note $\mathrm{N}(c_{12},\alpha)$ par la formule de Widom (3.59)

$$\mathrm{N}(c_{12},\alpha) = \frac{2c_{12}}{\pi} + \frac{1}{\pi^2}\log\left(\frac{1-\alpha}{\alpha}\right)\log(c_{12}).$$

Dans le cas de l'admissibilité de Fresnel, on a $c = 2\alpha$ avec $\alpha = k/2\mathrm{R}$. La largeur de bande c_{12} est alors donnée par

$$c_{12} = k\frac{d^2}{\mathrm{R}}, \tag{3.144}$$

avec $d = \sqrt{d_1d_2}$. Si $d_1 = d_2 = d$, on retrouve bien que le premier terme de l'expression de $\mathrm{N}(c_{12},\alpha)$ est $\frac{2}{\pi}\frac{kd^2}{\mathrm{R}}$ ce qui est en accord avec l'expression (3.33).

Proposition 3.30 (Rang de l'opérateur de Fox-Li, cas général)**.** *Pour $d_1 > 0$ et $d_2 > 0$, on considère l'opérateur de Fox-Li $\mathscr{F}_p$ défini de $\mathrm{L}^2([-d_1,d_1])$ dans $\mathrm{L}^2([-d_2,d_2])$ par*

$$\begin{aligned}\mathscr{F}_p : \mathrm{L}^2([-d_1,d_1]) &\mapsto \mathrm{L}^2([-d_2,d_2])\\ \lambda \mapsto [\mathscr{F}_p\lambda](x) &= \int_{-d_1}^{d_1} e^{ip(x-y)^2}\lambda(y)dy.\end{aligned}$$

On suppose que l'on se place dans le cas de l'admissibilité de Fresnel et ainsi, on a $p = k/2\mathrm{R}$. Soit une erreur relative $\epsilon \in]0,1[$, on note $\alpha = \epsilon^2$. On note r_ϵ le rang numérique à la précision ϵ de l'opérateur $\mathscr{F}_p$. Alors,

$$r_\epsilon \simeq \frac{2c}{\pi} + \frac{1}{\pi^2}\log\left(\frac{1-\alpha}{\alpha}\right)\log(c),$$

où c est la largeur de bande définie par

$$\begin{aligned}c &= 2pd^2\\ &= k\frac{d^2}{\mathrm{R}},\end{aligned}$$

avec $d = \sqrt{d_1d_2}$.

Remarque 3.31 (Minimisation de la section efficace). Dans le cas de deux objets quelconques, il est nécessaire de construire une base (u_1,u_2) du plan Π_{u_3} dans laquelle la section efficace est minimale. Plusieurs techniques sont envisageables et on présentera lors des applications une façon rapide de procéder. Si l'on ne minimise pas la section efficace, les largeurs de bande déterminées sont plus grandes et ceci conduit à une augmentation du nombre de nœuds par la formule de Widom.

3.5.2 Évolution du nombre de nœuds

Dans le cas de l'approximation du noyau oscillant, la largeur de bande c intervenant dans l'étude du terme du second ordre est définie par

$$c = \frac{(kd)^2}{k\mathrm{R}}, \tag{3.145}$$

où k est le nombre d'onde, d et R étant respectivement une dimension caractéristique de l'objet et R une distance carctéristique.

On rappelle dans le tableau suivant la relation entre kd et kR pour chaque zone d'admissibilité ainsi que la relation entre la largeur de bande c et kd. Pour l'admissibilité de Fresnel, on dissocie le cas où $d_{\parallel} = 0$ comme dans le cas de deux plans opposés.

Zone d'admissibilité	Relation entre kd et kR	Largeur de bande c
$1/\lvert x-y\rvert$	$kd \sim k\mathrm{R}$	$c \sim kd$
Fresnel	$(kd)^3 \sim (k\mathrm{R})^2$	$c \sim (kd)^{\frac{1}{2}}$
Fresnel (plans opposés)	$(kd)^4 \sim (k\mathrm{R})^3$	$c \sim (kd)^{\frac{2}{3}}$
Fraunhofer	$(kd)^2 \sim k\mathrm{R}$	$c \sim 1$

TABLEAU 3.1 – Comportement asymptotique de la largeur de bande en fonction de kd.

On rappelle que le nombre de nœuds d'interpolation choisi dans chaque direction du plan transverse est donné par

$$\mathrm{N}(c,\alpha) = \frac{2c}{\pi} + \frac{1}{\pi^2}\log\left(\frac{1-\alpha}{\alpha}\right)\log(c). \tag{3.146}$$

Ce nombre de points dépend à la fois de la précision et de la fréquence et on s'intéresse aux variations de $\mathrm{N}(c,\alpha)$ en fonction de la fréquence et de la précision.

3.5.2.1 Dépendance en fréquence

. D'après les différents critères d'admissibilité, on suppose que la largeur de bande suit une loi du type $c \sim (kd)^{\beta}, \beta > 0$. Dans le cas de l'électromagnétisme, on a $k = \frac{2\pi f}{c_0}$ où c_0 est la vitesse de la lumière et on a

$$c = \left(\frac{2\pi d}{c_0}\right)^{\beta} f^{\beta}. \tag{3.147}$$

On exprime alors le nombre de points $\mathrm{N}(c,\alpha)$ en fonction de la fréquence f.

$$\mathrm{N}(c,\alpha) = a_2 f^{\beta} + a_1 \log(f) + a_0,$$

avec

$$\begin{aligned} a_0 &= \frac{\beta}{\pi^2}\log\left(\frac{1-\alpha}{\alpha}\right)\log\left(\frac{2\pi d}{c_0}\right), \\ a_1 &= \frac{\beta}{\pi^2}\log\left(\frac{1-\alpha}{\alpha}\right), \\ a_2 &= \frac{2}{\pi}\left(\frac{2\pi d}{c_0}\right)^{\beta}. \end{aligned}$$

La dépendance en fréquence est présente par les deux termes f^{β} et $\log(f)$. Asymptotiquement, le terme en f^{β} prédomine et le nombre de points varie comme cette la puissance β de la fréquence. À cause du terme en $\log(f)$, il existe un régime transitoire pour

les fréquences moins élevées ce qui diminue la croissance dans cette zone transitoire. On note d'ailleurs que la formule de Widom est une formule asymptotique. Pour notre choix du nombres de nœuds, on a tronqué cette formule et on l'utilise pour toutes les fréquences.

On s'intéresse à la pente moyenne de cette croissance dans un intervalle de fréquence $[f_{\min}, f_{\max}]$. Pour ce faire, on considère la représentation en échelle logarithmique du nombre de points $\mathrm{N}(c,\alpha)$ en fonction de la fréquence f dont la pente fournit la puissance recherchée.

Considérons le changement de variable $u = \log_{10}(f)$. La pente en un point de l'intervalle $[\log_{10}(f_{\min}), \log_{10}(f_{\max})]$ est donné par la dérivée par rapport à u de la fonction

$$\mathrm{F}(u) = \log_{10}(\mathrm{N}(u)).$$

La pente asymptotique est $p_{as} = \beta$. Pour des fréquences comprises dans $[f_{\min}, f_{\max}]$, la pente moyenne p_{moy} peut être estimée par

$$p_{moy} = \frac{1}{\log_{10}(f_{\max}) - \log_{10}(f_{\min})} \int_{\log_{10}(f_{\max})}^{\log_{10}(f_{\max})} \mathrm{F}'(u)\,du. \tag{3.148}$$

3.5.2.2 Dépendance en précision

On suppose la fréquence (donc la largeur de bande c) fixée. On s'intéresse à la dépendance en précision du nombre de points $\mathrm{N}(c,\alpha)$ pour une précision ϵ donnée. Le nombre de nœuds d'interpolation est une fonction de $\alpha = \epsilon^2$ soit

$$\mathrm{N}(c,\epsilon) = \frac{2c}{\pi} + \frac{1}{\pi^2}\log\left(\frac{1-\epsilon^2}{\epsilon^2}\right)\log(c).$$

La dépendance en précision est logarithmique. Ainsi, à fréquence fixe, l'augmentation ou la diminution de la précision n'a pas une grande influence sur le nombre de points $\mathrm{N}(c,\epsilon)$.

Exemple 3.32 (Division par 10 de la précision)**.** On considère un objet dont la grandeur caractéristique d est de l'ordre de 100λ. En supposant que la largeur de bande varie comme la racine carrée de la fréquence, on obtient une largeur de bande de l'ordre de $c \approx 25$. Pour une précision $\epsilon' = \epsilon/10$, le nombre de points supplémentaires est d'environ 2.

3.5.3 Majoration du rang à l'aide de $\mathrm{N}(c,\alpha)$

On s'intéresse à déterminer une majoration du rang en fonction de $\mathrm{N}(c,\alpha)$ dans la zone d'admissibilité de Fresnel. Idéalement, on cherche une estimation également valable dans la zone d'admissibilité du noyau $1/|x-y|$. L'étude de la variation de $\mathrm{N}(c,\alpha)$ en fonction de la fréquence fournit le comportement asymptotique du rang. La majoration du rang permet dans la pratique de fournir des estimations pour l'allocation de la mémoire lors d'un code parallèle. Dans le cas général, le nombre de points dans la direction longitudinale u_3 entre en compte et on cherche à observer son influence sur le noyau. Dans cette partie, on se limite à étudier la partie oscillante du noyau de Green déconvolé par les ondes planes. On note alors $\mathrm{G}_e(x,y) = e^{ik(|x-y|-<u_3,x-y>)}$. On montre dans un premier une borne pour le rang numérique de G_e dans le cas de deux plans opposés puis sur le cas des sphères distantes.

3.5.3.1 Cas de deux plans opposés

Dans le cas de deux plans opposés de côté L, $d_{\parallel} = 0$ et la partie oscillante du noyau de Green est approchée par

$$\begin{aligned} e^{ik|x-y|} &\approx e^{ik\mathrm{R}} e^{ik\frac{|d_\perp|^2}{2\mathrm{R}}} e^{-ik\frac{|d_\perp|^4}{8\mathrm{R}^3}} \\ &\approx e^{ik\mathrm{R}} e^{ik\frac{|d_\perp|^2}{2\mathrm{R}}\left(1-\frac{|d_\perp|^2}{4\mathrm{R}^2}\right)} \end{aligned} \tag{3.149}$$

Dans l'expression (3.149), on note par $\mathrm{A}(d,\mathrm{R}) = 1 - \frac{|d_\perp|^2}{4\mathrm{R}^2}$ le coefficient venant modifié la phase. L'expression de $d_\perp$ dans la base (u_1, u_2, u_3) permet de donner le majorant suivant de $d_\perp$,

$$|d_\perp|^2 \leqslant 2\mathrm{L}^2.$$

Un majorant de $\mathrm{A}(d,\mathrm{R})$ est donné par

$$\mathrm{A}(d,\mathrm{R}) \leqslant 1 - \frac{\mathrm{L}^2}{2\mathrm{R}^2}.$$

Ce coefficient est donc toujours inférieur à 1 et la partie oscillante $\mathrm{G}_e(x,y)$ est alors approchée par

$$e^{ik|x-y|} \approx e^{ik\mathrm{R}} e^{ik_{eq}\frac{|d_\perp|^2}{2\mathrm{R}}}, \tag{3.150}$$

avec $k_{eq} = k.\left(1 - \frac{\mathrm{L}^2}{2\mathrm{R}^2}\right)$.

Quand la distance R est celle correspondant à l'admissibilité de $1/|x-y|$, on s'attend à un rang plus élevé que dans la zone de Fresnel. Dans ce cas, le facteur $\mathrm{A}(d,\mathrm{R})$ vient réduire le nombre d'onde et la phase se comporte comme k_{eq} et le rang est ainsi plus faible qu'escompté. On peut aussi définir une largeur de bande équivalente c_{eq} par

$$c_{eq} = k_{eq}\frac{d^2}{\mathrm{R}}, \tag{3.151}$$

et on peut déterminer le nombre de points $\mathrm{N}(c_{eq},\alpha)$ à l'aide de la formule de Widom.

Exemple 3.33 (Deux plaques carrées de coté $\mathrm{L} = 1m$)**.** On considère deux plaques carrées de côté $\mathrm{L} = 1m$ et on se donne une fréquence fixe $f = 11.43\mathrm{GH}z$. On effectue les tests avec une précision relative $\epsilon = 10^{-4}$ soit $\alpha = 10^{-8}$. Pour cette fréquence, on éloigne les deux plaques d'une distance R pour $\mathrm{R} \in [2m, 21m]$. Pour chaque distance R, on détermine le rang numérique r_ϵ à la précision ϵ du noyau $\mathrm{G}_e(x,y)$ ainsi que les nombres de points donnés par $\mathrm{N}(c_{eq},\alpha)$ et $\mathrm{N}(c,\alpha)$. La section efficace étant un carré, le nombre de points d'interpolation est donné par $\mathrm{N}(c,\alpha)^2$. On note que dans ce cas, on ne place pas de nœuds d'interpolation dans la direction u_3 car les plaques sont d'épaisseurs nulles. On représente alors les rapports $r_\epsilon(\mathrm{G}_e)/\mathrm{N}(c_{eq},\alpha)^2$ et $r_\epsilon(\mathrm{G}_e)/\mathrm{N}(c,\alpha)^2$ en fonction de la distance R.

Dans la zone d'admissibilité de $1/|x-y|$ (en vert clair sur la figure 3.5), on note que le nombre de points déterminé par $\mathrm{N}(c,\alpha)$ (*cf* courbe en bleu) est surestimé par rapport à celui obtenu à l'aide de $\mathrm{N}(c_{eq},\alpha)$ (*cf* courbe en noir). Cette surestimation perdure au début de la zone d'admissibilité de Fresnel (en vert foncé sur la figure 3.5). Quand la distance R augmente et que l'on se trouve dans la zone d'admissibilité de Fresnel, on a $\mathrm{A}(d,\mathrm{R}) \simeq 1$ et les estimations $\mathrm{N}(c,\alpha)$ et $\mathrm{N}(c_{eq},\alpha)$ deviennent similaires. Dans ce cas, il n'y a pas de variation suivant la direction u_3 car $d_{\parallel} = 0$. Le rang numérique r_ϵ est donc uniquement déterminé par le nombre de points utilisés pour décrire les variations dans le plan

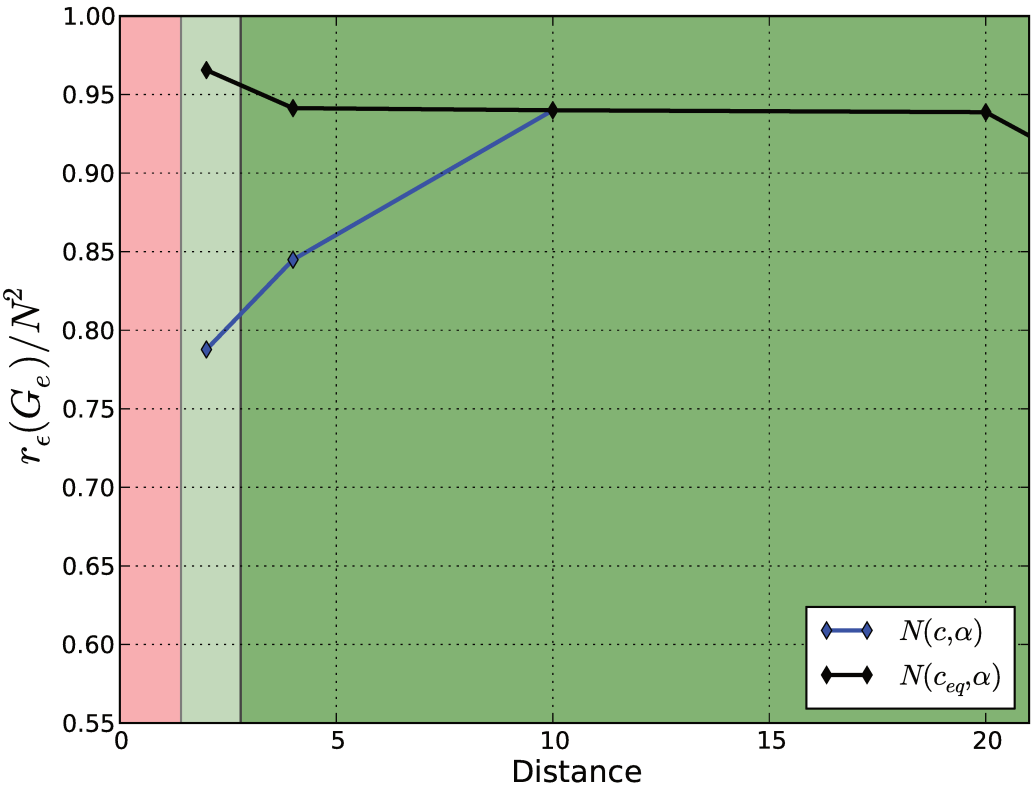


FIGURE 3.5 – Majoration du rang numérique $r_\epsilon(G_e)$ en fonction du nombre de points fournis par la formule de Widom

transverse Π_{u_3}.

Dans la pratique, les variations suivant Π_{u_3} sont correctement décrites à l'aide de $\mathrm{N}(c,\alpha)$ y compris dans la zone de d'admissibilité de $1/|x-y|$. Pour le choix d'une borne maximale sur le rang, on préfère utiliser $\mathrm{N}(c_{eq},\alpha)$. En effet, on a

$$r_\epsilon \leqslant \mathrm{C}.\mathrm{N}(c_{eq},\alpha)^2,$$

où $\mathrm{C} \approx 1$ pour une distance correspondant à l'admissibilité de $1/|x-y|$ et à l'admissibilité de Fresnel.

Remarque 3.34 (Condition d'admissibilité pour $d_\parallel = 0$). Le test précédent montre également qu'il est possible d'assouplir la condition sur γ dans la formule d'admissibilité (3.16). Par exemple, la distance correspondant au premier point des courbes correspond à $\gamma \simeq 2.5$.

3.5.3.2 Cas général : deux sphères distantes

Dans le cas général, on a $d_\parallel \neq 0$ et on écrit la phase en fonction de $d_\parallel$. La partie oscillante $e^{ik|x-y|}$ peut être approchée par

$$\begin{aligned} e^{ik|x-y|} &\approx e^{ik\mathrm{R}} e^{ikd_\parallel} e^{ik\frac{|d_\perp|^2}{2\mathrm{R}}} e^{-ik\frac{|d_\perp|^2 d_\parallel}{2\mathrm{R}^2}} e^{ik\frac{|d_\perp|^2\left(4|d_\parallel|^2-|d_\perp|^2\right)}{8\mathrm{R}^3}} \\ &\approx e^{ik\mathrm{R}} e^{ikd_\parallel} e^{ik\frac{|d_\perp|^2}{2\mathrm{R}}\left(1-\frac{d_\parallel}{\mathrm{R}}+\frac{|d_\parallel|^2}{\mathrm{R}^2}-\frac{|d_\perp|^2}{2\mathrm{R}^2}\right)} \end{aligned}$$

Comme pour le cas précédent, on s'intéresse au facteur $\mathrm{A}(d,\mathrm{R}) = 1-\frac{d_\parallel}{\mathrm{R}}+\frac{|d_\parallel|^2}{\mathrm{R}^2}-\frac{|d_\perp|^2}{2\mathrm{R}^2}$. On note que dans le cas où $d_\parallel = 0$, on retrouve le cas des plans opposés. Ici, $d_\parallel$ peut changer de signe et l'on doit se contenter de l'approximation pessimiste suivante,

$$\mathrm{A}(d,\mathrm{R}) \approx 1+\frac{|d_\parallel|}{\mathrm{R}}+\frac{|d_\parallel|^2}{\mathrm{R}^2}-\frac{|d_\perp|^2}{2\mathrm{R}^2}. \tag{3.152}$$

Dans le cas de sphères de rayon a, l'expression de $d_\parallel$ et $d_\perp$ dans la base (u_1,u_2,u_3) donne les majorations suivantes,

$$\begin{aligned} |d_\parallel| &\leqslant 2a, \\ |d_\perp|^2 &\leqslant 8a^2. \end{aligned}$$

Ainsi, un majorant de $\mathrm{A}(d,\mathrm{R})$ est donné par

$$\mathrm{A}(d,\mathrm{R}) \leqslant 1+\frac{2a}{\mathrm{R}}. \tag{3.153}$$

Ce facteur est cette fois-ci supérieur à 1 et le nombre d'onde équivalent défini par $k_{eq} = k.\left(1+\frac{2a}{\mathrm{R}}\right)$ est plus élevé que k. Avec la formule de Widom, on obtient donc un nombre de points plus élevé. Pour une dimension fixée, le facteur tend vers 1 lorsque la distance augmente. En utilisant la largeur de bande c_{eq} définie à l'aide de k_{eq}, le nombre de points $\mathrm{N}(c_{eq},\alpha)$ tend vers $\mathrm{N}(c,\alpha)$.

Exemple 3.35 (Deux sphères de rayon $a = 1m$)**.** On considère deux sphères de rayon $a = 1m$ et on se donne une fréquence fixe $f = 11.43\mathrm{GH}z$. On effectue les tests avec une précision relative $\epsilon = 10^{-4}$ soit $\alpha = 10^{-8}$. Pour cette fréquence, on éloigne les deux sphères

d'une distance R pour $R \in [10m, 60m]$. Pour chaque distance R, on détermine le rang numérique r_ϵ à la précision ϵ du noyau $G_e(x, y)$ ainsi que les nombre de points donnés par $N(c_{eq}, \alpha)$ et $N(c, \alpha)$. La section efficace étant un disque, le nombre de points d'interpolation est donné par $N(c, \alpha)^2$. Dans ce cas, on place $|\log_{10}(\epsilon)|$ nœuds dans la direction u_3 afin de prendre en compte les variations longitudinales. On représente alors les rapports $r_\epsilon(G_e)/N(c_{eq}, \alpha)^2$ et $r_\epsilon(G_e)/N(c, \alpha)^2$ en fonction de la distance R.

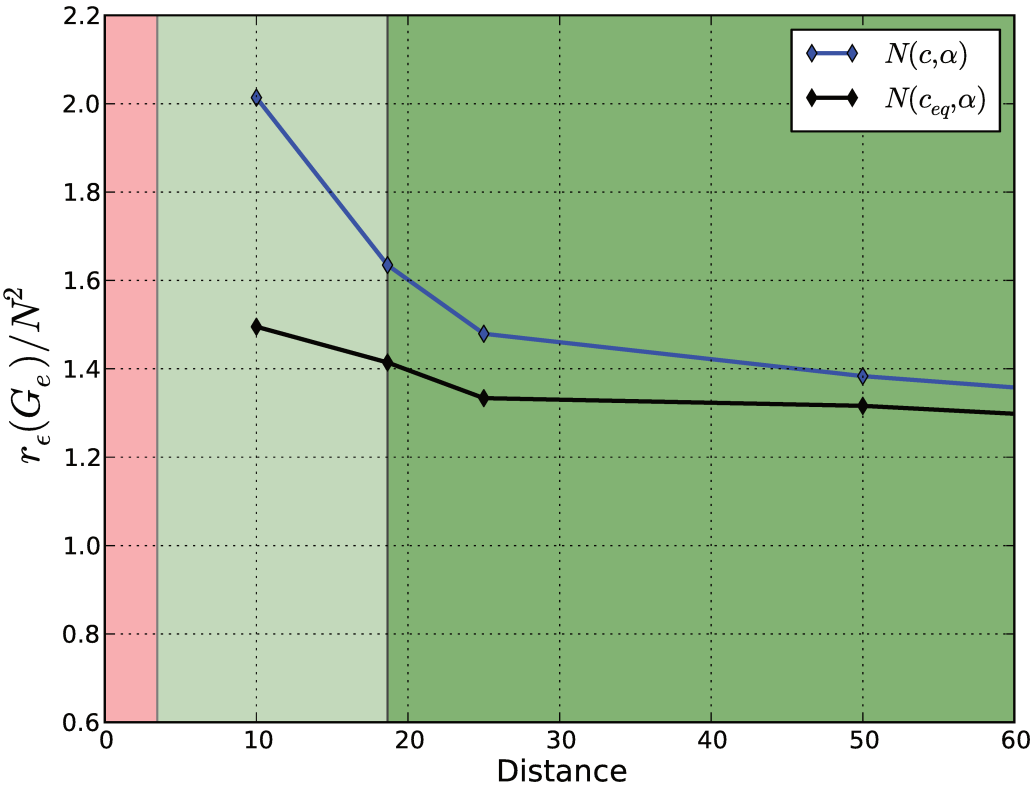


FIGURE 3.6 – Majoration du rang numérique $r_\epsilon(G_e)$ en fonction du nombre de points fournis par la formule de Widom

Dans les deux cas, le rapport est supérieur à 1. Ceci s'explique par le fait qu'il est nécessaire de prendre en compte les variations longitudinales selon u_3. L'estimation $N(c,\alpha)^2$ est correcte pour décrire les variations dans le plan transverse mais ne tient pas compte de la direction u_3. L'utilisation d'un nombre d'onde équivalent permet de rendre le rapport plus proche de 1 (*cf* courbe en noire sur la figure 3.6). Dans les deux cas, on constate que le rapport est majoré par deux.

D'un point de vue pratique, si $|d_\parallel| \neq 0$, on choisit le nombre de points dans le plan transverse à l'aide de $N(c,\alpha)$ et on obtient une borne maximale réaliste sur le rang à l'aide de $N(c_{eq},\alpha)$. Ce dernier est plus sûr car déterminé par une estimation pessimiste. On a alors dans ce cas général,

$$r_\epsilon \leqslant C.N(c_{eq},\alpha)^2,$$

où $1 \leqslant C \leqslant 2$ pour les distances correspondant à l'admissibilité de $1/|x-y|$ et à l'admissibilité de Fresnel.

3.6 Validation numérique

3.6.1 Géométries

Les géométries utilisées pour les tests sont les suivantes :

— deux sphères (*cf* développement limité de la phase),

— deux plaques opposées,
— deux plaques alignées.

Dans ces cas simples, la section efficace représentant le motif de la projection dans le plan Π_{u_3} peut être facilement déterminée et on peut en déterminer une base de façon triviale.

Pour chaque géométrie, on détermine les paramètres géométriques ainsi que la largeur de bande en fonction de la longueur d'onde du problème et des dimensions de la géométrie. La valeur de la largeur de bande permet d'obtenir une estimation du rang de l'opérateur de Fox-Li introduit précédemment. Ce rang permet alors de choisir le nombre de points d'interpolation dans la méthode HCA – II ci-dessus.

3.6.1.1 Calcul des paramètres géométriques et de la largeur de bande

On rappelle que dans la base d'interaction (u_1, u_2, u_3), on a

$$|d_\perp|^2 = (\xi_1 - \eta_1)^2 + (\xi_1 - \eta_1)^2,$$
$$|d_\parallel| = |(\xi_3 - \eta_3)|,$$

avec $\xi_1, \xi_2, \eta_1, \eta_2 \in [-d, d]$.

La distance centre à centre R est donnée alors par

$$k.\frac{|d_\perp|^2 |d_\parallel|}{2R^2} \leqslant 2\pi$$
$$R \geqslant$$

La largeur de bande utilisée pour la détermination du nombrede points dans la partie transverse est donné par

$$c = k.\frac{d^2}{R}.$$

Pour chaque géométrie, on détermine la valeur de la grandeur d et la distance R telle que l'on soit dans la zone d'admissibilité de Fresnel. Ainsi, on peut déterminer la valeur de la largeur de bande c en fonction de la fréquence.

3.6.1.2 Cas de la sphère

Valeur de d Dans le cas de deux sphères de rayon a, le motif obtenu par la projection sur le plan Π_{u_3} est un disque de rayon a que l'on peut inclure dans le carré $[-a, a] \times [-a, a]$ dans la base transverse (u_1, u_2) soit $d = a$. Pour $i = 1, 2, 3$, on a alors $\xi_i, \eta_i \in [-a, a]$. $d_\perp$ et $d_\parallel$ vérifient alors les majorations suivantes,

$$|d_\perp|^2 \leqslant (2a)^2 + (2a)^2 = 8a^2,$$
$$|d_\parallel| \leqslant 2a.$$

Valeur de la distance R En choisissant de borner le reste par une oscillation (ie $\alpha = 1$ dans (3.16)), le critère d'admissibilité de Fresnel est donné par

$$k.\frac{|d_\perp|^2 |d_\parallel|}{2R^2} \leqslant 2\pi.$$

En utilisant la relation $k = 2\pi/\lambda$ et les majorants de $|d_\perp|^2$ et $|d_\parallel|$, on obtient la condition suivante sur la distance centre à centre R,

$$R \geqslant \left(\frac{8a^3}{\lambda}\right)^{1/2}. \tag{3.154}$$

Afin de se placer au début de la zone de Fresnel, on choisit $R = \left(\frac{8a^3}{\lambda}\right)^{1/2}$.

Valeur de la largeur de bande c Dans ce cas, la largeur de bande est déterminée en fonction du rayon a et de la longueur d'onde λ par

$$\begin{aligned} c &= k.\frac{d^2}{\mathrm{R}} \\ &\approx 2.22\left(\frac{a}{\lambda}\right)^{1/2}. \end{aligned}$$

Valeurs numériques des paramètres Le calcul de la pente moyenne p_{moy} est effectué à l'aide de (3.148) avec les paramètres $d = a$ et $\beta = 1/2$.

Géométrie	2 sphères de rayon $a = 1.0(m)$
Physique	Électromagnétisme : $c_0 \simeq 3.10^8(m.s^{-1})$
Fréquence	$f \in [5.02.10^8(\mathrm{Hz}), 3.14.10^{10}(\mathrm{Hz})]$
Longueur d'onde	$\lambda \in [9.5.10^{-3}(m), 5.9.10^{-1}(m)]$
Motif de la projection dans (u_1, u_2, u_3)	disque de rayon $a = 1.0(m)$
Distance (limite basse Fresnel)	$\mathrm{R} \in [3.65(m), 28.93(m)]$
Largeur de bande	$c \in [2.87, 22.71]$

TABLEAU 3.2 – Configuration du cas test pour les plaques opposées

3.6.1.3 Cas des plaques opposées

Valeur de d Dans le cas de deux plaques carrées de côté L opposées, le motif obtenu par la projection sur le plan Π_{u_3} est le carré de côté L et dans la base du plan transverse, on a $d = \frac{\mathrm{L}}{2}$. Pour $i = 1, 2$, on a alors $\xi_i, \eta_i \in [-\frac{\mathrm{L}}{2}, \frac{\mathrm{L}}{2}]$. Comme il s'agit de plaques, les coordonnées ξ_3 et η_3 sont nulles. $d_\perp$ et $d_\parallel$ satisfont les conditions suivantes,

$$\begin{aligned} |d_\perp|^2 &\leqslant \mathrm{L}^2 + \mathrm{L}^2 = 2\mathrm{L}^2, \\ |d_\parallel| &= 0. \end{aligned}$$

Valeur de la distance R En choisissant de borner le reste par une oscillation (ie $\alpha = 1$ dans (3.16)), le critère d'admissibilité de Fresnel dans le cas où $d_\parallel = 0$ est donné par

$$k.\frac{|d_\perp|^4}{8\mathrm{R}^3} \leqslant 2\pi.$$

En utilisant la relation $k = 2\pi/\lambda$ et le majorant de $|d_\perp|^2$ on obtient

$$\mathrm{R} = \left(\frac{\mathrm{L}^4}{2\lambda}\right)^{1/3}.$$

Valeur de la largeur de bande c La largeur de bande est déterminée en fonction du côté L et de la longueur d'onde λ par

$$\begin{aligned} c &= k.\frac{d^2}{\mathrm{R}} \\ &\approx 1.97\left(\frac{\mathrm{L}}{\lambda}\right)^{2/3} \end{aligned}$$

Valeurs numériques des paramètres Le calcul de la pente moyenne p_{moy} est effectué à l'aide de (3.148) avec les paramètres $d = \mathrm{L}/2$ et $\beta = 2/3$.

Géométrie	2 carrés opposés de coté $\mathrm{L} = 1.0(m)$
Physique	Électromagnétisme : $c_0 \simeq 3.10^8(m.s^{-1})$
Fréquence	$f \in [2.38.10^9(\mathrm{Hz}), 4.10^{10}(\mathrm{Hz})]$
Longueur d'onde	$\lambda \in [7.5.10^{-3}(m), 1.2.10^{-1}(m)]$
Motif de la projection dans (u_1, u_2, u_3)	carré $[-0.5(m), 0.5(m)] \times [-0.5(m), 0.5(m)]$
Distance (limite basse Fresnel)	$\mathrm{R} \in [1.58(m), 4.05(m)]$
Largeur de bande	$c \in [7.8, 51.4]$

TABLEAU 3.3 – Configuration du cas test pour les plaques opposées

3.6.1.4 Cas des plaques coplanaires

Valeur de d Le motif obtenu par la projection sur le plan Π_{u_3} de deux plaques carrées de côté L est un segment de longueur $\frac{\mathrm{L}}{2}$ soit $d = \frac{\mathrm{L}}{2}$. Dans ce cas, dans le plan transverse, les coordonnées ξ_2 et η_2 sont nulles et on a $\xi_1, \eta_1 \in [-\frac{\mathrm{L}}{2}, \frac{\mathrm{L}}{2}]$. Les coordonnées longitudinales ξ_3 et η_3 vérifient $\xi_1, \eta_1 \in [-\frac{\mathrm{L}}{2}, \frac{\mathrm{L}}{2}]$. On a les estimations suivantes,

$$|d_\perp|^2 \leqslant \mathrm{L}^2,$$
$$|d_\parallel| \leqslant \mathrm{L}.$$

Valeur de la distance R En choisissant de borner le reste par une oscillation (ie $\alpha = 1$ dans (3.16)), le critère d'admissibilité de Fresnel est donné par

$$k.\frac{|d_\perp|^2 |d_\parallel|}{2\mathrm{R}^2} \leqslant 2\pi.$$

D'où,

$$\mathrm{R} = \left(\frac{\mathrm{L}^3}{2\lambda}\right)^{1/2}.$$

Valeur de la largeur de bande c La largeur de bande est déterminée en fonction du côté L et de la longueur d'onde λ par

$$c = k.\frac{d^2}{\mathrm{R}}$$
$$\approx 2.22\left(\frac{\mathrm{L}}{\lambda}\right)^{1/2}.$$

Valeurs numériques des paramètres Le tableau suivant résume la configuration du cas de test pour des plaques coplanaires.

Géométrie	2 carrés coplanaires de coté $L = 1.0(m)$
Physique	Électromagnétisme : $c_0 \simeq 3.10^8(m.s^{-1})$
Fréquence	$f \in [2.38.10^9(Hz), 4.10^{10}(Hz)]$
Longueur d'onde	$\lambda \in [7.5.10^{-3}(m), 1.2.10^{-1}(m)]$
Motif de la projection dans (u_1, u_2, u_3)	segment $[-0.5(m), 0.5(m)]$
Distance (limite basse Fresnel)	$R \in [1.99(m), 8.16(m)]$
Largeur de bande	$c \in [6.25, 25.63]$

TABLEAU 3.4 – Configuration du cas test pour les plaques coplanaires

3.6.2 Résultats numériques

Pour chaque géométrie, on détermine la largeur de bande ainsi que le nombre de points fourni par la formule de Widom nécessaires à l'étude du rang du noyau de Green. Nos formules étant basées sur un développement limité de la phase, on observe en premier lieu le rang du noyau $G_e(x,y) = e^{ik(|x-y|-<u_3,x-y>)}$ puis le rang du noyau $G_{u_3}(x,y) = e^{ik(|x-y|-<u_3,x-y>)}/|x-y|$. La croissance du rang en fonction de la fréquence est estimée grâce à la formule de Widom et à l'analyse de la section 3.5.2.1.

3.6.2.1 Valeurs numériques des approximations à $\epsilon = 1{,}0 \times 10^{-4}$

Sphères distantes Dans le cas des sphères, la section efficace est un disque que l'on inscrit dans une boite englobante carrée de section minimale. Dans chaque direction u_1, u_2, la formule de Widom estimant le rang de l'opérateur de Fox-Li fournit le nombre de points à utiliser pour l'approximation. On ajoute de l'ordre de $|\log_{10}(\epsilon)|$ points dans la direction longitudinale. Dans la pratique le rang est borné par $|\log_{10}(\epsilon)|N(c,\alpha)^2$ points.

Fréquence (Hz)	distance R	c	$N(c,\alpha)$	c_{eq}	$N(c_{eq},\alpha)$	$r_\epsilon(G_e(x,y))$	$r_\epsilon(G_{u_3}(x,y))$
$5{,}0 \times 10^8$	3.6	2.91	3.84	4.52	5.7	89	72
$1{,}0 \times 10^9$	5.16	4.06	5.20	5.63	6.81	117	91
$2{,}84 \times 10^9$	8.7	6.84	7.94	8.41	9.33	185	144
$4{,}98 \times 10^9$	11.54	9.04	9.86	10.60	11.15	252	192
$1{,}57 \times 10^{10}$	20.48	16.10	15.44	17.67	16.61	507	367
$3{,}14 \times 10^{10}$	28.92	22.75	20.31	24.32	21.44	806	582

TABLEAU 3.5 – Résultats numériques pour les sphères pour $\epsilon = 1.10^{-4}$

Plans opposés Dans le cas des plans carrés opposés, la section efficace est un carré. Ainsi, dans chaque direction u_1, u_2, la formule de Widom eestimant le rang de l'opérateur de Fox-Li fournit le nombre de points à utiliser pour l'approximation. Dans ce cas, l'épaisseur des plaques étant nulles, on n'ajoute pas de points supplémentaires dans la direction longitudinale. Dès lors, le rang est borné par le nombre de points utilisés dans le plan Π_{u_3}. Comme les plans sont carrés, il s'agit en tout de $N(c,\alpha)^2$ points.

Fréquence (Hz)	distance R	c	$\mathrm{N}(c,\alpha)$	c_{eq}	$\mathrm{N}(c_{eq},\alpha)$	$r_\epsilon(\mathrm{G}_e(x,y))$	$r_\epsilon(\mathrm{G}_{u_3}(x,y))$
$2{,}38\times10^9$	1.58	7.89	8.88	6.30	7.45	106	108
$5{,}02\times10^9$	2.02	13.02	13.07	11.42	11.81	194	204
$1{,}0\times10^{10}$	2.54	20.62	18.78	19.02	17.60	379	383
$2{,}0\times10^{10}$	3.22	32.54	27.21	30.97	26.12	753	765
$4{,}0\times10^{10}$	4.04	51.87	40.39	50.28	39.32	1582	1621

TABLEAU 3.6 – Résultats numériques pour les plans opposés pour $\epsilon = 1{,}0\times10^{-4}$

Plans coplanaires Dans le cas des plans coplanaires, la section efficace est un segment. Ainsi, la formule de Widom est utilisée dans une seule direction. Ce sont les valeurs de ces estimations qui sont représentées dans le tableau suivant. Ainsi, le nombre de points utilisés dans le plan transverse Π_{u_3} est déterminé par $\mathrm{N}(c,\alpha)$.

Fréquence (Hz)	distance R	c	$\mathrm{N}(c,\alpha)$	c_{eq}	$\mathrm{N}(c_{eq},\alpha)$	$r_\epsilon(\mathrm{G}_e(x,y))$	$r_\epsilon(\mathrm{G}_{u_3}(x,y))$
$2{,}38\times10^9$	1.98	6.29	7.44	8.66	9.54	23	26
$5{,}02\times10^9$	2.88	9.13	9.94	11.74	12.07	26	26
$1{,}0\times10^{10}$	4.08	12.84	12.94	15.60	15.05	30	30
$2{,}0\times10^{10}$	5.76	18.19	16.99	21.07	19.10	37	37
$4{,}0\times10^{10}$	8.16	25.68	22.40	28.63	24.49	46	46

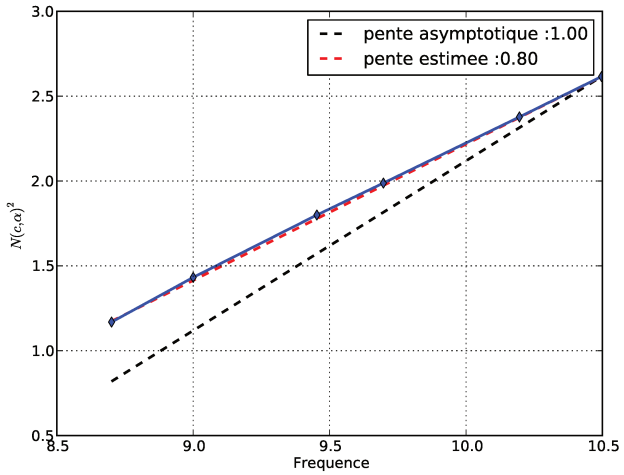
TABLEAU 3.7 – Résultats numériques pour les plans coplanaires pour $\epsilon = 1{,}0\times10^{-4}$

3.6.2.2 Croissance de $N(c,\alpha)$ en fonction de la fréquence

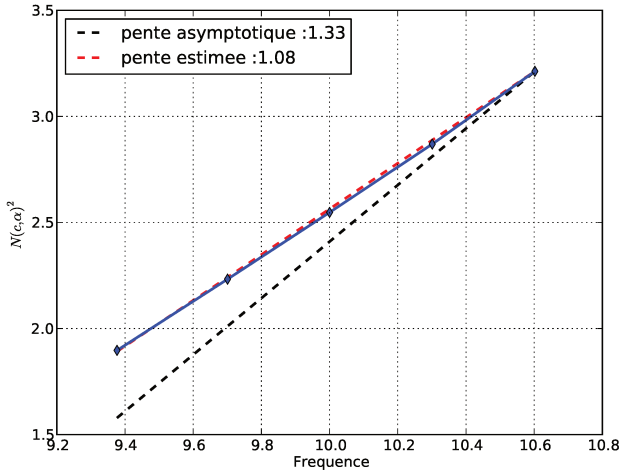
On utilise l'analyse sur le comportement de la pente développée auparavant. La dépendance en fréquence du nombre de points $N(c,\alpha)$ est explicité par la formule (3.148). Au debut (à basse fréquence), il y a une compétition entre les deux contributions et la pente asymptotique n'est visible que pour des hautes fréquences (*ie* de grands objets).

La figure 3.7 représente la croissance de $N(c,\alpha)$ en fonction de la fréquence pour les trois cas tests présentés (courbes bleues). Dans les cas des sphères et des plans opposés (3.7a et 3.7b), la section efficace est de dimension 2. On représente alors la croissance de $N(c,\alpha)^2$ en fonction de la fréquence. Le cas des plans coplanaires (3.7c) aboutit à une section efficace qui est un segment et on représente alors $N(c,\alpha)$ en fonction de la fréquence.

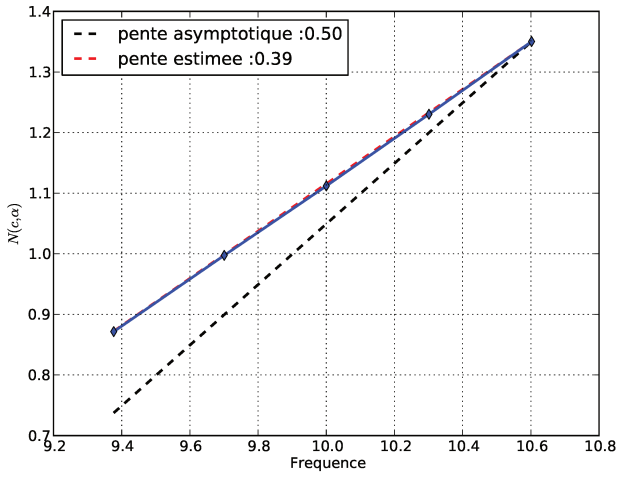
Pour chaque cas, la formule (3.148) fournit la pente moyenne pour la bande de fréquence testée. Cette pente moyenne est représenté par une ligne en pointillé rouge dans chaque cas. On représente également les pentes asymptotiques par une courbe en pointillé noir. On note que l'on n'atteint pas la croissance asymptotique quelque soit le cas test présenté pour les fréquences considérées. Cependant, dans le cas des plans opposés, à défaut d'apercevoir le régime asymptotique, on note que la pente est supérieure à 1. Ceci illustre que dans ce cas la croissance n'est pas linéaire. En effet, le critère de Fresnel dans le cas de plans opposés conduit à une croissance de pente asymptotique 4/3.



(a) Cas des sphères distantes (*cf* 3.2).



(b) Cas des plans opposés (*cf* 3.3).



(c) Cas des plans coplanaires (*cf* 3.4)

FIGURE 3.7 – Croissance du nombre de points dans le plan Π_{u_3} en fonction de la fréquence (échelle logarithmique).

3.6.2.3 Majoration du rang en fonction de $N(c,\alpha)$

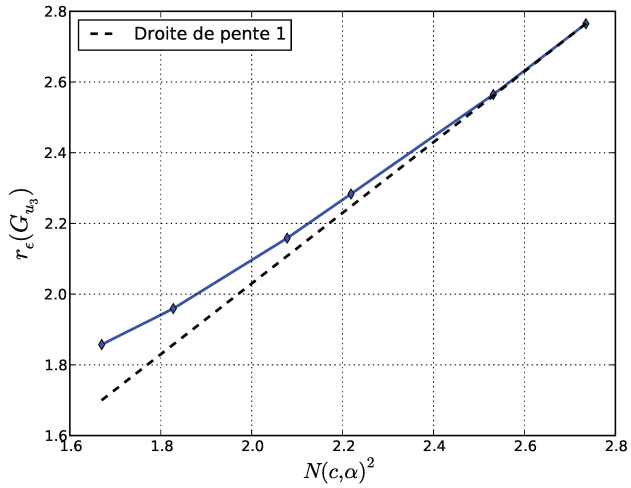
Ce test nous permet de chercher une borne maximale correcte afin d'estimer le rang a priori. Pour ce faire, on effectue le test sur le noyau de Green déconvolé par les ondes planes $G_{u_3}(x,y)$.

L'approximation du terme quadratique par un produit tensoriel de deux opérateurs de Fox-Li nécessite de prendre un nombre de points légérement supérieur à $N(c,\alpha)$ dans chaque direction. Dans la pratique, on constate qu'en prenant $N(c,\alpha')$ pour $\alpha' < \alpha$, ceci convient. L'exemple 3.32 de la section 3.5.2.2 montre qu'en prenant une précision relative $\epsilon' = \epsilon/10$ ceci rajoute de l'ordre de 3 points à $N(c,\alpha)$. Le nombre de points utilisé dans le calcul correspond à $N(c,\alpha)+3$. C'est ce nombre que l'on utilise pour la majoration du rang que l'on présente dans ce test. On veut obtenir une majoration du rang en fonction du nombre de points nécessaires à l'approximation des variations transverses dans Π_{u_3}. Dans les cas des sphères et des plans opposés (3.7a et 3.7b), la section efficace est de dimension 2 et le majorant recherché est fonction de $(N(c,\alpha)+3)^2$.

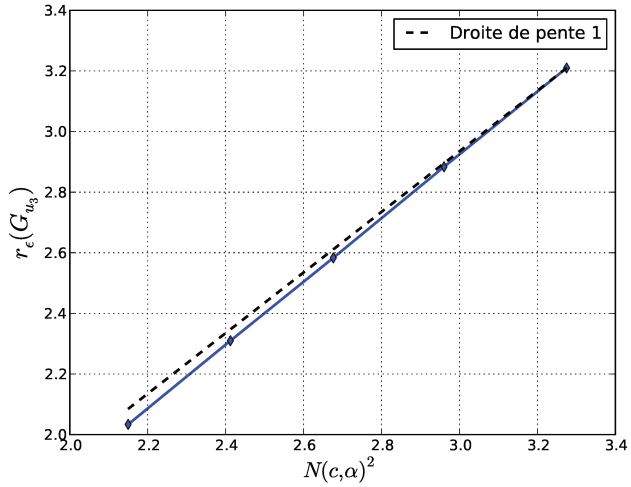
Asymptotiquement, la croissance du rang en fonction du nombre de points dans le plan transverse est linéaire comme le montre les figures suivantes. De plus, on constate numériquement que

De plus, dans le cas des plans opposés, on ne place pas de points dans la direction longitudinale car $d_{\parallel} = 0$. La borne sur le rang est donc simplement le nombre de points d'interpolation utilisé pour décrire les variations dans le plan Π_{u_3} et le rapport $r_\epsilon/N(c,\alpha)^2$ est inférieur à 1. L'utilisation d'un nombre d'onde équivalent permet d'avoir un nombre de points moins élevés dans ce cas. Dans les autres cas, l'utilisation d'un nombre d'onde équivalent n'amène pas d'informations supplémentaires car le majorant est trop grossier. Cependant, le rapport $r_\epsilon/N(c,\alpha)^2$ dans le cas des plans coplanaires ou $r_\epsilon/N(c,\alpha)^2$ dans le cas des sphères est majoré par 2 pour les grandes fréquences.

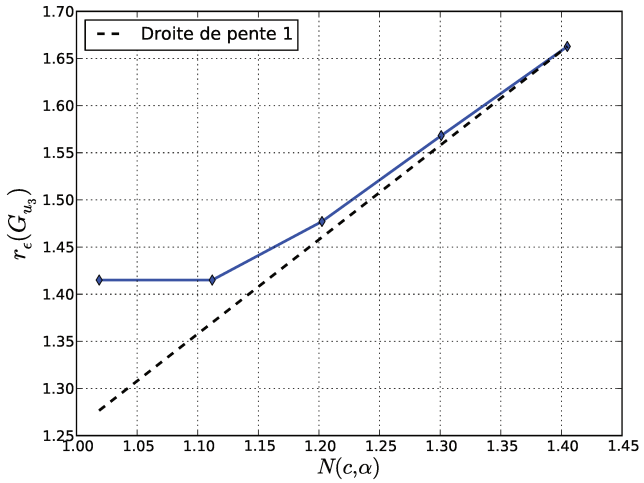
Le rang étant asymptotiquement linéaire en fonction du nombre de points dans Π_{u_3}, on peut en déduire que l'étude de la croissance de $N(c,\alpha)$ en fonction de la fréquence fournit le comportement asymptotique du rang en fonction de la fréquence. C'est le but du test suivant.



(a) Cas des sphères distantes (*cf* 3.2).



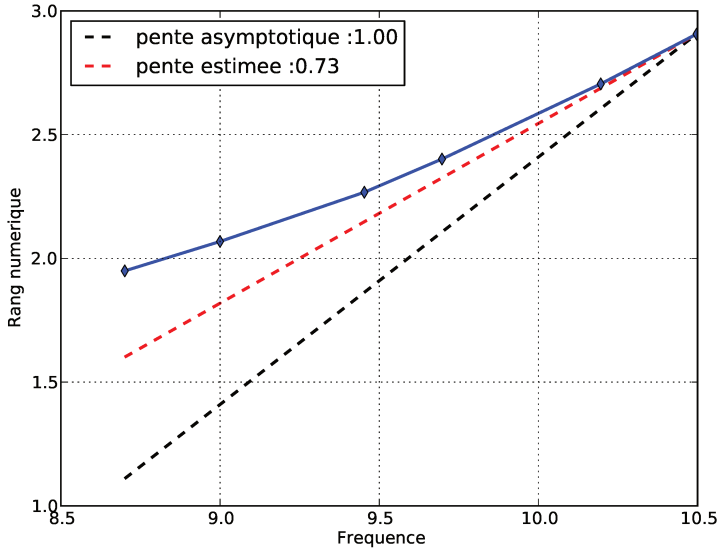
(b) Cas des plans opposés (*cf* 3.3).



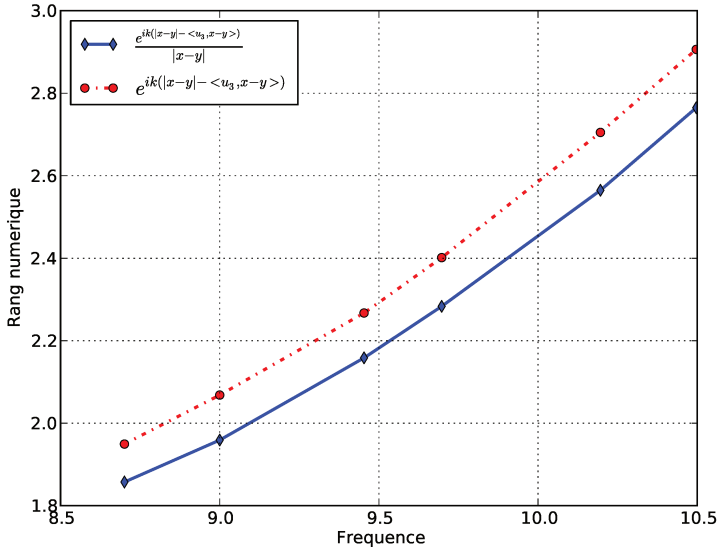
(c) Cas des plans coplanaires (*cf* 3.4)

FIGURE 3.8 – Croissance du rang de G_{u_3} en fonctiondu nombre de points dans le plan Π_{u_3} (échelle logarithmique).

3.6.2.4 Croissance du rang en fonction de la fréquence



(a) Rang de $G_e(x,y) = e^{ik(|x-y|-<u_3,x-y>)}$



(b) Rang de $G_{u_3}(x,y) = e^{ik(|x-y|-<u_3,x-y>)}/|x-y|$

FIGURE 3.9 – Croissance du rang en fonction de la fréquence (échelle logarithmique). Cas des sphères (*cf* 3.2)

La figure 3.9a représente la croissance du rang du noyau $G_e(x,y) = e^{ik(|x-y|-<u_3,x-y>)}$ en fonction de la fréquence (courbe bleue en trait plein) en échelle logarithmique. On note que cette courbe est constituée de deux parties où la croissance est différente. Pour les basses fréquences, la croissance est plus lente tandis qu'elle augmente à haute fréquences. Dans le cas des sphères, le calcul explicite du motif de la projection dans Π_{u_3} ainsi que la formule d'admissibilité fournissent les paramètres suivants, que l'on peut

utiliser dans la formule (3.148) :

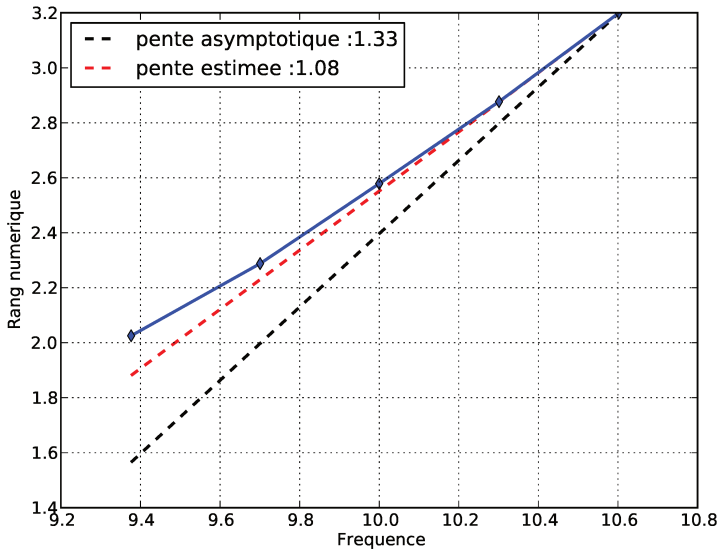
$$d = 1.0, \tag{3.155}$$

$$\beta = 1/2. \tag{3.156}$$

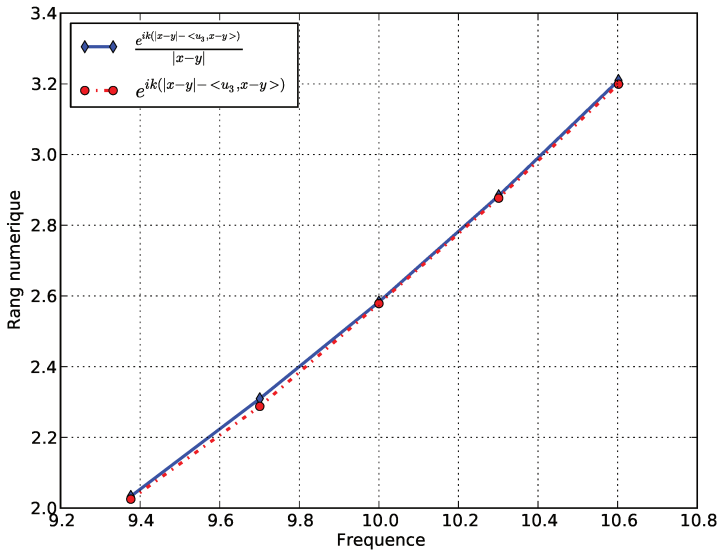
Cette formule permet de déterminer la pente moyenne de la croissance pour une dimension d'espace. La section efficace étant de dimension 2, on obtient la pente moyenne en prenant le double de l'estimation 1D. Cette pente est représentée en rouge sur la figure 3.9a et est déterminée a priori. C'est la conséquence de l'analyse de la section 3.5.2.1 qui permet d'estimer le comportement du rang. La pente moyenne estimée est d'environ $p_{moy} \approx 0.73$. D'après les résultats de la section 3.5.2.1, on sait que la pente asymptotique est $p_{as} = 1$ (droite représentée en noir sur la figure).

La figure 3.9b représente la croissance du rang du noyau de Green déconvolé par les ondes planes en fonction de la fréquence (courbe bleue en trait plein) en échelle logarithmique. On effectue une comparaison avec le rang de la partie oscillante G_e représentée en rouge. On note que le rang du noyau G_{u_3} est inférieur à celui de la partie oscillante mais que le profil de croissance en fonction de la fréquence est le même. La partie en $1/|x-y|$ ne joue que par un facteur multiplicatif et ne change pas la croissance.

Dans le cas des plans opposés, on sait que la section efficace est également de dimension 2 et on exprime le nombre de points dans Π_{u_3} par un produit cartésien. Cependant, dans ce cas, on a $d_{\parallel} = 0$ et le critère d'admissibilité de Fresnel est différent. La largeur de bande en une dimension est en $(kd)^{2/3}$ donc pour le terme quadratique complet, le nombre de points est en $(kd)^{4/3}$. La pente asymptotique attendu lorsqu'on observe la croissance du rang est donc $p_{as} = 4/3$. La formule (3.148) avec $\beta = 2/3$ et $d = 1/2$ (la plaque est un carré unitaire) fournit une pente moyenne $p_{moy} \approx 0.54$ dans la bande de fréquence testée. Afin d'obtenir la pente en dimension 2, il suffit de doubler cette pente. Ainsi, au lieu d'une pente asymptotique de 1.33 (droite noire en pointillé sur la figure XX), on a une pente d'environ 1.08 (droite rouge en pointillé).
Comme $d_{\parallel} = 0$, on ne place pas de points dans la direction longitudinale u_3 et le rang est uniquement déterminé par le nombre de points mis dans Π_{u_3}. Ainsi, le rang du noyau de Green déconvolé par les ondes planes est étudié avec les mêmes nœuds d'interpolation que la partie oscillante. À un faible pourcentage près, le rang de G_{u_3} est approximativement le même que celui de G_e (*cf* table 3.6).



(a) Cas des plans opposés (*cf* 3.3)

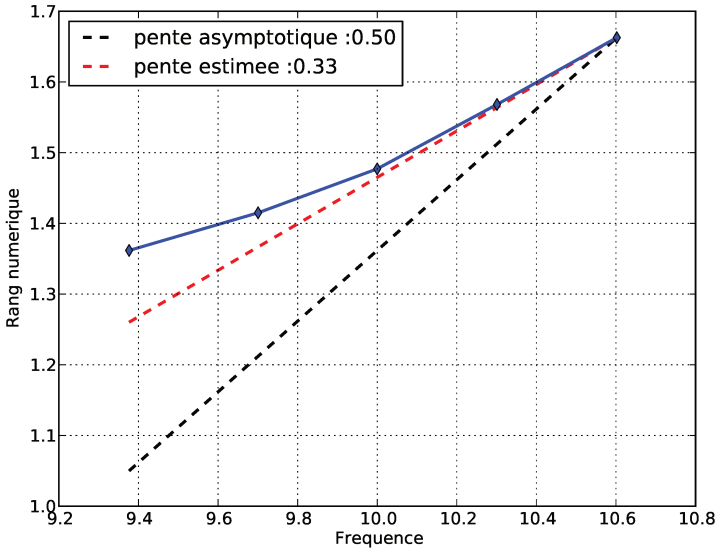


(b) Cas des plans opposés (*cf* 3.3)

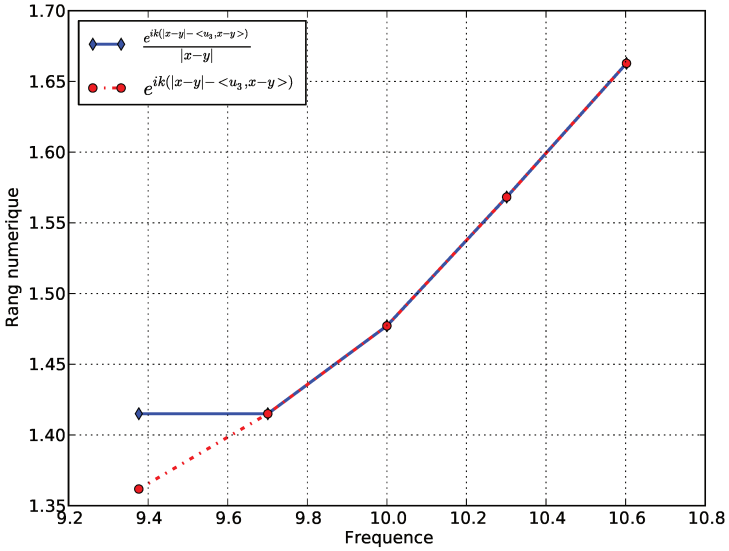
FIGURE 3.10 – Croissance du rang en fonction de la fréquence (échelle logarithmique). Cas des plans opposés (*cf* 3.3)

Pour deux plans coplanaires (*cf* 3.11), la section efficace se résume à un segment. Dans ce cas particulier, on doit s'attendre à un rang très faible car à haute fréquence, deux telles géométries ne se "voient" pas. La profondeur de la plaque nécessite que l'on dispose des points d'interpolation afin de correctement décrire les variations longitudinales (selon l'axe u_3). Cependant, comme nous nous restreignons à l'étude du noyau déconvolépar les ondes planes, il est moral d'obtenir un rang qui ne dépende que très faiblement du nombre de points placé dans la direction u_3. Dans ce cas, le critère d'admissibilité de Fresnel prend la même forme que celui utilisé dans le cas des sphères mais le terme $|d_\perp|^2$ est moins important du fait de la section efficace qui est un segment et non une surface. Dans ce cas, on vise à comparer le rang du noyau à l'estimation par la formule de Widom. La section efficace étant de dimension 1, il s'agitr d'une validation directe de la formule de Widom. La pente asymptotique attendue dans le cas de la partie oscillante G_e est $p_{as} = 1/2$ (droite noire en pointillé sur la figure YY). On détermine la pente moyenne sur la bande de fréquence testée à l'aide de la formule (3.148) avec $\beta = 1/2$ et $d = 1/2$. Dans ce cas, on obtient une pente moyenne $p_{moy} = 0.33$ (droite rouge). Le noyau de Green déconvolé par les ondes planes G_{u_3} possède ici un comportement similaire à celui de G_e et on constate que l'asymptotique déterminée reste valable pour G_{u_3}.

Résumé sur les trois cas Dans ces trois cas, on note que la bande de fréquence testée correspond à une zone où l'on atteint pas encore la pente asymtotique. Ceci suggère que l'on atteint l'asymtotique lorsque l'on traite des objets de très grandes dimensions (en terme de longueurs d'ondes). Dans le cas d'un objet de plusieurs millions de degrés de liberté, il est possible d'atteindre la croissance asymptotique.



(a) Cas des plans coplanaires (*cf* 3.4)



(b) Cas des plans coplanaires (*cf* 3.4)

FIGURE 3.11 – Croissance du rang en fonction de la fréquence (échelle logarithmique). Cas des plans coplanaires (*cf* 3.4)

3.7 Conclusion

Ce chapitre représente la contribution majeure de cette thèse. Les résultats de la littérature sont donnés d'après le développement limité à l'ordre 2. On en déduit alors une condition pour le terme de reste ; c'est la condition de Fraunhofer. Cette condition d'admissibilité est d'un intérêt pratique car seules les ondes planes contribuent à la croissance du rang de la partie oscillante du noyau. Nous avons alors utilisé le terme du second ordre pour obtenir des interactions plus proches. La condition que nous introduisons ici est celle de Fresnel, bien connue en optique. Il s'agit de la première contribution de cette thèse sur le noyau oscillant. Nous montrons que sous réserve de satisfaire cette condition, nous pouvons alors exprimer le terme de second ordre comme le produit cartésien de deux opérateurs 1D quitte à considérer une base d'interaction adaptée.

Sous la condition de Fresnel, on peut "à la main" obtenir une première estimation du rang à l'aide d'un raisonnement sur le sinus cardinal et le critère de Shannon. La seconde contribution, sur le rang du noyau, est obtenue grâce aux résultats théoriques sur l'opérateur de Fox-Li $\mathscr{F}_c$. En effet, ces résultats améliorent l'estimation naïve du rang obtenue par le critère de Shannon. Pour ce faire, nous avons utilisé la formule asymptotique de Landau-Widom pour la détermination du rang du terme quadratique. Notre approche a permis de mettre en lumière l'importance de la largeur de bande : c'est le principal facteur intervenant dans la formule de Landau-Widom. Ainsi, nous pouvons obtenir des estimations asymptotiquement précises de la croissance du rang du noyau oscillant. Notons que l'on retrouve pour le critère de Fraunhofer des résultats sur le rang comparables à ceux de la littérature. Dans cette zone la largeur de bande est proche de l'unité et le terme quadratique possède un rang numérique proche de l'unité également.

Nous validons ces estimations sur plusieurs géométries canoniques. Le cas où $|d_{\|}| = 0$ fournit une asymptotique en $4/3$. Néanmoins, dans le cas d'un objet de la vie courante comme un fuselage d'avion, ce cas n'est pas fréquent et la croissance asymptotique ne s'observe que pour de grands objets. La plupart des cas vont se ramener à l'étude des deux sphères et ont une croissance asymptotique linéaire en la fréquence.
Pour une erreur relative ϵ, on note $\alpha = \epsilon^2$ et on obtient la majoration suivante sur le rang du noyau de Green déconvolué par les ondes planes $\mathrm{G}_{u_3}(x, y)$:

$$r_\epsilon(\mathrm{G}_{u_3}(x,y)) \leqslant \mathrm{C} \cdot \mathrm{N}_{\Pi_{u_3}},$$

avec

$$\mathrm{N}_{\Pi_{u_3}} = \begin{cases} \mathrm{N}(c_1,\alpha)\mathrm{N}(c_2,\alpha) & \text{si } c_1 \neq 0, c_2 \neq 0 \\ \mathrm{N}(c_1,\alpha) & \text{si } c_1 \neq 0, c_2 = 0 \\ \mathrm{N}(c_2,\alpha) & \text{si } c_2 \neq 0, c_1 = 0 \end{cases}$$

et $\mathrm{C} \lesssim 2$ dans la zone d'admissibilité de Fresnel.

Ces résultats montrent que dans la zone de Fresnel, l'emploi d'une méthode de type HCA-II est pleinement justifiée. Dans la suite, nous utilisons la largeur de bande pour déterminer le nombre de points d'interpolation à utiliser dans le cas d'une approche par HCA-II. *Le cas idéal où le terme de reste est nul* pourrait se traiter par une approche de type HCA-I. En effet, les fonctions d'onde sphéroïdales, qui sont les fonctions propres de l'opérateur $\mathscr{F}_c$ sont de bons candidats pour remplacer les polynômes de Chebyshev. Cependant, les fonctions propres peuvent tendre exponentiellement vers zéro sur les bords

et les méthodes numériques employées doivent impérativement en tenir compte. Le chapitre 4 montre l'emploi des résultats de ce chapitre sur des cas plus complexes comme la résolution d'un problème d'onde haute fréquence avec une formulation EFIE. Nous montrons qu'une modification légère de la méthode des $\mathscr{H}$-matrices pour tenir compte de ces résultats théoriques est possible.

3.8 Références

[BG92] E. L. Basor and I. Gohberg. *Toeplitz Operators and Related Topics, The Harold Widom Anniversary Volume*. Springer Basel AG, 1992. 131

[Fla57] C. Flammer. *Spheroidal wave functions*. Stanford University Press Stanford, 1957. 124, 126, 129, 136, 137

[LP61] H. J. Landau and H. O.Slepian Pollak. Prolate spheroidal wave functions, fourier analysis and uncertainty-ii. *The Bell System Technical Journal*, 1961. 129

[LP62] H. J. Landau and H. O.Slepian Pollak. Prolate spheroidal wave functions, fourier analysis and uncertainty-iii :the dimension of the space of essentially time- and band-limited signals. *The Bell System Technical Journal*, 1962. 129, 131

[LW80] H. J. Landau and H. Widom. Eigenvalue distribution of time and frequency limiting. *Journal of Mathematical Analysis and Applications*, 77 :469–481, 1980. 124, 128

[ORX13] A. Osipov, V. Rokhlin, and H. Xiao. *Prolate Spheroidal Wave Functions of Order Zero*. Springer, 2013. 127, 129, 133, 134, 136, 140

[PTVF92] W. H. Press, S. A. Teukolsky, W. T. Vetterling, and B. P. Flannery. *Numerical Recipes in C*. Cambridge University Press, 1992. 135

[Sle64] D. Slepian. Prolate spheroidal wave functions, fourier analysis and uncertainty-iv :extensions to many dimensions ; generalized prolate spheroidal functions. *The Bell System Technical Journal*, 1964. 131

[Sle65] D. Slepian. Some asymptotic expansions for prolate spheroidal wave functions. *Journal of Mathematics and Physics*, 44(1-4) :99–140, 1965. 131, 134

[Sle78] D. Slepian. Prolate spheroidal wave functions, fourier analysis and uncertainty-v :the discrete case. *The Bell System Technical Journal*, 1978. 131

[Sle83] D. Slepian. Some comments on fourier analysis, uncertainty and modeling. *SIAM Review*, 25(3) :379–393, 1983. 124, 130

[SP61] D. Slepian and H. O. Pollak. Prolate spheroidal wave functions, fourier analysis and uncertainty-i. *The Bell System Technical Journal*, 1961. 129, 130

[Var93] J. M. Varah. The prolate matrix. *Linear Algebra and its Applications*, 187 :269–278, 1993. 131

Chapitre 4

Applications aux $\mathscr{H}$- matrices

Sommaire

On illustre dans ce chapitre les résultats de l'approximation du noyau oscillant effectuée précédemment sur un cas de diffraction d'une onde électromagnétique. Signalons que ces résultats ne sont pas spécifiques à la formulation étudié dans ce chapitre ni à l'électromagnétisme, on pourrait traiter de la même façon d'autres cas issus de problèmes d'ondes.
Dans la suite, on choisira comme modèle l'équation intégrale du champ électrique dite *EFIE* qui résout le problème de diffraction d'une onde électromagnétique par un conducteur parfait. Les opérateurs intégraux mis en jeu s'expriment à partir du noyau de Green scalaire $G(x,y)$ que l'on approche. Il serait également possible de traiter d'autres conditions aux limites par les mêmes techniques que nous présenterons. En effet, celles-ci s'appuient sur l'approximation de rang faible du noyau de Green scalaire et s'adaptent aux différents opérateurs intervenant dans les problèmes d'ondes.

La première partie de ce chapitre décrit le contexte d'utilisation de nos approximations, la diffraction d'une onde électromagnétique. En particulier, on établira une équation intégrale au paragraphe 4.1.2 que l'on discrétisera et dont la matrice de discrétisation sera approchée par une $\mathscr{H}$-matrice.
La partie 4.2 sera consacrée aux modifications apportées à la méthode des $\mathscr{H}$-matrices pour traiter le cas du noyau oscillant à haute fréquence. Il y sera notamment question de la séparation des interactions proches et lointaines au paragraphe 4.2.1 ainsi qu'à l'assemblage compressé des interactions admissibles au paragraphe 4.2.3.
La dernière partie 4.3.3 de ce chapitre sera consacrée aux tests et validations numériques des modifications évoquées. Suivant le même plan, on effectuera les validations sur le critère d'admissibilité et la compression. Ainsi, le paragraphe 4.4 prouvera qu'il est nécessaire de travailler avec un critère fréquentiel dès lors que la section efficace mise en jeu est élevée. Cet exemple sera également mis en valeur au paragraphe 4.4.6.2 sur un exemple inspiré du test présenté dans [Mes11]. Le paragraphe 4.5 décrit l'utilisation pratique que l'on fera de la formule de Landau-Widom pour l'approximation du noyau oscillant.

4.1 Diffraction d'une onde électromagnétique

4.1.1 Équations de Maxwell – Représentation intégrale des champs

4.1.1.1 Description du problème

Une onde électromagnétique est caractérisée par les deux champs vectoriels électrique $\mathbf{E}$ et magnétique $\mathbf{H}$. On s'intéresse à la diffraction par un objet Ω d'une onde incidente notée $(\mathbf{E}^{\text{in}}, \mathbf{H}^{\text{in}})$. Cet objet Ω dont la frontière $\partial\Omega$ est noté Γ est supposé régulier et borné. La figure 4.1 décrit la position du problème. On note Ω' l'extérieur de l'objet soit $\Omega' = \mathbb{R}^3 \setminus \overline{\Omega}$.

On note $\vec{n}$ la normale sortante de Ω, ϵ_0 la permittivité électrique et μ_0 la perméabilité magnétique du milieu Ω' supposé homogène. L'objet est de plus supposé parfaitement conducteur (aussi appelé PEC pour *Perfectly Electrically Conducting*). On note $Z_0 = \sqrt{\frac{\mu_0}{\epsilon_0}}$ l'impédance du milieu. La célérité des ondes dans le milieu est donnée par $c_0 = \frac{1}{\epsilon_0\mu_0}$ et on note k le nombre d'onde donné par

$$k = \frac{\omega}{c_0} = \frac{2\pi}{\lambda}$$

où ω est la pulsation ou fréquence angulaire et λ la longueur d'onde.

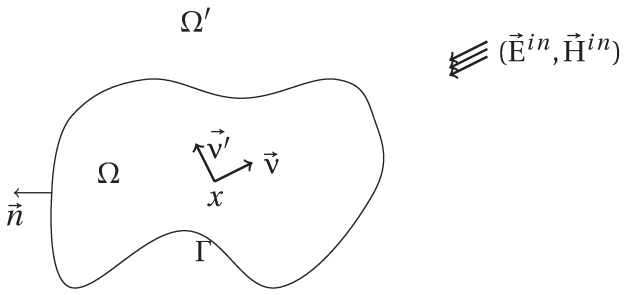


FIGURE 4.1 – Diffraction d'une onde électromagnétique par un obstacle.

4.1.1.2 Équations de Maxwell en domaine fréquentiel

On s'intéresse au problème harmonique et pour ce faire, on suppose une dépendance en temps en $e^{-i\omega t}$ des solutions temporelles des équations de Maxwell et l'on considère ici les équations de Maxwell en domaine fréquentiel obtenues par une transformée de Fourier des équations temporelles. Les champs $\mathbf{E}(\omega, x)$ et $\mathbf{H}(\omega, x)$ dépendent de la pulsation ω et du point x. Dans la suite, comme on travaille à fréquence fixée, on écrit de manière plus légère $\mathbf{E} = \mathbf{E}(x)$ et $\mathbf{H} = \mathbf{H}(x)$.
On résout dans Ω' le système des équations de Maxwell

$$\begin{cases} \overrightarrow{\mathrm{rot}}\,\mathbf{E} - i\omega\mu_0\mathbf{H} = & 0, \\ \overrightarrow{\mathrm{rot}}\,\mathbf{H} + i\omega\epsilon_0\mathbf{E} = & 0, \end{cases} \tag{4.1}$$

On considère un champ émectromagnétique incident de type onde plane :

$$\mathbf{E}^{in} = \mathbf{E}_0^{in} e^{-ik\vec{\nu}\cdot\vec{x}} \tag{4.2}$$

$$\mathbf{H}^{in} = \mathbf{H}_0^{in} e^{-ik\vec{\nu}\cdot\vec{x}} \tag{4.3}$$

où $\mathbf{E}^{in}$ et $\mathbf{E}^{in}$ sont les amplitudes des champs (vecteurs complexes constants) et $-\vec{\nu}$ est un vecteur unitaire représentant la direction de propagation de l'onde plane ($\vec{\nu}$ est la direction d'où vient l'onde vu de l'objet). Ces vecteurs vérifient les relations :

$$\mathbf{H}_0^{in} = \frac{\mathbf{E}_0^{in}}{Z_0} \wedge \vec{\nu} \tag{4.4}$$

$$\mathbf{E}_0^{in} = -Z_0\mathbf{H}_0^{in} \wedge \vec{\nu} \tag{4.5}$$

de sorte que le champ incident $(\mathbf{E}^{\mathrm{in}}(x), \mathbf{H}^{\mathrm{in}}(x))$ vérifie les équations de Maxwell homogènes dans $\mathbb{R}^3$.
La condition aux limites sur un conducteur parfait s'écrit

$$\mathbf{E} \wedge \overrightarrow{\mathbf{n}} = 0 \qquad \text{sur } \Gamma. \tag{4.6}$$

On appelle $(\mathbf{E}, \mathbf{H})$ le champ électromagnétique total et $(\mathbf{E}^{\mathrm{diff}}, \mathbf{H}^{\mathrm{diff}})$ le champ électromagnétique diffracté défini par :

$$\mathbf{E}^{\mathrm{diff}} = \mathbf{E} - \mathbf{E}^{\mathrm{in}},$$
$$\mathbf{H}^{\mathrm{diff}} = \mathbf{H} - \mathbf{H}^{\mathrm{in}}.$$

Pour fermer le problème, il est nécessaire d'ajouter aux équations précédentes une « condition aux limites à l'infini ». C'est la condition de radiation dite de Silver-Müller (ou condition d'ondes sortantes) :

$$\lim_{|x|\to\infty} |x|.\left(\mathbf{E}^{\text{diff}} - Z_0\,\mathbf{H}^{\text{diff}} \wedge \frac{x}{|x|}\right) = 0. \tag{4.7}$$

Il s'agit de l'équivalent fréquentiel de la condition de causalité dans le domaine temporel.

4.1.1.3 Représentations intégrales des champs diffractés

En prolongeant le champ diffracté dans le domaine Ω par une solution quelconque des équations de Maxwell homogènes, on obtient grâce au théorème de représentation intégrale une expression de ce champ en fonction des potentiels vecteurs $\vec{A}_E$ et $\vec{A}_M$ et scalaires Φ_E et Φ_M :

$$\mathbf{E}^{\text{diff}} = -Z_0\nabla\Phi_E + ikZ_0\vec{A}_E - Z_0\overrightarrow{\text{rot}}\vec{A}_M, \tag{4.8}$$

$$\mathbf{H}^{\text{diff}} = -Z_0\nabla\Phi_M + ikZ_0\vec{A}_M + Z_0\overrightarrow{\text{rot}}\vec{A}_E, \tag{4.9}$$

Ces potentiels sont donnés par les formules suivantes :

$$\begin{cases} \phi_E(x) = \dfrac{1}{ik}\displaystyle\int_\Gamma G(x,y)\,\text{div}_\Gamma\,\overrightarrow{\mathbf{j}}\,(y)\,d\Gamma(y), \\ \phi_M(x) = \dfrac{1}{ik}\displaystyle\int_\Gamma G(x,y)\,\text{div}_\Gamma\overrightarrow{\mathbf{m}}(y)\,d\Gamma(y), \\ \vec{A}_E(x) = \displaystyle\int_\Gamma G(x,y)\,\overrightarrow{\mathbf{j}}\,(y)\,d\Gamma(y), \\ \vec{A}_M(x) = \displaystyle\int_\Gamma G(x,y)\,\overrightarrow{\mathbf{m}}(y)\,d\Gamma(y), \\ G(x,y) = \dfrac{e^{ik|x-y|}}{4\pi|x-y|}, \end{cases} \tag{4.10}$$

où $G(x,y)$ est la fonction de Green de l'équation de Helmholtz satisfaisant la condition de radiation à l'infini :

$$G(x,y) = \frac{e^{ik|x-y|}}{4\pi|x-y|},$$

et $\overrightarrow{\mathbf{j}}$ et $\overrightarrow{\mathbf{m}}$ respectivement les courants électrique et magnétique équivalents donnés par les sauts suivants à travers Γ :

$$\overrightarrow{\mathbf{j}} = [\mathbf{H}^{\text{diff}} \wedge \overrightarrow{\mathbf{n}}] = \mathbf{H}^{\text{diff}}|_\Omega \wedge \overrightarrow{\mathbf{n}} - \mathbf{H}^{\text{diff}}|_{\Omega'} \wedge \overrightarrow{\mathbf{n}},$$
$$\overrightarrow{\mathbf{m}} = [\overrightarrow{\mathbf{n}} \wedge \mathbf{E}^{\text{diff}}] = \overrightarrow{\mathbf{n}} \wedge \mathbf{E}^{\text{diff}}|_\Omega - \overrightarrow{\mathbf{n}} \wedge \mathbf{E}^{\text{diff}}|_{\Omega'}.$$

où dans ces formules la notation $u_{|\Omega}$ (resp. $u_{|\Omega'}$) désigne la trace de u sur Γ venant de Ω (resp. de Ω').
Si l'on prolonge le champ diffracté dans Ω par

$$\mathbf{E}^{\text{diff}} = -\mathbf{E}^{\text{in}},$$
$$\mathbf{H}^{\text{diff}} = -\mathbf{H}^{\text{in}},$$

on obtient en utilisant la continuité du champ incident à travers Γ :

$$\vec{\mathbf{j}} = -\mathbf{H}^{\text{in}}|_{\Omega} \wedge \vec{\mathbf{n}} - \mathbf{H}^{\text{diff}}|_{\Omega'} \wedge \vec{\mathbf{n}} = -\mathbf{H}|_{\Omega'} \wedge \vec{\mathbf{n}},$$
$$\vec{\mathbf{m}} = \vec{\mathbf{n}} \wedge \mathbf{E}^{\text{in}}|_{\Omega} - \vec{\mathbf{n}} \wedge \mathbf{E}^{\text{diff}}|_{\Omega'} = -\vec{\mathbf{n}} \wedge \mathbf{E}|_{\Omega'} = \vec{0}.$$

Les potentiels $\vec{\mathrm{A}}_{\mathrm{M}}$ et Φ_{M} sont donc nuls et l'expression du champ diffracté se simplifie en :

$$\mathbf{E}^{\text{diff}} = -\mathrm{Z}_0 \nabla \Phi_{\mathrm{E}} + ik\mathrm{Z}_0 \vec{\mathrm{A}}_{\mathrm{E}}, \tag{4.11}$$
$$\mathbf{H}^{\text{diff}} = \mathrm{Z}_0 \overrightarrow{\mathrm{rot}} \vec{\mathrm{A}}_{\mathrm{E}}. \tag{4.12}$$

4.1.2 Formulation EFIE – Discrétisation par éléments finis

En traduisant les conditions aux limites, on trouve une équation intégrale à résoudre. Plusieurs choix sont possibles. Nous utilisons la représentation du champ électrique ce qui aboutit à ce qu'on appelle l'équation intégrale du champ électrique ou EFIE pour l'acronyme anglais de *Electric-Field Integral Equation*.

La condition aux limites sur le champ diffracté est donnée par :

$$\mathbf{E}^{\text{diff}}(x) \wedge \vec{\mathbf{n}} = -\mathbf{E}^{\text{in}}(x) \wedge \vec{\mathbf{n}} \qquad \text{sur } \Gamma. \tag{4.13}$$

Elle s'écrit aussi

$$\Pi_x \mathbf{E}^{\text{diff}}(x) = -\Pi_x \mathbf{E}^{\text{in}}(x) \qquad \text{sur } \Gamma, \tag{4.14}$$

où Π_x est la projection orthogonale sur le plan tangent à Γ au point x définie par :

$$\Pi_x \vec{v}(x) = \vec{v}(x) - \left(\vec{v}(x) \cdot \vec{\mathbf{n}}(x)\right) \vec{v}(x),$$

pour un champ vectoriel $\vec{v}$ quelconque. Notons que pour un champ scalaire u régulier près de Γ, le gradient de ce champ se décompose en une partie tangentielle et une partie normale de la sorte

$$\nabla u = \nabla_\Gamma u + \frac{\partial u}{\partial n} \vec{\mathbf{n}},$$

$\nabla_\Gamma u$ est le gradient tangentiel de u :

$$\Pi_x \nabla u = \nabla_\Gamma u.$$

L'équation intégrale du champ électrique s'obtient en injectant la représentation (4.11) dans (4.14) :

$$ik\mathrm{Z}_0 \left(\Pi_x \int_\Gamma \mathrm{G}(x,y) \vec{\mathbf{j}}(y)\, \mathrm{d}\Gamma(y) + \frac{1}{k^2} \nabla_\Gamma \int_\Gamma \mathrm{G}(x,y) \mathrm{div}_\Gamma \vec{\mathbf{j}}(y)\, \mathrm{d}\Gamma(y) \right) = -\Pi_x \mathbf{E}^{\text{in}}. \tag{4.15}$$

On peut calculer les champs $\mathbf{E}^{\text{diff}}$ et $\mathbf{H}^{\text{diff}}$ en tout point de l'espace à l'aide des formules de représentation (4.11) et (4.12).

Remarque 4.1 (Autres formulations). Il existe d'autres formulations amenant à d'autres équations intégrales, notamment celle décrivant le champ magnétique $\mathbf{H}$ désignée par « MFIE » dans la littérature (*Magnetic Field Integral Equation*). De plus, on peut considérer la formulation combinée, dite CFIE, définie par la combinaison convexe

$$\mathrm{CFIE} = \alpha \mathrm{EFIE} + (1-\alpha) \frac{i}{k} \mathrm{MFIE}.$$

Ces formulations ne sont néanmoins pas développées dans la suite de ce manuscrit mais pourraient se traiter de la même façon.

4.1.2.1 Discrétisation de l'EFIE

Pour une analyse rigoureuse de l'EFIE, nous renvoyons le lecteur à la thèse d'état de A. Bendali [Ben84]. Dans ce paragraphe, nous rappelons succintement les étapes permettant d'aboutir à la formulation du problème discret en vue d'une résolution par la méthode des $\mathscr{H}$-matrices.

Formulation variationnelle Pour introduire le cadre fonctionnel, on définit d'abord l'espace

$$\mathrm{TH}^{-\frac{1}{2}}(\Gamma) = \left\{ \vec{\mathbf{j}} \in \left(\mathrm{H}^{-\frac{1}{2}}(\Gamma)\right)^3 : \vec{\mathbf{j}} \cdot \vec{\mathbf{n}} = 0 \right\}. \tag{4.16}$$

Le courant électrique $\vec{\mathbf{j}}$ appartient à l'espace $\mathrm{H}^{-\frac{1}{2}}(\mathrm{div}, \Gamma)$ suivant

$$\mathrm{H}^{-\frac{1}{2}}(\mathrm{div}, \Gamma) = \left\{ \vec{\mathbf{j}} \in \mathrm{TH}^{-\frac{1}{2}}(\Gamma), \mathrm{div}_\Gamma \vec{\mathbf{j}} \in \mathrm{H}^{-\frac{1}{2}}(\Gamma) \right\}. \tag{4.17}$$

Le problème variationnel associé à (4.15) est le suivant,

$$\begin{cases} \text{Trouver } \vec{\mathbf{j}}\,(x) \in \mathrm{V} = \mathrm{H}^{-\frac{1}{2}}(\mathrm{div}, \Gamma) \text{ tq :} \\ \\ ikZ_0 \displaystyle\int_\Gamma \int_\Gamma \mathrm{G}(x,y)\, \vec{\mathbf{j}}\,(y)\, \vec{\mathbf{j}}^{\,t}(x)\, \mathrm{d}\Gamma(y)\, \mathrm{d}\Gamma(x) \\ +ikZ_0 . \dfrac{1}{k^2} \displaystyle\int_\Gamma \int_\Gamma \mathrm{G}(x,y)\, \mathrm{div}_\Gamma \vec{\mathbf{j}}\,(y) \mathrm{div}_\Gamma \vec{\mathbf{j}}^{\,t}(x)\, \mathrm{d}\Gamma(y)\, \mathrm{d}\Gamma(x) \quad = -\displaystyle\int_\Gamma \mathbf{E}^{in} \vec{\mathbf{j}}^{\,t}(x)\, \mathrm{d}\Gamma(x) \\ \forall\, \vec{\mathbf{j}}^{\,t}(x) \in \mathrm{V} \end{cases} \tag{4.18}$$

Approximation variationnelle On approche la solution de ce problème à l'aide d'une méthode d'éléments finis. Il s'agit donc de remplacer dans (4.18) l'espace continu V par un espace de dimension finie V_h ayant de bonnes propriétés. Pour ce faire on approche la surface Γ de l'objet par une triangulation Γ_h où h représente typiquement la finesse de la triangulation (ex : la longueur d'une arête). On travaille avec l'espace des éléments finis de Raviart-Thomas [RT77] : les degrés de liberté sont les flux à travers les arêtes.

Pour h fixé, l'espace vectoriel V_h ainsi obtenu a généralement pour dimension N_h le nombre d'arêtes de la triangulation (hors bords libres, arêtes multiples, ...). Dans le cas où Γ n'est pas un polyhèdre, $\Gamma_h \neq \Gamma$ et V_h n'est pas inclus dans V : approximation non conforme.

Le problème variationnel discret s'écrit :

$$\begin{cases} \text{Trouver un courant } \vec{\mathbf{j}}_h(x) \in \mathrm{V}_h \text{ tq :} \\ \\ ikZ_0 \displaystyle\int_{\Gamma_h} \int_{\Gamma_h} \mathrm{G}(x,y)\, \vec{\mathbf{j}}_h(y)\, \vec{\mathbf{j}}_h^{\,t}(x)\, \mathrm{d}\Gamma_h(y)\, \mathrm{d}\Gamma_h(x) \\ +ikZ_0 . \frac{1}{k^2} \displaystyle\int_{\Gamma_h} \int_{\Gamma_h} \mathrm{G}(x,y) \mathrm{div}_{\Gamma_h} \vec{\mathbf{j}}_h(y) \mathrm{div}_{\Gamma_h} \vec{\mathbf{j}}_h^{\,t}(x)\, \mathrm{d}\Gamma_h(y)\, \mathrm{d}\Gamma_h(x) \quad = -\displaystyle\int_{\Gamma_h} \mathbf{E}^{\mathrm{in}} \vec{\mathbf{j}}_h^{\,t}(x)\, \mathrm{d}\Gamma_h(x) \\ \forall\, \vec{\mathbf{j}}_h^{\,t}(x) \in \mathrm{V}_h. \end{cases} \tag{4.19}$$

Notons $(\vec{\mathbf{w}}_i)_{1\le i\le N_h}$ une base de V_h. Le courant discret j_h s'écrit alors

$$\vec{\mathbf{j}}_h(x) = \sum_{j=1}^{N_h} \lambda_j \vec{\mathbf{w}}_j(x), \tag{4.20}$$

où les coefficients $\{\lambda_j\}_{1\le i\le N_h}$ sont les degrés de liberté.

Éléments finis de Raviart-Thomas Les fonctions de base des éléments de Raviart-Thomas sont associés aux arêtes du maillage. La fonction de base $\vec{\mathbf{w}}_i$ associée à l'arête A_i possède un flux nul à travers les autres arêtes et un flux unitaire à travers A_i ce qui suppose d'avoir choisi arbitrairement une orientation. Nous supposons pour simplifier que A_i est uniquement partagée par deux triangles notés T_{i_1} et T_{i_2}. Le support de $\vec{\mathbf{w}}_i$ est $T_{i_1} \cup T_{i_2}$. On choisit par exemple d'orienter le flux : le flux de $\vec{\mathbf{w}}_i$ sortant de T_{i_1} vaut 1 tandis que celui sortant de T_{i_2} vaut -1.
On note S_{i_1} et S_{i_2} les sommets opposés à l'arête et appartenant respectivement aux triangles T_{i_1} et T_{i_2}.

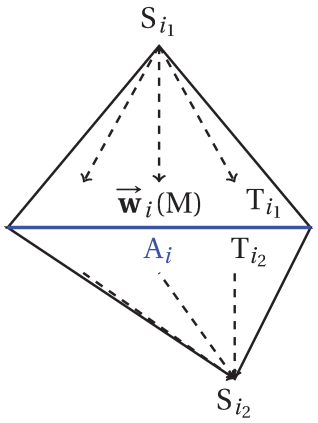


FIGURE 4.2 – Représentation des éléments finis de Raviart-Thomas.

Pour un point $M \in T_{i_1}$, $\vec{\mathbf{w}}_i(M)$ est donnée par :

$$\vec{\mathbf{w}}_i(x) = \frac{\overrightarrow{S_{i_1}M}}{2|T_{i_1}|}, \tag{4.21}$$

où $|T|$ désigne l'aire du triangle T. De même, pour $M \in T_{i_2}$, $\vec{\mathbf{w}}_i(M)$ on a :

$$\vec{\mathbf{w}}_i(x) = -\frac{\overrightarrow{S_{i_2}M}}{2|T_{i_2}|}. \tag{4.22}$$

Par ailleurs, la divergence surfacique de la fonction de base $\vec{w}_i$ est constante par triangle :

$$\mathrm{div}_\Gamma(\vec{w}_i(M)) = \begin{cases} \dfrac{1}{|T_{i_1}|} & M \in T_{i_1} \\ \dfrac{-1}{|T_{i_2}|} & M \in T_{i_2} \\ 0 & \text{ailleurs.} \end{cases} \tag{4.23}$$

Écriture matricielle On rappelle que l'on cherche une approximation du courant $\vec{\mathbf{j}}$ sous la forme

$$\vec{\mathbf{j}}_h = \sum_{j=1}^{\mathrm{N}_h} \lambda_j \vec{\mathbf{w}}_j, \tag{4.24}$$

où les $\{\lambda_j\}_{j=1}^{\mathrm{N}_h}$ sont les inconnues du problème variationnel discret et les $\{\vec{\mathbf{w}}_j\}_{j=1}^{\mathrm{N}}$ sont les fonctions de base décrites précédemment. Comme les $\vec{\mathbf{w}}_i$ forment une base de V_h, il suffit de tester par les fonctions de base et la discrétisation de la formulation variationnelle conduit ainsi à trouver les coefficients $\{\lambda_j\}_{j=1}^{\mathrm{N}_h}$ satisfaisant pour tout $i \in \{1,\dots,\mathrm{N}_h\}$,

$$\sum_{j=1}^{\mathrm{N}_h} \lambda_j \left(\int_{\Gamma_h}\int_{\Gamma_h} \mathrm{G}(x,y) \left(\vec{\mathbf{w}}_i^t(x)\cdot\vec{\mathbf{w}}_j(y) - \frac{1}{k^2}\mathrm{div}_{\Gamma_h}\vec{\mathbf{w}}_i^t(x)\,\mathrm{div}_{\Gamma_h}\vec{\mathbf{w}}_j(y)\right) \mathrm{d}\Gamma_h(x)\,\mathrm{d}\Gamma_h(y)\right)$$
$$= -\frac{1}{ikZ_0}\int_{\Gamma_h} \mathbf{E}^{\mathrm{in}}(x)\vec{\mathbf{w}}_i^t(x)\,\mathrm{d}\Gamma_h(x). \tag{4.25}$$

Matriciellement, cela correspond à résoudre le système d'équations

$$\mathrm{A}^h \Lambda^h = b^h, \tag{4.26}$$

avec pour $i \in \{1,\dots,\mathrm{N}_h\}$ et $j \in \{1,\dots,\mathrm{N}_h\}$

$$\mathrm{A}^h_{ij} = \int_{\Gamma_h}\int_{\Gamma_h} \mathrm{G}(x,y)\left(\vec{\mathbf{w}}_i^t(x)\cdot\vec{\mathbf{w}}_j(y) - \frac{1}{k^2}\mathrm{div}_{\Gamma_h}\vec{\mathbf{w}}_i^t(x)\,\mathrm{div}_{\Gamma_h}\vec{\mathbf{w}}_j(y)\right)\mathrm{d}\Gamma_h(x)\,\mathrm{d}\Gamma_h(y), \tag{4.27}$$

$$b_i^h = -\frac{1}{ikZ_0}\int_{\Gamma} \mathbf{E}^{\mathrm{in}}(x)\vec{\mathbf{w}}_i(x)\,\mathrm{d}\Gamma_h(x), \tag{4.28}$$

$$\Lambda^h = (\lambda_1,\dots,\lambda_{\mathrm{N}_h})^{\mathrm{T}}. \tag{4.29}$$

4.1.3 Commentaires

Pour le problème de la diffraction d'une onde plane par un obstacle (4.1), nous avons d'abord représenté le champ diffracté sous une forme intégrale (4.11). La condition aux limites conduit à une équation intégrale (4.15) sur la surface Γ de l'obstacle. La résolution de celle-ci permet de retrouver les courants équivalents sur l'obstacle. Les formules de représentation permettent ensuite de retrouver les champs en tout point de l'espace. L'approximation numérique de l'équation intégrale à l'aide d'une méthode d'éléments finis de frontière conduit après discrétisation à la résolution d'un système linéaire de taille $\mathrm{N}_h \times \mathrm{N}_h$.

La matrice A^h du système linéaire est complexe, symétrique et pleine. En effet, les opérateurs intégraux mis en jeu ne sont pas locaux ce qui rend les coeffcients généralement non nuls. Dans la pratique, on considère cette matrice de discrétisation pour des tailles allant de quelques dizaines de milliers d'inconnues à plusieurs millions d'inconnues. Des matrices pleines de cette taille sont évidemment délicates à manipuler dans la pratique. En effet, le temps d'assemblage et l'espace mémoire nécessaire pour une telle matrice augmentent comme N_h^2. Le temps de résolution du système linéaire par une méthode directe augment lui comme N_h^3. Une méthode itérative aurait un coût en $\mathrm{N}_h^2 \times \mathrm{N}_{\mathrm{iter}}$ où le nombre d'itérations $\mathrm{N}_{\mathrm{iter}}$ dépend à la fois de la géométrie, du maillage, de la fréquence ainsi que du second membre (sans compter le temps d'un éventuel préconditionneur).

D'où une forte activité dans les communautés académiques et industrielles autour du développement de méthodes d'assemblage et de résolution rapides pour de tels systèmes.

Dans la suite, on s'intéressera donc à une approximation favorable de la matrice A^h. Plus particulièrement, on l'approchera par une $\mathscr{H}$-matrice. On effectuera la même approximation pour sa factorisée. Les points délicats de la méthode des $\mathscr{H}$-matrices à prendre en compte sont d'une part le critère d'admissibilité utilisé pour séparer les interactions proches et lointaines et d'autre part l'algorithme d'assemblage compressé utilisé pour construire les approximations de rang faible. Les méthodes présentées au premier chapitre de ce manuscrit ne prennent pas en compte le paramètre fréquence ce qui peut avoir des conséquences négatives sur la précision et le coût de calcul. Nous proposons une méthode permettant d'une part de contrôler :

— le niveau d'erreur demandé quelle que soit la fréquence ;

— l'évolution du taux de compression et du temps de calcul en fonction de la la fréquence.

Cette méthode est une adaptation fréquentielle de l'algorithme HCA – II. Elle est basée sur les éléments suivants :

un critère d'admissibilité dépendant de la fréquence qui garantit des interactions dans la zone de Fresnel ;

une approximation directionnelle du noyau de Green qui permet de factoriser les oscillations dans une direction de propagation privilégiée ;

une estimation asymptotique du rang basée sur des résultats de la littérature sur les propriétés des filtres à bande limitée (formule de Landau-Widom) ;

un choix automatique et optimal du nombre de nœuds d'interpolation de l'algorithme HCA – II appliqué au noyau modifié.

4.2 Approximation $\mathscr{H}$- matrice fréquentielle de l'EFIE

On décrit dans cette partie les modifications apportées à la méthode d'assemblage d'une $\mathscr{H}$-matrice permettant l'approximation de l'opérateur de l'EFIE à partir des résultats précédents. Le développement limité de la phase à l'ordre 5 permet d'écrire une nouvelle condition d'admissibilité fréquentielle. La condition d'admissibilité de Fresnel est moins restrictive que celle de Fraunhofer et « rapproche » les interactions lointaines. Le rang numérique à une précision donnée des blocs admissibles est plus élevé que dans le cas d'une admissibilité du type Fraunhofer mais la formule de Landau-Widom permet d'obtenir une bonne estimation de ce rang pour chaque composante de l'EFIE. On montre que ce rang croît au plus linéairement avec la fréquence et cela permet d'adapter l'algorithme HCA – II présenté dans [BG05] au cas fréquentiel.

4.2.1 Construction d'une base orientée

4.2.1.1 Position du problème

On rappelle que dans toute la suite et sauf mention contraire, les degrés de liberté mentionnés sont ceux de l'EFIE *i.e* localisés aux centres des arêtes de la triangulation Γ_h. Les éléments du support d'un degré de liberté sont des triangles et les matrices considérées sont des blocs matriciels issus de la discrétisation de l'EFIE par la méthode des

éléments finis de frontières. Le critère d'admissibilité permet la séparation des interactions entre les interactions proches et lointaines.
On considère deux nuages distincts de degrés de liberté $\left\{x_{i_p}\right\}_{p=1}^{m}$ et $\left\{y_{j_q}\right\}_{q=1}^{n}$ identifiés de façon unique par les indices i_p et j_q. Conformément à l'usage, t et s désignent des ensembles d'indices consécutifs dans la numérotation $\mathscr{H}$-matrice définis par

$$s = \{i_1, \ldots, i_m\} \tag{4.30}$$
$$t = \{j_1, \ldots, j_n\} \tag{4.31}$$

Remarque 4.2 (Numérotation). Les indices originaux i_p et j_q dépendent de la façon dont on a construit la triangulation Γ_h et ne sont pas nécessairement consécutifs. La renumérotation locale présente dans la méthode des $\mathscr{H}$-matrices permet d'utiliser des indices contigüs permettant de manipuler aisément des sous-blocs matriciels.

On note X_s et Y_t les groupes d'inconnues de taille respective m et n définis par

$$\mathrm{X}_s = \left\{x_{i_k}, k = 1, \ldots, m\right\} \tag{4.32}$$
$$\mathrm{Y}_t = \left\{y_{j_k}, k = 1, \ldots, n\right\} \tag{4.33}$$

Enfin, on note $\mathrm{A}_{s\times t}$ la matrice d'interaction entre ces deux groupes. La figure 4.3 illustre les notations précédentes.

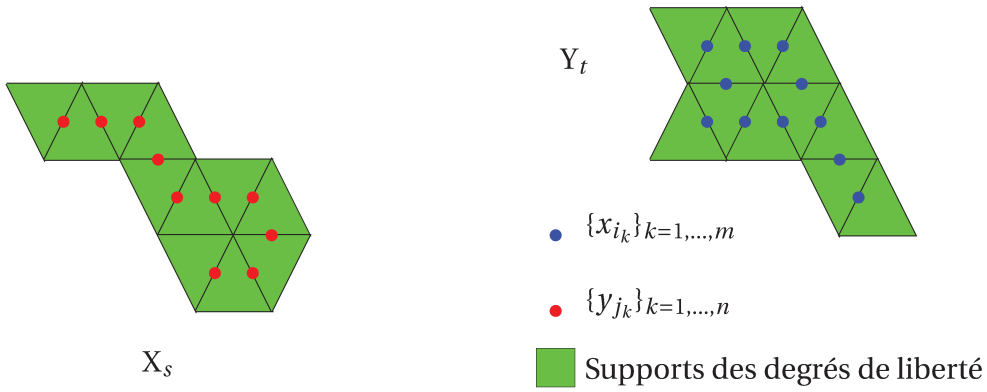


FIGURE 4.3 – Interaction de deux groupes de degrés de liberté X_s et Y_t. Les supports des degrés de libertés (ici des triangles) sont représentés en vert.

4.2.1.2 Boîtes englobantes et supports des degrés de liberté

Plutôt que de travailler sur l'intégralité des nuages de degrés de liberté, la méthode des $\mathscr{H}$-matrices, comme d'autres méthodes rapides, travaille sur des boîtes englobantes. Ceci permet le calcul rapide de quantités géométriques comme les distances et les diamètres.
Comme l'illustre la figure 4.4, on rappelle que la boîte englobante des degrés de liberté contenus dans X_s (respectivement Y_t) est noté B_s (resp. B_t) tandis que la boîte englobante du support de X_s est quant à elle notée Q_s (resp. Q_s). En l'absence de précision supplémentaire, les boîtes sont données dans la base canonique.

Dans toute la suite, sauf mention contraire, on considère la boîte englobante contenant le support des degrés de liberté. Ce choix *a priori* plus conservateur est motivé par l'emploi d'un schéma d'interpolation polynomiale. Avec ce choix, on s'assure que l'on place des nœuds d'interpolation sur l'intégralité du support. On rappelle également que

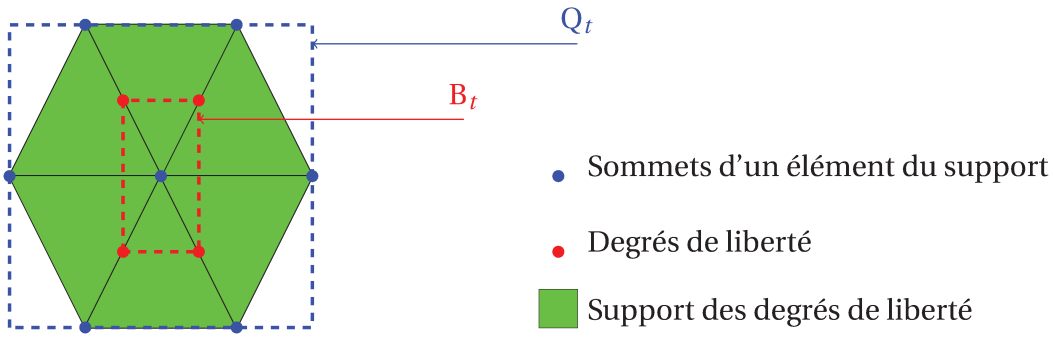


FIGURE 4.4 – Boîtes englobantes B et Q pour des degrés de liberté localisés aux centres des arêtes.

dans le cas de la condition d'admissibilité statique, si les boites Q_t et Q_s satisfont la condition, alors cette condition est également réalisée pour B_t et B_s.

La suite de cette partie fournit une description détaillée de l'implémentation du critère d'admissibilité de Fresnel développé au chapitre précédent. Notons que le critère d'admissibilité de Fraunhofer utilisé dans la littérature nécessite le même travail et peut être obtenu « en passant ». On reprend les mêmes notations qu'au chapitre précédent. (e_1, e_2, e_3) désigne la base canonique de $\mathbb{R}^3$ et (u_1, u_2, u_3) désigne une base orientée de $\mathbb{R}^3$ que l'on se propose de construire.

4.2.1.3 Calcul de la base orientée (u_1, u_2, u_3)

Choix de la direction u_3 La direction u_3 est choisie comme étant la direction entre les centres de gravité des nuages de degrés de liberté X_s et Y_t. Dans la pratique, on a également pris pour u_3 la direction liant les centres géométriques des boîtes englobantes B_s et B_t sans que cela ne change les effets. Le choix de la direction est relativement souple ce qui s'avère être un plus pour l'implémentation. C'est la construction de la base (u_1, u_2) qui représente le point important. Une fois choisie la direction u_3, on considère les projetés orthogonaux des degrés de libertés des groupes X_s et Y_t selon u_3. On note Π_{u_3} le plan normal à u_3. Les degrés de libertés $\left\{x_{i_p}\right\}_{p=1}^{m}$ et $\left\{y_{j_q}\right\}_{q=1}^{n}$ sont projetés dans le plan Π_{u_3} (*cf* 4.5) par la projection suivante,

$$\begin{aligned} P : \mathbb{R}^3 &\mapsto \Pi_{u_3} \\ P(x) &= x - < x, u_3 > u_3. \end{aligned} \tag{4.34}$$

Remarque 4.3 (Projection des boites englobantes). Une piste d'amélioration serait de ne projeter que les boîtes englobantes Q_s et Q_t dans le plan transverse. Ceci aboutit au calcul d'une base orientée en un nombre constant d'opérations. Cependant, on peut augmenter ainsi la section efficace car la base (u_1, u_2) du plan Π_{u_3} est déterminée seulement à partir des projections des extrêmités des boîtes. Par exemple, dans le cas d'un *clustering* par composantes principales, on dispose pour chaque groupe d'inconnues X_s et Y_t d'une boîte englobante orientée. Les projections de ces boîtes fournissent une base (u_1, u_2) dans laquelle la section efficace est plus petite.

Calcul de la base orientée (u_1, u_2) Afin d'utiliser la formule de Landau-Widom, on souhaite déterminer une base (u_1, u_2) de Π_{u_3} de section efficace la plus petite possible. Déterminer une telle base de Π_{u_3} revient à déterminer la boîte englobante d'aire la plus petite

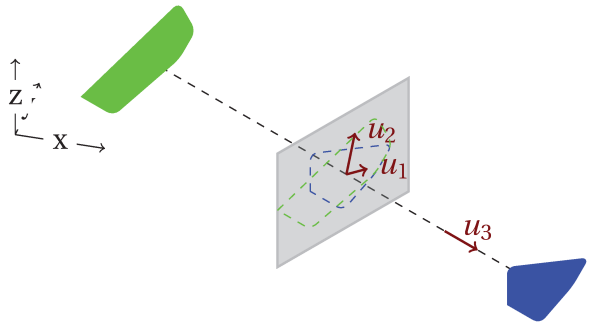


FIGURE 4.5 – En pointillé, la projection des groupes sur le plan transverse Π_{u_3} représenté en gris. La base (u_1, u_2, u_3) est la base liée à l'interaction des deux groupes X_s et Y_t représentés en bleu et en vert.

possible contenant les projections des inconnues. Plusieurs méthodes existent et nous choisissons une méthode rapide et efficace dans la pratique.

Analyse en composantes principales Les sommets des éléments du support de chaque groupe d'inconnues forment un nuage de points dans le plan Π_{u_3}. Une solution acceptable pour déterminer une base orientée est d'utiliser l'analyse en composantes pricipales sur le nuage de points projetés dans le plan transverse Π_{u_3}. On choisit u_1 comme étant la direction principale du nuage et on forme le vecteur restant u_2 par $u_2 = u_3 \wedge u_1$. Cependant, cette technique n'est pas la meilleure. En effet, selon la finesse du maillage, pour des plaques carrées en opposition, la direction principale choisie peut être selon la diagonale. Ceci double la section efficace.

Technique du « pied à coulisse » En géométrie algorithmique, la méthode des *rotating callipers* permet la détermination rapide d'une boîte englobante en dimension 2. Cette méthode consiste à déterminer l'enveloppe convexe des points (par exemple avec le parcours de Graham) puis à déterminer la boite englobante d'aire minimale contenant l'enveloppe. Pour ce faire, on imagine un pied à coullisse dont l'une des parties est tangente à une arête de l'enveloppe convexe. On itère pour chaque arête de l'enveloppe et on sélectionne l'orientation donnant l'aire minimale.

Une méthode heuristique Dans la pratique, la méthode heuristique suivante est suffisante pour obtenir de bons résultats et c'est la méthode que nous privilégions. Elle s'apparente à la méthode du pied à coulisse décrit auparavant. On note $(u_1^{(0)}, u_2^{(0)})$ une base quelconque du plan Π_{u_3} dans laquelle on détermine la boîte englobante minimale. On se donne n angles $\theta_k \in]0, \frac{\pi}{2}[$ et pour chaque rotation d'angle θ_k, on définit la base $(u_1^{(k)}, u_2^{(k)})$ comme la rotation d'angle θ_k de la base originale $(u_1^{(0)}, u_2^{(0)})$. La figure 4.6 illustre cette construction pour un angle $\theta_k = \frac{\pi}{4}$.

À chaque itération n, on détermine la boîte englobante des projections dans la base $\left(u_1^{(n)}, u_2^{(n)}\right)$ et on calcule l'aire de cette boite. On retient comme base acceptable (u_1, u_2) celle dont l'aire calculée est la plus petite. Dans la pratique, nous utilisons N = 10 angles. On retrouve parfois dans la littérature la dénomination de *bounding diamond* dans le cas où l'on utilise que l'angle $\frac{\pi}{4}$.

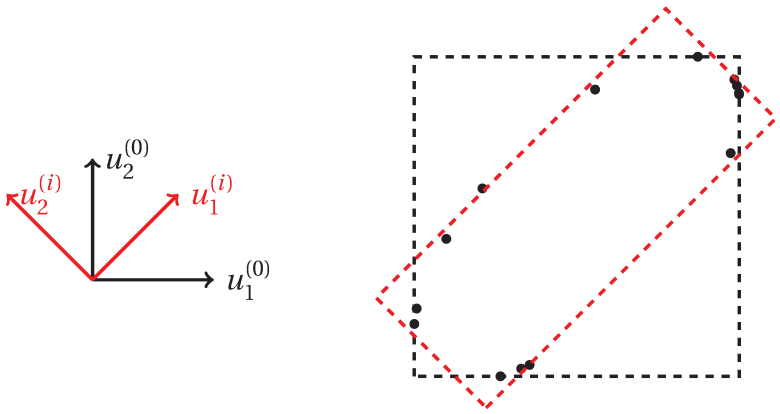


FIGURE 4.6 – Construction de la base (u_1, u_2) du plan transverse Π_{u_3}. Exemple pour $\theta_i = \frac{\pi}{4}$,l'aire de la boite rouge est bien inférieure à celle en noir.

4.2.2 Admissibilité fréquentielle de Fresnel

On conserve les notations du chapitre précédent, et on rappelle que sous la condition nécessaire,

$$\frac{\left|\overrightarrow{\mathbf{d}}\right|}{\mathrm{R}} < 1 \qquad \text{(condition statique)} \tag{4.35}$$

le développment limité de la phase à l'ordre 5 effectué au chapitre précédent permet d'obtenir le critère d'admissibilité fréquentiel de Fresnel (en supposant que $|d_{/\!/}| \neq 0$) suivant

$$k\frac{|\overrightarrow{\mathbf{d}}_{/\!/}||\overrightarrow{\mathbf{d}}_{\perp}|^2}{2\mathrm{R}^2} \leqslant \nu 2\pi \qquad \text{(Critère de Fresnel),} \tag{4.36}$$

ν étant un petit paramètre fixé de l'ordre de l'unité contrôlant le nombre d'oscillations du terme de reste. Dans la pratique, on choisit ν dans l'intervalle $[1,4]$, $\nu = 1$ étant le cas le plus restrictif dans lequel la formule de Landau-Widom fournit une bonne estimation du rang.

4.2.2.1 Évaluation de la condition d'admissibilité statique

La condition nécessaire (4.35) traduit l'admissibilité de la partie statique. Dans la pratique, on se donne un petit paramètre $\alpha < 1$ et l'on cherche à vérifier la condition $|d| < \alpha\mathrm{R}$. Conformément à nos notations, R désigne ici une distance centre-à-centre. Dans un contexte $\mathscr{H}$-matrice, on privilégie d'ordinaire la plus petite distance réalisée entre les boîtes englobantes.
On détermine les boîtes englobantes orientées $\mathrm{Q}_s^{(u)}$ et $\mathrm{Q}_t^{(u)}$ respectivement des groupes X_s et Y_t dans la base (u_1, u_2, u_3). Cette étape est d'une complexité linéaire et consiste en un changement de base. La condition (4.35) est alors considérée satisfaite si l'on réalise la condition d'admissibilité forte entre les boîtes $\mathrm{Q}_s^{(u)}$ et $\mathrm{Q}_t^{(u)}$,

$$\max\left(\mathrm{diam}(\mathrm{Q}_t^{(u)}), \mathrm{diam}(\mathrm{Q}_s^{(u)})\right) \leqslant \eta\,\mathrm{dist}(\mathrm{Q}_t^{(u)}, \mathrm{Q}_s^{(u)}), \tag{4.37}$$

avec η un petit paramètre d'admissibilité. Dans la pratique, on se limite à $\eta \leqslant 2$.

4.2.2.2 Évaluation de la partie oscillante

L'évaluation de la condition d'admissibilité pour la partie oscillante décrite par l'expression (4.36) requiert le calcul de deux majorants de $|\overrightarrow{\mathbf{d}_{/\!/}}|$ et $|\overrightarrow{\mathbf{d}_{\perp}}|$. Dans le cas des sphères, ce calcul est immédiat car ces deux quantités s'expriment naturellement en fonction du rayon de la sphère.

Remarque 4.4 (Cas $\overrightarrow{\mathbf{d}_{/\!/}} = 0$). On rappelle que dans le cas où $\overrightarrow{\mathbf{d}_{/\!/}} = 0$, le critère de Fresnel est la donnée de la majoration du prochain terme non nul dans le développement limité de la phase, soit

$$k\frac{|\overrightarrow{\mathbf{d}_{\perp}}|^4}{8\mathrm{R}^3} \leqslant \nu 2\pi. \tag{4.38}$$

Dans le contexte des $\mathscr{H}$-matrices, on travaille avec des boîtes englobantes parallélépipédiques plutôt qu'avec des sphères englobantes. Il est tout à fait possible de travailler avec des sphères englobantes mais dans le cas d'un morceau de géométrie étiré (un morceau de cylindre par exemple), le volume de la sphère englobante serait plus grand que celui d'une boite parallélépipédique. Le diamètre serait du même ordre mais la distance entre deux sphères englobantes serait plus petite que dans le cas de boite carrés. On adapte alors le critère vu précédemment à des boîtes parallélépipédiques de tailles différentes et de même orientation.

4.2.2.3 Calcul d'un majorant de $|\overrightarrow{\mathbf{d}_{\perp}}|$ et de $|\overrightarrow{\mathbf{d}_{/\!/}}|$

On note $c_x^{(u)}$ et $c_y^{(u)}$ les centres des nouvelles boîtes $\mathrm{Q}_s^{(u)}$ et $\mathrm{Q}_t^{(u)}$. Ces derniers ne sont pas les mêmes que ceux des boîtes Q_s et Q_t déterminées dans la base canonique. On effectue le développement limité de la phase avec les anciens centres car c'est à partir de ceux-ci que l'on a fixé la direction u_3 dans laquelle on retranche les ondes planes. Le développement limité fait intervenir le vecteur $\overrightarrow{\mathbf{d}}$ défini par

$$\overrightarrow{\mathbf{d}} = \overrightarrow{\mathbf{d}}_x - \overrightarrow{\mathbf{d}}_y \tag{4.39}$$

$$\overrightarrow{\mathbf{d}}_x = x - c_x \tag{4.40}$$

$$\overrightarrow{\mathbf{d}}_y = y - c_y. \tag{4.41}$$

Dans la base (u_1, u_2, u_3), on effectue la décomposition suivante,

$$\overrightarrow{\mathbf{d}} = \overrightarrow{\mathbf{d}_{\perp}} + \overrightarrow{\mathbf{d}_{/\!/}}, \tag{4.42}$$

$$\overrightarrow{\mathbf{d}_{\perp}} = < \overrightarrow{\mathbf{d}}, u_1 > u_1 + < \overrightarrow{\mathbf{d}}, u_2 > u_2, \tag{4.43}$$

$$\overrightarrow{\mathbf{d}_{/\!/}} = < \overrightarrow{\mathbf{d}}, u_3 > u_3. \tag{4.44}$$

Par inégalité triangulaire, on a

$$|\overrightarrow{\mathbf{d}_{\perp}}| \leqslant | < \overrightarrow{\mathbf{d}}_x, u_1 > u_1 + < \overrightarrow{\mathbf{d}}_x, u_2 > u_2| \tag{4.45}$$

$$+ | < \overrightarrow{\mathbf{d}}_y, u_1 > u_1 + < \overrightarrow{\mathbf{d}}_y, u_2 > u_2| \tag{4.46}$$

Par orthogonalité, on a

$$| < \overrightarrow{\mathbf{d}}_x, u_1 > u_1 + < \overrightarrow{\mathbf{d}}_x, u_2 > u_2|^2 = | < \overrightarrow{\mathbf{d}}_x, u_1 > |^2 + | < \overrightarrow{\mathbf{d}}_x, u_2 > |^2 \tag{4.47}$$

$$| < \overrightarrow{\mathbf{d}}_y, u_1 > u_1 + < \overrightarrow{\mathbf{d}}_y, u_2 > u_2|^2 = | < \overrightarrow{\mathbf{d}}_y, u_1 > |^2 + | < \overrightarrow{\mathbf{d}}_y, u_2 > |^2 \tag{4.48}$$

Pour $i = 1,2,3$, les majorants recherchés sont donnés par

$$\left| < \overrightarrow{\mathbf{d}}_x, u_i > \right| \leqslant \sup_{x \in \mathrm{Q}_s^{(u)}} \| < c_x - x, u_i > \|_2,$$
$$| < \overrightarrow{\mathbf{d}}_y, u_i > | \leqslant \sup_{y \in \mathrm{Q}_t^{(u)}} \| < c_y - y, u_i > \|_2.$$

Remarque 4.5 (Implémentation pratique). On constate dans la pratique que l'on peut utiliser les dimensions des boîtes englobantes. Pour la boîte $\mathrm{Q}_s^{(u)}$ (respectivement $\mathrm{Q}_t^{(u)}$) , on note $\sigma_i^{(s)}$ (respectivement $\sigma_i^{(t)}$) les dimensions de la boîte englobante dans la base (u_1, u_2, u_3). On définit alors les quantités suivantes,

$$d_{/\!/}^{(s)} = \frac{1}{2}\sigma_3^{(s)} \qquad d_{\perp}^{(s)} = \frac{1}{2}\sqrt{\left(\sigma_1^{(s)}\right)^2 + \left(\sigma_2^{(s)}\right)^2} \tag{4.49}$$
$$d_{/\!/}^{(t)} = \frac{1}{2}\sigma_3^{(t)} \qquad d_{\perp}^{(t)} = \frac{1}{2}\sqrt{\left(\sigma_1^{(t)}\right)^2 + \left(\sigma_2^{(t)}\right)^2} \tag{4.50}$$

que l'on utilise alors pour les majorations de $|\overrightarrow{\mathbf{d}_{/\!/}}|$ et $|\overrightarrow{\mathbf{d}_{\perp}}|$,

$$|\overrightarrow{\mathbf{d}_{/\!/}}| \leqslant d_{/\!/}^{(s)} + d_{/\!/}^{(t)}, \tag{4.51}$$
$$|\overrightarrow{\mathbf{d}_{\perp}}| \leqslant d_{\perp}^{(s)} + d_{\perp}^{(t)}. \tag{4.52}$$

Résumé de la construction d'un critère d'admissibilité

Pour donner une vision plus claire des étapes décrites ci-dessus, on explicite les étapes de l'implémentation de notre critère d'admissibilité :

1. Détermination d'une direction principale u_3 en $\mathscr{O}(1)$ opérations.
2. Projection des inconnues et/ou des boîtes englobantes sur le plan normal à u_3 en $\mathscr{O}(m+n)$ opérations.
3. Construction d'une base orientée (u_1, u_2) de Π_{u_3} minimisant la section efficace en $\mathscr{O}(m+n)$ opérations.
4. Construction des boîtes orientées $\mathrm{Q}_t^{(u)}$ et $\mathrm{Q}_s^{(u)}$ en $\mathscr{O}(m+n)$ opérations.
5. Détermination des majorants de $|d_{/\!/}|$ et $|d_{\perp}|$ en $\mathscr{O}(1)$ opérations.
6. Évaluation de la condition d'admissibilité statique (admissibilité forte) en $\mathscr{O}(1)$ opérations.
7. Évaluation de la condition d'admissibilité fréquentielle (condition de Fresnel) en $\mathscr{O}(1)$ opérations.

L'évaluation de ce critère est d'une complexité linéaire en nombre d'inconnues, ce qui ne nous pénalise pas pour la suite des calculs. Par exemple, l'assemblage des blocs représentant les interactions proches possède une complexité plus élevée.

4.2.3 Algorithme HCA-II fréquentiel

Le premier chapitre de ce manuscrit présente de nombreuses méthodes algébriques afin de construire une approximation de bonne qualité d'un bloc admissible. Ces méthodes fonctionnent également dans le cadre de l'EFIE. Cependant, on souhaite adapter la méthode d'*Hybrid Cross Approximation* (HCA – II) initialement présentée dans [BG05]

pour un bloc de l'EFIE. L'emploi de cette méthode analytique permet un meilleur contrôle du rang de l'approximation. On va en présenter ici une version adaptée au noyau oscillant.
On se place dans le cas où l'on satisfait le critère d'admissibilité fréquentiel.

4.2.3.1 Écriture de la méthode HCA pour l'EFIE à quatre composantes

Z_0 désigne l'impédance du milieu intervenant dans l'équation intégrale (4.15). k est le nombre d'onde, lié à la longueur d'onde λ par la relation $k = 2\pi/\lambda$. Pour deux groupes de degrés de liberté X_s et Y_t respectivement de taille m et n, on considère la matrice $A_{s\times t}$ de discrétisation de l'EFIE correspondant à l'interaction entre ces deux groupes. Pour alléger les notations et en l'absence de confusion possible, on notera la matrice simplement par A. On rappelle également que $\vec{w}_i^t$ et $\vec{w}_j$ désignent les fonctions de base associées aux éléments finis de Raviart-Thomas (*cf* 4.2). Pour $i \in \{1,\dots,m\}$ et $j \in \{1,\dots,n\}$, les coefficients A_{ij} de la matrice sont donnés par

$$A_{ij} = -ikZ_0 \int_{\Gamma_h}\int_{\Gamma_h} G(x,y)\left[\vec{\mathbf{w}}_i^t(x)\vec{\mathbf{w}}_j(y) - \frac{1}{k^2}\mathrm{div}_{\Gamma_h}\vec{\mathbf{w}}_i^t\,\mathrm{div}_{\Gamma_h}\vec{\mathbf{w}}_j\right]\mathrm{d}\Gamma_h(y)\,\mathrm{d}\Gamma_h(x). \tag{4.53}$$

On se place dans le cadre de l'hypothèse d'admissibilité de Fresnel (**??**). Dès lors, la base de $\mathbb{R}^3$ considérée est la base d'interaction mutuelle (u_1,u_2,u_3) décrite en 4.2.1.3 et déterminée lors de l'évaluation de la condition d'admissibilité. Pour une fonction de base $\vec{\mathbf{w}}_j(x)$ donnée, on considère le quadri-vecteur $\overrightarrow{W_j}$

$$\overrightarrow{W_j} = \left(w_j^{(1)}, w_j^{(2)}, w_j^{(3)}, \frac{1}{ik}\mathrm{div}_\Gamma\vec{\mathbf{w}}_j\right)^{\mathrm{T}}, \tag{4.54}$$

où pour $q = 1,2,3$, $w_j^{(q)}$ est la composante suivant la direction u_q. On désigne par $W_j^{(\alpha)}$ la composante α de ce vecteur. Avec cette notation, on a

$$\begin{aligned} A_{ij} &= -ikZ_0 \int_{\Gamma_h}\int_{\Gamma_h} G(x,y)\,\overrightarrow{W_i^t}(x)\cdot\overrightarrow{W_j}(y)\,\mathrm{d}\Gamma_h(y)\,\mathrm{d}\Gamma_h(x),\\ &= -ikZ_0 \sum_\alpha \int_{\Gamma_h}\int_{\Gamma_h} G(x,y) W_i^{t,(\alpha)} W_j^{(\alpha)}\,\mathrm{d}\Gamma_h(y)\,\mathrm{d}\Gamma_h(x),\\ &= -ikZ_0 \sum_{\alpha=1}^{4} A_{ij}^{(\alpha)}, \end{aligned} \tag{4.55}$$

où $A_{ij}^{(\alpha)}$ est le coefficient (i,j) de la composante α. L'expression de la composante α est donnée par

$$A_{ij}^{(\alpha)} = \int_{\Gamma_h}\int_{\Gamma_h} G(x,y) W_i^{(\alpha)}(x)\, W_j^{(\alpha)}(y)\,\mathrm{d}\Gamma_h(y)\,\mathrm{d}\Gamma_h(x). \tag{4.56}$$

L'application de l'algorithme HCA – II à l'approximation de la matrice de l'EFIE correspond à la somme de quatre composantes et l'on parle alors de formulation à quatre composantes de l'EFIE. La formulation continue du problème variationnel (4.18) représente la somme de quatre formes bilinéaires. Une fois discrétisée cette formulation donne lieu à une somme de quatre matrices. Dans le contexte des $\mathscr{H}$-matrices, il s'agit de la somme approchée de quatre matrices de rang faible.

Remarque 4.6 (Formulations à deux et trois composantes). On trouve dans la littérature et particulièrement dans le cadre de la méthode FMM (voir par exemple [Syl02]), des formulations à trois et même deux composantes pour l'EFIE. Ces formulations sont rendues possibles par la décomposition du noyau de Green à l'aide du développement multipolaire. Un axe d'amélioration intéressant pour l'approximation de l'EFIE par l'algorithme HCA – II serait d'exhiber une méthode à deux ou trois composantes. Ceci permettrait d'améliorer la méthode à la fois sur le plan de l'implémentation mais aussi dans l'analyse numérique du schéma d'interpolation employé.

L'application de l'algorithme HCA – II à chacune des composantes α permet d'en obtenir une représentation de rang faible. Le rang obtenu est le même pour chaque composante car le rang est uniquement déterminé par le comportement du noyau de Green scalaire dans le domaine $X_s \times Y_t$. Plus particulièrement, le rang est majoré par le nombre de total de nœuds d'interpolation utilisés lors de l'approximation de $G(x,y)$ dans $Q_s^{(u)} \times Q_t^{(u)}$ à l'aide d'un schéma d'interpolation polynomiale. Seules les intégrales simples faisant intervenir les fonctions de base $\vec{w}_i^t$ et $\vec{w}_j$ sont différentes d'une composante à une autre. En toute rigueur, le rang numérique $r_\epsilon(A)$ de la matrice de l'EFIE est borné par celui de la somme de ses quatre composantes. En notant $r_{epsilon}(A^\alpha)$ le rang numérique à la précison ϵ d'une composante on ne dispose pas de borne plus précise que

$$r_\epsilon(A) \leqslant 4 r_\epsilon(A^\alpha). \tag{4.57}$$

Dans la pratique, le rang de cette somme n'atteint pas la borne maximale et l'étape de sommation approchée fournit un rang plus proche de $r_\epsilon(A^\alpha)$. Pareillement à la remarque précédente, l'amélioration de cette borne est une piste d'amélioration de la méthode.

Pour des cas particuliers comme des plaques en opposition ou coplanaires, cette borne est ramenée à $3r_\epsilon(A^\alpha)$. En effet, quitte à effectuer un changement d'origine, on peut toujours fixer une composante d'une des plaques à zéro. Par suite, cette composante « disparaît » du produit scalaire en ce sens où elle fournit une contribution nulle à ce produit scalaire. La somme (4.55) est donc ramenée à une somme de trois termes.

4.2.3.2 Approximation d'une composante par HCA-II

Interpolation du noyau de Green oscillant Dans le cadre du développement limité du chapitre précédent et on écrit le noyau de Green sous la forme

$$G(x,y) = e^{ik\vec{u}_3\cdot x}\, G_{u_3}(x,y)\, e^{-ik\vec{u}_3\cdot y}. \tag{4.58}$$

Ainsi, dans la direction u_3, on réalise directement la séparation des variables en considérant le noyau de Green déconvolué par les ondes planes dans cette direction. On utilise alors les résultats et les notations des paragraphes 2.3.3 et 2.3.4.2 du chapitre 1 et on note $\tilde{G}_{u_3}(x,y)$ l'approximation polynomiale du noyau de Green déconvolué par les ondes planes dans la direction u_3 et définie dans $Q_s^{(u)} \times Q_t^{(u)}$. On note $N_1^{(s)}, N_2^{(s)}$ et $N_3^{(s)}$ (respectivement $N_1^{(t)}, N_2^{(t)}$ et $N_3^{(t)}$) le nombre de nœuds d'interpolation dans les directions u_1, u_2 et u_3 dans la boîte englobante $Q_s^{(u)}$ (resp. $Q_t^{(u)}$). On rappelle qu'avec les notations du chapitre précédent, les directions portées par u_1 et u_2 sont les directions du plan transverse Π_{u_3}.

L'approximation $\tilde{G}_{u_3}(x,y)$ est définie par

$$\tilde{G}_{u_3}(x,y) = \sum_\nu \sum_\mu G_{u_3}\left(\xi_\nu^{Q_s^{(u)}}, \xi_\mu^{Q_t^{(u)}}\right) \mathscr{L}_\nu^{Q_s^{(u)}}(x)\, \mathscr{L}_\mu^{Q_t^{(u)}}(y), \tag{4.59}$$

où comme auparavant, $\nu = (\nu_1, \nu_2, \nu_3)$ et $\mu = (\mu_1, \mu_2, \mu_3)$ sont des multi-indices de $\mathbb{N}^3$ vérifiant $1 \leqslant \nu_i \leqslant N_i^{(s)}$ et $1 \leqslant \mu_i \leqslant N_i^{(t)}$ pour $i = 1,2,3$. Ces multi-indices sont associés aux nœuds et aux polynômes correspondants.

Remarque 4.7 (Rang de l'approximation). On note $N^{(s)}$ et $N^{(t)}$ le nombre de points d'interpolation dans les boîtes $Q_s^{(u)}$ et $Q_t^{(u)}$ respectivement. Par construction, le rang de l'approximation est directement majoré par

$$\min(N^{(s)}, N^{(t)}).$$

Approximation d'une composante En utilisant l'écriture déconvoluée par les ondes planes du noyau de Green (*cf* (4.58)) ainsi que l'approximation $\tilde{G}_{u_3}(x,y)$ (*cf* (4.59)) dans (4.56), on écrit la décomposition suivante de la composante α,

$$A_{ij}^{(\alpha)} = (U^{(\alpha)} C^T . (V^{(\alpha)} D^T))_{ij}$$

où C et D sont des matrices de coefficients calculées par un pseudo-inverse au lieu de l'algorithme 12 et les matrices d'intégrales simples suivantes, propres à la composante considérée,

$$U_{iq}^{(\alpha)} = \int_\Gamma e^{ik\vec{u}_3.x} W_i^{(\alpha)}(x).G_{\vec{u}}(x, \xi_{\mu_q}^s) d\Gamma(x), \tag{4.60}$$

$$V_{jq}^{(\alpha)} = \int_\Gamma e^{-ik\vec{u}_3.y} W_j^{(\alpha)}(y).G_{\vec{u}}(\xi_{\nu_q}^t, y) d\Gamma(y). \tag{4.61}$$

Nœuds d'interpolation Pour une surface X_s fixe, le rang numérique de son interaction avec une autre surface Y_t également fixée est constant pour le noyau de Laplace. On considère les boîtes englobantes Q_s et Q_t contenant respectivement les surfaces X_s et Y_t. Dans le cas du noyau oscillant, le nombre de degrés de liberté considérés sur ces surfaces va croître avec les dimensions en longueurs d'ondes des boîtes englobantes. En l'absence d'indication particulière, le nombre de points d'interpolation serait proportionnel à la longueur d'onde λ dans chacune des directions de l'espace. Ainsi, le nombre total de nœuds d'interpolation croît comme $\mathscr{O}((kd)^3)$! Ce schéma d'interpolation naïf est d'un intérêt nul dans la pratique et il convient d'utiliser un nombre de points dont la croissance avec la fréquence est moindre.

La qualité et la complexité de l'approximation dépend du choix du nombre de nœuds d'interpolation. Contrairement au noyau $\frac{1}{|x-y|}$, la dépendance en fréquence a une influence sur le choix du nombre nœuds. La première modification que l'on a utilisée est de manipuler le noyau de Green déconvolué par les ondes planes dans la direction u_3. Grâce à cette réécriture du noyau, on peut se contenter de choisir un nombre de points dans la direction des ondes planes u_3 suffisament petit. En effet, il suffit de pouvoir décrire les variations dues à la partie en $\frac{1}{|x-y|}$ dans cette direction. Pour les deux autres directions, on peut obtenir une estimation de l'ordre d'interpolation grâce à la formule de Landau-Widom précédemment mentionnée. Le paragraphe suivant explicite cette construction.

4.2.3.3 Choix des ordres d'interpolation

Avec les mêmes notations, u_3 désignant la direction des ondes planes, on rappelle que le développement limité de la phase permet d'écrire le noyau de Green déconvolué par les

ondes planes sous la forme,

$$\mathrm{G}_{u_3}(x,y) = e^{ik\mathrm{R}} e^{\frac{ik|d_\perp|^2}{2\mathrm{R}}} . \frac{e^{\Phi(x,y)}}{|x-y|}, \tag{4.62}$$

où $\Phi(x,y)$ est un terme de reste dans le développement limité de la phase et R la distance entre les centres de X_s et Y_t. L'utilisation de la base (u_1,u_2,u_3) permet d'écrire le terme du second ordre sous la forme d'un produit tensoriel comme suit,

$$e^{\frac{ik|d_\perp|^2}{2\mathrm{R}}} = e^{\frac{ik(\tilde{x}_1-\tilde{y}_1)^2}{2\mathrm{R}}} e^{\frac{ik(\tilde{x}_2-\tilde{y}_2)^2}{2\mathrm{R}}}, \tag{4.63}$$

où les quantités $\tilde{x}_i$ et $\tilde{y}_i$ ne dépendent respectivement que de x dans la direction u_i et y dans la direction u_i. On regroupe les termes décrivant le noyau G_{u_3} de la façon suivante,

$$\mathrm{G}_{u_3}(x,y) = \frac{e^{ik\mathrm{R}}}{|x-y|} . e^{\frac{ik(\tilde{x}_1-\tilde{y}_1)^2}{2\mathrm{R}}} e^{\frac{ik(\tilde{x}_2-\tilde{y}_2)^2}{2\mathrm{R}}} . e^{\Phi(x,y)}. \tag{4.64}$$

La détermination du nombre de nœuds d'interpolation requis est effectuée selon la nature des termes de l'écriture (4.64). Le premier terme (ainsi que le dénominateur) correspond à la partie statique du noyau. Il s'agit du cas du noyau de l'équation de Laplace déjà mentionné lors de la première exposition de la méthode HCA – II. Les deux termes exponentiels suivants sont deux opérateurs de Fresnel pour lesquels on a individuellement une bonne estimation du rang grâce à la formule de Landau-Widom à partir de laquelle on choisit les ordres d'interpolation nécessaires à leurs approximations respectives. Contrairement au premier terme, le rang de ces opérateurs dépend de la fréquence. Le dernier terme est un terme de reste dont le nombre d'oscillations est contrôlé par le paramètre ν dans la condition d'admissibilité fréquentielle de Fresnel (4.36).

Choix de l'ordre d'interpolation pour la partie statique $\frac{e^{ik\mathrm{R}}}{|x-y|}$ Pour une précision relative ϵ voulue, les résultats numériques et la littérature [BG05; BGH12] sur le noyau de l'équation de Laplace indiquent que l'ordre d'interpolation N_ϵ est de l'ordre de $\log_{10}(\epsilon)$,

$$\mathrm{N}_\epsilon = \mathscr{O}\left(|\log_{10}(\epsilon)|\right). \tag{4.65}$$

Dans le cas de l'approximation du noyau de Laplace ou des interactions basses fréquences (petites boîtes en terme de longueurs d'onde) pour le noyau oscillant, le nombre de points d'interpolation requis est alors de l'ordre de N_ϵ^3.

Choix de l'ordre d'interpolation pour le terme de reste $e^{\Phi(x,y)}$ Le paramètre d'admissibilité fréquentielle ν mesure le nombre d'oscillations du terme dominant dans le reste du développement limité de la phase. Ainsi, la condition d'admissibilité (4.36) avec $\nu < 1$ assure que les termes d'ordres supérieurs à trois ne dépendent plus de la fréquence. La fonction $e^{\Phi(x,y)}$ est une fonction trigonométrique prenant ses valeurs sur une période et son rang, proportionnel à $\log_{10}(\epsilon)$ ne dépend pas de la fréquence. Ainsi, le nombre de points d'interpolation requis $\mathrm{N}_{\epsilon,\nu}$ pour décrire ce terme est

$$\mathrm{N}_{\epsilon,\nu} = \mathscr{O}(\nu|\log_{10}(\epsilon)|). \tag{4.66}$$

Choix de l'ordre d'interpolation pour la partie $e^{\frac{ik(\bar{x}_i-\bar{y}_i)^2}{2R}}$ Il s'agit de la modification majeure que l'on apporte à la méthode HCA – II. On dispose grâce à la formule de Landau-Widom d'une bonne estimation du rang de chacun des opérateurs de Fresnel dans les directions (u_1, u_2) du plan transverse ce qui nous permet un choix anisotrope des ordres d'interpolation dans ces directions. On rappelle que pour une largeur de bande c fixée, le rang numérique à la précision ϵ de l'opérateur de Fresnel vérifie l'estimation

$$r_\epsilon = \frac{2c}{\pi} + \frac{1}{\pi^2}\log\left(\frac{1-\epsilon^2}{\epsilon^2}\right)\log(c) + o(\log(c)). \tag{4.67}$$

La largeur de bande est liée à une dimension caractéristique d et une distance caractéristique R ainsi qu'à la fréquence par la relation

$$c = k\frac{d^2}{R}. \tag{4.68}$$

En substituant dans (4.69) la largeur de bande par son expression (4.68), le rang r_ϵ s'écrit

$$r_\epsilon = \frac{2k}{\pi}\frac{d^2}{R} - \frac{2}{\pi^2}\log(\epsilon)\log\left(k\frac{(d)^2}{R}\right) + o\left(\log\left(k\frac{d^2}{R}\right)\right). \tag{4.69}$$

On détermine alors deux largeurs de bande c_1 et c_2 propres aux directions u_1 et u_2 à partir des dimensions des boîtes englobantes $Q_t^{(u)}$ et $Q_s^{(u)}$.
On note par σ_i la dimension d'une boîte $Q^{(u)}$ dans la direction u_i. On rappelle que la distance centre-à-centre des boîtes est notée par R et que les largeurs de bande c_1 et c_2 sont définies par

$$c_1 = k\frac{d_1^2}{R}, \qquad c_2 = k\frac{d_2^2}{R}, \tag{4.70}$$

avec

$$d_1 = \frac{1}{2}\sqrt{\sigma_1^{(s)}\sigma_1^{(t)}}, \qquad d_2 = \frac{1}{2}\sqrt{\sigma_2^{(s)}\sigma_2^{(t)}}. \tag{4.71}$$

Les ordres d'interpolation N_1^{LW} et N_2^{LW} respectivement dans les directions u_1 et u_2 sont alors les suivants,

$$N_1^{LW} = \frac{2c_1}{\pi} - \frac{2}{\pi^2}\log(\epsilon)\log(c_1) \tag{4.72}$$

$$N_2^{LW} = \frac{2c_2}{\pi} - \frac{2}{\pi^2}\log(\epsilon)\log(c_2). \tag{4.73}$$

Remarque 4.8. On note que dans chaque direction du plan transverse, l'ordre d'interpolation est le même pour les deux boîtes $Q_t^{(u)}$ et $Q_s^{(u)}$ car la largeur de bande prend déjà en compte les dimensions des deux boîtes. Ces ordres d'interpolation ne concernent que les directions u_1 et u_2 et n'interviennent pas dans le schéma d'interpolation dans la direction u_3 des ondes planes.

4.2.3.4 Ordre d'interpolation total

Rang d'un produit d'opérateurs de rangs finis En toute rigueur le rang numérique de l'opérateur (4.62) est le produit des rangs de chaque opérateur le constituant. D'où

$$r_\epsilon(G_{u_3}(x,y)) = \mathscr{O}(N_\epsilon.N_{v,\epsilon}.N_1^{LW}.N_2^{LW}). \tag{4.74}$$

Dans la pratique il n'en est rien et l'on constate sans avoir de preuve formelle que le terme dominant est celui dû au second ordre soit $r_\epsilon(G_{u_3}(x,y)) = \mathscr{O}(N_1^{LW}.N_2^{LW})$. Ceci sera montré plus loin lors des tests numériques.

Heuristique pour le nombre de points dans chaque direction On utilise le même nombre de points dans chaque boîte englobante et l'on note N_i le nombre de points d'interpolation dans la direction u_i. On effectue le choix suivant pour les nombre de points

$$N_1 = N_1^{LW}, \tag{4.75}$$

$$N_2 = N_2^{LW}, \tag{4.76}$$

$$N_3 = N_\epsilon. \tag{4.77}$$

C'est la remarque précédente sur le rang du produit des contributions qui motive le choix du nombre de points dans les directions transverses u_1 et u_2. Le choix dans la direction longitudinale u_3 est quant à lui motivé par l'emploi des ondes planes dans cette direction qui atténuent ainsi les oscillations. Pour les petites boîtes, la formule de Landau-Widom fournit un nombre presque constant de points et l'on est dans la situation de l'approximation du noyau de Lapalce. Pour de plus grandes boîtes, les termes N_1^{LW} et N_2^{LW} rendent la grille d'interpolation plus fine dans les directions transverses comme l'illustre la figure 4.7.

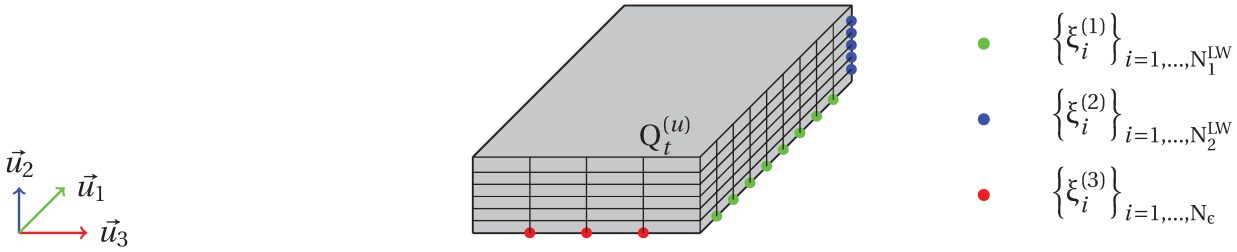


FIGURE 4.7 – Placement des nœuds d'interpolation dans la boîte englobante $Q_t^{(u)}$.

La densité de points n'est pas la même suivant la direction. En effet, dans la direction u_3, les ondes planes portent l'information donc il n'est pas nécessaire de mettre autant de points que dans les directions transverses. Seule la partie en $1/|x-y|$ varie dans la direction u_3. Par ailleurs, le deux groupes sont admissibles pour le critère d'admissibilité de Fresnel donc la distance est supérieure à celle obtenue avec juste un critère statique. Dans ce cas, on constate numériquement que l'approximation est efficace avec peu de points. Ainsi dans le plan (u_2, u_3) on place moins de points que dans le plan (u_1, u_2) décrivant la section efficace où les oscillations sont plus importantes.

Le nombre total de points dans chaque boîte est $N_{tot} = N_\epsilon N_1^{LW} N_2^{LW}$. Cependant, les boîtes peuvent être de dimensions hétérogènes et une amélioration serait de construire un schéma anisotrope sur le même modèle que décrit dans la référence [BGH12] dans le cas du noyau de Laplace $1/|x-y|$ afin de réduire le nombre de points utilisés. De plus, on peut également souhaiter un nombre de points équilibré dans chaque boîte afin d'avoir un schéma plus stable. Il s'agit de pistes à explorer pouvant apporter de bons résultats dans la pratique.

Remarque 4.9 (Cas d'une dimension dégénérée). Dans le cas où une dimension est nulle, on n'utilise qu'un seul point d'interpolation et l'on considère une approximation d'ordre zéro.

4.2.3.5 Résumé

Le choix du nombre de points d'interpolation dépend de la base d'interaction mutuelle (u_1, u_2, u_3) ainsi que des dimensions des boîtes englobantes $Q^{(u)}$ dans cette base. Dans la pratique, on détermine le nombre de points d'interpolation de la sorte :

1. Le nombre de points dans la direction longitudinale u_3 est déterminé par (4.65) :

$$\mathrm{N}_3 = \mathrm{N}_\epsilon.$$

2. Pour chaque direction u_1, u_2 du plan transverse, le nombre de points d'interpolation est donné par la formule de Landau-Widom (4.72)),

$$\mathrm{N}_1 = \mathrm{N}_1^{\mathrm{LW}},$$
$$\mathrm{N}_2 = \mathrm{N}_2^{\mathrm{LW}}.$$

Le choix du nombre de points est donc adapté à l'écriture choisie du noyau G_{u_3}. Notre modification de la méthode originale consiste en deux points particuliers. Le premier point est de se placer dans une base « adaptée » à l'interaction de X_s et Y_t. Le second point est de construire une grille d'interpolation polynomiale anisotrope. En effet, dans la direction portée par u_3, il n'est nécessaire que de capter les variations de la partie statique ce qui requiert peu de points. Les directions transverses sont elles maillées de telle sorte à ce que le nombre de points dans chacune des directions soit en accord avec la formule de Landau-Widom. Ce choix est adapté aux dimensions de la section efficace exprimée par les directions u_1 et u_2.

4.3 Influence du *clustering* sur le taux de compression

Ce paragraphe décrit l'influence du *clustering* - la création de l'arbre de blocs - sur le taux de compression d'un bloc matriciel. Il s'agit d'une étape importante de la construction d'une $\mathscr{H}$-matrice. En effet, la condition d'admissibilité permet de détérminer *a posteriori* si les interactions produites par l'étape de *clustering* sont proches ou lointaines.

4.3.1 De l'intérêt de subdiviser un bloc

On s'intéresse à un bloc matriciel A correspondant à une interaction admissible pour le critère d'admissibilité statique. On note r le rang numérique de ce bloc et $\mathscr{C}_{\mathrm{COEF}}^{(0)}$ le nombre de coefficients de l'approximation de rang faible construite à ce niveau initial. Ainsi, on a

$$\mathscr{C}_{\mathrm{COEF}}^{(0)} = 2\mathrm{N}r.$$

Dans le cas d'une découpe de ce bloc matriciel en une matrice blocs 2×2 de la sorte,

$$\mathrm{A} = \begin{bmatrix} \mathrm{A}_1 & \mathrm{A}_3 \\ \mathrm{A}_2 & \mathrm{A}_4 \end{bmatrix},$$

on note r_i le rang du sous-bloc A_i. Le nombre total de coefficients $\mathscr{C}_{\mathrm{COEF}}^{(1)}$ construits pour l'ensemble des sous-blocs est alors

$$\mathscr{C}_{\mathrm{COEF}}^{(1)} = \frac{\mathrm{N}}{2}\left(\sum_{i=1}^{4} r_i\right) \times 2.$$

Du point de vue de la mémoire requise pour le stockage, la moyenne des rangs des sous-blocs $(\mathrm{A}_i)_{1 \leqslant i \leqslant 4}$ doit être inférieure à la moitié du rang initial r :

$$\frac{1}{4}\sum_{i=1}^{4} r_i \leqslant \frac{r}{2}. \tag{4.78}$$

Dans le cas de l'admissibilité fréquentielle, une condition nécessaire est que l'interaction soit admissible pour le critère statique. Pour le critère d'admissibilité fréquentiel de Fresnel, la section efficace entre les objets. Pour rappel, la largeur de bande est

$$c = \frac{kd^2}{\mathrm{R}}, \tag{4.79}$$

avec la longueur efficace d et R est la distance des centres des groupes, k le nombre d'onde. On constate d'après l'expression (4.79) que si la dimension d est divisée par 2 alors la largeur de bande est elle divisée par 4. Ce comportement est celui attendu dans le cas d'objet présentant une grande section efficace. On cherche absolument à réduire la section efficace lors des découpes hiérarchiques effectuées. Selon l'arbre de blocs créé initialement, on peut observer deux comportements distincts selon que l'on réduit avec succès ou non la section efficace. On met en valeur ces comportements en calculant l'interaction de sous-domaines représentés par des plaques coplanaires distantes.

Description de du cas test La géométrie utilisée dans les deux sous-paragraphes suivants consiste en l'interaction de deux plaques rectangulaires identiques. Celles-ci sont caractérisées par les paramètres suivants :

dimensions : $l \times \mathrm{L} = 1m \times 4.2m$;

maillage : taille moyenne des arêtes $h = 8.10^{-3}m$;

nombre de degrés de liberté : N = 226177 degrés de libertés.

La fréquence à laquelle a été effectuée les calculs correspond à celle satisfaisant une règle de 5 points par longueur d'onde soit $\lambda = 5h$ et une fréquence $f = 7.48\mathrm{GH}z$.
Par ailleurs, les deux plaques sont **éloignées d'une distance** $\mathrm{D} = 4.316m$ afin que l'interaction soit considérée **admissible pour le critère statique de Hackbusch avec** $\eta = 2$. On effectue les calculs suivants à la précision relative de $\varepsilon = 1.10^{-4}$.

4.3.2 Amélioration de la compression : réduction de la section efficace

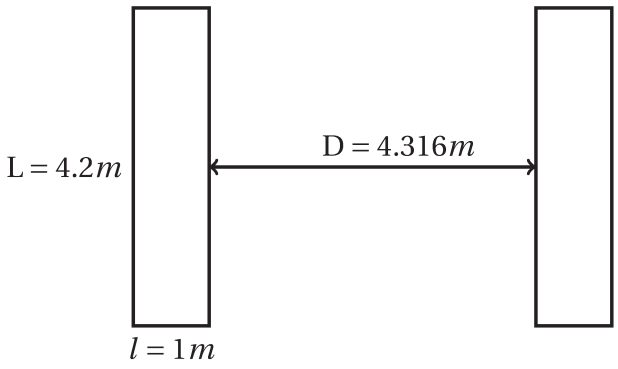
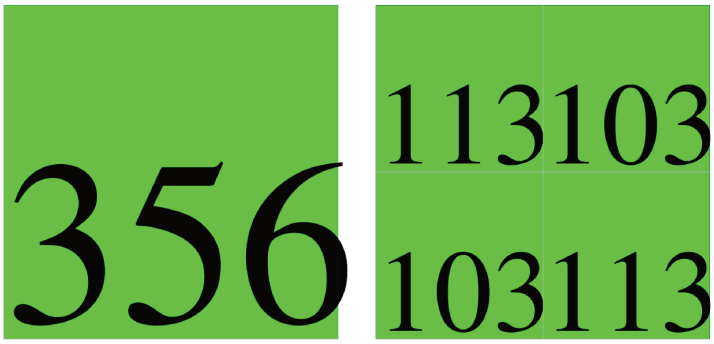


FIGURE 4.8 – Cas test présentant une forte section efficace.

Dans ce cas, la section efficace est maximale. Pour la fréquence considérée, l'admissibilité fréquentielle n'est pas réalisée pour une valeur faible du paramètre d'admissibilité ν. Il est nécessaire de découper le bloc matriciel décrivant l'interaction de ces deux plaques. Pour le passage au niveau hiérarchique supérieur, chaque plaque est ainsi coupée en deux dans le sens normal à la plus grande dimension (*i.e* par un plan horizontal sur la figure (4.8)). La largeur de bande est donc elle est asymptotiquement divisée par 4 pour



(a) Rang de l'interaction initiale.

(b) Rangs des interactions obtenues par une seule découpe hiérarchique.

les interactions de *clusters* (groupes) en vis-à-vis après cette première découpe comme mentionné plus haut.

L'interaction complète possède un rang numérique $r = 356$ (4.9a). Pour que le gain en mémoire soit substantiel, la condition (4.78) stipule que le rang moyen des sous-blocs doit être inférieur à $r/2 = 178$ ce qui est le cas d'après la représentation des rangs de la figure (4.9b) où le rang moyen des sous-blocs est de 108. Dans ce cas, la réduction de la section efficace permet un gain mémoire et améliore le taux de compression : 0.31% de la taille mémoire de la matrice originale sans découpe et 0.19% dans le cas d'une seule découpe.Gain de 50%.

4.3.3 Déterioration du taux de compression

La deuxième configuration que l'on souhaite traiter dans ce paragraphe est celle de deux plaques rectangulaires coplanaires décrites par la figure (4.10). La configuration

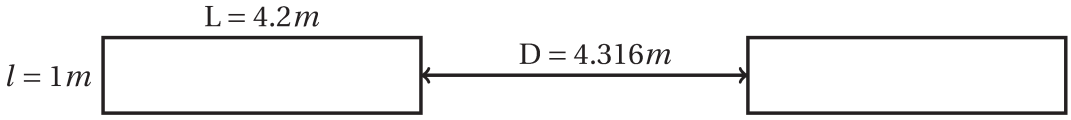
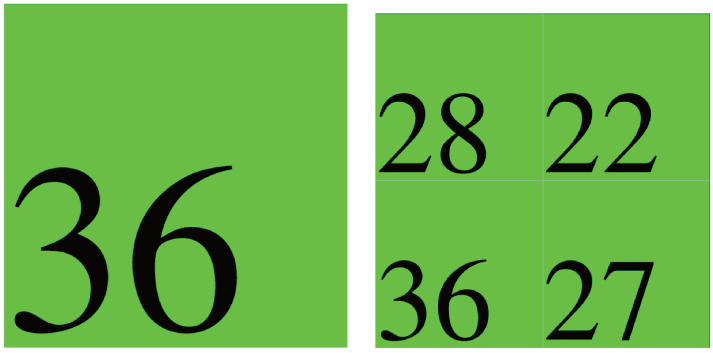


FIGURE 4.10 – Cas de plaques orientées suivant une direction privilégiée.

choisie ici est plus défavorable d'un point de vue du *clustering*. Comme précédemment, on découpe le bloc original en une matrice 2×2. Pour ce faire, les plaques sont coupées en deux suivant un plan vertical. Dans ce cas, la section efficace ne change pas ! Dans l'expression de la largeur de bande (4.79), la variable d demeure la même et seule la distance R varie. Entre les deux extrémités les plus lointaines, la distance n'a pas augmenté suffisament pour réduire significativement la largeur de bande. La distance entre les centres des plaques est de $R = 8.516m$ pour l'interaction originale et $R' = 10.616m$ pour les extrémités les plus éloignées. Ce rapport de 1.24 entre ces deux distances ne conduit qu'à réduire la largeur de bande de 20%.

La largeur de bande n'est donc presque pas modifiée en effectuant la découpe en une matrice blocs 2×2. Ceci est visible sur les représentations des rangs des figures (4.11a) et (4.11b). Le rang moyen de la découpe n'est pas suffisament faible (28 au lieu de 18



(a) Rang de l'interaction initiale.

(b) Rangs des interactions obtenues par une seule découpe hiérarchique.

pour obtenir un gain) pour respecter la condition (**??**) et la mémoire requise est ainsi augmentée de près de 50% ! Il s'agit d'un cas où la découpe s'avère être inadaptée et on souhaiterait plutôt traiter l'interaction originale. En effet, l'interaction originale est orientée suivant une direction de propagation privilégiée qui peut être traitée efficacement avec l'écriture modifiée du noyau de Green de l'équation de Helmholtz.

Commentaires

Les deux configurations présentées ci-dessus soulignent l'importance de l'étape de *clustering*. En effet, la première étape de la construction d'une $\mathscr{H}$-matrice est la construction de l'arbre de groupes (le *clustering*). L'évaluation de l'admissibilité fréquentielle d'un bloc matriciel correspond à l'interaction de deux éléments de l'arbre de groupes. Dans le cas où la condition d'admissibilité n'est pas réalisée, il est alors nécessaire d'étudier l'admissibilité des fils de ces groupes lesquels ont été déterminés d'après une découpe indépendante de la condition d'admissibilité ou de la fréquence.
Les comportements limites des interactions des fils sont alors donnés par les exemples ci-dessus. On peut améliorer la mémoire et le taux de compression si la découpe réduit significativement la section efficace. *A contrario*, si la section efficace demeure la même lors de la découpe, on dégrade alors le taux de compression. Il s'agit d'un thème de recherche actuel et une piste d'amélioration est de construire l'arbre de groupes de façon semi-automatique. Les premières bissections peuvent être réalisées en fonction de la géométrie du problème afin de garantir une minimisation de la section efficace pour les blocs matriciels les plus gros, susceptibles de poser problème. En-deçà d'une certaine taille on peut utiliser les techniques habituelles de la littérature.

La méthode des $\mathscr{H}$-matrices a été implémentée avec succès dans [Liz14]. Cette implémentation efficace et parallèle est fortement utilisée dans la pratique pour de nombreuses applications allant jusqu'au million de degrés de liberté.
Les problèmes actuels auquel la communauté s'intéresse sont ceux de l'ordre de plusieurs dizaines de millions de degrés de liberté et le projet HIBOX est un pas dans cette direction et les tests suivants s'inscrivent dans les objectifs de ce projet.

[Liz14] contient une implémentation parallèle et efficace de la méthode des $\mathscr{H}$-matrices appliquée à de nombreux cas pratiques de l'ordre de la centaine de milliers d'inconnues.

Un résultat majeur de cette implémentation est d'avoir montré la bonne scalabilité du parallélisme de la méthode. Pour les cas à plusieurs millions d'inconnues, soit de plus hautes fréquences, on s'intéresse également à la scalabilité en fréquence de la méthode et ce notament car le cadre théorique de la méthode des $\mathscr{H}$-matrices décrite dans la littérature ([BGH12; Beb00]) n'est pas adapté aux hautes fréquences. L'admissibilité fréquentielle de Fresnel ainsi que l'adaptation de la méthode HCA – II sont des éléments de réponse au traitement des cas visés.

Points d'intérêt Les validations numériques des développements précédents portent sur deux points principaux :

— La construction d'un critère d'admissibilité fréquentiel.

— La construction d'un algorithme d'approximation du noyau de Green.

Pour le premier point, l'on souhaite montrer qu'à haute fréquence, le critère d'admissibilité statique peut conduire à une croissance de la mémoire comme le carré de la fréquence pour certains blocs. Les tests effectués sont donc une comparaison entre les deux critères d'admissibilité. Le second point aborde la validation de l'approximation du noyau de Green par un algorithme HCA – II modifié. On montre que les modifications apportées sont maitrisées, robustes et fournissent des résultats corrects.

Station de travail Les tests numériques présentés dans la suite de ce chapitre ont été effectués avec une station de travail dont les caractéristiques sont résumées dans la table 4.1.

Processeur	Intel(R) Core(TM) i7-3930K
Fréquence	3.20GHz
Mémoire	64Go

TABLEAU 4.1 – Configuration de la machine de calcul utilisée.

4.4 Analyse du taux de compression

4.4.1 Conditions des tests

4.4.1.1 Simulation d'une interaction haute fréquence

La figure **??** représentant les différentes zones d'admissibilité du chapitre précédent montre qu'il est nécessaire d'avoir de grands objets afin que le critère de Fresnel puisse s'exprimer. Pour exploiter au mieux nos ressources informatiques, nous simulons une interaction à haute fréquence en considérant seulement l'interaction d'un groupe d'inconnues sur un autre. On construit ces inconnues de façon à ce que leurs supports respectifs possèdent une taille très grande par rapport à la longueur d'onde. Ce type d'interaction sera amené à être traiter sur un cas pratique de l'ordre de plusieurs millions de degrés de liberté. En effet, pour un objet maillé uniformement avec des arêtes d'une dimension de $\lambda/5$ le nombre d'inconnues N_{inc} est lié à la surface S de l'objet par l'approximation suivante,

$$S \approx \frac{N_{inc}}{150}\lambda^2. \tag{4.80}$$

Ainsi, pour un problème de l'ordre de $N_{inc} = 1.5.10^6$ degrés de liberté, la surface S est de l'ordre de $10^4\lambda^2$ et le diamètre de l'objet est ainsi de l'ordre de la centaine de longueurs d'onde. Les premières interactions vérifiant le critère d'admissibilité statique sont celles d'objet dont le diamètre est de l'ordre de la dizaine de longueurs d'onde. La figure **??** indique que pour une telle taille, le critère de Fresnel n'est pas encore très déterminant.

Nos moyens informatiques ne sont pas encore suffisants pour observer l'influence du critère de Fresnel sur un objet dans son intégralité. Pour pallier cette difficulté, on préfere construire « à la main » des interactions entre des groupes contenant un grand nombre de longueurs d'onde. Pour une fréquence de 15GHz, une plaque carrée de côté $a = 1m$ maillée en $\lambda/5$ comporte environ 215000 degrés de liberté et possède un diamètre d'environ 70λ. On simule alors une interaction à haute fréquence etn considérant l'interaction de deux plaques distantes. Cette géométrie est facilement mainpulable avec les ressources à notre disposition et permet de tester efficacement l'influence du critère d'admissibilité fréquentiel.

4.4.1.2 Stratégies de maillage dans la pratique

La génération d'un maillage pour un cas pratique présentant des singularités ou des zones plus raffinées nécessite un soin particulier. Il est fréquent qu'un maillage de meilleur qualité procure des résultats numériques beaucoup plus satisfaisants et ce avant même d'optimiser la méthode numérique. Dans la pratique, une étude fréquentielle (un balayage en fréquence) est réalisée suivant trois stratégies possibles :

— Emploi d'un maillage fixe, adapté à la fréquence maximale ;

— Emploi d'un maillage par fréquence d'étude ;

— Emploi d'un maillage pour une bande de fréquences donnée.

Il n'y a pas de méthode universelle et l'on trouve dans la pratique les trois stratégies suivant le cas étudié. Le cas d'un maillage fixe adapté à la plus grande fréquence correspond à un cas où la plage de fréquence n'est pas trop étendu (par exemple une décade). Dans le cas où la génération du maillage est couteuse ou peu aisée, il s'agit d'une stratégie très confortable. Cependant, pour les plus basses fréquences, nous sommes surmaillés et cela peut nous donner une fausse impression. En effet, une méthode approchée telle que la méthode des $\mathscr{H}$-matrices donnera des résultats très impressionnants. Ces bons résultats sont directement liés à l'utilisation d'un maillage trop fin. Si possible, le cas idéal est d'utiliser un maillage adapté à la fréquence d'étude. Cela permet de pouvoir comparer efficacement les résultats sans devoir tenir compte du phénomène de surmaillage. Dans la pratique néanmoins, le coût de l'étape de maillage peut s'avérer être un obstacle. La dernière stratégie est un mixte des deux autres et est fréquemment utilisées lorsque la plage de fréquences est étendue.
Les tests numériques présentés ont été effectués sur un maillage fixe et adapté à la fréquence maximale. La longueur moyenne des arêtes est de l'ordre de $\lambda/5$, λ étant la longueur d'onde associée à la fréquence maximale. Dans un premier temps, les tests numériques visent à montrer l'intérêt d'utiliser un critère d'admissibilité fréquentiel. Nous montrons cet intérêt en représentant la croissance de la mémoire ainsi que le taux de compression d'un bloc matriciel issu de l'EFIE à quatre composantes.

4.4.1.3 Choix des quantités observées

Pour illustrer la croissance en mémoire d'un bloc d'une $\mathscr{H}$-matrice en fonction de la fréquence, on utilise deux mesures différentes. Dans un premier cas, on considère un

maillage fixe de l'objet sur lequel on effectue un balayage en fréquence. La seconde méthode consiste à remailler l'objet pour chaque fréquence afin d'être maillé en $\lambda/5$ quelle que soit la fréquence.

Nos développements théoriques portent sur une estimation du comportement du rang numérique en fonction de la fréquence. La croissance du rang se traduit directement par la croissance de la mémoire et la mesure du taux de compression de la méthode. On rappelle que pour un problème fréquentiel, le nombre de degrés de liberté N d'un maillage varie de manière quadratique avec la fréquence, soit

$$\mathrm{N} = \mathscr{O}(f^2).$$

La quantité mesuré par notre code est le pourcentage de compression réalisé du bloc matriciel. Le choix des géométries est effectué de sorte que le bloc matriciel ne soit pas subdivisé dans le cas d'un critère d'admissibilité statique. Dans ce cas, le taux de compression permet directement d'obtenir le rang numérique à la précision ϵ (noté r_ϵ) de l'approximation construite à la fréquence f_i. On note $\tau_{\max}(f)$ le taux de compression mesuré pour une fréquence $f \leqslant f_{\max}$ sur le maillage adapté à la plus grande fréquence. Ce taux de compression correspond au rapport entre le nombre de coefficients de la représentation de rang faible et le nombre de coefficients nécessaires au stockage plein soit,

$$\begin{aligned} \tau_{\max}(f) &= \frac{2 r_\epsilon(f) \mathrm{N}_{\max}}{\mathrm{N}_{\max}^2} \\ &= \frac{2 r_\epsilon(f)}{\mathrm{N}_{\max}}. \end{aligned} \tag{4.81}$$

Ce taux mesure le gain d'espace mémoire à maillage fixe. On note que ce taux est d'autant meilleur que le rang r_ϵ a une croissance plus faible que f^2. Cependant, comme nous l'avons mentionné, aux plus basses fréquences on observe un phénomène de surmaillage. Les résultats peuvent être jugés pertinents non pas grâce à la méthode employée mais grâce au fait que l'on soit surmaillé. Pour avoir une mesure plus satisfaisante du gain obtenu, on représente le taux de compression dans le cas d'un maillage adapté à la fréquence de façon à ce que les arêtes soient de l'ordre de $\lambda/5$. On note τ_i le taux de compression pour le maillage composé de N_i degrés de liberté et associé à la fréquence f_i. Pour une interaction donnée, le rang numérique r_ϵ ne dépend pas de la finesse du maillage. La seule condition est que le maillage résolve la longueur d'onde. Dans le cas extrême où le nombre de degrés de liberté devient très petit, le rang peut varier de façon significative mais ce cas limite n'est pas considéré ici.

Exemple 4.10 (Stabilité du rang & finesse du maillage)**.** On considère le cas de deux plaques carrées opposées de côté $a = 1m$. Le maillage le plus grossier comporte $\mathrm{N}_0 = 95885$ DDLs, le maillage intermédiaire comporte $\mathrm{N}_1 = 215506$ DDLs tandis que le maillage le plus fin contient $\mathrm{N}_2 = 384469$ DDLs.

Fréquence (Hz)	r_ϵ (N_0)	r_ϵ (N_1)	r_ϵ (N_2)
$1.25.10^8$	53	53	53
$3.12.10^8$	56	55	56
$6.25.10^8$	92	92	93
$1.25.10^9$	184	184	184
$2.5.10^9$	429	429	429
$5.0.10^9$	1132	1135	1136

TABLEAU 4.2 – Rang numérique r_ϵ à la précision $\epsilon = 1.10^{-4}$

La table 4.2 contient les résultats de cette expérience sur la stabilité du rang selon la finesse du maillage. Les variations du rang sont de l'ordre de l'unité entre le maillage le plus fin et le plus grossier.

La constance du rang numérique permet de définir τ_i à partir de $\tau_{\max}$ de la sorte

$$\begin{aligned} \tau_i &= \tau_{max}(f_i).\frac{N_{\max}}{N_i} \\ &\approx \tau_{max}(f_i).\left(\frac{f_i}{f_{\max}}\right)^{-2} \end{aligned} \tag{4.82}$$

Le taux de compression τ_i est plus représentatif de la performance de la compression. À basses fréquences, le nombre de degrés de liberté requis afin de décrire correctement l'interaction est plus faible qu'à la plus haute fréquence. On s'attend donc à ce que le taux de compression adapté à la fréquence soit dércoissant avec la fréquence. Le rang étant lui constant, le taux le plus mauvais est celui associé à la plus basse fréquence. À l'opposé, le taux de compression pour la plus haute fréquence doit tendre vers zéro avec la fréquence. Cette décroissance est une façon de caractériser une méthode rapide.

Remarque 4.11 (Caractérisation d'une méthode rapide). Pour qu'une méthode soit qualifiée de rapide, il est nécessaire que le taux de compression tende vers zéro avec la fréquence. Dans le cas où ce taux devient constant, la méthode perd son caractère rapide.

La deuxième quantité que l'on souhaite représenter est l'espace mémoire requis pour traiter l'interaction. Plus particulièrement, on veut maîtriser la croissance de la mémoire en fonction de la fréquence. Il s'agit d'un point particulièrement important dans la pratique. Selon la taille du bloc, la mémoire requise pour la fréquence f_i que l'on note $\text{mem}(f_i)$ correspond simplement à $\tau_{\max}(f).N_{\max}^2$.

4.4.1.4 Choix des paramètres

La comparaison des deux critères d'admissibilité s'effectue avec le choix de paramètres suivant.

Algorithme de compression	ACA+
Erreur relative d'assemblage	$\epsilon_{\text{ACA}} = 1e-4$
Erreur relative de recompression	$\epsilon_{r\text{SVD}} = 1e-4$
Critère d'admissibilité statique	condition statique avec $\eta = 2$
Critère d'admissibilité fréquentiel	condition de Fresnel avec $\eta = 2$, $\nu = 2$
Représentation informatique des coefficients	complexe double précision

TABLEAU 4.3 – Paramètres pour les tests de croissance de la mémoire.

4.4.2 Rappels sur la croissance du rang

Sous la condition d'admissibilité statique (la condition usuelle pour la construction d'un bloc d'une $\mathscr{H}$-matrice), la largeur de bande c des opérateurs de Fresnel (le second ordre dans le développement limité) varie comme la fréquence. En effet, la condition d'admissibilité impose que le rapport entre le diamètre et la distance de deux blocs soit borné,

$$\frac{|d|}{\mathrm{R}} \leqslant \mathscr{O}(1).$$

Par exemple, le critère d'admissibilité statique dans la méthode des $\mathscr{H}$-matrices est couramment utilisé avec le rapport

$$\eta := \frac{|d|}{\mathrm{R}}$$

tel que $\eta \leqslant 2$.
La largeur de bande dans chaque direction direction du plan transverse vérifie ainsi

$$\begin{aligned} c &= \frac{(kd)^2}{k\mathrm{R}}, \\ &\approx \eta(kd). \end{aligned}$$

Le rang $\mathrm{N}(c,\epsilon)$ des termes du second ordre dans le développement limité de la phase du noyau de Green scalaire est donné dans chaque direction du plan transverse par la formule de Landau-Widom,

$$\mathrm{N}(c,\epsilon^2) = \frac{2c}{\pi} - \frac{2}{\pi^2}\log(\epsilon)\log(c). \tag{4.83}$$

Asymptotiquement, le rang se comporte comme la largeur de bande. Ainsi, dans le cas où la section efficace est réduite à un segment, on s'attend à une croissance des termes du second odre comme la fréquence car une seule direction agit. Dans le cas où la section efficace est un rectangle non dégénéré, la croissance du rang est plus rapide jusqu'à devenir quadratique dans le cas d'un carré par exemple.
On illustre ces deux cas extremaux en s'intéressant à des plaques distantes coplanaires et en opposition.

4.4.3 Cas de deux plaques coplanaires

4.4.3.1 Géométrie

On considère deux plaques carrées de côté $a = 1m$ coplanaires et appartenant au plan d'équation $z = 0$. Ces plaques possèdent de plus la même orientation dans le plan $(\mathrm{O}xy)$.

Ce cas est réputé favorable pour une approche directionnelle. En effet, en considérant le noyau déconvolué par les ondes planes dans la direction liant les centres des plaques, la contribution majeure au rang du bloc matriciel est due à la section efficace qui est ici réduite à un segment. Les deux plaques sont espacées de telle sorte à ce que la distance minimale soit $R = \sqrt{2}/2$. Ainsi, ces deux plaques sont admissibles pour la condition d'admissibilité statique telle qu'utilisé dans les $\mathscr{H}$-matrices pour le paramètre d'admissibilité $\eta = 2$.

4.4.3.2 Évolution de la mémoire selon la fréquence

Cas des basses fréquences Sur l'intervalle $[312\text{MH}z, 1.25\text{GH}z]$, la croissance de la mémoire est illustré par la figure 4.12. Dans ce cas, les courbes représentant la croissance

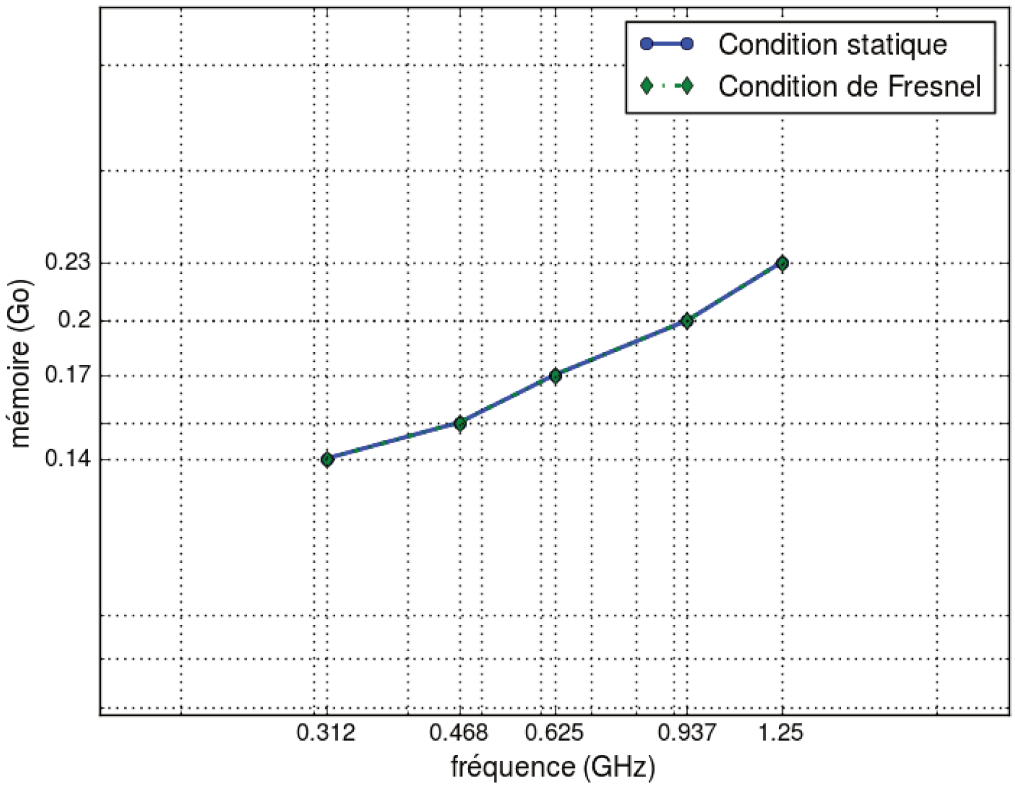


FIGURE 4.12 – Croissance de la mémoire d'un bloc de la matrice de discrétisation en fonction de la fréquence (échelle logarithmique).

de la mémoire pour la condition statique et la condition fréquentielle de Fresnel sont confondues. En effet, à basse fréquence le critère fréquentiel n'intervient pas. Cette croissance est d'une pente inférieure à l'unité et possède une croissance très lente. On poursuit l'expérience en utilisant des fréquences plus élevées afin de s'assurer que la croissance asymptotique de la mémoire est correcte.

Cas des hautes fréquences On observe la croissance de la mémoire sur une plage de fréquences plus large. La figure 4.13 illustre cette croissance entre $312\text{MH}z$ et $15\text{GH}z$. Par rapport aux basses fréquences, la pente a augmenté et l'on note une croissance asymptotique linéaire aux hautes fréquences. Ce comportement est celui attendu d'après l'analyse sur la largeur de bande effectuée au paragraphe 4.4.2.

Sans compression, le bloc matriciel correspondant à cette interaction aurait une taille mémoire d'environ $743\text{G}o$.

Le critère fréquentiel donne une moins bonne approximation pour la fréquence $f = 2.5\text{GH}z$ ce qui est dû à la façon dont on découpe hiérarchiquement les blocs de la matrice. En effet, la méthode de *clustering* utilisée dans les $\mathscr{H}$-matrices est indépendante de la fréquence et du critère directionnel. Ainsi, il se peut que l'on soit amené à découper une boîte dans une mauvaise direction à cause du critère.

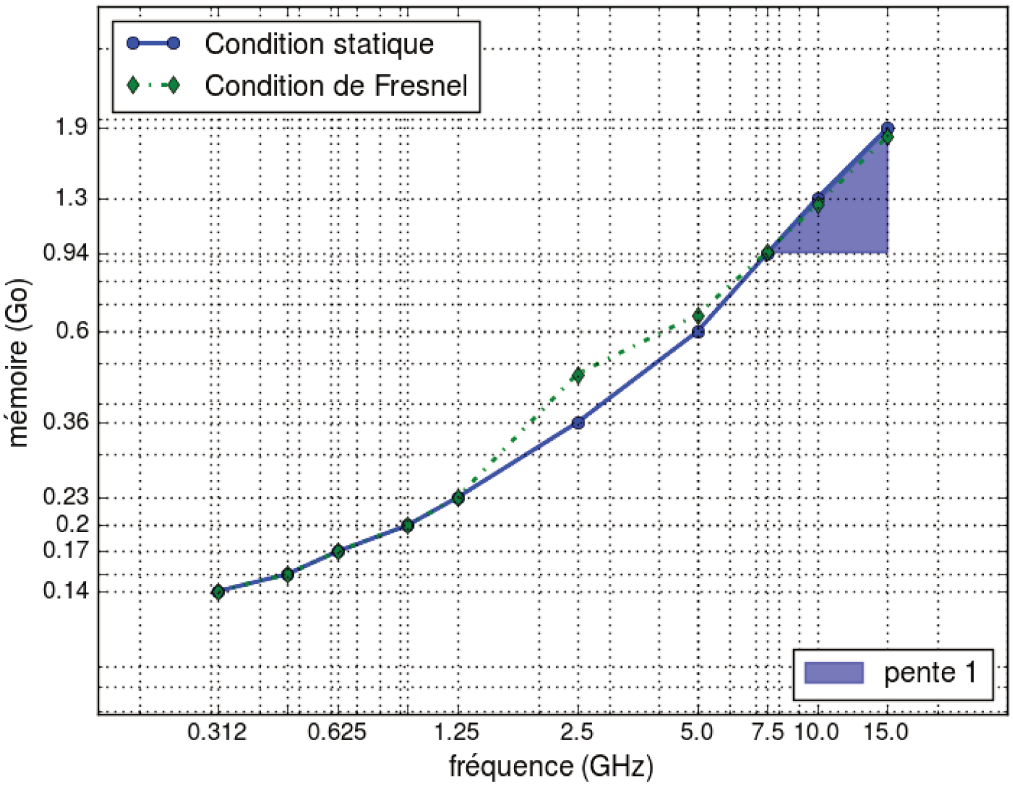


FIGURE 4.13 – Croissance de la mémoire d'un bloc de la matrice de discrétisation en fonction de la fréquence (échelle logarithmique). Le triangle bleu illustre une croisance linéaire.

Si une interaction n'est pas admissible pour le critère d'admissibilité fréquentiel directionnel il se peut que l'on soit obligé de traiter les interactions d'un niveau plus haut (des blocs matriciels plus petits) car

Pour les basses fréquences, les courbes sont confondues car le critère fréquentiel est conforme au critère statique : la $\mathscr{H}$-matrice fréquentielle a la même structure. Dans une zone pré-asymptotique, le critère de Fresnel est moins efficace que le critère statique car il a découpé de manière abusive des blocs matriciels en fonction des directions du *clustering* et non dans une direction qui tend à réduire la section efficace. Cette dernière n'est donc pas forcément réduite en un seul niveau et plusieurs découpes successives conduisent à une approximation moins bonne. C'est le cas du point correspondant à la fréquence $f = 2.5\text{GH}z$.

Un bloc à la limite de la zone d'admissibilité sera mal découpé plutôt que conservé intact ce qui augmente la mémoire. Il s'agit d'une piste d'amélioration supplémentaire de la méthode. Dans le cas présent, la mémoire des sous-blocs (37Mo) est plus importante que celle du bloc intact (24Mo). Par rapport à la taille originale, ces tailles sont néanmoins très faibles et le critère de Fresnel fournit un résultat similaire au critère d'admissibilité statique.
Pour la plus haute fréquence la compression par l'algorithme ACA+ avec le critère statique fournit une approximation d'environ 152Mo. La découpe du bloc à l'aide du critère de Fresnel et l'assemblage des sous-blocs également par ACA+ fournit une mémoire totale de 145Mo. Dans ce cas, le gain apporté par le critère fréquentiel est infime.

Remarque 4.12 (Sorties avec les rangs). Ne pas mettre en remarque. Comparaison entre les rang et une decoupe. On peut ajouter ceci au critère d'admissibilité.

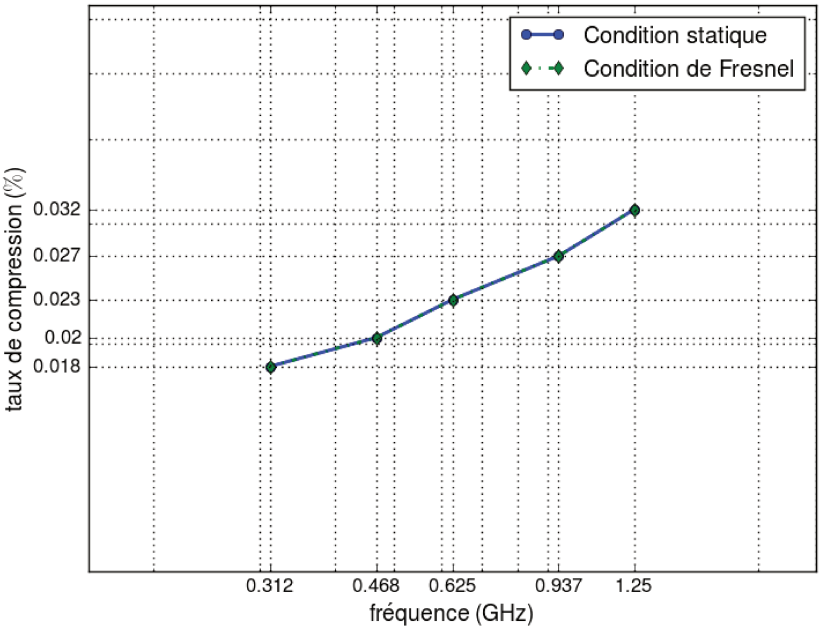
4.4.3.3 Taux de compression selon la fréquence

On représente les taux de compressions (4.81) et (4.82) décrits précédemment, pour une plage limitée aux basses fréquences puis pour les hautes fréquences.

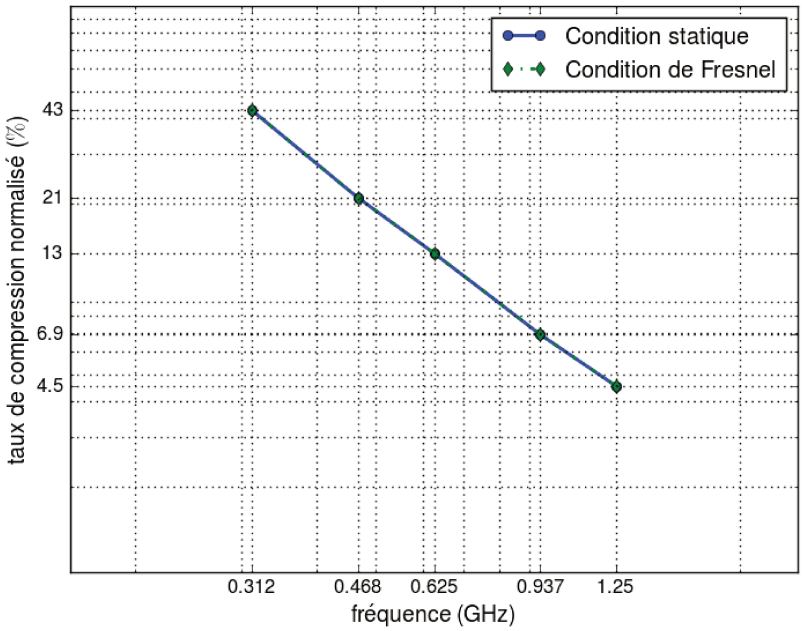
Cas des basses fréquences Les figures 4.14a et 4.14b montrent les taux de compressions pour chaque critère en fonction de la fréquence. À basse fréquence, les deux critères sont confondus. Le taux de compression défini par (4.81) reste petit mais n'est pas représentatif de la performance de la compression à cause du phénomène de surmaillage. Le taux normalisé (4.82) est un meilleur indicateur. Pour cet indicateur, le taux de compresson pour la fréquence la plus basse est de 43% et atteint 4.5% pour une fréquence de 1.25GHz.

Cas des hautes fréquences Pour les hautes fréquences, la croissance asymptotique attendue et mesurée est la même que celle de la mémoire. On observe une croissance linéaire avec la fréquence pour les deux critères d'admissibilité sur la figure 4.15a. La figure 4.15b présente une décroissance en $\mathscr{O}(k^{-1})$ du taux de croissance normalisé, conformément aux estimations théoriques.

Remarque 4.13 (Choix du paramètre ν). Dans le cas des plaques coplanaires, le choix du paramètre ν dans le critère de Fresnel paraît peu important. On peut choisir un paramètre loin de la valeur optimale et obtenir des résultats corrects. En réalité, tout se vaut et le rang de l'interaction est intrinsèquement faible. Pour observer une différence notable, il faudrait considérer un objet énorme en terme de longueurs d'onde. Pour les applications

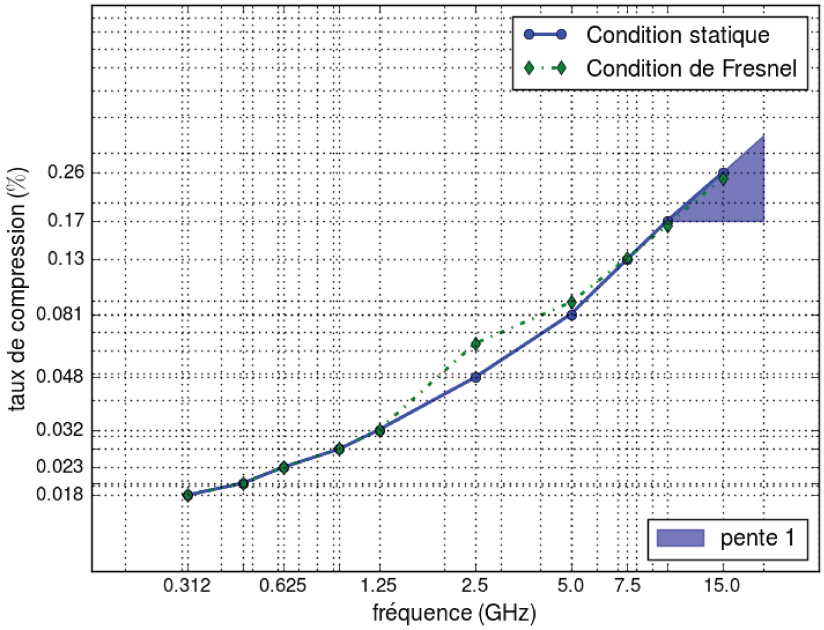


(a) Taux de compression en fonction de la fréquence (échelle logarithmique).

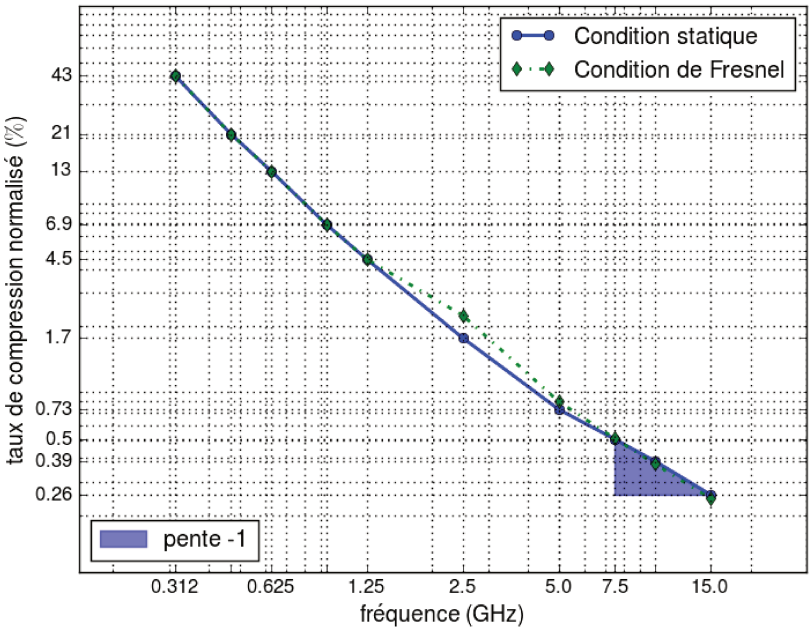


(b) Taux de compression en fonction de la fréquence (échelle logarithmique).

FIGURE 4.14 – Comportement à basse fréquence des différents taux de compression.



(a) Taux de compression en fonction de la fréquence (échelle logarithmique). Le triangle bleu illustre une croissance linéaire.



(b) Taux de compression normalisé en fonction de la fréquence (échelle logarithmique). Le triangle bleu illustre une décroissance linéaire avec la fréquence.

FIGURE 4.15 – Comportement à haute fréquence des différents taux de compression.

traitées à ce jour, on ne peut exhiber de cas où le critère de Fresnel améliore considérablement les choses. Dans la pratique, on peut choisir une valeur de ν plus grande pour « coller » le plus possible au critère statique.

4.4.4 Cas de deux plaques opposées

4.4.4.1 Géométrie

On considère à présent deux plaques carrées de côté $a = 1m$ opposées comme deux faces opposées d'un cube. Comme dans le cas des plaques coplanaires, ces deux plaques sont espacées de telle sorte à ce que la distance minimale soit R = $\sqrt{2}/2$. Ainsi, ces deux plaques sont admissibles pour la condition d'admissibilité statique avec $\eta = 2$.

4.4.4.2 Évolution de la mémoire selon la fréquence

Cas des basses fréquences Sur l'intervalle $[156\text{MH}z, 1.25\text{GH}z]$, la croissance de la mémoire est la suivante Dans ce cas, la condition fréquentielle de Fresnel s'exprime plus

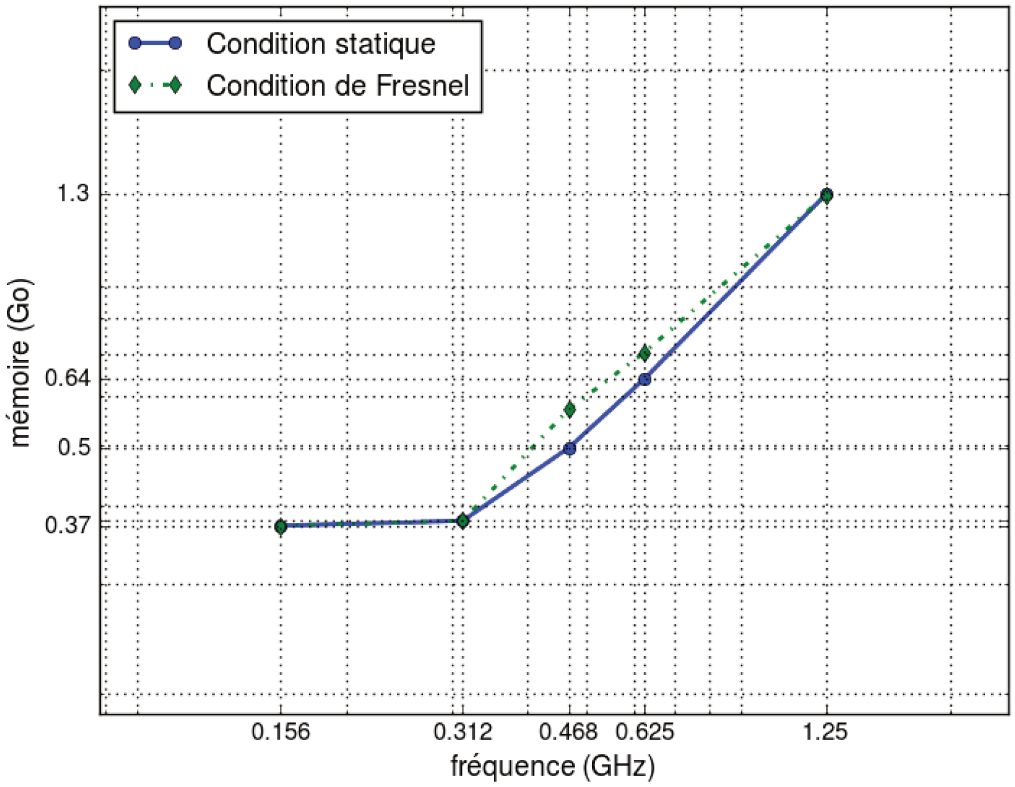


FIGURE 4.16 – Croissance de la mémoire d'un bloc de la matrice de discrétisation en fonction de la fréquence (échelle logarithmique).

rapidement que dans le cas précédent mais pour des basses fréquences, le critère statique fournit une approximation dont la mémoire est légèrement mieux. Ceci est dû à une découpe du bloc matriciel infructueuse pour les fréquences considérées. Pour la fréquence de $1.25\text{GH}z$, les résultats obtenus à l'aide des deux critères sont les mêmes et le désavantage observé auparavant n'apparaît plus. La croissance de la mémoire pour les deux critères semble être linéaire ce qui est un résultat mieux que ce que l'on avait prévu avec le calcul de la largeur de bande au paragraphe 4.4.2. Une étude plus minutieuse en augmentant la fréquence conduit aux résultats suivants,

Cas des hautes fréquences La figure 4.17 suivante illustre la croissance sur l'intervalle de fréquence plus large $[156\text{MH}z, 15\text{GH}z]$. À partir de $1.25\text{GH}z$, on constate que les deux cri-

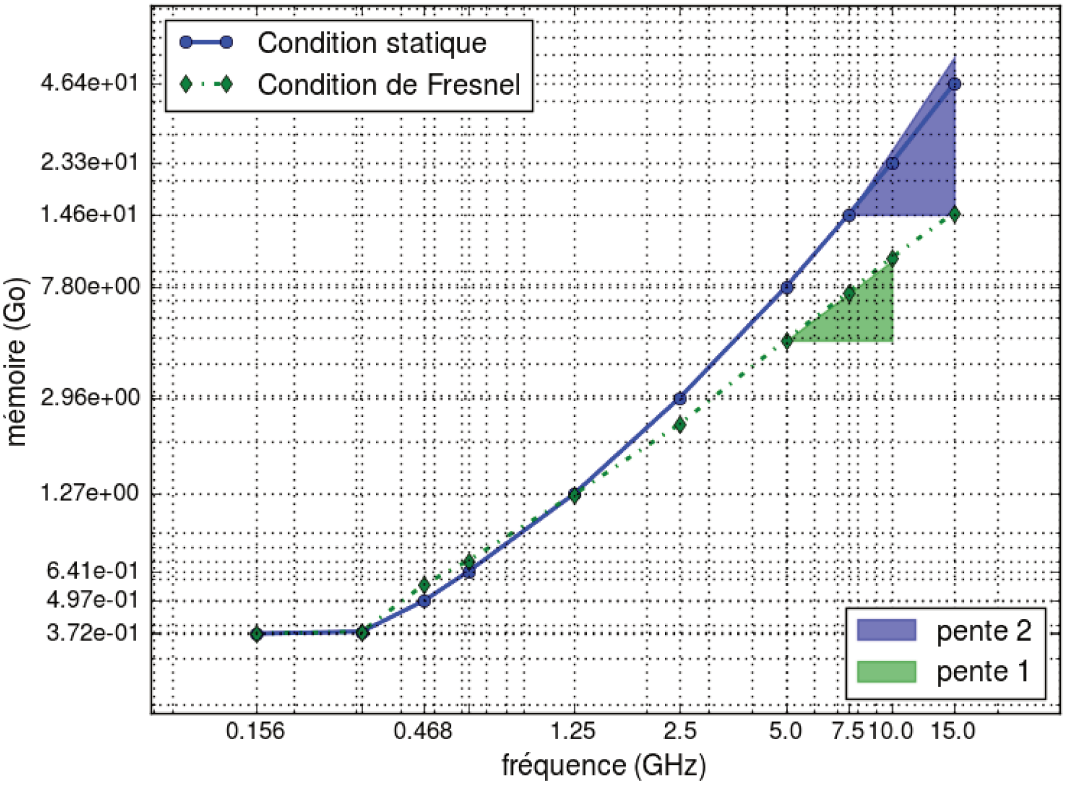


FIGURE 4.17 – Croissance de la mémoire d'un bloc de la matrice de discrétisation en fonction de la fréquence (échelle logarithmique). Le triangle bleu illustre une croisance quadratique tandis que le triangle vert illustre une croissance linéaire.

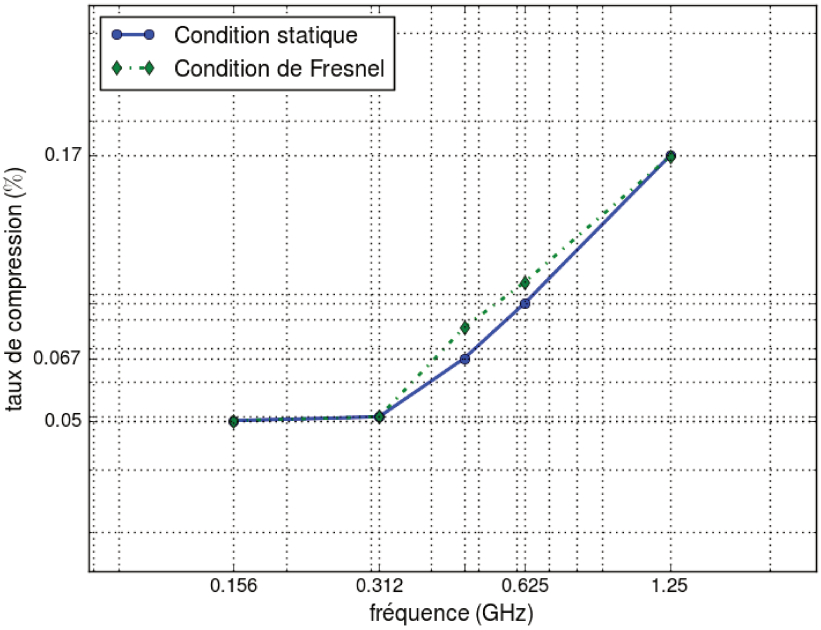
tères fournissent des résultats différents. Sur la figure 4.17, pour la plus haute fréquence, la mémoire requise par l'approximation avec le critère statique est d'environ $45\text{G}o$ tandis que le bloc est représenté seulement à l'aide de $15\text{G}o$ avec le critère de Fresnel. Le rapport 1/3 est plus conséquent dans ce cas que dans le cas des plaques coplanaires.

Le comportement du critère statique est bien celui prévu par la théorie soit une croissance en $\mathscr{O}(k^2)$ tandis que l'emploi du critère de Fresnel conduit à une croissance en $\mathscr{O}(k)$ qu'elle que soit la fréquence. *Dans le cas du critère statique, cela conduit à une temps d'assemblage ainsi qu'un croissance de la mémoire en $\mathscr{O}(\text{N}^2)$ ce qui met en échec le caractère rapide de la méthode lorsqu'employée hiérarchiquement.*

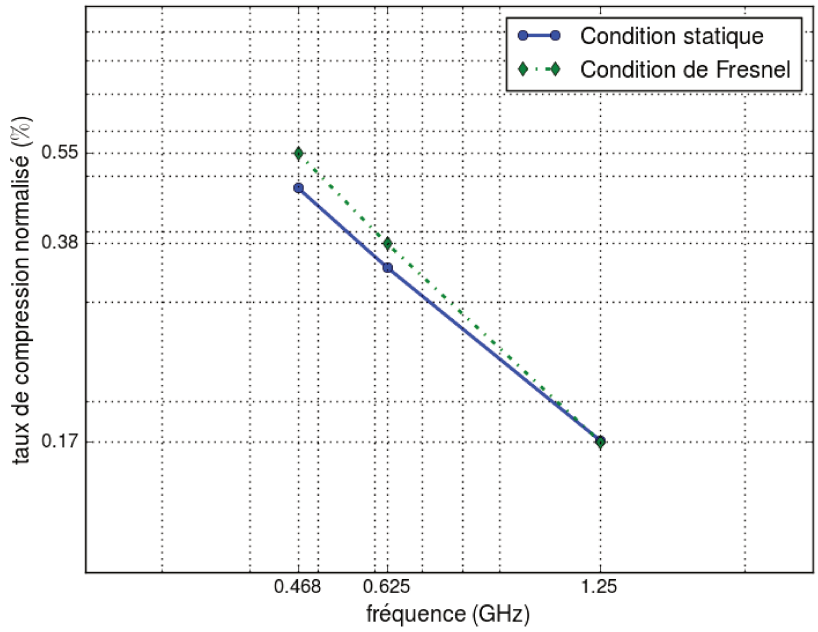
4.4.5 Taux de compression selon la fréquence

Cas des basses fréquences Pour les basses fréquences, on ne représente pas le taux normalisé pour les plus basses fréquence car la formule que l'on utilise perd de son sens pour un nombre de degrés de liberté très petit. Les taux de compression se comportent comme attendu et comme nous l'avons déjà observé, le critère statique semble légèrement favorable à basse fréquence. En réalité, le critère de Fresnel est très proche du critère statique à basse fréquence et la décision de subdiviser ou non le bloc est prise par rapport à un critère binaire. Aussi, une valeur légéèrement différente du paramètre d'admissibilité peut changer la découpe à basse fréquence. Dans la pratique, on se concentre sur des applications à haute fréquence.

Cas des hautes fréquences Pour les hautes fréquences, le taux de compression normalisé décroit comme $\mathscr{O}(k^{-1})$ ce qui est un élément important d'une méthode rapide. Le critère statique conduit à un taux normalisé qui devient constant. En d'autres termes, au-delà d'une certaine fréquence, le taux de compression est constant et la méthode cesse d'être intéressante car l'approximation est une partie constante d'une quantité augmentant asymptotiquement comme $\mathscr{O}(k^4)$.
On peut également observer une comparaison de ces deux décroissances sur le tableau 4.4 qui illustre la pente glissante déterminée pour chaque intervalle de croissance. Dans le cas statique, cette pente augmente et tend vers zéro avec la fréquence tandis que pour le critère de Fresnel la pente reste constamment égale à -1.

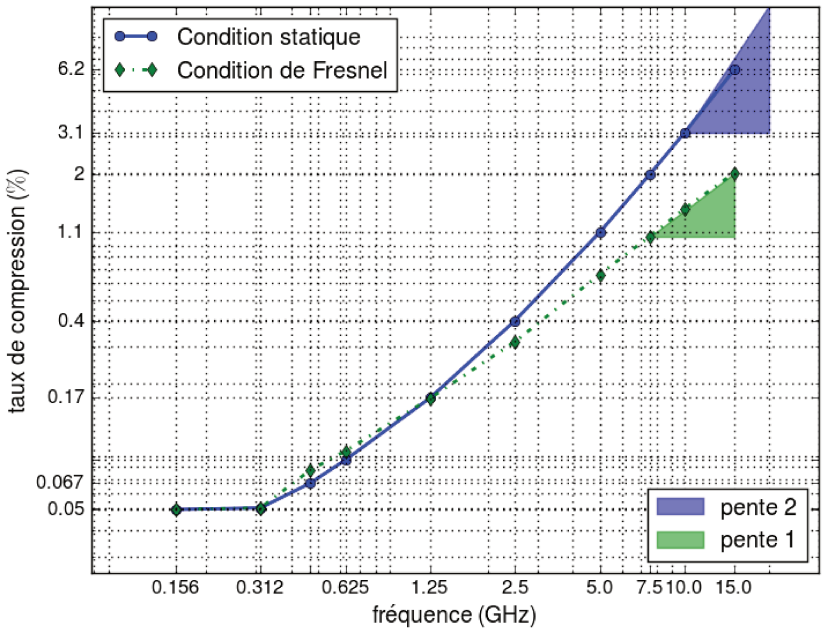


(a) Taux de compression en fonction de la fréquence (échelle logarithmique).

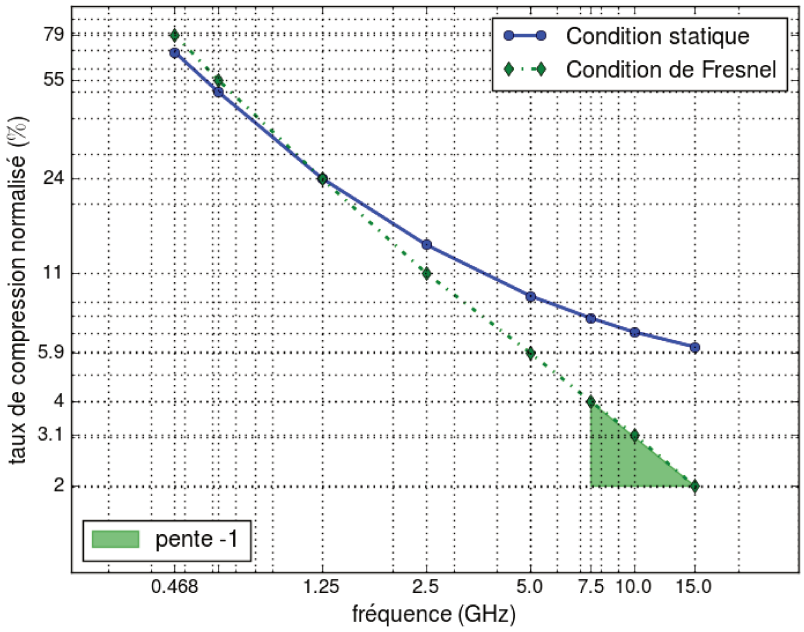


(b) Taux de compression normalisé en fonction de la fréquence (échelle logarithmique).

FIGURE 4.18 – Comportement à basse fréquence des différents taux de compression.



(a) Taux de compression en fonction de la fréquence (échelle logarithmique). Le triangle bleu illustre une croissance quadratique tandis que le triangle vert représente une criossance linéaire avec la fréquence.



(b) Taux de compression normalisé en fonction de la fréquence (échelle logarithmique). Le triangle vert illustre une décroissance linéaire avec la fréquence.

FIGURE 4.19 – Comportement à haute fréquence des différents taux de compression.

Intervalle (GHz)	Condition statique	Condition de Fresnel
[0.468, 0.625]	-1.11	-1.27
[0.625, 1.25]	-1.01	-1.16
[1.25, 2.5]	-0.77	-1.10
[2.5, 5.0]	-0.60	-0.93
[5.0, 7.5]	-0.44	-0.97
[7.5, 10.0]	-0.39	-0.93
[10.0, 15.0]	-0.29	-1.03

TABLEAU 4.4 – Pente glissante mesurée par intervalle de fréquence.

4.4.6 Conséquences sur les $\mathscr{H}$- matrices à haute fréquence

Les deux configurations de plaques testées montrent l'influence d'un critère d'admissibilité fréquentiel. Les résultats sont de nature très différentes selon la section efficace. Dans le cas des plaques coplanaires, la section efficace se résume à un segment et la croissance du rang est en $\mathscr{O}(k)$ et croît très lentement. Par conséquent, le taux de compression est généralement bon pour ces interactions et nous n'avons pu observer de cas où cette croissance devenait un frein à la compression. Les tests effectués sur cette géométrie suggèrent que les interactions planes gagneraient à être traitées seulement à l'aide d'une admissibilité statique. Le critère fréquentiel n'apporte pas d'amélioration dans ce cas.

Au contraire, le cas des plaques opposées illustre parfaitement le gain apporté par un critère fréquentiel dès que la section efficace est grande. Pour les problèmes de petite taille (*cf* 4.16) la mémoire croît de manière linéaire avec la fréquence comme le cas des plaques coplanaires. Cependant, il ne s'agit ici que d'un comportement transitoire et la croissance devient quadratique avec la fréquence (*cf* 4.17). Cette croissance conduit à une importante dégradation des performances de compression à haute fréquence. En effet, le taux de compression tend vers une constante car le rang possède la même asymptotique que le nombre de degrés de liberté (*cf* 4.19b). Il s'agit d'un contre-exemple pour une méthode rapide et dans ce cas, l'introduction d'un critère d'admissibilité fréquentiel permet d'obtenir une croissance linéaire en la fréquence. Les tests réalisés sur une géométrie de taille modeste avec seulement 215506 degrés de liberté illustre parfaitement ce gain. On s'attend à observer des gains plus importants pour des objets de taille plus importante grâce au critère d'admissibilité.

4.4.6.1 Cas limites, section efficace

Les deux cas des plaques ci-dessus représentent des cas limites pour lesquels le critère d'admissibilité de Fresnel fournit un élément de réponse sur la croissance du rang. La notion fondamentale est celle de section efficace. La croissance en mémoire en fonction de la fréquence de l'interaction de deux objets dépend de la section efficace avec laquelle ils se regardent. Le cas des plaques coplanaires illustre le cas d'une section efficace dégénérée et faible. Au contraire, le cas très défavorable des plaques en opposition illustre les interactions d'objets se voyant avec une section efficace maximale. L'exemple suivant montre l'importance de ces cas limites pour comprendre le comportement général lors de l'assemblage complet d'une $\mathscr{H}$-matrice.

4.4.6.2 Exemple d'un ellipsoïde

On illustre la distribution des interact ions admissibles présentes dans le cas général de l'approximation de la matrice de l'EFIE par une $\mathscr{H}$-matrice sur un ellipsoïde décrit par la figure 4.20. L'assemblage de l'approximation $\mathscr{H}$-matrice est effectuée avec les

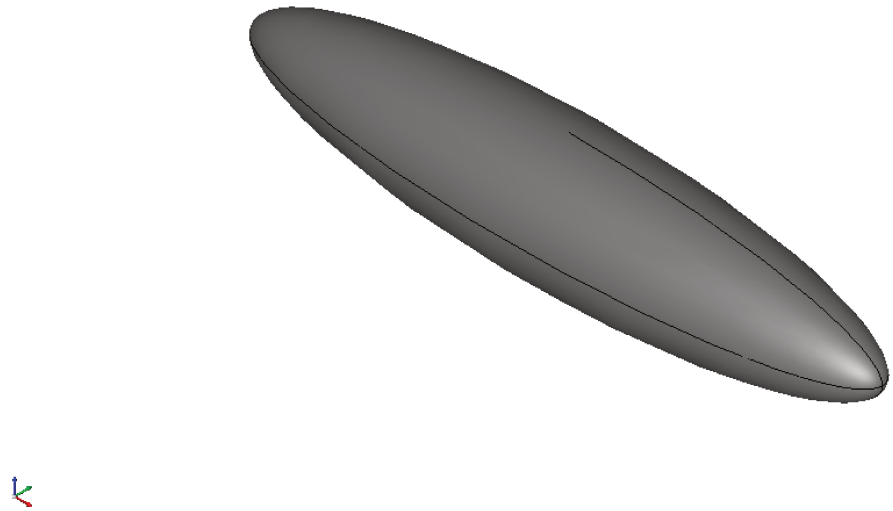


FIGURE 4.20 – Cas de l'ellipsoïde allongé.

paramètres suivants.

Géométrie	Ellispoïde allongé de taille $10m \times 2m \times 2m$.
Nombre de degrés de liberté	N = 193707
Algorithme de compression	ACA+
Assemblage	$\epsilon_{ACA} = 10^{-4}$
Recompression	$\epsilon = 10^{-4}$
Admissibilité	Condition statique avec $\eta = 2$

TABLEAU 4.5 – Paramètres de l'ellipsoïde et de l'assemblage $\mathscr{H}$-matrice.

4.4.6.3 Comportement des blocs admissibles

Dans le cadre de l'admissibilité de Fresnel, la section efficace est portée par les directions u_1 et u_2 de la base d'interaction mutuelle. On note d et D respectivement la plus petite et la plus grande des dimensions de la section efficace. On appelle rapport d'aspect la quantité d/D. Par section efficace on entend l'aire dD. Un rapport d'aspect proche de zéro correspond à une interaction modélisée par les plaques coplanaires tandis qu'un rapport d'aspect proche de l'unité correspond aux plaques opposées. La croissance en fréquence de la mémoire dépend également de la section efficace. En effet, une interaction avec un rapport d'aspect proche de l'unité mais de section eficace faible correspond à une croissance en fréquence modérée. Les blocs dont la croissance est très importante (*ie* quadratique) correspondent à des rapports d'aspect proches de l'unité ainsi qu'une importante section efficace.

La figure 4.21 décrit le rang d'un bloc admissible en fonction du rapport d'aspect et de la section efficace.

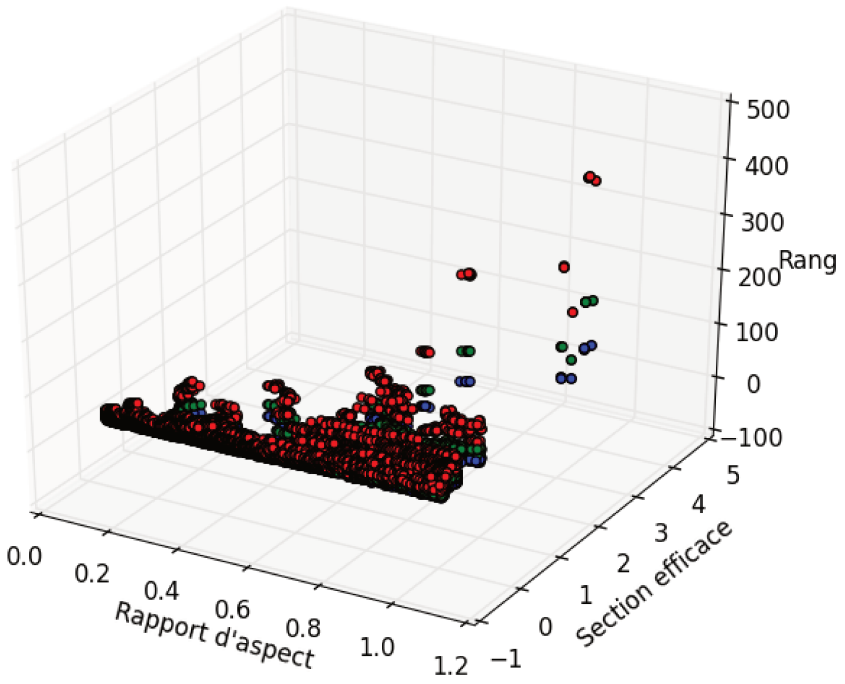
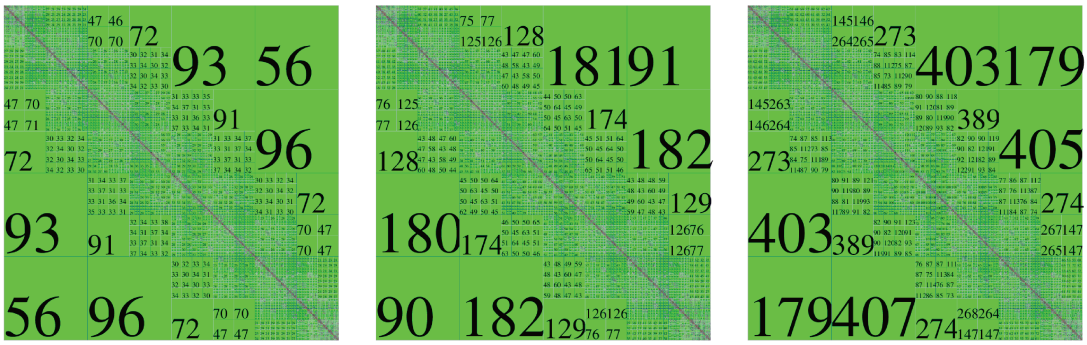


FIGURE 4.21 – Distribution des rangs en fonction du rapport d'aspect et de la section efficace pour trois fréquences différentes. En rouge, la fréquence 2GHz. En vert, la fréquence 1GHz et en bleu la fréquence 500MHz.

On constate sur la figure 4.21 toute la diversité des interactions retrouvées au sein de la méthode des $\mathscr{H}$-matrices. Tous les rapports d'aspect sont présents sur cette géométrie et l'on note la présence d'interactions proches du cas favorable des plaques coplanaires. Il s'agit surtout d'interactions entre des petites boîtes et sont situées en bas dans l'arbre des blocs.
Les gros blocs présents dans cette géométrie ont un rapport d'aspect proche de l'unité ainsi qu'une section efficace importante ce qui correspond au cas des plaques opposées. Les moyens informatiques à disposition actuellement ne permettent pas de montrer l'explosion du rang pour les gros blocs. On s'attend à plus haute fréquence à une croissance quadratique du rang et donc à une dégradation des performances de la méthode des $\mathscr{H}$-matrices. La présence de ces gros blocs est due au critère employé pour l'étape de *clustering*. Une piste d'amélioration de la méthode serait de construire une découpe plus adaptée à un critère fréquentiel. La complexité de ces gros blocs représentant l'interaction de deux grandes parties de la géométrie est cruciale pour l'évaluation de la complexité générale de la méthode des $\mathscr{H}$-matrices sur le cas complet. L'analyse effectuée dans le cas des plaques opposées suggère que la méthode générale ne peut avoir une complexité logarithmique.
Dans la pratique et pour des cas modérés, ces blocs sont néanmoins dans une phase pré-asymptotique et leur croissance en fréquence est plus linéaire que quadratique. C'est

d'ailleurs ce que l'on constate sur la figure 4.22 qui représente l'approximation $\mathscr{H}$-matrice du problème pour le critère statique pour diverses fréquences.



(a) $f = 500\text{MH}z$ (b) $f = 1\text{GH}z$ (c) $f = 2\text{GH}z$

FIGURE 4.22 – Représentation graphique du rang des blocs de l'approximation $\mathscr{H}$-matrice de la matrice de l'EFIE sur l'ellipsoïde pour des fréquences différentes.

La fréquence de $2\text{GH}z$ correspond à la fréquence maximale que l'on peut utiliser sur ce maillage contenant N = 193707 degrés de liberté. Sur la figure 4.22, le bloc dans le coin supérieur droit présente une croissance linéaire en la fréquence malgré une section efficace maximale. Il s'agit bien des résultats constatés dans la phase pré-asymptotique pour les plaques opposées. Le bloc voisin à gauche possède lui une croissance légèrement plus rapide. Ceci est du au fait que ce bloc décrit une interaction entre des blocs géométriquement plus proches. Le rang est ainsi plus élevé et possède une croissance plus rapide car plus proche du critère statique que l'autre bloc.

4.5 Contrôle de l'erreur d'approximation $\mathscr{H}$- matrice

On observe le comportement de l'algorithme HCA – II décrit précédemment sur un cas pratique. On souhaite notament pouvoir contrôler l'erreur d'approximation d'un bloc ainsi que le nombre de points d'interpolation utlisé. Enfin, on observe sur un produit matrice/vecteur l'influence de l'erreur commise par l'algorithme HCA – II .

4.5.1 Erreur relative commise par blocs

Le test consiste à représenter l'erreur relative commise lors de l'assemblage par HCA – II et calculée à l'aide d'une approximation de la norme spectrale. Pour chaque bloc admissible $\text{A}_{s\times t}$ de taille $m \times n$, on peut utilise à la fois les dimensions de la matrice ainsi que les diamètres des boîtes englobantes associées aux supports des inconnues constituant l'interaction $s \times t$. Par commodité, on utilise la taille moyenne $\sqrt{mn}$ d'un bloc de taille $m \times n$. Pour les diamètres, on utilise de la même façon le diamètre moyen normalisé par la longueur d'onde, soit $\sqrt{\text{diam}(\text{Q}_s^{(u)}).\text{diam}(\text{Q}_s^{(u)})}/\lambda$. Le cas test présenté ci-après est celui d'une sphère avec les paramètres suivants.

Géométrie	Sphère de rayon $a = 1m$
Nombre d'inconnues	N = 192000
Admissibilité	Condition de Fresnel avec $\eta = 2$ et $\nu = 2$
Algorithme de compression	HCA – II fréquentiel
Précision d'assemblage	$\epsilon_{HCA} \in \{10^{-3}, 10^{-4}, 10^{-5}, 10^{-6}\}$
Recompression	$\epsilon = \epsilon_{HCA}$
Fréquence	$f = 3.98Ghz$ (correspondant à une taille d'arête en $\lambda/5$).

TABLEAU 4.6 – Configuration du cas test du contrôle de l'erreur relative d'assemblage par blocs.

Les figures 4.23 illustrent l'erreur relative commise sur chaque bloc approché par l'algorithme HCA – II. On note que l'erreur commise respecte la tolérance que l'on s'impose. On se doit de garder à l'esprit que l'erreur représentée est une erreur approchée. Ainsi, le fait que l'erreur maximale commise soit légèrement supérieure à la tolérance spécifiée n'est pas un frein pour l'utilisation de l'algorithme HCA – II fréquentiel décrit.
Pour chaque tolérance spécifiée, on note que l'amplitude de l'estimation d'erreur des blocs correspondant à des petites boîtes, on note que l'amplitude de l'estimation d'erreur est plus large que pour les grands diamètres. Les blocs correspondant à des grands diamètres représentent des interactions à haute fréquence. Pour ces blocs, le choix du nombre de points par la formule de Landau-Widom est de plus en plus précise et l'on observe une bonne estimation de l'erreur. Pour ce test, les plus grands blocs représentent les premières interactions admissibles pour le critère fréquentiel et sont situés à une profondeur de 4 dans l'arbre des blocs. Notre algorithme est donc capable d'approcher correctement les interactions haute fréquence pour une tolérance donnée.

4.5.2 Nombre de nœuds d'interpolation

Le test précédent montre que pour un bloc admissible pour le critère de Fresnel, l'algorithme HCA – II fréquentiel fournit une approximation dont on contrôle correctement l'erreur. Cette approximation utilise un schéma d'interpolation polynomiale et l'on souhaite utiliser un nombre de nœuds le plus petit possible d'une part pour des raisons de complexité et d'autre part car le nombre de points disposés dans les boîtes englobantes $\mathrm{Q}_t^{(u)}$ et $\mathrm{Q}_t^{(u)}$ fournit par construction une borne sur le rang. La maîtrise du nombre de points d'interpolation utilisés en fonction de la fréquence est donc un point important de la méthode.

4.5.2.1 Nombre de nœuds

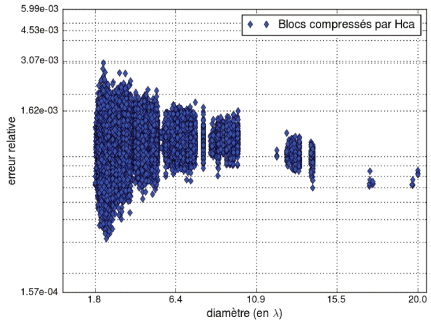
Pour chaque bloc $\mathrm{A}_{s\times t}$ admissible, l'algorithme HCA – II requiert le calcul d'une base d'interaction mutuelle (u_1,u_2,u_3). Le noyau de Green est alors déconvolué par les ondes planes dans la direction u_3 et l'on exploite les propriétés des termes du second ordre à travers la formule de Landau-Widom. Pour la direction u_1 (resp. u_2), on détermine la largeur de bande c_1(resp. c_2) puis l'ordre d'interpolation $\mathrm{N}_1(c_1,\epsilon)$ (resp. $\mathrm{N}_2(c_2,\epsilon)$) dans cette direction donné par la formule de Landau-Widom,

$$\mathrm{N}_1(c_1,\epsilon) = \frac{2c_1}{\pi} + \frac{1}{\pi^2}\log\left(\frac{1-\epsilon^2}{\epsilon^2}\right)\log(c_1). \tag{4.84}$$

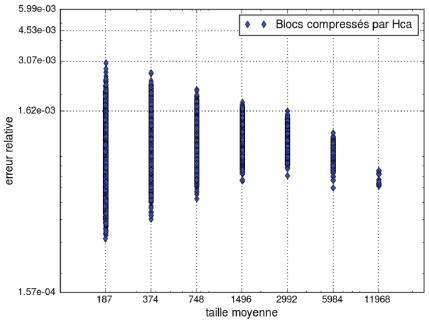
L'ordre choisi dans la direction u_1 (resp. u_2) est alors $\max(\mathrm{N}_\epsilon,\mathrm{N}_1)$ (resp. $\max(\mathrm{N}_\epsilon,\mathrm{N}_2)$.

4.5.2.2 Description du test

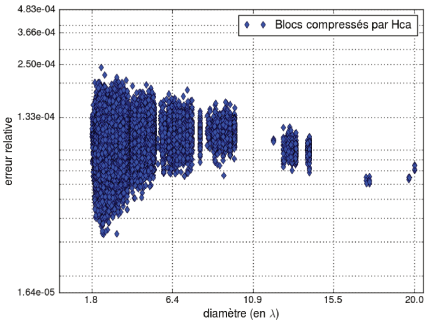
Le test consiste à représenter le rapport entre le rang numérique de l'approximation et le nombre total de points $\mathrm{M} = \mathrm{N}_\epsilon\mathrm{N}_1(c_1,\epsilon)\mathrm{N}_2(c_2,\epsilon)$ en fonction de la largeur de bande moyenne $\sqrt{c_1c_2}$. Pour le cas de la sphère, on assemble une approximation de la matrice de l'EFIE avec les paramètres suivants.



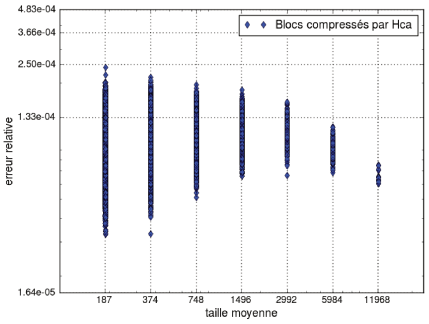
(a) $\epsilon = 10^{-3}$



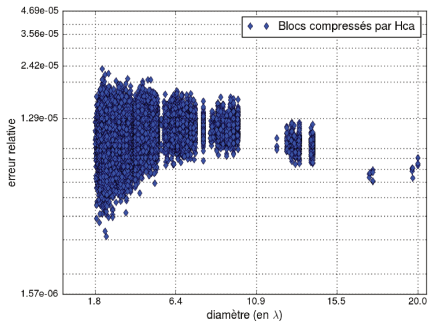
(b) $\epsilon = 10^{-3}$



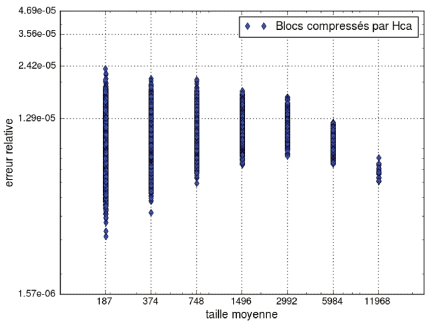
(c) $\epsilon = 10^{-4}$



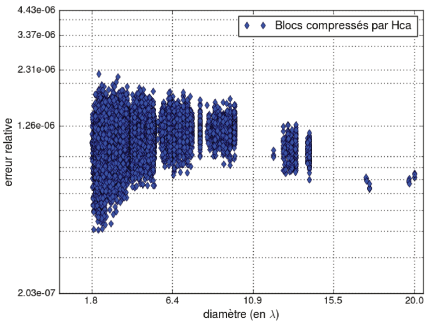
(d) $\epsilon = 10^{-4}$



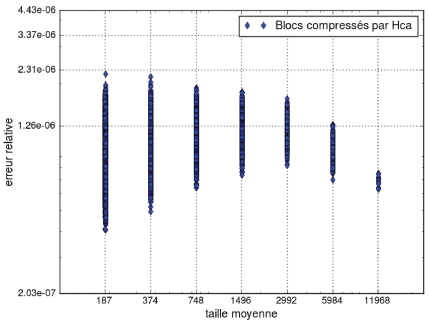
(e) $\epsilon = 10^{-5}$



(f) $\epsilon = 10^{-5}$



(g) $\epsilon = 10^{-6}$



(h) $\epsilon = 10^{-6}$

FIGURE 4.23 – Représentation des erreurs relatives par blocs sur la sphère en fonction du diamètre en longueurs d'onde (colonne de gauche) et de la taille moyenne (colonne de droite).

Géométrie	Sphère de rayon $a = 1m$
Nombre d'inconnues	N = 300000
Admissibilité	Condition de Fresnel avec $\eta = 2$ et $\nu = 2$
Algorithme de compression	HCA – II fréquentiel
Précision d'assemblage	$\epsilon_{HCA} = 10^{-5}$
Recompression	$\epsilon = 10^{-5}$
Fréquence	$f = 4.98Ghz$ (correspondant à une taille d'arête en $\lambda/5$).

TABLEAU 4.7 – Configuration du cas test du contrôle du nombre de nœuds d'interpolation.

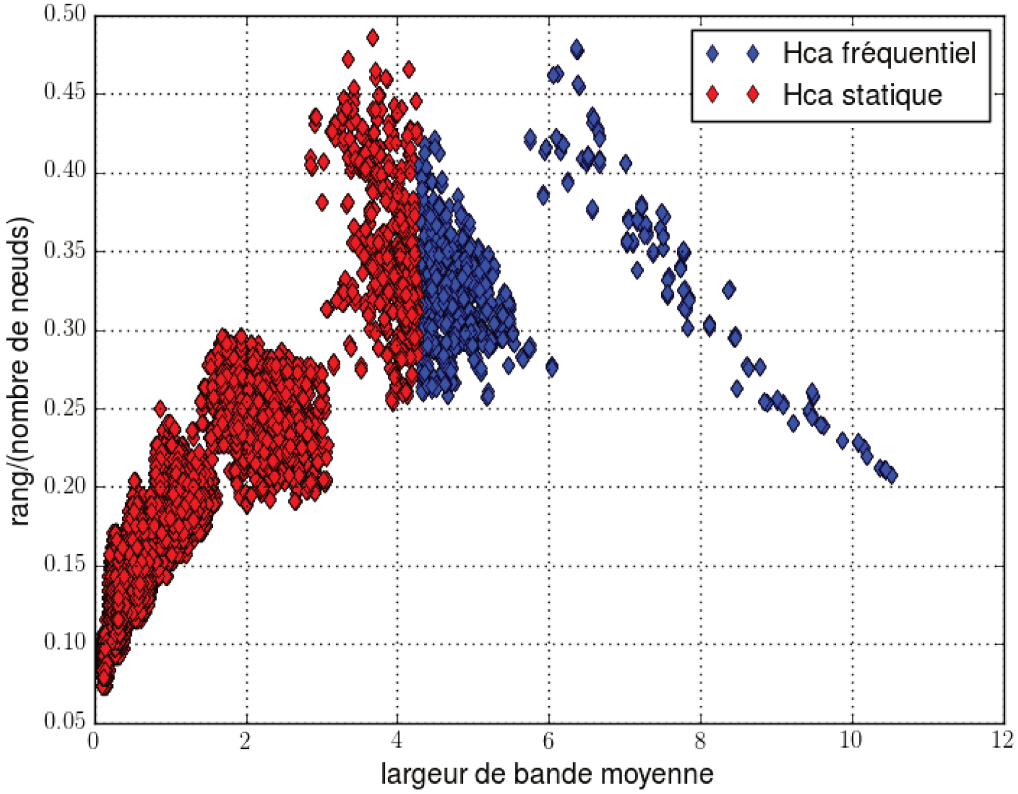


FIGURE 4.24 – Rapport entre le rang numérique à la précision $\epsilon = 10^{-5}$ et le nombre total de nœuds d'interpolation utilisés en fonction de la largeur de bande moyenne. En rouge les interactions correspondant à un choix des nœuds par la partie statique seule. En bleu, les interactions pour lesquelles on utilise la formule de Landau-Widom.

Sur le graphe 4.24, on retrouve deux comportements attendus, ceux pour les petites et les grandes largeurs de bande. Les petites largeurs de bande correspondent géométriquement à des petites boîtes ou encore à des basses fréquences. Pour ces interactions (en rouge sur la figure), le nombre de points sélectionnés correspond purement à la condition statique.
Pour $\epsilon = 10^{-5}$, cela revient à considérer $N_\epsilon = 6$ dans chaque direction soit au total 216 points d'interpolation dans chaque boîte. Pour des interactions à basse fréquence, le caractère dominant du noyau est celui de $1/|x-y|$ et le rang attendu est faible, de l'ordre de N_ϵ. Ainsi, le rapport observé est d'autant plus faible que la largeur de bande est petite. Ce rapport croît avec la largeur de bande (donc avec la fréquence) jusqu'à atteindre la valeur

1/2. L'algorithme dispose donc de suffisament de points pour produire une approximation de qualité. Lorsque ce rapport se rapproche de l'unité, il est possible que le nombre de points choisi soit trop faible et l'approximation de qualité inférieure.
Pour cet exemple, la formule de Landau-Widom s'exprime quand la largeur de bande est supérieure à $c = 4.26$ (en bleu sur la figure). Pour les grandes largeurs de bande, on s'attend à ce que le rang soit déterminé principalement par les oscillations dans les directions u_1 et u_2. Grâce à la déconvolution par les ondes planes dans la direction u_3, il n'est pas nécessaire de placer un grand nombre de points dans cette direction et le choix de N_ϵ points est plus que nécessaire pour capter les oscillations. Le rapport observé décroît et devient plus faible ce qui confirme ce comportement. La zone intermédiaire correspond à une zone de transition où la partie oscillante n'est pas encore dominante et la formule de Landau-Widom n'est pas encore pleinement utilisée.

4.5.3 Erreur commise sur le produit matrice-vecteur

4.5.3.1 Description du test

On mesure l'erreur de l'algorithme HCA−II fréquentiel sur le produit matrice/vecteur. On assemble l'approximation $\mathscr{H}$-matrice $\tilde{\mathrm{A}}$ de la matrice A de l'EFIE. Pour un vecteur x_i aléatoire donné, on effectue le produit $y_i = \tilde{\mathrm{A}}.x_i$. Pour ce produit, on mesure l'erreur relative E_i par

$$\mathrm{E}_i = \frac{\|\mathrm{A}.x - \tilde{\mathrm{A}}.x\|_2}{\|\mathrm{A}.x\|_2}. \tag{4.85}$$

Pour un nombre n_{rhs} de vecteurs, on observe l'erreur maximale commise $\mathrm{E}_{\max}$, définie par

$$\mathrm{E}_{\max} = \max_{i=1,\ldots,n_{rhs}} \mathrm{E}_i. \tag{4.86}$$

On utilise plusieurs raffinements d'une même sphère de rayon $a = 1m$ avec les paramètres suivants pour l'assemblage $\mathscr{H}$-matrice.

Géométrie	Sphères de rayon $a = 1m$
Nombre d'inconnues	$\mathrm{N} \in \{12000, 48000, 97470, 192000\}$
Admissibilité	Condition de Fresnel avec $\eta = 2$ et $\nu = 2$
Algorithme de compression	HCA − II fréquentiel
Précision d'assemblage	$\epsilon_{\mathrm{HCA}} \in \{10^{-3}, 10^{-4}, 10^{-5}, 10^{-6}\}$
Recompression	$\epsilon = \epsilon_{\mathrm{HCA}}$
Fréquence	$f \in \{1.0\mathrm{GH}z, 1.99\mathrm{GH}z, 2.84\mathrm{GH}z, 3.98\mathrm{GH}z\}$ (correspondant à une taille d'arête en $\lambda/5$).
Nombre de seconds membres	$n_{rhs} = 1000$.

TABLEAU 4.8 – Configuration du cas test pour le contrôle de l'erreur sur le produit matrice/vecteur.

On effectue deux tests distincts. Le premier consiste à utiliser un maillage adapté à la fréquence d'étude. C'est le cas idéal dans la pratique mais l'on ne peut pas toujours avoir le maillage parfaitement adapté. Pour cette raison, le second test consiste à utiliser un seul maillage pour plusieurs fréquences d'étude et observer le comportement de l'algoritme HCA − II lors du produit matrice/vecteur.

4.5.3.2 Maillage adapté

On mesure l'erreur relative maximale commise sur le produit matrice/vecteur en fonction de la fréquence. Pour un chaque maillage, on se place à la fréquence correspondante à une finesse du maillage en $\lambda/5$ et l'on effectue n_{rhs} produits matrice/vecteur. La figure 4.25 montre que l'erreur commise respecte la tolérance que l'on s'est fixé pour l'assemblage de l'approximation $\mathscr{H}$-matrice. L'algorithme HCA−II nous permet d'obtenir de bons résultats stables en fréquence pour des précisons fines comme $\epsilon = 10^{-6}$.

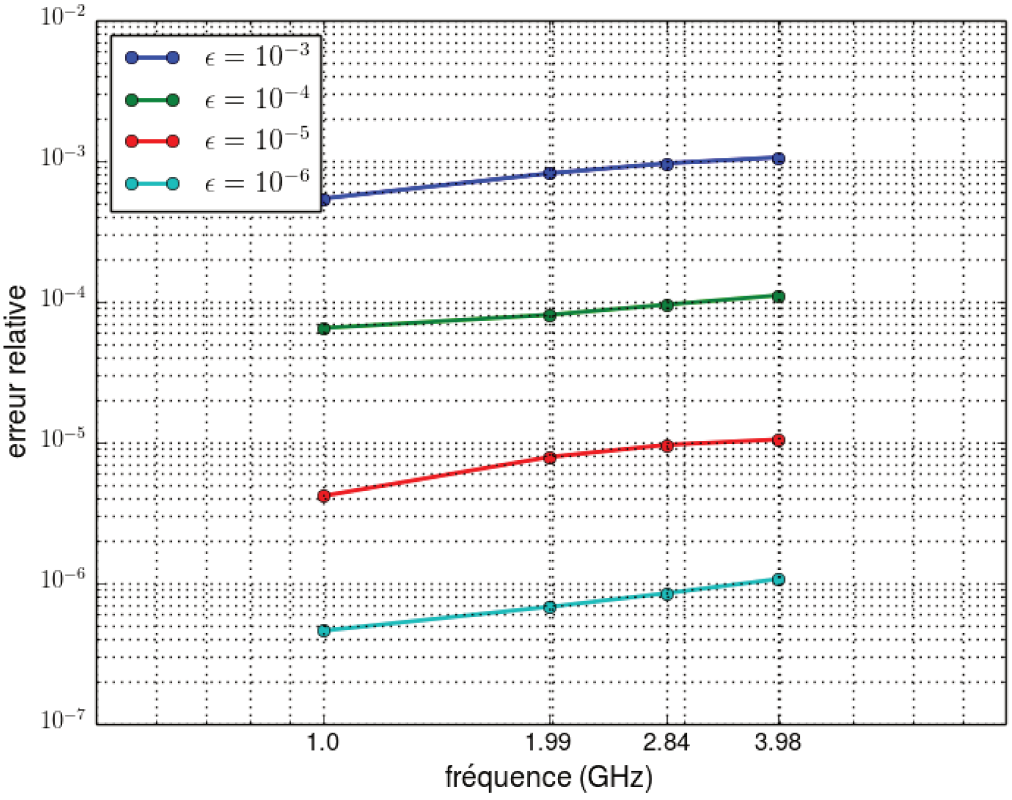


FIGURE 4.25 – Représentation de l'erreur relative maximale commise pour $n_{rhs} = 1000$ vecteurs en fonction de la fréquence. Pour chaque fréquence, on utilise un maillage adapté à $\lambda/5$.

4.5.3.3 Maillage fixé

Pour la sphère à N = 192000 degrés de liberté, on effectue un balayage en fréquence. Pour chaque fréqence, on effectue n_{rhs} produits matrice/vecteur et l'on mesure l'erreur relative maximale commise.

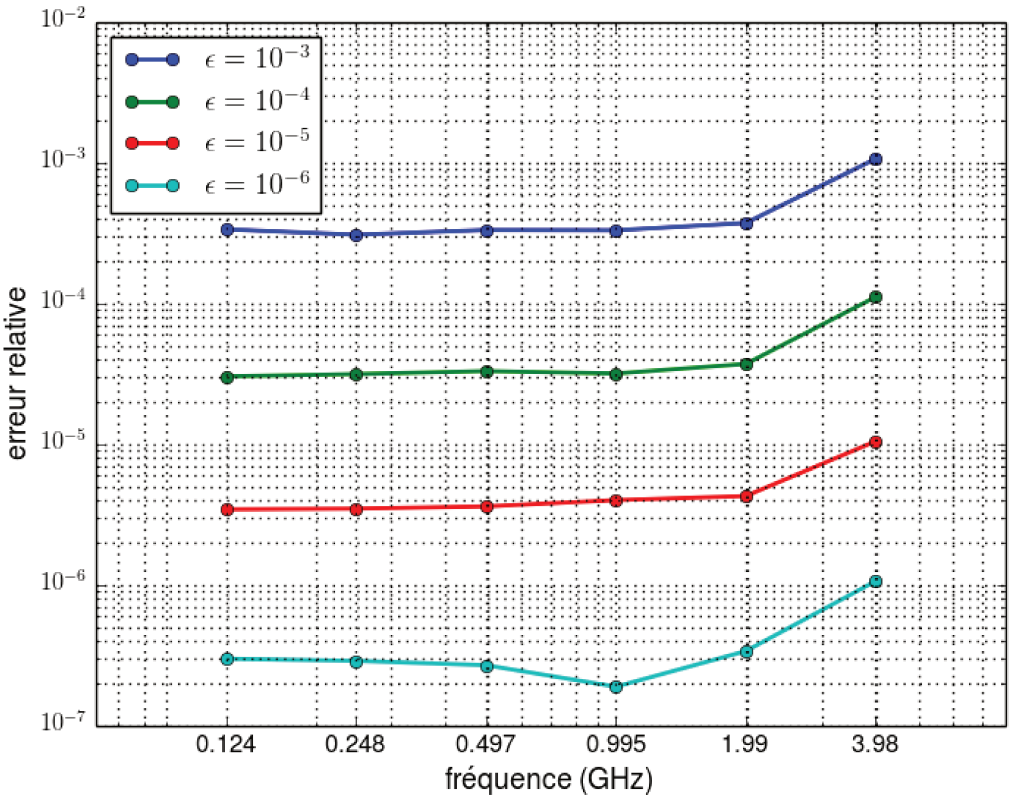


FIGURE 4.26 – Représentation de l'erreur relative maximale commise pour n_{rhs} = 1000 vecteurs en fonction de la fréquence dans le cas de la sphère à N = 192000 degrés de liberté.

On remarque sur la figure 4.26 que l'erreur mesurée est stable jusqu'à une fréquence correspondant à une finesse de maillage en $\lambda/10$. La dernière fréquence correspondant à $\lambda/5$ fournit également une erreur correcte mais légèrement supérieure aux autres erreurs. Il s'agit de la fréquence maximale que l'on peut utiliser pour ce maillage et cela indique que notre maillage n'est plus adapté plutôt qu'un défaut du produit matrice/vecteur.

4.5.4 Commentaires

Les tests précédents montrent que l'algorithme HCA – II est performant dans la pratique. Sous réserve d'admissibilité de Fresnel, les blocs admissibles sont approchés par l'algorithme HCA – II et l'on maîtrise l'erreur commise sur tous les blocs. L'approximation est par ailleurs de bonne qualité pour toutes les dimensions (diamètres et tailles) traitées. Les blocs de grandes dimensions nécessitent un nombre de points suffisant pour approcher les oscillations du noyau de Green. Ce nombre de points est également contrôlé de manière performante dans l'implémentation décrite et testée dans cette partie. Enfin, le bon comportement de l'algorithme HCA – II sur un bloc matriciel conduit à une bonne approximation du produit de la matrice de discrétisation approchée par un vecteur. Le produit matrice/vecteur présenté dans cette partie est stable par rapport à la fréquence et satisfait la tolérance sur l'erreur reltive fournie.

4.6 Conclusion

Ce chapitre présente un cadre rigoureux pour l'approximation de l'opérateur de l'EFIE à l'aide de la méthode des H-matrices. En effet, cette dernière a été introduite dans la pratique pour le noyau statique de l'équation de Laplace $G(x,y) = 1/|x-y|$. Les estimations de complexité que l'on trouve dans la littérature sont données dans ce cas. La pratique montre cependant que l'approximation du noyau oscillant $G(x,y) = \frac{e^{ik|x-y|}}{|x-y|}$ par les $\mathscr{H}$-matrices fournit de bons résultats. L'exemple de la figure 4.22 en est une bonne illustration. En effet, on peut constater numériquement sur des exemples que la méthode produit une bonne approximation avec une croissance du rang apparemment maîtrisée.

Nous avons introduit un critère fréquentiel original qui prend à la fois en compte la déconvolution par les ondes planes réalisée par le critère de Fraunhofer ainsi que le terme du second ordre. À l'aide d'un changement de base adapté, on peut exprimer ce terme du second ordre comme le produit de deux opérateurs 1D, dits opérateurs de Fresnel. Cette décomposition nous fournit un critère fréquentiel moins restrictif que le critère de Fraunhofer et nous avons validé sur des cas pratiques pour des hautes fréquences que ce critère améliore la compression d'un bloc par rapport au critère statique usuel.

Sous réserve de satisfaire ce critère, nous disposons également d'une formule asymptotique décrivant le rang des opérateurs de Fresnel, c'est la formule de Landau-Widom. Ce résultat nous permet de modifier la méthode d' *Hybrid Cross Approximation* (HCA – II) de la littérature afin d'approcher rigoureusement le noyau oscillant tout en contrôlant à la fois l'erreur et le nombre de nœuds d'interpolation. Ce nombre de points d'interpolation représente une borne maximale sur le rang de l'approximation construite et l'on montre que cette borne croît au pire linéairement avec la fréquence comme les tests numériques sur l'opérateur de l'EFEIE à quatre composantes le confirme.

Cependant, ces outils sont particulièrement bien adaptés pour les hautes fréquences et il convient de souligner que le critère statique peut donner dans la pratique de bons résultats (*cf* [Liz14]). En effet, pour des problèmes courants de l'ordre du million d'inconnues, le critère de Fresnel n'amène pas nécessairement de gros gains dans la pratique. Les tests effectués montrent que pour les petites boîtes (*ie* les basses fréquences, le critère statique est aussi performant que le critère de Fresnel. Dans le cas des hautes fréquences, les interactions mettant en jeu une faible section efficace peuvent également être trai-

tées par le critère statique. Ce sont les interactions pour lesquelles la section efficace est grande qui sont la cible du critère que l'on a développé. Pour ces interactions, on a prouvé en l'absence de critère fréquentiel que la croissance de la mémoire varie comme le carré de la fréquence et ainsi que le taux de compression devient constant à haute fréquence. Le critère de Fresnel examiné ici permet de réduire cette croissance à une complexité en $\mathscr{O}(k)$ pour la mémoire. L'exemple de l'ellipsoïde à la figure 4.21 montre qu'il ne s'agit pas d'un cas purement théorique et que pour toutes les interactions de ce type, on s'attend à une dégradation des performances de la méthode des $\mathscr{H}$-matrices si elle est utilisée avec un critère statique.

On peut pallier ces difficultés en effectuant un découpage « informatique ». En effet, de nombreux codes limitent la taille des blocs matriciels admissibles à quelques dizaines de milliers d'inconnues. Grossièrement, à l'aide d'une analyse similaire à celle effectuée à l'aide de (4.80), cette découpe ne considère que des interactions où le critère de Fresnel est proche de la condition statique. En imposant une telle limite, on n'observe donc pas la croissance quadratique dans le cas des plaques opposées car par construction ces interactions sont dans le régime pré-asymptotique. Par ailleurs, les performances de l'algorithme ACA+ ne sont pas les meilleures pour des blocs de plusieurs centaines de milliers d'inconnues et une telle découpe améliore également les performances de l'ACA. Il s'agit d'un bon compromis pratique pour des problèmes de taille modérée mais cela ne peut s'appliquer pour les problèmes de très grandes tailles que nous visons par la suite.

Enfin, on notera tout particulièrement que les développements effectués sur le noyau de Green ne sont pas spécifiques aux $\mathscr{H}$-matrices et que les résultats sur la croissance du rang et/ou l'approximation du noyau de Green peuvent être utilisés de manière indépendante dans d'autres méthodes rapides, par exemple la méthode FMM directionnelle.

4.7 Références

[Beb00] M Bebendorf. *Hierarchical Matrices.* Springer, 2000. 191

[Ben84] A. Bendali. *Approximation par éléments finis de surface de problèmes de diffraction des ondes électromagnétiques.* PhD thesis, Université Pierre et Marie Curie Paris 6, 1984. 171

[BG05] S. Börm and L. Grasedyck. Hybrid cross approximation for integral operators. *Numerische Mathematik,* 2005. 174, 180, 184

[BGH12] Steffen Börm, Lars Grasedyck, and Wolfgang Hackbusch. Hierarchical matrices. Technical report, 2012. 184, 186, 191

[BKV15] M. Bebendorf, C. Kuske, and R. Venn. Wideband nested cross approximation for helmholtz problems. *Numerische Mathematik*, 130(1) :1–34, 2015.

[Liz14] B. Lizé. *Résolution Directe Rapide pour les Éléments Finis de Frontière en Électromagnétisme et Acoustique : H-Matrices. Parallélisme et Applications Industrielles.* PhD thesis, Université Paris 13, 2014. 190, 219

[Mes11] M. Messner. *Fast Boundary Element Methods in Acoustics.* PhD thesis, Graz University of Technology,Institute of Applied Mechanics, 2011. 167

[RT77] P.A. Raviart and J.M. Thomas. A mixed finite element method for 2-nd order elliptic problems. *Mathematical Aspects of Finite Element Methods*, vol. 606 de Lecture Notes in Mathematics :292–315, 1977. 171

[Syl02] G. Sylvand. *La Méthode Multipôle Rapide en Électromagnétisme. Per- formances, Parallélisation, Applications.* PhD thesis, ENPC/CERMICS, 2002. 182

Conclusion

La motivation principale de cette thèse était de comprendre le bon fonctionnement de la méthode des $\mathscr{H}$-matrices dans le cas d'une matrice BEM pour un problème d'onde haute fréquence.

Nous avons ainsi présenté la méthode des $\mathscr{H}$-matrices en soulignant les points délicats lors de l'utilisation pour un problème d'onde, contrairement au cas de l'équation de Laplace. Nous avons effectué une revue d'articles de la littérature concernant la compression d'un bloc de rang faible. Plusieurs méthodes, algébriques ou analytiques, ont alors été présentées. Nous avons de plus souligné les idées communes reliant ces méthodes ; l'algorithme HCA-II par exemple, fait intervenir à la fois l'interpolation du noyau ainsi que des compressions algébriques. Outre la compression, nous avons montré une condition qui garantit l'obtention d'une approximation de rang faible : c'est la condition d'admissibilité statique.
La croissance quadratique du rang d'un bloc matriciel BEM dans la zone d'admissibilité statique nous a conduit à introduire un critère fréquentiel d'admissibilité , *le critère de Fresnel*, expliquant les bons résultats de la littérature quant à l'application des $\mathscr{H}$-matrices pour des problèmes d'ondes relativement haute fréquence. Sous réserve de satisfaire le critère de Fresnel, nous avons établi -par un changement de base adapté- une estimation asymptotique originale et précise du rang du noyau oscillant. Cette estimation nous a permis de développer un algorithme d'approximation compressée rapide, fiable et robuste du noyau oscillant.

Les perspectives liées à cette estimation du rang du noyau oscillant sont nombreuses et de natures différentes. D'un point de vue théorique, le terme du second ordre dans le développement limité correspond à la partie transverse de la phase. Nous avons choisi d'exploiter la forme de produit tensoriel de deux opérateurs 1D afin de développer notre estimation du rang. Cependant, Slepian montre qu'il existe des approximations utilisant les fonctions d'onde sphéroïdales en dimension supérieure. Une piste d'amélioration serait d'étudier ce terme d'ordre deux à l'aide de ces résultats afin d'obtenir une estimation plus précise. Les fonctions d'onde sphéroïdales elles, interviennent en tant que fonctions propres des opérateurs 1D composant le terme d'ordre 2. Le comportement asymptotique de ces fonctions propres à haute fréquence est connu et utilisé dans un contexte d'apprentissage statistique. Le théorème de Mercer permet alors de décomposer le noyau oscillant sur cette base de fonctions propres.
D'un point de vue algorithmique, plusieurs modifications sont envisageables. Nous avons montré que le *clustering* est une étape d'une grande importance. La réduction de la section efficace au cours du *clustering* permet à la condition d'admissibilité fréquentielle

d'être satisfaite pour de plus gros blocs matriciels. Un *clustering* fréquentiel et directionnel serait alors une nette amélioration. Pour le calcul de la base orientée, une amélioration possible est de travailler sur les boîtes englobantes plutôt que sur les supports des degrés de liberté afin d'effectuer les calculs plus rapidement. Le critère fréquentiel est composé d'une partie statique ainsi que d'une partie oscillante. Dans un premier temps, la condition statique fournit une découpe en blocs de la matrice telle que décrite pour les $\mathscr{H}$-matrices et le noyau $1/|x-y|$. Le test d'admissibilité sur la partie oscillante est effectué à partir du calcul de la section efficace. Le calcul de la section efficace permet également de déterminer les largeurs de bande associées au terme d'ordre 2. Il est alors possible, outre le test d'admissibilité fréquentielle, d'obtenir une estimation *a priori* du rang du bloc. L'estimation *a priori* sera d'autant plus fiable que le bloc est haute fréquence. Une amélioration possible est alors d'utiliser cette information pour piloter la découpe de la matrice.
Les estimations du rang en fonction de la section efficace peuvent être à la source d'améliorations de la méthode des $\mathscr{H}$-matrices d'un point de vue industriel. En effet, la connaissance du rang permet d'estimer le nombre de coefficients à garder en mémoire pour un bloc admissible. Il est possible d'utiliser ces estimations pour améliorer le parallélisme de la méthode pour des calculs distribués ainsi que la gestion de la mémoire dans le cas de calculs *out-of-core* dans le cas de matrices ne tenant pas en mémoire.

Par ailleurs, nous envisageons plusieurs contextes d'utilisation pour l'estimation de rang développée dans ce manuscrit. Bebendorf a utilisé un critère d'admissibilité fréquentiel, le critère de Fraunhofer, pour l'assemblage et le produit matrice-vecteur d'une $\mathscr{H}^2$-matrice. Le critère de Fresnel pourrait se révéler efficace dans ce contexte pour obtenir moins de blocs matriciels tout en contrôlant la croissance de leurs rangs. Les estimations étudiées dans le cas du noyau oscillant n'étant pas spécifiques à la méthode des $\mathscr{H}$-matrices, l'utilisation de ces résultats dans un contexte FMM est une application possible. Enfin, une autre application digne d'intérêt est de combiner ces estimations avec des techniques employées dans le domaine de l'apprentissage statistique (*machine learning*). En effet, le calcul de la base d'interaction orientée fait intervenir plusieurs paramètres géométriques : la distance entre les deux groupes, la section efficace, les diamètres. Pour plusieurs blocs, ces paramètres ainsi que le rang numérique pour une précision donnée peuvent servir à construire une base d'apprentissage et déterminer des heuristiques sur le rang à l'aide de machines à vecteurs de support (*Support Vector Machine* ou SVM).